IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 25, 2019, accepted March 17, 2019, date of publication March 27, 2019, date of current version April 16, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2907798

Optimized Energy Aware 5G Network Function Virtualization

AHMED N. AL-QUZWEENI, AHMED Q. LAWEY, TAISIR E. H. ELGORASHI, AND JAAFAR M. H. ELMIRGHANI, (Senior Member, IEEE)
School of Electronic and Electrical Engineering, University of Leeds, Leeds LS2 9JT, U.K.

Corresponding author: Ahmed N. Al-Quzweeni (ml13anaq@leeds.ac.uk)

This work was supported in part by the Engineering and Physical Sciences Research Council (ESPRC), in part by the INTERNET under Grant EP/H040536/1, and in part by the STAR Projects under Grant EP/K016873/1.

ABSTRACT In this paper, network function virtualization (NFV) is identified as a promising key technology, which can contribute to energy-efficiency improvement in 5G networks. An optical network supported architecture is proposed and investigated in this paper to provide the wired infrastructure needed in 5G networks and to support NFV toward an energy efficient 5G network. In this paper, the mobile core network functions, as well as baseband function, are virtualized and provided as VMs. The impact of the total number of active users in the network, backhaul/fronthaul configurations, and VM inter-traffic are investigated. A mixed integer linear programming (MILP) optimization model is developed with the objective of minimizing the total power consumption by optimizing the VMs location and VMs servers' utilization. The MILP model results show that virtualization can result in up to 38% (average 34%) energy saving. The results also reveal how the total number of active users affects the baseband virtual machines (BBUVMs) optimal distribution whilst the core network virtual machines (CNVMs) distribution is affected mainly by the inter-traffic between the VMs. For real-time implementation, two heuristics are developed, an energy efficient NFV without CNVMs inter-traffic (EENFVnoITr) heuristic and an energy efficient NFV with CNVMs inter-traffic (EENFVwithITr) heuristic, both produce comparable results to the optimal MILP results. Finally, a genetic algorithm is developed for further verification of the results.

INDEX TERMS 5G networks, backhaul, BBU, energy efficiency, fronthaul, genetic algorithm, IP over WDM, network function virtualization, NFV.

I. INTRODUCTION

According to Cisco Visual Networking Index, mobile data traffic will witness seven pleats between 2016 and 2021 and will grow at a Compound Annual Growth Rate (CAGR) of 46% reaching 48.3 exabytes per month by 2021 [2]. This growth is driven by a number of factors such as the enormous amount of connected devices and the development of data-greedy applications [3]. With such a tremendous amount of data traffic, a revolutionary mobile network architecture is needed. Such a network (5G) will contain a mix of a multiple access technologies supported by a significant amount of new spectrum to provide different services to a massive number of different types of users (eg., IoT, personal, industrial) at high data rate, any time with potentially less than 1 ms latency [4]. 5G networks are expected to be operational by 2020 where a huge number of devices and application will use it [5]. Users, applications, and devices of different kinds and purposes need to send and access data from distributed and centralized servers and databases using public and/or private networks and clouds. To support these requirements, 5G mobile networks have to possess intelligence, flexible traffic management, adaptive bandwidth assignment, and at the forefront of these traits is energy efficiency. Information and Communication Technology (ICT) including services and devices are responsible for about 8% of the total world energy consumption [6] and contributed about 2% of the global carbon emissions [7]. It is estimated that, if the current trends continue, the ICT energy consumption will reach about 14% of the total worldwide consumption by 2020 [6].

There have also been various efforts from researchers on reducing the power consumption and improving the energy-efficiency in mobile networks [8] and 5G networks. For

The associate editor coordinating the review of this manuscript and approving it for publication was Irfan Ahmed.

instance, the authors in [9] focused on the power consumption of base stations. They proposed a time-triggered sleep mode for future base stations in order to reduce the power consumption. The authors in [10] investigated the base stations computation power and compared it to the transmission power. They concluded that the base station computation power will play an important role in 5G energy-efficiency. The authors of [11] developed an analytical model to address the planning and the dimensioning of 5G Cloud RAN (C-RAN) and compared it to the traditional RAN. They showed that C-RAN can improve the 5G energy-efficiency. The research carried out in [12] focused on offloading the network traffic to the mobile edge to improve the energy-efficiency of 5G mobile networks. The authors developed an offloading mechanism for mobile edge computing in 5G where both file transmission and task computation were considered.

Virtualization has been proposed as an enabler for the optimum use of network resources, scalability, and agility. In [13] the authors stated that NFV is the most important recent advance in mobile networks where among its key benefits is the agile provisioning of mobile functions on demand. The fact that it is now possible to separate the functions form their underlying hardware and transfer them into software-based mobile functions as well as provide them on demand, presents opportunities for optimizing the physical resources and improving the network energy efficiency.

In this paper, network function virtualization is identified as a promising key technology that can contribute to the energy-efficiency improvement in 5G networks. In addition, an optical network architecture is proposed and investigated in this paper to provide the wired infrastructural needed in 5G networks, and to support NFV and content caching. In the literature, NFV was investigated either in mobile core networks [14]–[16] or in the radio access network [17]–[19] of the mobile network and mostly using pooling of resources such as the work in [20], [21]. In contrast, virtualization in this paper is not limited to a certain part in the mobile network, but is applied in both the mobile core network and the radio access network. Moreover, it is not confined to pooling the network resources, but is concerned with mobile functions-hardware decoupling and considers converting these functions into software-based functions that can be placed optimally. For instance, instead of having a dedicated hardware in RAN to implement baseband processing for each remote radio head and another for user policy in the core network, two virtual machines can share a single general purpose node to implement these functions. In such way the amount of hardware could be reduced to save energy. In addition, packing and hosting VMs close to the user results in short traffic path which results in low traffic induced power.

A Mixed Integer Linear Programming model, real-time heuristics and Genetic Algorithm are developed in this paper with the goal of improving the energy-efficiency in 5G mobile networks.

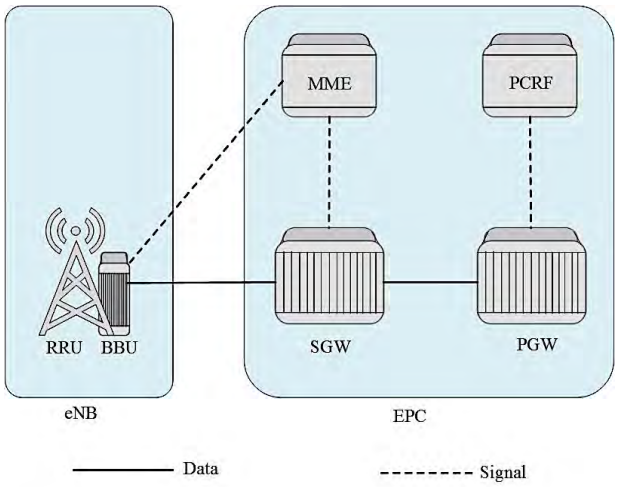

FIGURE 1. Evolved packet system architecture.

II. NFV IN 5G NETWORKS

According to the third generation partnership project (3GPP) the evolved packed core (EPC) is an important step change [22]. There are four main functions in the EPC [23], [24] illustrated in Fig. 1: the packet data network gateway (PGW), the serving gateway (SGW), the mobility and management entity (MME), and the policy control and charging role function (PCRF).

The work in this paper extends our work in [25], [26] to include a number of factors such as the total number of active users in the network during the day, the backhaul and fronthaul configuration and the required workload for baseband processing. It introduces an optical-based framework for energy efficient NFV deployment in 5G networks and provides full MILP details and associated heuristics. In this framework, the functions of the four entities of mobile core network are virtualized and provided as one virtual machine, which is dubbed ''core network virtual machine'' (CNVM). For the radio access side, the BBU and RRU are split and the function of the BBU is virtualized and provisioned as a ''BBU virtual machine'' (BBUVM). Consequently, the wireless access network of the mobile system will encompass only the RRU that remain after the RRU-BBU decoupling. RRU is referred to here as ''RRH'' (as in a number of studies [27]–[29]) after it is separated from BBU. The traffic from CNVM to RRH is compelled to pass through BBU-VMs for baseband processing, as in Fig. 2. Moreover, the capabilities of Passive Optical Networks (PON) are leveraged as an energy-efficient broadband access network to connect the IP over WDM core network to RRH nodes, and to represent the wired access network of our proposed system. Fig. 3 shows three locations that can accommodate virtual machines (VMs) of any type (BBUVMs or CNVMs), which are the optical network unit (ONU), optical line terminator (OLT), and the IP over WDM nodes. For simplicity, the nodes where the hosted servers are accommodated are referred to as ''Hosting Nodes''.

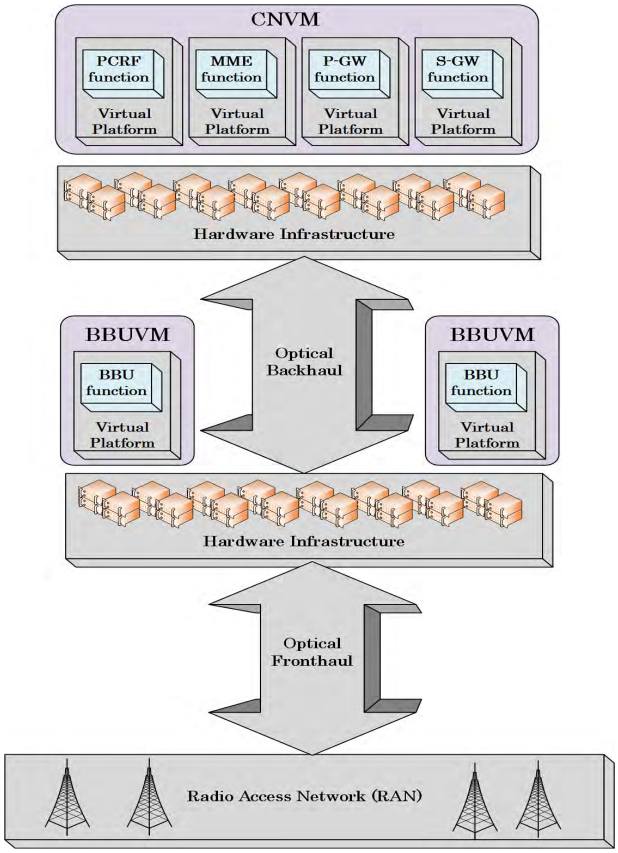

FIGURE 2. The proposed architecture for Energy Efficient NFV in 5G.

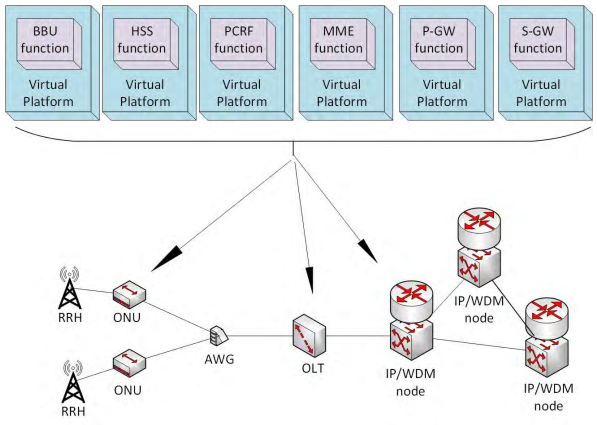

FIGURE 3. The candidate locations for hosting VM in the proposed architecture.

The hosting nodes (ONU, OLT and IP over WDM nodes) might host one VM or more than one VM of the same or different types, bringing forth the creation of small clouds, or ''Cloudlets''. Therefore, the proposed architecture will provide an agile allotment of services and processes through flexible distribution of VMs over the optical network (PON and IP over WDM network), which is one of the main concerns of this work in minimizing the total power consumption. Based on this architecture, a MILP formulation has been developed with the overall aim of minimizing power consumption.

III. FRONTHAUL AND BACKHAUL CONFIGURATION AND THE AMOUNT OF BASEBAND PROCESSING WORKLOAD

This section illustrates the configuration of the fronthaul and backhaul used in the proposed network; so that the ratio of the backhaul to the fronthaul data rate could be calculated. Fronthaul is the network segment that connects the remote radio head (RRH) to the baseband unit (BBU) [30], whilst the network segment that connects the BBU to the mobile core network (CN) is called ''backhaul'' [31]. The internal interface of the fronthaul is defined as a result of the digitization of the radio signal according to a number of specifications. The well-known and most used specification among radio access network (RAN) vendors is the Common Public Radio Interface (CPRI) specification [32] which is implemented using digital radio over fiber (D-RoF) techniques. On the other hand, the backhaul interface leverages Ethernet networks as they are the most cost effective network for transporting the backhaul IP packets [33], [34].

In order to adequately determine the data rate in each network segment (backhaul and fronthaul), we will start with the physical layer of the current mobile network which is the Long-Term Evolution (LTE) network. The LTE network uses single-carrier frequency-division multiple access (SC-FDMA) uplink (UL), whilst orthogonal frequency-division multiple access (OFDM) is used in the downlink (DL) [35]. In both techniques, the transmitted data are turbo coded and modulated using one of the following modulation formats: QPSK, 16QAM, or 64QAM with 15 kHz subcarriers spacing [36]. A generic frame is defined in LTE which has 10 ms duration and 10 equal-sized subframes. Each subframe is divided into two slot periods of 0.5 ms duration [37]. Depending on the cyclic prefix (CP) used, slots in OFDMA have either 7 symbols for normal CP or 6 symbols for extended CP [38]. Fig. 4 illustrates an LTE downlink frame with normal CP. In the LTE frames, a resource element (RE) is the smallest modulation structure which has one subcarrier of 15 kHz by one symbol [39]. Resource elements are grouped into a physical resource block (PRB) which has dimensions of 12 consecutive subcarriers by one slot (6 or 7 symbols). Therefore, one PRB has a bandwidth of 180 kHz (12×15 kHz). Different transmission bandwidths use different number of physical resource blocks (PRBs) per time slot (0.5 ms) which are defined by 3GPP [40]. Fig. 5 illustrates the LTE downlink resource grid. For instance, 10 MHz transmission bandwidth has 50 PRBs whilst 20 MHz has 100 PRBs [41]. If 10 MHz bandwidth is used with 16 QAM (6 bits/symbol) and 7 OFDM symbols (short CP), we have

$$\frac{\left(50RB \times \frac{12subcarriers}{RB} \times \frac{7\ symbols}{subcarrier} \times \frac{6\ bits(QAM)}{symbol}\right)}{0.5\ ms\ times\ lot} = 50.4\ Mbps \quad (1)$$

$$50.4\ Mbps \times 0.874\ system\ efficiency = 44.0496\ Mbps \quad (2)$$

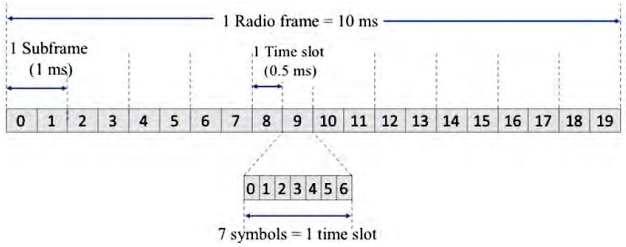

FIGURE 4. LTE downlink frame with normal CP.

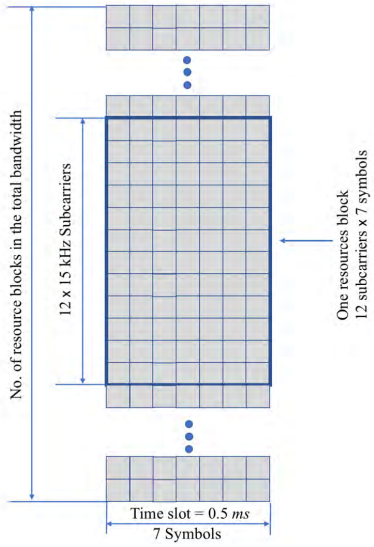

FIGURE 5. LTE downlink resources grid.

It is worth mentioning that for each transmission antenna, there is one resource grid (50 PRBs for 10 MHz); therefore in 2×2 MIMO the previous data rate is doubled (100.8 Mbps) [42].

The transmission of user plane data is achieved in the form of In-phase and quadrature (IQ) components that are sent via one CPRI physical link where each IQ data flow represents the data of one carrier for one antenna that is called Antenna-Carrier (AxC) [43]. Two main parameters affect the data carried by AxC, [42]:

Sampling frequency which is calculated as: subcarrier BW (15 kHz) times the FFT window (size). The FFT size is chosen to be the least multiple of 2 that is greater than the ratio of the radio signal bandwidth to the subcarrier BW. For instance, if the radio bandwidth is 10 MHz, the FFT size is the least multiple of 2 that is greater than 666.67 (10 MHz / 15 kHz) which is 1028 (2^{10}). In this case the sampling frequency is calculated as $150\ kHz \times 1024 = 15.36\ MHz$. Using the same approach, the sampling frequency at 20 MHz radio bandwidth system is 30.72 MHz.

IQ sample width (M-bits per sample): According to the CPRI specification, the IQ sample width supported by CPRI is between 4 and 20 bits per sample for I and Q in the uplink and it is between 8 and 20 in the downlink [43]. For instance, with M = 15 bits per sample; one AxC contains 15 bits per sample for I and 15 bits per sample for Q which are 30 ($2 \times M$) bits per sample I and Q which are transported in sequence: $I_0Q_0I_1Q_1 \ldots I_{14}Q_{14}$. The IQ sample data rate can be calculated by multiplying the number of bits per sample by the sampling frequency. For instance; for a radio bandwidth of 10 MHz (fs = 15.36 MHz) and IQ samples 15 (M = 15) the IQ data rate is:

$$(2 \times M) \times f_s = \left(30\ bits/sample\right) \times 15.36\ MHz = 0.4608\ Gbps \quad (3)$$

CPRI data rate is designed based on the Universal Mobile Telecommunications System (UMTS) chip rate [43] which is 3.84 Mbps [44], [45]. Therefore, one basic CPRI frame is created every Tc = 260.416 ns (1/3.84 MHz) and this duration should remain constant for all CPRI options and data rates. According to CPRI specification in [43], one basic CPRI frame consists of 16 words indexed (W = 0...15), where the first word is reserved for control. The length of the frame word (T) depends on the CPRI line rate as specified by CPRI specification in [43]. Accordingly, the transmission of AxC data will be expanded by a factor of 16/15 (15 bits payload, 1 bit control and management). In addition to the sampling rate fs that is calculated earlier, AxC data needs to be coded using either 8B/10B or 64B/66B.

To put all these calculations together, let's start with the number of bits per word in the CPRI frame. The number of bits per word is equal to the total number of bits per frame divided by the frame payload words (15 words). Recall that the frame duration should be constants (206.416 ns); therefore:

$$\frac{(no\ of\ bits\ per\ word \times 15\ words)}{samples\ of\ IQ\ f_{IQ}} = 260.416\ ns \quad (4)$$

$$no\ of\ bits\ per\ word\ \left(N^{bpw}\right) = \frac{f_{IQ} \times 260.416\ ns}{15}. \quad (5)$$

One CPRI frame word has N bpw bits, since the CPRI frame has 16 words:

$$N^{bpF} = N^{bpw} \times (15\ payload\ words) + N^{bpw} \times 1\ control\ word \quad (6)$$

$$N^{bpF} = N^{bpw} \times (15+1) = \frac{f_{IQ} \times 260.416\ ns}{15} \times 16 \quad (7)$$

To calculate the data rate in one CPRI frame

$$\frac{N^{bpF}}{260.416ns} = \frac{\frac{f_{IQ} \times 260.416ns}{15} \times 16}{260.416ns} = f_{IQ} \times \frac{16}{15} \quad (8)$$

By replacing f_{IQ} with $2 \times M \times f_s$ where M is defined earlier as the number of IQ bits.

In addition, AxC data are coded by either 8B/10B or 64B/66B. By putting these together, the CPRI data rate is calculated as

$$2 \times M \times f_s \times \frac{16}{15} \times L_{coding} \quad (9)$$

Finally, the ratio of the backhaul to fronthaul data rate is calculated as:

$$\frac{backhaul\ \ (IP)\ data\ rate}{fronthaul\ (CPRI)\ data\ rate} = \frac{44.0496\ Mbps}{327.68\ Mbps} \times 100\% = 13.44\% \quad (10)$$

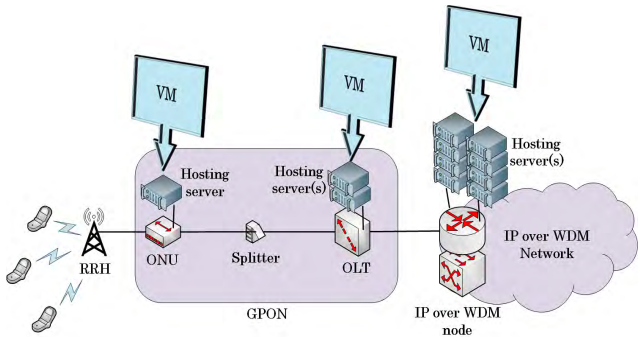

FIGURE 6. Candidate locations for hosting VMs in the proposed architecture.

Therefore, depending on coding, sampling, quantization, and other parameters; the baseband processing adds overheads to the backhaul traffic as it passes through the BBU. In this work the ratio (13.44%) calculated in (10) is used in our model, whilst the amount of workload in Giga Operation Per Second (GOPS) needed to process one user traffic is used based on the following relation which is explained in [46]:

$$wl = \left(30 \cdot A \; + \; 10 \cdot A^2 \; + \; 20\frac{M}{6} \cdot C \cdot L\right) \cdot \frac{R}{50} \quad (11)$$

where:

wl: is the baseband workload in (GOPS) needed to process one user traffic,

A: number of antennas used,

M: modulation bits,

C: the code rate,

L: number of MIMO layers,

R: number of physical resource blocks allocated for the user.

IV. MILP MODEL

This section introduces the MILP model that has been developed to minimize the power consumption due to both processing by virtual machines (hosting servers) and the traffic flow through the network. As mentioned in the previous section, the MILP model considers an optical-based architecture with two types of VMs (BBUVM and CNVMs) that could be accommodated in ONU, OLT and/or IP over WDM as in Fig. 6. The maximum number of VM-hosting servers considered was 1, 5, and 20 in ONU, OLT, and IP over WDM nodes respectively, which is commensurate with the node size and its potential location and hence space limitations (together with the size of exemplar network considered in the MILP). All VM-hosting servers were considered as sleep-capable servers for the purpose of VM consolidation (bin packing).

For a given request, the MILP model responds by selecting the optimum number of virtual machines and their location so that the total power consumption is minimized.

The developed MILP model is defined by a set of indices, parameters and variable which are listed in Table 1, Table 2 and Table 3 respectively.

TABLE 1. Energy efficient NFV MILP model indices.

Indices	Comments
x, y	Indices of any two nodes in the proposed model
m, n	Indices of any two nodes in the physical layer of the IP over WDM network
i, j	Indices of any two nodes in the IP layer of the IP over WDM network.
r	Index of RRH node
h, u, p, q	Indices of the nodes where the VM could be hosted

The total power consumption is composed of:

1) The power consumption of RRHs and ONUs

$$\sum_{x\in U}\left[\Omega^R_x + \frac{\Omega^U}{C^U}\cdot\left(\sum_{h\in H}\sum_{r\in R}\sum_{y\in T^N_x}\lambda^R_{h,r,x,y} + \sum_{p\in H}\sum_{q\in H:p\neq q}\sum_{y\in T^N_x\cap H}\lambda^T_{p,q,x,y}\right)\right] \quad (12)$$

2) The power consumption of the OLTs

$$\sum_{x\in L}\left[\Omega^{Ld}Ld + \frac{\Omega^L - \Omega^{Ld}}{C^L}\cdot\left(\sum_{h\in H}\sum_{r\in R}\sum_{y\in T^N_x}\lambda^R_{h,r,x,y} + \sum_{p\in H}\sum_{q\in H:p\neq q}\sum_{y\in T^N_x\cap H}\lambda^T_{p,q,x,y}\right)\right] \quad (13)$$

3) The power consumption of the IP over WDM network

$$\left(\Omega^{RP}\cdot\sum_{m\in N}\Lambda_m\right) + \left(\Omega^{RP}\cdot\sum_{m\in N}\sum_{n\in N^N_m}W_{m,n}\right) + \left(\Omega^T\cdot\sum_{m\in N}\sum_{n\in N^N_m}W_{m,n}\right) + \left(\Omega^E\cdot\sum_{m\in N}\sum_{n\in N^N_m}A_{m,n}\cdot f_{m,n}\right) + \left(\Omega^G\cdot\sum_{m\in N}\sum_{n\in N^N_m}N^G_{m,n}\cdot W_{m,n}\right) \quad (14)$$

4) The total power consumption of VMs and hosting servers

$$\sum_{h\in H}\left(\Omega^{Sd}\cdot\left(\Psi^i_h + \sigma^\chi_h\right) + \Psi^f_h\cdot\left(\Omega^S - \Omega^{Sd}\right)\right) \quad (15)$$

The model objective is to minimize the total power consumption as follows:

$$\text{Minimize}\sum_{x\in U}\left[\Omega^R_x + \frac{\Omega^U}{C^U}\cdot\left(\sum_{h\in H}\sum_{r\in R}\sum_{y\in T^N_x}\lambda^R_{h,r,x,y} + \sum_{p\in H}\sum_{q\in H:p\neq q}\sum_{y\in T^N_x\cap H}\lambda^T_{p,q,x,y}\right)\right]$$

TABLE 2. Energy-efficient NFV MILP model parameters.

Parameters	Comments
R	Set of RRH nodes
U	Set of ONU nodes
L	Set of OLT nodes
N	Set of IP over WDM nodes
T	Set of all nodes (RRH, ONU, OLT, and IP over WDM nodes)
N_m^N	Set of neighbors of node m in the IP over WDM network, $\forall\ m\ \in N$
T_x^N	Set of neighbors of node x, $\forall\ x\ \in T$
H	Set of hosting nodes (ONU, OLT, and IP over WDM nodes)
l	Line coding rate (bits per sample)
y	Number of MIMO layers (ie number of data streams)
q^{am}	Number of bits used in QAM modulation
a	Number of antennas in a cell
c^p	CPRI link data rate
Ψ^χ	Maximum BBU workload needed for fully loaded RRH (GOPS); calculated as: $30 \cdot a + 10 \cdot a^2 + 20 \cdot q^{am} \cdot l \cdot y$
Ψ^S	Server CPU maximum workload (GOPS)
Ψ_h^C	Workload needed for hosting one CNVM (GOPS)
ρ_r	Number of active users connected to RRH node r
n	Maximum number of physical resource blocks for cell (r)
p^b	Number of physical resource blocks per user
λ_r^R	RRH node r traffic demand (Gbps); calculated as: $[(pb/n) \cdot c^p \cdot \rho_r]$, where $r\ \in R$
$\nabla_{p,q}$	Inter-traffic between core network VMs (CNVM) at hosting nodes p, and q (Gbps)
α	The ratio of the backhaul to the fronthaul traffic (unitless)
Ω^U	ONU maximum power consumption (W)
Ω^L	OLT maximum power consumption (W)
Ω^{Ld}	OLT idle power (W)
C^L	OLT maximum capacity (Gbps)
C^U	ONU maximum capacity (Gbps)
Ω_x^R	Power consumption of the Remote Radio Head (RRH) connected to ONU node x (W)
Ω^S	Server maximum power consumption (W)
Ω^{Sd}	Server idle power (W)
Ω_h^H	Maximum power consumption of hosting VMs at node h
β	Large number (unitless)
η	Very small number (unitless)
B	Capacity of the wavelength channel (Gbps)
w	Number of wavelengths per fiber
Ω^T	Transponder power consumption (W)
Ω^{RP}	Router power consumption per port (W)
Ω^G	Regenerator power consumption (W)
Ω^E	EDFA power consumption (W)
$N_{m,n}^G$	Number of regenerators in the optical link (m,n)
S	Maximum span distance between EDFAs (km)
$D_{m,n}$	Distance between node pair (m,n) in the IP over WDM network (km)
$A_{m,n}$	Number of EDFAs between node pair (m,n) calculated as $A_{m,n} = \left((D_{mn}/S) - 1\right) + 2$

TABLE 3. Energy-efficient NFV MILP model variables.

Variables	Comments
$\lambda_{p,h}^B$	Traffic from CNVMs in node p to the BBUVMs in node h (Gbps)
$\lambda_{h,r}^R$	Traffic from BBUVMs in node h to the RRH node r (Gbps)
$\sigma_{h,r}^B$	Binary indicator, set to 1 if the node h hosts BBUVMs to serve the RRH node r, 0 otherwise
σ_h^B	Binary indicator, set to 1 if the node h hosts a BBUVM, 0 otherwise
$\sigma_{p,h}^E$	Binary indicator, set to 1 if the node h hosts CNVMs to serve the BBUVMs at hosting node h, 0 otherwise
σ_p^E	Binary indicator, set to 1 if the hosting node p hosts CNVMs is, 0 otherwise
$\psi_{p,q}$	Binary indicator, set to 1 if two different hosting nodes p and q host CNVMs, 0 otherwise. It is equivalent to the ANDing of the two binary variables (σ_p^E, σ_q^E).
σ_h^χ	Binary indicator, set to 1 if the hosting node h hosts any virtual machine of any type, 0 otherwise. It is equivalent to the ORing of the two binary variables (σ_h^B, σ_h^E).
$\lambda_{p,q}^E$	Traffic between hosting nodes due to CNVMs communication (Gbps)
$\lambda_{p,q}^T$	Total traffic from node p to node q caused by CNVM to CNVM traffic and CNVM to BBUVM traffic (Gbps)
$\lambda_{h,r,x,y}^R$	Traffic from hosting node h to RRH node r that traverses the link between the nodes (x,y) in the network in Gb/s
$\lambda_{p,q,x,y}^T$	Total traffic from node p to node q that traverses the link between the nodes (x,y) in the network (Gbps)
Ψ_h^B	BBU workload at node h (GOPS)
Ψ_h^i	The integer part of the total normalized workload at node h.
Ψ_h^f	The fractional part of the total normalized workload at node h.
$W_{i,j}$	Number of wavelength channels in the virtual link (i,j)
$W_{i,j,m,n}$	Number of wavelength channels in the virtual link (i,j) that traverse the physical link (m,n)
$f_{m,n}$	Number of fibers in the physical link (m,n)
$W_{m,n}$	Total number of wavelengths in the physical link (m,n)
Λ_m	Number of aggregation ports of the router at node m

$$+\sum_{x\in L}\left[\Omega^{Ld}+\frac{\Omega^L-\Omega^{Ld}}{C^L}\cdot\left(\sum_{h\in H}\sum_{r\in R}\sum_{y\in T_x^N}\lambda_{h,r,x,y}^R+\sum_{p\in H}\sum_{q\in H:p\neq q}\sum_{y\in T_x^N\cap H}\lambda_{p,q,x,y}^T\right)\right]$$

$$+\left(\Omega^{RP}\cdot\sum_{m\in N}\Lambda_m\right)+\left(\Omega^{RP}\cdot\sum_{m\in N}\sum_{n\in N_m^N}W_{m,n}\right)$$

$$+\left(\Omega^{T}\cdot\sum_{m\in N}\sum_{n\in N_m^N}W_{m,n}\right)+\left(\Omega^{E}\cdot\sum_{m\in N}\sum_{n\in N_m^N}A_{m,n}\cdot f_{m,n}\right)$$

$$+\left(\Omega^{G}\cdot\sum_{m\in N}\sum_{n\in N_m^N}N_{m,n}^G\cdot W_{m,n}\right)$$

$$+\sum_{h\in H}\left(\Omega^{Sd}\cdot\left(\Psi_h^i+\sigma_h^\chi\right)+\Psi_h^f\cdot\left(\Omega^S-\Omega^{Sd}\right)\right)$$

$$\forall r\in R,\forall h\in H \tag{16}$$

Subject to the following constraints:

1) Traffic from CNVM to BBUVM

$$\sum_{p\in H}\lambda_{p,h}^B=\alpha\cdot\sum_{r\in R}\lambda_{h,r}^R\quad\forall h\in H \tag{17}$$

2) Traffic to RRH nodes

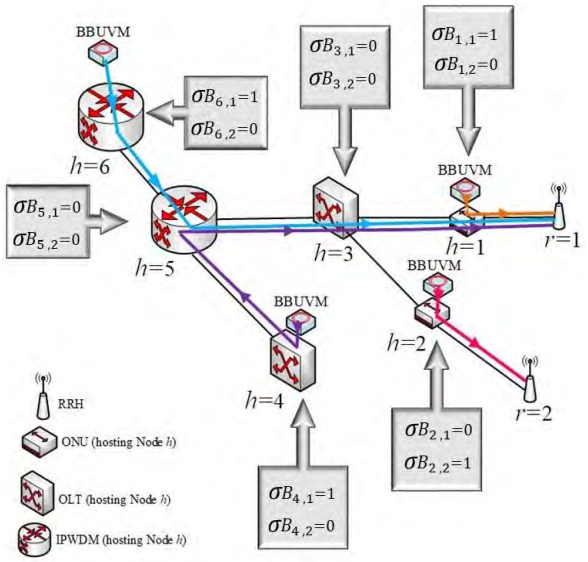

FIGURE 7. BBUVM and the traffic toward RRH nodes.

$$\sum_{h\in H} \lambda^R_{h,r} = \lambda^R_h \quad \forall r \in R \tag{18}$$

Constraint (17) represents the traffic from CNVMs to the BBUVM in node h where α is a unitless quantity which represents the ratio of backhaul to fronthaul traffic. Note that this constraint allows a BBUVM to receive traffic from more than a single CNVM, which may occur for example in network slicing.

Constraint (18) represents the traffic to RRH nodes from all BBUVMs that are hosted in hosting nodes. This enables an RRH to receive traffic from more than a single BBUVM (network slicing).

3) The served RRH nodes and the location of BBUVM

$$\beta \cdot \lambda^R_{h,r} \geq \sigma^B_{h,r} \quad \forall r \in R, \forall h \in H \tag{19}$$

$$\lambda^R_{h,r} \leq \beta \cdot \sigma^B_{h,r} \quad \forall r \in R, \forall h \in H \tag{20}$$

$$\beta \cdot \sum_{\forall r\in R} \lambda^R_{h,r} \geq \sigma^B_h \quad \forall h \in H \tag{21}$$

$$\sum_{\forall r\in R} \lambda^R_{h,r} \leq \beta \cdot \sigma^B_h \quad \forall h \in H \tag{22}$$

Constraint (19) and (20) ensure that the RRH node r is served by the BBUVM that is hosted at node h as illustrated in Fig. 7. Constraints (21) and (22) determine the location of BBUVM; β is equal to 10^6 which is large enough to ensure that $\sigma^B_{h,r}$ and σ^B_h are equal to 1 when $\sum_{\forall r\in R} \lambda^R_{h,r} > 0$. They are equivalent to the logical operation $\sigma^B_{h,r} = (\sum_{\forall r\in R} \lambda^R_{h,r} > 0 ? 1 : 0)$ which could be expressed as:

$$\sigma^B_{h,r} = \begin{cases} 1, & if \sum_{\forall r\in R} \lambda^R_{h,r} > 0 \\ 0, & otherwise \end{cases} \tag{23}$$

TABLE 4. BBUVM constraints operation.

Input	Constraint	Outcome	σ^B_h	Value of σ^B_h that satisfies both constraints
$\sum_{\forall r\in R} \lambda^R_{h,r} > 0$	$\beta \cdot \sum_{\forall r\in R} \lambda^R_{h,r} \geq \sigma^B_h$	$\beta \cdot \sum_{\forall r\in R} \lambda^R_{h,r} \gg 1$	0 or 1	1
	$\sum_{\forall r\in R} \lambda^R_{h,r} \leq \beta \cdot \sigma^B_h$	$\beta \cdot \sigma^B_h \gg 1$	1	
$\sum_{\forall r\in R} \lambda^R_{h,r} = 0$	$\beta \cdot \sum_{\forall r\in R} \lambda^R_{h,r} \geq \sigma^B_h$	$\beta \cdot \sum_{\forall r\in R} \lambda^R_{h,r} = 0$	0	0
	$\sum_{\forall r\in R} \lambda^R_{h,r} \leq \beta \cdot \sigma^B_h$	$\beta \cdot \sigma^B_h = 0$	0 or 1	

In constraint (21) there are two possibilities for the value of $(\sum_{\forall r\in R} \lambda^R_{h,r})$ which are either zero (no traffic from h to r) or greater than zero (there is a traffic from h to r). When the value of $\sum_{\forall r\in R} \lambda^R_{h,r}$ is zero, the left-hand side of the inequality $(\beta \cdot \sum_{\forall r\in R} \lambda^R_{h,r})$ should be zero and this sets the value of σB_h to zero. In the second case when the value of $\sum_{\forall r\in R} \lambda^R_{h,r}$ is greater than zero, the left-hand side of the inequality $(\beta \cdot \sum_{\forall r\in R} \lambda^R_{h,r})$ will be much greater than 1 because of the large value β. In this, the value of σ^B_h may be set to 1 or zero. In the same way constraint (22) sets the value of σ^B_h. Table 4 illustrates the operation of constraints (21) and (22).

4) CNVM locations

$$\beta \cdot \lambda^B_{p,h} \geq \sigma^E_{p,h} \quad \forall p, q \in H, p \neq q \tag{24}$$

$$\lambda^B_{p,h} \leq \beta \cdot \sigma^E_{p,h} \quad \forall p, q \in H, p \neq q \tag{25}$$

$$\sigma^E_p \geq \eta \cdot \sum_{h\in H} \lambda^B_{p,h} \quad \forall p \in H \tag{26}$$

$$\sigma^E_p \leq 1 + \sum_{h\in H} \lambda^B_{p,h} - \eta \quad \forall p \in H \tag{27}$$

$$\psi_{p,q} \leq \sigma^E_p \quad \forall p, q \in H, p \neq q \tag{28}$$

$$\psi_{p,q} \leq \sigma^E_p \quad \forall p, q \in H, p \neq q \tag{29}$$

$$\psi_{p,q} \geq \sigma^E_p + \sigma^E_q - 1 \quad \forall p, q \in H, p \neq q \tag{30}$$

5 Hosting any VM of any type

$$\sigma^\chi_h \leq \sigma^B_h + \sigma^E_h \quad \forall h \in H \tag{31}$$

$$\sigma^\chi_h \geq \sigma^B_h \quad \forall h \in H \tag{32}$$

$$\sigma^\chi_h \geq \sigma^E_h \forall h \in H \tag{33}$$

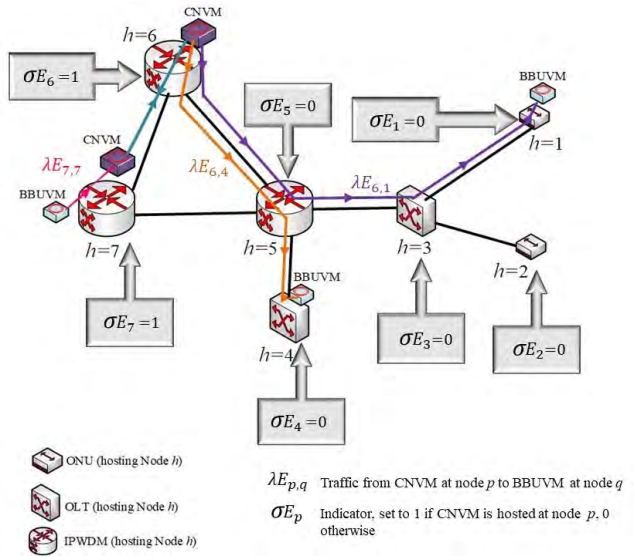

FIGURE 8. CNVM and the traffic toward BBUVMs nodes.

Constraints (24) and (25) ensure that the BBUVMs at node h are served by the CNVMs that are hosted at the node p. Constraints (26) and (27) determine the location of CNVMs by setting the binary variable σ_p^E to 1 if there is a CNVM hosted at node p, where η is equal to 10^{-9} which is small enough to ensure that the value of σ_p^E is equal to 1 when $\sum_{h\in H}\lambda^B_{p,h} > 0$ and ensure $\sigma_p^E = 0$ when $\sum_{h\in H}\lambda^B_{p,h} = 0$. They are equivalent to the logical operation $\sigma_p^E = (\sum_{h\in H}\lambda^B_{p,h} > 0?1:0)$ explained previously in constraint (23). Fig. 8 illustrates the functions of constraints (26) and (27) whilst Table 5 illustrates their operation. Constraints (28)–(30) ensure that the CNVMs communicate with each other if they are hosted at different nodes p and q, and this is equivalent to the logical operation $\psi_{p,q} = \sigma_p^E \; AND \; \sigma_q^E$. Fig. 9 illustrates the function of constraints (28)–(30). Constraints (31) - (38) determine if the hosting node h hosts any VM of any type (BBUVM or CNVM). It is equivalent to the logical operation $\sigma_h^X = \sigma_p^E \; AND \; \sigma_q^E$

6) Communication traffic between CNVMs

$$\lambda^E_{p,q} = \nabla_{p,q}\cdot\psi_{p,q} \quad \forall p,q\in H : p\neq q \tag{34}$$

7) Total traffic between two hosting nodes

$$\lambda^T_{p,q} = \lambda^E_{p,q} + \lambda^B_{p,q} \quad \forall p,q\in H : p\neq q \tag{35}$$

8) Flow conservation of the total traffic to the RRH nodes

$$\sum_{y\in T_x^N}\lambda^R_{h,r,x,y} - \sum_{y\in T_x^N}\lambda^R_{h,r,y,x} = \begin{cases}\lambda^R_{h,r} & if\ x = h\\ -\lambda^R_{h,r} & if\ x = r\\ 0 & otherwise\end{cases} \quad \forall r\in R, \forall h\in H, \forall x\in T \tag{36}$$

TABLE 5. CNVM constraints operation.

Input	Constraint	Outcome	σ_p^E	Value of σ_p^E that satisfies both constraints
$\sum_{h\in H}\lambda^B_{p,h} > 0$	$\sigma_p^E \geq \eta\cdot\sum_{h\in H}\lambda^B_{p,h}$	$\eta\cdot\sum_{h\in H}\lambda^B_{p,h} \ll 1$	1	1
	$\sigma_p^E \leq 1+\sum_{h\in H}\lambda^B_{p,h}-\eta$	$1+\sum_{h\in H}\lambda^B_{p,h}-\eta > 1$	0 or 1	
$\sum_{h\in H}\lambda^B_{p,h} = 0$	$\sigma_p^E \geq \eta\cdot\sum_{h\in H}\lambda^B_{p,h}$	$\eta\cdot\sum_{h\in H}\lambda^B_{p,h} = 0$	0 or 1	0
	$\sigma_p^E \leq 1+\sum_{h\in H}\lambda^B_{p,h}-\eta$	$1+\sum_{h\in H}\lambda^B_{p,h}-\eta < 1$	0	

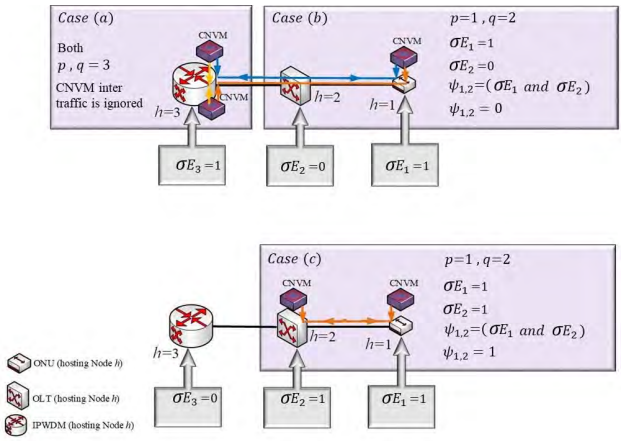

FIGURE 9. CNVM and the common locus.

9) Flow conservation of hosting nodes communication traffic

$$\sum_{y\in T_x^N\cap H}\lambda^T_{p,q,x,y} - \sum_{y\in T_x^N\cap H}\lambda^T_{p,q,y,x} = \begin{cases}\lambda^T_{p,q} & if\ x = p\\ -\lambda^T_{p,q} & if\ x = q\\ 0 & otherwise\end{cases} \quad \forall p,q,x\in H : p\neq q \tag{37}$$

Constraint (34) represents the traffic between CNVMs at hosting nodes p and q. Constraint (35) represents the total traffic between any two hosting nodes (pq) which is caused by virtual machines communication. Constraint (36) represents the flow conservation of the total fronthaul traffic to the RRH

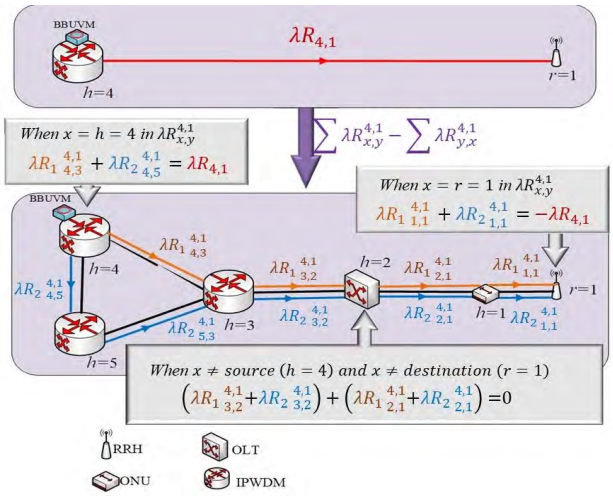

FIGURE 10. Flow conservation principle.

nodes. Fig. 10 illustrates the principle of flow conservation, and for clarification purposes, it is applied to constraint (36). Constraint (37) represents the flow conservation of the total traffic between any two hosting nodes that might host virtual machines of any type (BBUVM or CNVM).

10) Total BBU workload at any hosting node h

$$\Psi_h^B = \left(\left(\sum_{\forall r \in R} \lambda_{h,r}^R \right) \Big/ cp \cdot \right) \Psi^\chi \quad \forall h \in H \tag{38}$$

11) Total normalized workload at hosting node h

$$\Psi_h^i + \Psi_h^f = \left(\Psi_h^B + \Psi_h^C \right) \Big/ \Psi^S \quad \forall h \in H \tag{39}$$

12) Hosting node capacity

$$\left(\Omega_d^S \cdot \left(\Psi_h^i + \sigma_h^\chi \right) + \Psi_h^f \cdot \left(\Omega^S - \Omega_d^S \right) \right) \leq \Omega_h^H \quad \forall h \in H \tag{40}$$

13) GPON link constraints

$$\sum_{h \in H} \sum_{r \in R} \sum_{j \in T_i^N \cap L} \lambda_{h,r,i,j}^R \leq 0 \quad \forall i \in U \tag{41}$$

$$\sum_{p \in H} \sum_{q \in H, q \neq p} \sum_{j \in T_i^N \cap L} \lambda_{p,q,i,j}^T \leq 0 \quad \forall i \in U \tag{42}$$

$$\sum_{h \in H} \sum_{r \in R} \sum_{j \in T_i^N \cap N} \lambda_{h,r,i,j}^R \leq 0 \quad \forall i \in L \tag{43}$$

$$\sum_{p \in H} \sum_{q \in H, q \neq p} \sum_{j \in T_i^N \cap N} \lambda_{p,q,i,j}^T \leq 0 \quad \forall i \in L \tag{44}$$

Constraint (38) represents the total BBU workload at any hosing node h. Constraint (39) calculates the total BBU and CNVM normalized workload at any hosting node. The workload is scaled and normalized relative to the server CPU workload and is separated into integer and fractional parts. Constraint (40) ensures that the total power consumption of hosting VMs does not exceed the maximum power consumption allocated for each host. Constraints (41) – (44) ensure that the total PON downlink traffic does not flow in the opposite direction.

14) Virtual Link capacity of the IP over WDM network

$$\sum_{p \in H} \sum_{q \in H, q \neq p} \lambda T_{i,j}^{p,q} + \sum_{h \in H} \sum_{r \in R} \lambda R_{i,j}^{h,r} \leq W_{i,j} \cdot B \; \forall i, j \in N, i \neq j. \tag{45}$$

15) Flow conservation in the optical layer of IP over WDM network

$$\sum_{n \in N_m^N} W_{i,j,m,n} - \sum_{n \in N_m^N} W_{i,j,n,m} = \begin{cases} W_{i,j} & if \; n = i \\ -W_{i,j} & if \; n = j \\ 0 & otherwise \end{cases} \quad \forall i, j, m \in N, i \neq j \tag{46}$$

Constraint (45) ensures that the total traffic traversing the virtual link (i, j) does not exceed its capacity, in addition it determines the number of wavelength channels that carry the traffic burden of that link. Constraint (46) represents the flow conservation in the optical layer of the IP over WDM network. It ensures that the total expected number of incoming wavelengths for the IP over WDM nodes of the virtual link (i, j) is equal to the total number of outgoing wavelengths of that link.

16) Number of wavelength channels

$$\sum_{i \in N} \sum_{j \in N: i \neq j} W_{i,j,m,n} \leq w \cdot f_{m,n} \quad \forall m \in N, \; \forall n \in N_m^N \tag{47}$$

17) Total number of wavelength channels

$$W_{m,n} = \sum_{i \in N} \sum_{j \in N: i \neq j} W_{i,j,m,n} \quad \forall m \in N, \; \forall n \in N_m^N \tag{48}$$

18) Number of aggregation ports

$$\Lambda_i = \left(\sum_{j \in L \cap TN_i} \left(\sum_{p \in H} \sum_{q \in H, q \neq p} \lambda_{p,q,i,j}^T + \sum_{h \in H} \sum_{r \in R} \lambda_{h,r,i,j}^R \right) \right) \Big/ B \quad \forall i \in N \tag{49}$$

Constraints (47) and (48) are the constraints of the physical link (mn). Constraint (47) ensures that the total number of wavelength channels in the logical link (ij) that traverse the physical link (mn) does not exceed the fiber capacity. Constraint (48) determines the number of wavelength channels in the physical link and ensures that it is equals to the total number of wavelength channels in the virtual link traversing that physical link. Constraint (49) determines the required number of aggregation ports in each IP over WDM router.

V. MILP MODEL SETUP AND RESULTS

All MILP models were run on high performance computing nodes (HPC) provided through a partnership of the most research-intensive universities in the North of England. The used HPC has four computation modes, the standard mode provides clusters of 252 nodes with up to 6048 cores in total and supports 65 TFLOPS peak using IBM's iDataPlex hardware; and includes a high throughput cluster with 1 GB RAM per nodes, 2900 cores based on twin nodes with 6-core Westmere processors supported by 1GE connectivity between nodes. IBM ILOG CPLEX (12.7) optimization studio is used as an optimization software package where it uses simplex algorithm [47] to solve the developed MILP models.

Five IP over WDM nodes are considered constituting the optical backbone network of the proposed architecture. The distribution and topology of the IP over WDM nodes have been built upon the NSFNET network described in [48]–[53]. Each IP over WDM node in turn is attached to two GPONs with one OLT and two ONUs for each GPON. Accordingly, the network topology has 10 OLTs and 20 ONUs. In addition, each ONU is connected to one RRH node as shown in Fig. 11. Two GPONs for each IP over WDM node are enough to investigate the VM response for demands and power savings. To finalize the portrait of the network topology, we have concentrated on the distribution of the hosting nodes and the way in which they are connected to each other and for this reason the GPON splitters are not shown.

As alluded to earlier, two types of VMs have been considered: BBUVM, which realizes the functions of the BBU, and CNVM to achieve the functions of the mobile core network. The amount of workload needed for BBUVMs is calculated in GOPS according to (11)[46] and based on the calculated workload, the hosting server CPU utilization due to hosting BBUVMs is determined. On the other hand, the total workload needed for CNVMs is calculated based on the number BBUVMs group in each hosting node since we have allocated one CNVM for each group of BBUVMs in one hosting node. A single VM consumes around 18W [54] and by knowing the hosting server maximum power consumption 365 (W), idle power 112 (W) and the maximum workload 368 (GOPS), Ψ_h^C can calculated for a single VM. Therefore $\Psi_h^C =$ corresponds is $(18 \times 368)/(365 - 112) = 26$ (GOPS).

We have investigated the effect of the inter-traffic between CNVMs which is needed to maintain the communication from one side of the network to another such as a call that is held between two mobiles belonging to two different CNVMs. The distribution of CNVMs is sensitive to the inter-traffic flows between them. However, we chose small values for CNVMs inter-traffic in the investigated range. The first value in the range is zero which represents the case where no inter-traffic flows between CNVMs. The maximum value in the range is 16% of the total backhaul traffic which the minimum value at which the MILP model tends to host (pack) all CNVMs at the same node to eliminate the impact of inter-traffic.

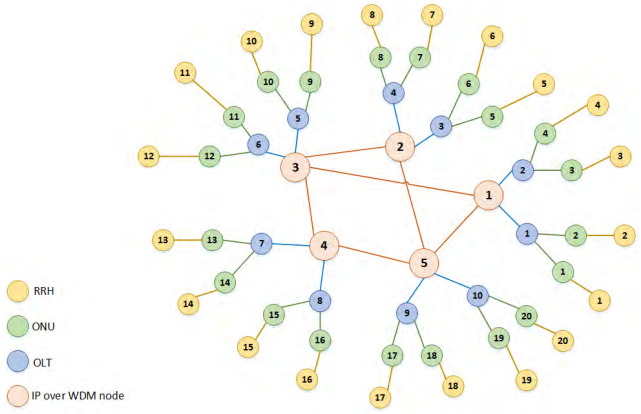

FIGURE 11. Tested network topology.

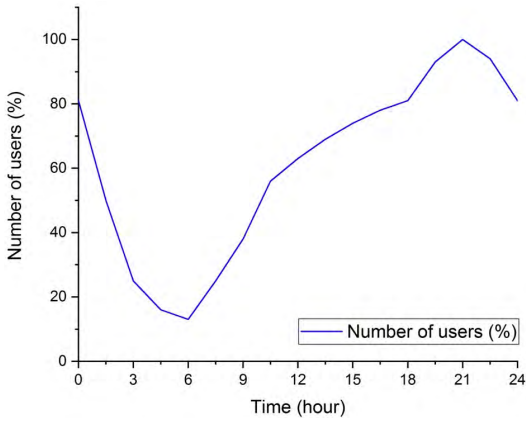

FIGURE 12. Average number of users daily profile [1].

Moving towards the access network, each RRH node is considered to serve a small cell that operates on 10 MHz bandwidth and with a maximum capacity of 10 users. Each user in the small cell is allocated 5 physical resources blocks (PRB) as the users are assumed to request the same task from the network. Accordingly, the total downlink traffic to the RRH node depends on the total number of active users in the small cell. The input parameters to the developed MILP model are listed in Table 6. We have considered 17 time slots over all the day from 00:00 hours to 24:00 hours in steps of 1.5 hours using the average number of users daily profile shown in Fig. 12. The MILP results are compared with the case where there is no NFV deployment. In the "no virtualization" scenario, the BBU is located close to the RRH where they are attached to each other, whilst the integrated platform ASR5000 is deployed to realize mobile core network functionalities and it is connected directly to the IP over WDM network. The ASR5000 maximum power consumption, idle power, and maximum capacity are 5760 (W), 800 (W), and 320 (Gbps) respectively [55], whilst the BBU maximum power consumption, idle power, and maximum capacity are 531 (W), 51 (W), 9.8 (Gbps) respectively [56].

The results in Fig. 13 show the total power consumption of the of the case where no virtualization is deployed

TABLE 6. MILP model input parameters.

Parameters	Comments
Line coding rate for 8B/10B line coding (l)	10/8 (bit / sample)
Number of MIMO layers (y)	2
Number of bits used in QAM modulation for 64 QAM modulation (q)	6 (bits)
Number of antennas in a cell (a)	2
Maximum fronthaul (CPRI) data rate for CPRI line rate option 7 (cp)	9.8304 (Gbps) [43]
Maximum baseband processing workload needed for fully loaded RRH (Ψ^X) given by: $30 \cdot a + 10 \cdot a^2 + 20 \cdot q \cdot l \cdot y$	400 (GOPS)
Server CPU maximum workload (Ψ^S)	368 (GOPS) [57]
Workload needed for hosting one CNVM (Ψ_h^C)	26.17 (GOPS)
Number of active users in a small cell (ρ_r)	Uniformly distributed (1-10 users)
Maximum number of users per cell (n)	10 (users)
Number of physical resource blocks per user (pb)	5 (PRB)
The ratio of the backhaul to the front haul traffic (α)	0.1344 (unitless)
ONU maximum power consumption (Ω^U)	15 (W) [58]
OLT maximum power consumption (Ω^L)	1940 (W) [59]
OLT idle power (Ω^{Ld})	60 (W) [59]
OLT maximum capacity (C^L)	8600 (Gbps) [59]
ONU maximum capacity (C^U)	10 (Gbps) [58]
RRH node power consumption (Ω_x^R)	1140 (W) [60]
Hosting server maximum power consumption (Ω^S)	365 (W) [61]
Hosting server idle power consumption (Ω^{Sd})	112 (W) [61]
Capacity IP over WDM wavelength channel (B)	40 (Gbps) [62-64]
Number of wavelengths per fiber in IP over WDM (w)	32 [62]
Transponder power consumption (Ω^T)	167 (W) [65]
Router port power consumption (Ω^{RP})	825 (W) [66]
Regenerator power consumption (Ω^G)	334 (W) [66]
EDFA power consumption (Ω^E)	55 (W) [66]
Maximum span distance between EDFAs (S)	80 (km) [62, 63]

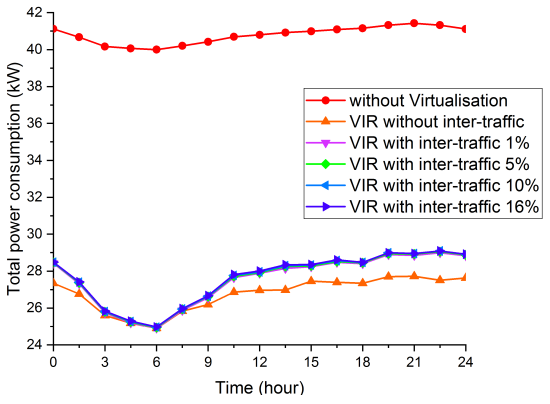

FIGURE 13. Total power consumption without and with virtualization under different CNVMs inter-traffic at different time slots of a day.

(standard model) as well as the cases where the virtualization is deployed under different CNVMs inter-traffic for different time slots in a day. Fig. 14 shows the total power consumption of the same scenarios versus the total number of active users

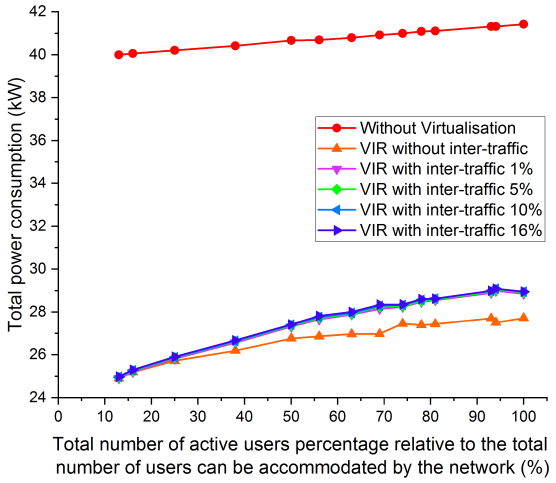

FIGURE 14. Total power consumption without and with virtualization under different CNVMs inter-traffic versus total active users in the network.

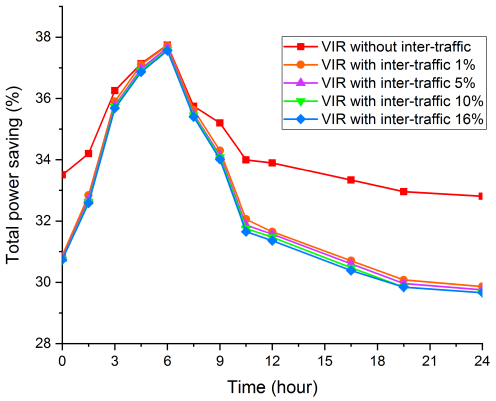

FIGURE 15. Power saving comparison of virtualization under different CNVMs inter-traffic for a day.

in the networks. The virtualization model has resulted in less power consumption compared to the no virtualization model (standard model) as it optimizes the processing locations of the downlink traffic through optimum placement and consolidation of VMs.

Fig. 15 compares the total power saving of the virtualization model under different CNVMs inter-traffic for one day while Fig. 16 show the virtualization power saving under different CNVMs inter-traffic versus total number of active users. Compared to other virtualization cases, virtualization without CNVMs inter-traffic has saved a maximum of 38% (average 34%). This is because there is no power consumed by the CNVMs inter-traffic as this traffic is zero. The total power saving decreases as the CNVMs inter-traffic increases to reach its lowest value in the case of virtualization with 16% CNVMs inter-traffic which is 37% (average 32%).

Virtualization in the presence of CNVMs inter-traffic resulted in comparable values of total power consumption (and power saving) for all values of CNVMs inter-traffic greater than zero. The main reason behind this is that the

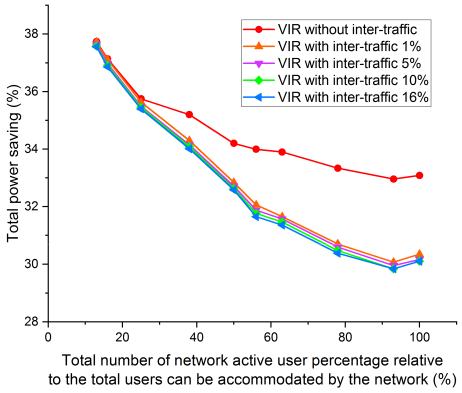

FIGURE 16. Power saving comparison of virtualization under different CNVMs inter-traffic versus total number of active users.

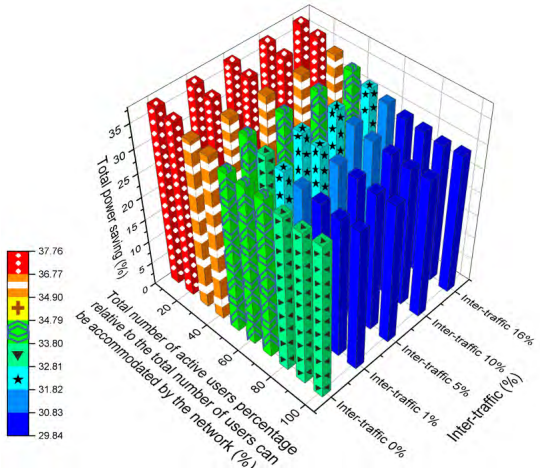

FIGURE 17. 3-Dimensional presentation of the total power saving for virtualization under different CNVMs inter-traffic versus different number of active users in the network.

CNVMs inter-traffic produces relatively small amount of power consumption compared to the power consumption induced by the fronthaul traffic and hosting server as shown in Fig. 17. As the inter-traffic increases, the MILP model tends to eliminate its effect by consolidating CNVMs in one place.

Although virtualization has saved a maximum of 38% (without CNVMs inter-traffic) and 37% (with 16% CNVMs inter-traffic) of the total power consumption, it cannot provide such level of power saving over the entire day. As the number of active users varies with the time of day (as in Fig. 12), the power saving achieved by virtualization varies accordingly. The results in Figs. 15 and 16 show that a high-power saving is achieved when the total number of active users is around 20% (around 4 am to 8 am) while the lowest power saving is recorded at high number of active users (during the day rush hours). At small number of active users, the MILP model tends to consolidate all the VMs in the IP over WDM network to minimize the number of servers hosting VMs to reduce the total power consumption.

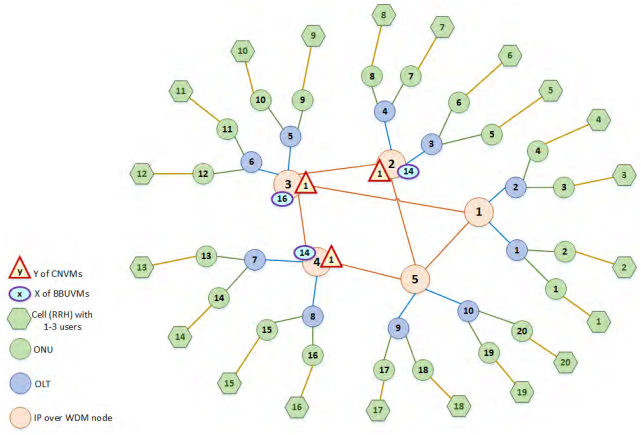

FIGURE 18. VMs distribution over network under active users 13% of the total network capacity without CNVMs inter-traffic.

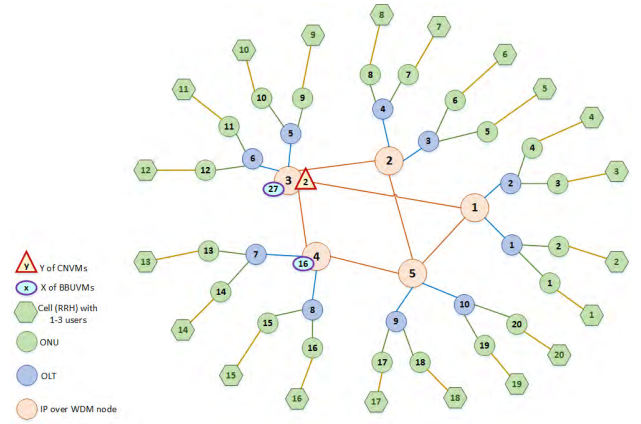

FIGURE 19. VMs distribution over network under active users 13% of the total network capacity and 16% CNVMs inter-traffic.

Figs. 18 and 19 show the VMs consolidation and distribution over the network at low number of active users (13%) under CNVMs of 0% and 16% respectively. At low number of active users and 0% inter-traffic, the MILP model consolidates the VMs at the IP over WDM network. Since the total number of active users is low, the fronthaul traffic is relatively low and consequently the power consumption induced by the fronthaul traffic is low compared to the hosting power consumption (servers power). For this reason, the MILP model tends to pack BBUVMs in the IP over WDM network as much as possible to reduce the power consumed by the hosting servers. Also, the MILP model tends to host CNVMs close to the BBUVMs as the inter-traffic between CNVMs is zero. Once the inter-traffic is greater than zero, the MILP model consolidates the CNVMs at one location as in Fig. 19.

Fig. 20 and Fig. 21 show the VMs consolidation and distribution over the network with high number of active users (around 100%) under 0% and 16% CNVMs inter-traffic. When the number of active users is high, the amount of fronthaul traffic is high, for that reason the MILP model tends to distribute the BBUVMs at the closest centralized location

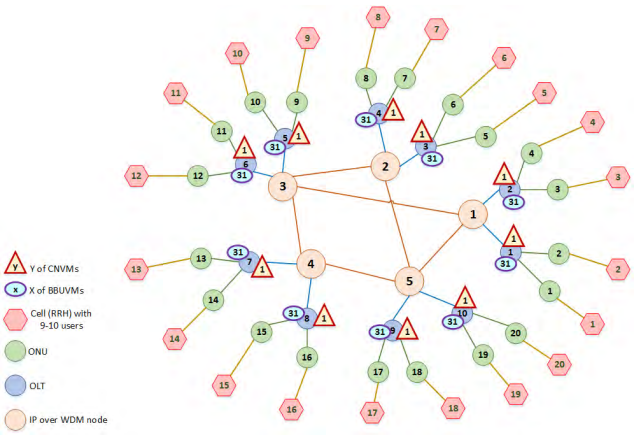

FIGURE 20. VMs distribution over network under active users 100% of the total network capacity without CNVMs inter-traffic.

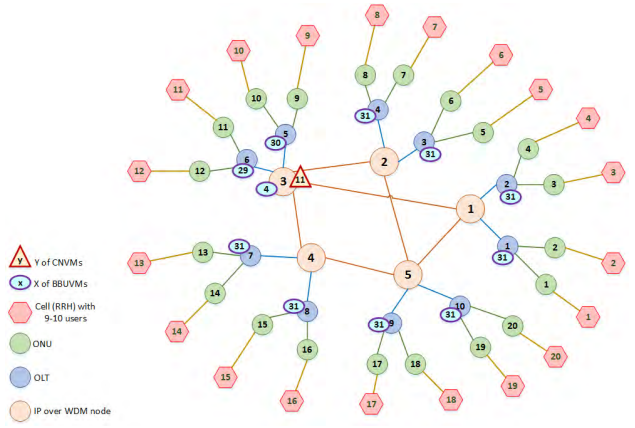

FIGURE 21. VMs distribution over network under active users 100% of the total network capacity and 16% CNVMs inter-traffic.

to the users which are the OLTs, while CNVMs inter-traffic has no effect on the distribution of BBUVMs.

Hosting BBUVMs in OLTs when the number of users is high ensures shorter paths for the fronthaul traffic than hosting BBUVMs in the IP over WDM networks and consequently, the power consumed by this traffic is less. For CNVMs, the MILP model tends to distribute them close to the BBUVM when there is no inter-traffic between them, and this is clearly seen in Fig. 20. In contrast, when the inter-traffic between CNVMs is greater than zero, the MILP model tends to centralize the location of CNVMS in the IP over WDM network to reduce the power consumption induced by the inter-traffic and the power of the hosting servers as shown in Fig. 21.

VI. REAL-TIME HEURISTICS IMPLEMENTATION AND RESULTS

A. ENERGY EFFICIENT NFV WITH NO CNVMs INTER-TRAFFIC (EENFVnoITr) HEURISTIC

The EENFVnoITr provides real-time implementation of the MILP model without CNVMs inter-traffic. The pseudocode

Algorithm 1 Energy Efficient NFV without CNVMs inter-traffic (EENFVnoITr) Heuristic

INPUT: $G = (NE, L)$, $G_p = (N, L_p)$
OUTPUT: VMs location, workloads, and distribution

1: $\forall r \in RRH$ determine number of users and calculate node demand (rD_r); where $D_r \in D$ /$*D$ is the total demands$*$/
2: $\forall D_r \in D$ determine BBUVM workload Ψ_r
3: $\forall r \in RRH$ find $(r, h) = min(shortestPath\,(r, \{h \in NE \cap OLT\})$
4: **if** total workload of $h >> \Psi_r$
5: host BBUVM in h
6: update workload of h
7: $D_r \in D_{served}$
8: **end if**
9: $\forall D_r \notin D_{served}$ find $(r, h) = \min\,(shortestPath\,(r, \{h \in N\}))$ /$*$ where N is the IP over WDM nodes $*$/
10: host BBUVM in h
11: update workload of h
12: $D_r \in D_{served}$
13: Route the fronthaul traffic from BBUVMs to RRH nodes
14: $N^{'} \leftarrow DESCEND_SORT(N)$ and set $i = 1$
15: Host CNVM in $N^{'}(i)$, $N^{'}(i-1)$,...$N^{'}(1)$
16: $\forall CNVM\ in\ n \in N^{'}\ and\ \forall BBUVM\ in\ h \in NE$ find $(n, h) = \min\,(shortest\ Path\,(n, h))$
17: Route the traffic from CNVMs to BBUVMs
18: Determine the IP over WDM network configuration
19: **if** $i = 1$
20: Determine total power consumption as (minTPC)
21: **end if**
22: Determine the total power consumption as TPC(i)
23: **if** minTPC$\geq$ TPC(i)
24: minTPC = TPC(i)
25: $i = i + 1$
26: goto 15
27: **else**
28: minimum power consumption is minTPC
29: EXIT
30: **end if**

of the heuristic is shown in Algorithm 1. The network is modelled by sets of network elements NE, and links L. The heuristic obtains the network topology $G = (NE, L)$ and the physical topology of the IP over WDM network $G_p = (N, L_p)$ where N is the set of IP over WDM nodes and L_p is the set of physical links. The total download request (fronthaul traffic) of each RRH node is calculated based on the total number of active users in each cell (RRH). The heuristic determines the amount of baseband workload needed to process each RRH download request. According to the baseband workload for each requested download traffic and the available capacity of the hosting VM server, the EENFVnoITr heuristic chooses the closest place to accommodate BBUVM in such

a way that it serves as many RRH requests as possible. The EENFVnoITr heuristic may host a BBUVM in an OLT node if it has enough processing capacity to serve all the requests from the closest RRH nodes. In this way, the heuristic exploits bin packing techniques to reduce the processing power consumption. The amount of fronthaul traffic delivered by each BBUVM determines the backhaul traffic flows from each CNVMs toward BBUVMs. The EENFVnoITr heuristic determines the total amount of backhaul traffic that may flow from each IP over WDM node and sorts them in a descending order. The nodes in the top of the sorted list of IP over WDM nodes represent highly recommended nodes to host CNVMs. In such a scenario, the EENFVnoITr heuristic ensures less of the backhaul traffic flows in the IP over WDM network. The EENFVnoITr heuristic uses the sorted list to accommodate CNVMs. Once the VMs are distributed and the logical traffic is routed, the EENFVnoITr heuristic obtains the physical graph $G_p = (N, L_p)$ and determines the traffic in each network segment. The IP over WDM network configuration such as the number of fibers, router ports, and the number of EDFA is determined the total power consumption is evaluated. The heuristic algorithm continuously increases the number of CNVMs candidate locations by one, re-configures the IP over WDM network, and re-evaluates the power consumption and compare it to the power consumption in the previous iteration until it determines the best number and location of CNVMs for minimum power consumption. The heuristic model

B. ENERGY EFFICIENT NFV WITH CNVMs INTER-TRAFFIC (EENFVwithITr) HEURISTIC

This section describes the energy efficient NFV with CNVMs inter-traffic heuristic (EENFVwithITr). The EENFVwithITr heuristic extends the EENFVnoITr heuristic to provide real-time implementation of the MILP model where the CNVMs are considered. The pseudocode of the heuristic is shown in algorithm 2. It uses the same approach used by EENFVnoITr, but it evaluates the CNVMs inter-traffic after the locations of CNVMs are determined.

Algorithm 2 Energy Efficient NFV With CNVMs Inter-Traffic (EENFVwithITr) Heuristic

INPUT: $G = (NE, L)$, $G_p = (N, L_p)$
OUTPUT: VMs location, workloads, and distribution

1: $\forall r \in RRH$ determine number of users and calculate node demand (rD_r); where $D_r \in D$ /$*D$ is the total demands$*$/
2: $\forall D_r \in D$ determine BBUVM workload Ψ_r
3: $\forall r \in RRH$ find $(r, h) = min(shortestPath\,(r, \{h \in NE \cap OLT\})$
4: **if** total workload of $h >> \Psi_r$
5: host BBUVM in h
6: update workload of h
7: $D_r \in D_{served}$
8: **end if**
9: $\forall D_r \notin D_{served}$ find $(r, h) = \min\,(shortestPath\,(r, \{h \in N\}))$ /$*$ where N is the IP over WDM nodes $*$/
10: host BBUVM in h
11: update workload of h
12: $D_r \in D_{served}$
13: Route the fronthaul traffic from BBUVMs to RRH nodes
14: $N^{'} \leftarrow DESCEND_SORT(N)$ and set $i = 1$
15: Host CNVM in $N^{'}(i)$,$N^{'}(i-1)$,$\ldots N^{'}(1)$
16: $\forall$CNVM$inn \in N^{'}$ and $\forall$BBUVM$inh \in NE$ find $(n, h) = \min\,(shortestPath\,(n, h))$
17: Route the traffic from CNVMs to BBUVMs
18: $\forall$CNVM$inn_x n_y \in N^{'}; x \neq y$ find $(n_x, n_y) = \min\left(shortestPath\left(n_x, n_y\right)\right)$
19: Route the traffic from CNVMs to CNVMs
20: Determine the IP over WDM network configuration
21: **if** $i = 1$
22: Determine total power consumption as (minTPC)
23: **end if**
24: Determine the total power consumption as TPC(i)
25: **if** minTPC$\geq$ TPC(i)
26: minTPC = TPC(i)
27: $i = i + 1$
28: goto 15
29: **else**
30: minimum power consumption is minTPC
31: EXIT
32: **end if**

C. EENFVnoITr AND EENFVwithITr HEURISTICS RESULTS

In order to verify the results of the proposed MILP model, the network topology in Fig. 11 used for the MILP model is also used to evaluate the heuristics. All the parameters considered in the MILP model such as the wireless bandwidth, number of resources blocks per user, and the parameters in Table 6 are considered in the evaluation of both EENFVnoITr and EENFVwithITr heuristics. The number of users allocated to each cell in the heuristics is the same as in the MILP model to ensure the requested traffic by each RRH node is the same in all models. Fig. 22 compares the total power consumption of MILP with EENFVnoITr model at different times of the day when the CNVMs inter-traffic is not considered. It is clearly seen that there is a small difference in the total power consumption of the two models and it varies over the day according to the total number of active users. The total power consumption of the MILP model is less than the EENFVnoITr heuristic with a maximum of 9% (average 5%) drop in the total power consumption. This is mainly caused by the distribution of CNVMs in the EENFVnoITr heuristic. As there is no traffic flowing between CNVMs, the EENFVnoITr accommodates them close to the BBUVMs wherever the VM servers have enough capacity. To accommodate the CNVMs, the heuristic sequentially examines the capacity of the VM servers in the OLT nodes that are close to the BBUVMs before investigating other servers in the IP over WDM network. As the distance and capacity requirements of

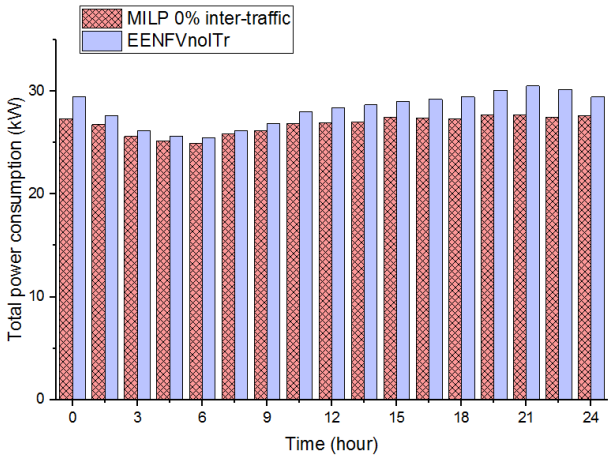

FIGURE 22. Total power consumption of MILP with without CNVMS inter-traffic compared with EENFVnoITr heuristic model.

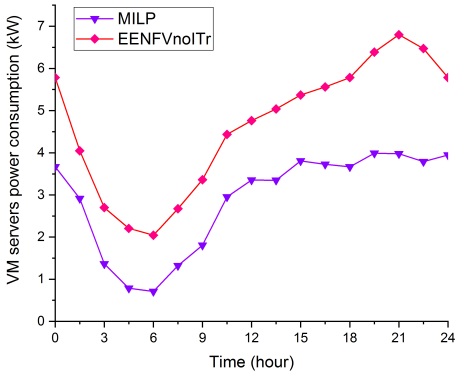

FIGURE 23. VM servers power consumption of MILP model compared with EENFVnoITr heuristic.

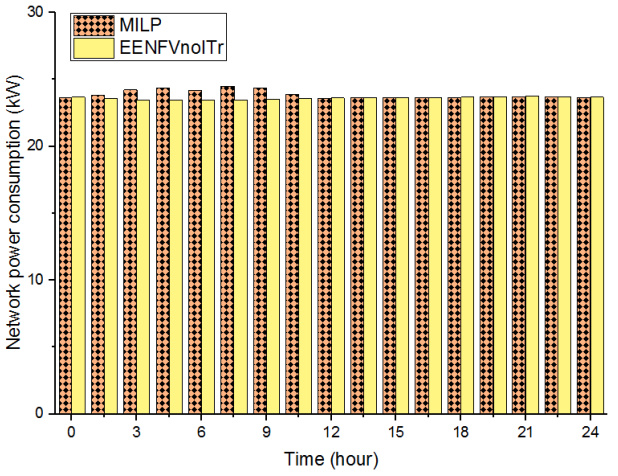

FIGURE 24. Network power consumption of MILP model compared with EENFVnoITr heuristic.

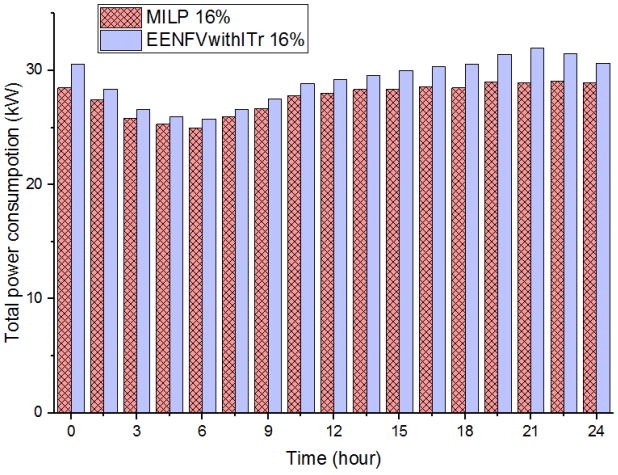

FIGURE 25. Total power consumption of MILP model compared with EENFVwithITr heuristic at CNVMs inter-traffic 16% of the total backhaul traffic.

the VM servers are met, the heuristic accommodates a CNVM in the server. This case results in high EENFVnoITr VM server power consumption compared with MILP model. This is clearly seen in Fig. 23 where the VM servers power consumption of the MILP model and the EENFVnoITr heuristic are compared. The total network power consumption of both EENFVnoITr heuristic and the MILP model are the same for most of the time of the day. Fig. 24 shows the network power consumption of the MILP model compared with EENFVnoITr heuristic.

It shows that there is a small difference in the network power consumption between the two models during the time of the day when the total number of active users is low. This is driven by the approach of the MILP model where it tends to accommodate the CNVMs at the IP over WDM nodes rather than OLT at the time of the day where the total number of users is low. In contrast, the heuristic tends to accommodate the CNVMs wherever the VM server is close to the BBUVMs and it has enough capacity. Fig. 25 compares the total power consumption of EENFVwithITr with the MILP model when the CNVMs inter-traffic is 16% of the total backhaul traffic. It is clearly seen that there is a small difference in the total power consumption of the two models and this varies over the day according to the total number of active users. The total power consumption of the MILP model is less than the EENFVnoITr model with a maximum drop of 9.5% (average 5%) in the total power consumption. This is mainly driven by the distribution of both CNVMs and BBUVM over the network nodes. The MILP model tends to accommodate BBUVMs and CNVMs at the IP over WDM network during times of the day when there is a small number of active users.

This causes more traffic from BBUVMs and CNVMs to flow in the IP over WDM network which eventually increases the IP over WDM network power consumption as shown in Fig. 26 which compares the IP over WDM network power consumption of both MILP model and EENFVwithITr heuristic when CNVMs inter-traffic is considered 16% of the total backhaul traffic. In contrast, the IP over WDM network power consumption of EENFVwithITr varies according to the total number of active users during the day. The sequential examination by EENFVwithITr of VM servers,

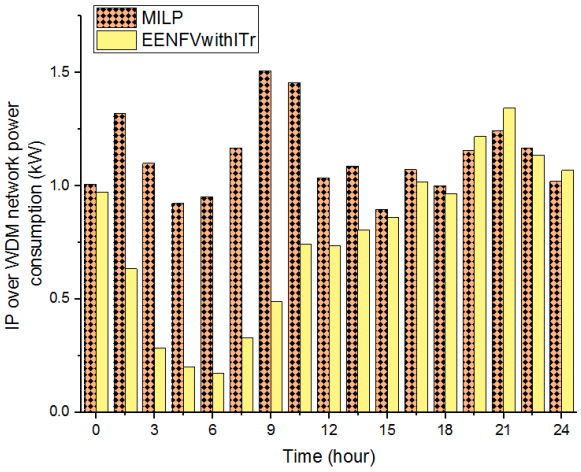

FIGURE 26. IP over WDM network power consumption of MILP model compared with EENFVwithITr heuristic at CNVMs inter-traffic 16% of the total backhaul traffic.

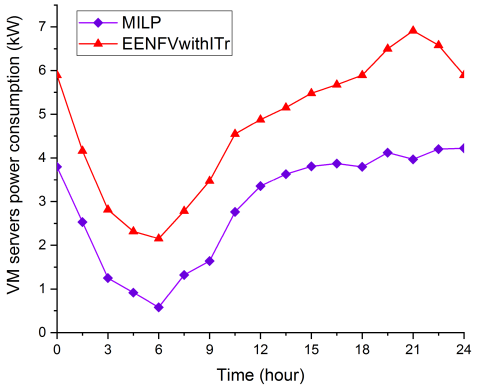

FIGURE 27. VM servers power consumption of MILP model compared with EENFVwithITr heuristic at CNVMs inter-traffic 16% of the total backhaul traffic.

their location, and available capacity increases the processing distribution of VMs in the network which leads to a high VM servers power consumption compared with the MILP model as shown in Fig. 27 which compares the VM servers power consumption of the MILP model with EENFVwithITr heuristic during different times of the day.

D. COMPLEXITY ANALYSIS

Both EENFVnoITr and EENFVwithITr heuristic algorithms were developed to tackle the non-deterministic-polynomial time hardness (NP-hard) problem of the linear programming model proven in [67]. Two main processes form the core of both heuristics: hosting VMs in the network nodes and routing the traffic between VMs; and between VMs and RRH nodes. Hosting VMs in network nodes is done in two stages: hosting BBUVMs and hosting CNVMs. To host BBUVMs in the network nodes, two nested loops are needed with (R × N) iterations where R is the number of RRH nodes and N is the number of hosting nodes, while (N × N =N^2) iterations are needed to host the CNVMs. As the number of users increases,

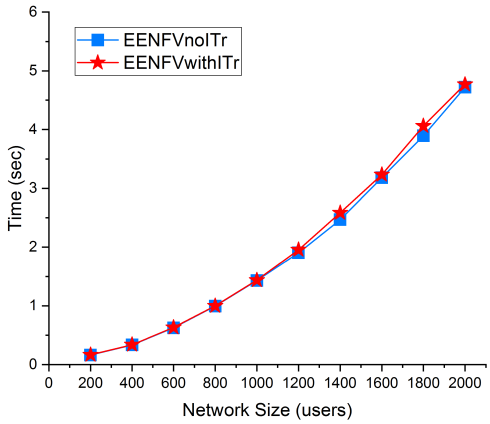

FIGURE 28. VM servers power consumption of MILP model compared with EENFVwithITr heuristic at CNVMs inter-traffic 16% of the total backhaul traffic.

the network size increases and the number of RRH nodes will be approximately equal to the number of hosting nodes N ($R = N$). However, hosting VMs in the network nodes has complexity of (Big O) O($k \times N^2$) time, where k is a constant equals to 2. On the other hand, routing the traffic between VMs and also between VMs and RRH nodes is based on the minimum hop algorithm which has complexity of order O(N) [68]. Thus, the overall time complexity is a polynomial time complexity expressed as $O(kN^2) + O(N)$. However, EENFVnoITr and EENFVwithITr algorithms time complexity is approximately $O(N^2)$ as it is the dominant term in the aforementioned complexity expression. EENFVnoITr and EENFVwithITr algorithms running time versus the network size is shown in Fig. 28 where it is evaluated up to 2000 users. The results in Fig. 28 were obtained when the EENFVnoITr and EENFVwithITr algorithms were executed in an Intel Core i5, 3.00 GHz processor with 16 GB RAM.

VII. GENETIC ALGORITHM IMPLEMENTATION

In this section, a genetic algorithm (GA) is introduced as an alternative approach to validate the results. The main difference between EENFVnoITr and EENFVwithITr algorithms is the added consideration of the traffic between CNVMs carried out by EENFVwithITr where this traffic is considered zero in EENFVnoITr. The genetic algorithm is developed here for the case where traffic flows between CNVMs which is the general case, thus avoiding redundant verification of the results.

The principle of GA is to let a certain population of several candidates or individuals to evolve through a number of generations in the search of the individual that has the optimum fitness. The optimum fitness in this work is the minimum.

The principle of GA is based on enabling a certain population of several candidates or individuals to evolve through a number of generations in the search for the individual that has the optimum fitness. The optimum fitness in this work is the minimum power consumption while the distribution of the VMs in the network represents the candidates or individuals.

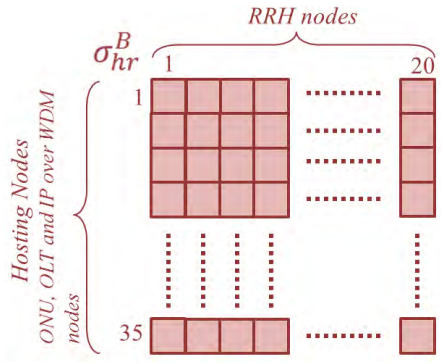

FIGURE 29. BBUVMs chromosome structure.

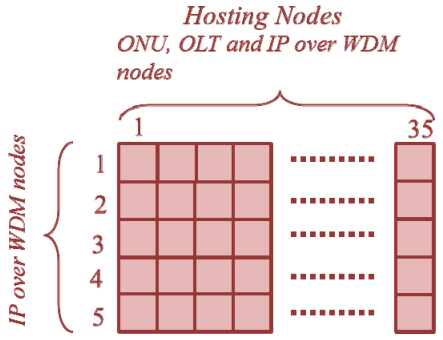

FIGURE 30. CNVMs chromosome structure.

The first step in the GA is to define chromosomes for the problem under study. Two types of chromosomes are defined in this work: BBUVMs chromosome, and CNVMs chromosome. The BBUVMs chromosome has 700 genes corresponding to the binary decision (σ^B_{hr}) that is introduced in the MILP model. Each gene represents a BBUVM hosted in a specific node to serve an RRH node. If the hosting node h1 has a BBUVM that serves a RRH node r1 then the corresponding gene is set to 1, otherwise it is set to zero. A two dimensional array representation of BBUVM chromosome is shown in Fig. 29.

The same methodology is applied to construct the CNVMs chromosome. The CNVMs chromosome has 175 genes where each gene represents a CNVM at an IP over WDM node that serves a BBUVM at a specific hosting node. If a node h2 has a CNVM that serves a BBUVM at node h5 then the corresponding chromosome is set to 1. Otherwise, it is set to zero. The two dimensional array representation of CNVMs chromosome is shown in Fig. 30.

The fitness function in the developed algorithm consists of evaluating the power consumption associated with traffic routing and VM processing for each individual. After that selection, crossover, and mutation are applied to create a new generation of individuals as shown in Algorithm 3.

A. GA SETUP AND RESULTS

We have considered two types of chromosomes: BBUVMs chromosome and CNVMs chromosome. The

Algorithm 3 Energy Efficient NFV With CNVMs Inter-Traffic (EENFVwithITr) Heuristic

INPUT: initial populations
OUTPUT: minimum power consumption

```
Set a counter for the number of generation i= 1
Get the initial population
Evaluation of the population fitness
Select parents
Crossover
Mutation
Elitism
Get a new population
if the termination criteria is not satisfied
    i = i + 1
    goto 3
else
      Optimal Power Consumption
      EXIT
end if
```

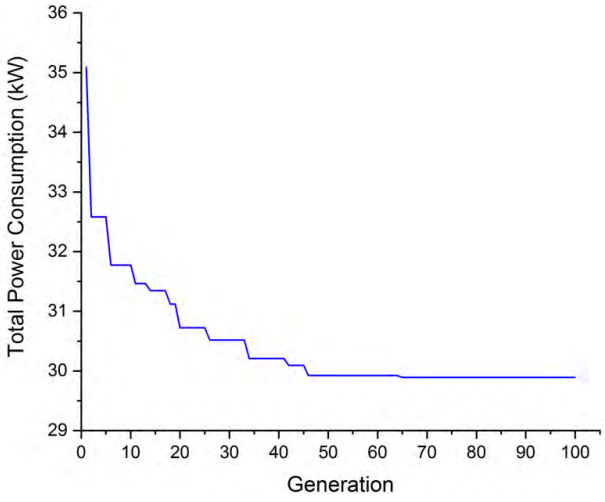

FIGURE 31. Fitness history.

BBUVMs chromosome has 700 genes corresponding to 35 candidates that can host BBUVM (ONU+OLT+IP over WDM nodes) multiplied by 20 RRH nodes as proposed in the topology used. CNVMs chromosome has 175 genes corresponding to 5 candidates IP over WDM nodes to host CNVMs multiplied by 35 candidates to host BBUVMs. We have considered 10 chromosomes for each type (BBUVMs and CNVMs) to be evolved throughout 100 generations with probabilities of crossover 95%, mutation 0.01% and elitism 20%. The fitness function (power consumption) was recorded at each generation and is illustrated in Fig. 31. The fitness history graph shows that the solution (minimum power consumption) is obtained beyond generation 65 where the graph shows no further change.

As alluded to earlier, the developed GA considers the case where there are traffic flows between CNVMs and the results are compared with EENFVwithITr and MILP model with

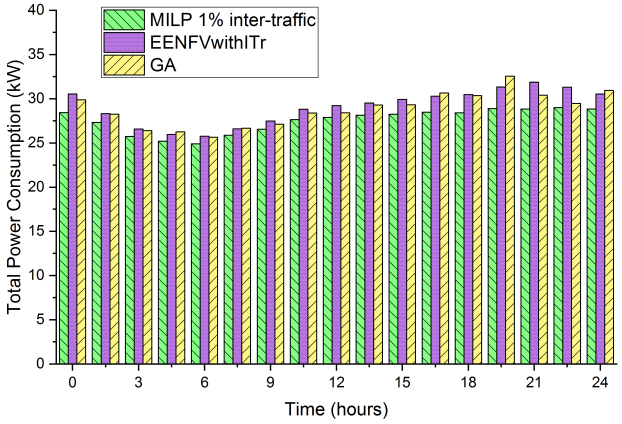

FIGURE 32. Total power consumption of GA compared with EENFVwithITr heuristic and MILP model at CNVMs inter-traffic 1% of the total backhaul traffic.

CNVMs inter-traffic 1% of the total backhaul traffic. The results are illustrated in Fig. 32 and they are comparable.

VIII. CONCLUSIONS

This paper has investigated network function virtualization in 5G mobile networks with the impact of total number of active users in the network, the backhaul/fronthaul configurations, and the inter-traffic between VMs. A MILP optimization model was developed with the objective of minimizing the total power consumption by optimizing the VMs locations and VM servers' utilization. The MILP model results have been investigated under the impact of CNVMs traffic variation, and variation in the total number of active users during different times of the day. The MILP model results show that virtualization can save up to 38% (average 34%) of the total power consumption, also the results reveal how the total number of active users affects the BBUVMs distribution while CNVMs distribution is affected mainly by the inter-traffic between them. For real-time implementation, this paper has introduced two heuristics: Energy Efficient NFV without CNVMs inter-traffic and Energy Efficient NFV with CNVMs inter-traffic. The results obtained through the use of the heuristics were compared with the MILP model results. The comparisons showed that the total power consumption when the heuristics are used is higher than the total power consumption when the MILP optimization model is used by a maximum of 9% (average 5%). Finally, a Genetic algorithm has been introduced to for further results verification.

ACKNOWLEGMENT

The authors would like to acknowledge them for funding this work. Mr. Ahmed Al-Quzweeni would like to acknowledge HCED for funding his Ph.D. All data are provided in full in the results section of this paper.

REFERENCES

[1] G. Auer *et al.*, "How much energy is needed to run a wireless network?" *IEEE Trans. Wireless Commun.*, vol. 18, no. 5, pp. 40–49, Oct. 2011.
[2] *Cisco Visual Networking Index: Forecast and Methodology 2016–2021*, Cisco, San Jose, CA, USA, Jun. 2017.
[3] I. Neokosmidis, T. Rokkas, P. Paglierani, C. Meani, K. M. Nasr, and K. Moessner, "Techno economic assessment of immersive video services in 5G converged optical/wireless networks," in *Proc. Opt. Fiber Commun. Conf. Expo. (OFC)*, Mar. 2018, pp. 1–3.
[4] V. G. Nguyen *et al.*, *A Comprehensive Guide to 5G Security*, 1st ed. Hoboken, NJ, USA: Wiley, 2018, pp. 31–57.
[5] Z. Zhang, Y. Gao, Y. Liu, and Z. Li, "Performance evaluation of shortened transmission time interval in LTE networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2018, pp. 1–5.
[6] L. Belkhir and A. Elmeligi, "Assessing ICT global emissions footprint: Trends to 2040 & recommendations," *J. Cleaner Prod.*, vol. 177, pp. 448–463, Mar. 2018.
[7] A. Z. Aktas, "Could energy hamper future developments in information and communication technologies (ICT) and knowledge engineering?" *Renew. Sustain. Energy Rev.*, vol. 82, pp. 2613–2617, Feb. 2018.
[8] P. Liu, G. Xu, K. Yang, K. Wang, and X. Meng, "Jointly optimized energy-minimal resource allocation in cache-enhanced mobile edge computing systems," *IEEE Access*, vol. 7, pp. 3336–3347, 2019.
[9] P. Lähdekorpi, M. Hronec, P. Jolma, and J. Moilanen, "Energy efficiency of 5G mobile networks with base station sleep modes," in *Proc. IEEE Conf. Standards Commun. Netw. (CSCN)*, Sep. 2017, pp. 163–168.
[10] X. Ge, J. Yang, H. Gharavi, and Y. Sun, "Energy efficiency challenges of 5G small cell networks," *IEEE Commun. Mag.*, vol. 55, no. 5, pp. 184–191, May 2017.
[11] R. Bassoli, M. Di Renzo, and F. Granelli, "Analytical energy-efficient planning of 5G cloud radio access network," in *Proc. Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–4.
[12] K. Zhang *et al.*, "Energy-efficient offloading for mobile edge computing in 5G heterogeneous networks," *IEEE Access*, vol. 4, pp. 5896–5907, 2016.
[13] J. G. Andrews *et al.*, "What will 5G be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.
[14] T. Choi, T. Kim, W. TaverNier, A. Korvala, and J. Pajunpaa, "Agile management of 5G core network based on SDN/NFV technology," in *Proc. Int. Conf. Inf. Commun. Technol. Converg. (ICTC)*, Oct. 2017, pp. 840–844.
[15] H. Hawilo, L. Liao, A. Shami, and V. C. M. Leung, "NFV/SDN-based vEPC solution in hybrid clouds," in *Proc. IEEE Middle East North Africa Commun. Conf. (MENACOMM)*, Apr. 2018, pp. 1–6.
[16] S. H. Won *et al.*, "Development of 5G CHAMPION testbeds for 5G services at the 2018 Winter Olympic Games," in *Proc. IEEE 18th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jul. 2017, pp. 1–5.
[17] V. Q. Rodriguez and F. Guillemin, "Cloud-RAN modeling based on parallel processing," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 3, pp. 457–468, Mar. 2018.
[18] G. C. Valastro, D. Panno, and S. Riolo, "A SDN/NFV based C-RAN architecture for 5G mobile networks," in *Proc. Int. Conf. Sel. Topics Mobile Wireless Netw. (MoWNeT)*, Jun. 2018, pp. 1–8.
[19] A. Tzanakaki *et al.*, "Wireless-optical network convergence: Enabling the 5G architecture to support operational and end-user services," *IEEE Commun. Mag.*, vol. 55, no. 10, pp. 184–192, Oct. 2017.
[20] M. Riva *et al.*, "An elastic optical network-based architecture for the 5G fronthaul," in *Proc. Simpósio Brasileiro Redes Comput. (SBRC)*, 2018, pp. 1–11.
[21] J. Luo, Q. Chen, and L. Tang, "Reducing power consumption by joint sleeping strategy and power control in delay-aware C-RAN," *IEEE Access*, vol. 6, pp. 14655–14667, 2018.
[22] 3GPP. *Releases*. Accessed: Mar. 3, 2016. [Online]. Available: http://www.3gpp.org/specifications/67-releases.
[23] C. D. Monfreid. (2012). The LTE Network Architecture–A Comprehensive Tutorial. Alcatel-Lucent, Paris, France. Accessed: Oct. 11, 2018. [Online]. Available: http://www.cse.unt.edu/~rdantu/FALL_2013_WIRELESS_NETWORKS/LTE_Alcatel_White_Paper.pdf
[24] Alcatel-Lucent. *Interworking LTE EPC With W-CDMA Packet Switched Mobile Cores*. Accessed: Sep. 20, 2017. [Online]. Available: http://www.alcatel-lucent.com
[25] A. Al-Quzweeni, T. E. H. El-Gorashi, L. Nonde, and J. M. H. Elmirghani, "Energy efficient network function virtualization in 5G networks," presented at the 17th Int. Conf. Transparent Opt. Netw. (ICTON), Jul. 2015.
[26] A. Al-Quzweeni, A. Lawey, T. El-Gorashi, and J. M. H. Elmirghani, "A framework for energy efficient NFV in 5G networks," in *Proc. 18th Int. Conf. Transparent Opt. Netw. (ICTON)*, Jul. 2016, pp. 1–4.
[27] M. H. Alsharif and R. Nordin, "Evolution towards fifth generation (5G) wireless networks: Current trends and challenges in the deployment of millimetre wave, massive MIMO, and small cells," in *Telecommunication Systems: Modelling, Analysis, Design and Management*. New York, NY, USA: Springer, 2016, pp. 1–21.

[28] M. Jaber, M. A. Imran, R. Tafazolli, and A. Tukmanov, "5G backhaul challenges and emerging research directions: A survey," *IEEE Access*, vol. 4, pp. 1743–1766, Apr. 2016.
[29] P. Chanclou *et al.*, "Optical fiber solution for mobile fronthaul to achieve cloud radio access network," in *Proc. Future Netw. Mobile Summit*, Jul. 2013, pp. 1–11.
[30] Z. Tayq, "Fronthaul integration and monitoring in 5G networks," Ph.D. dissertation, Univ. de Limoges, Limoges, France, Tech. Rep., 2017.
[31] S. Little, "Is microwave backhaul up to the 4G task?" *IEEE Microw. Mag.*, vol. 10, no. 5, pp. 67–74, Aug. 2009.
[32] A. Pizzinat, P. Chanclou, T. Diallo, and F. Saliou, "Things you should know about fronthaul," *J. Lightw. Technol.*, vol. 33, no. 5, pp. 1077–1083, Mar. 1, 2015.
[33] R. Chundury, "Mobile broadband backhaul: Addressing the challenge," *Planning Backhaul Netw., Ericsson Rev.*, vol. 3, no. 3, pp. 4–9, 2008.
[34] Alcatel-Lucent. *LTE Mobile Transport Evolution*. Accessed: Dec. 4, 2017. [Online]. Available: http://www.alcatel-lucent.com
[35] R. Kwan and C. Leung, "A survey of scheduling and interference mitigation in LTE," *J. Elect. Comput. Eng.*, vol. 2010, Jan. 2010, Art. no. 1.
[36] H. G. Myung, (2008). *Technical Overview of 3GPP LTE*. Accessed: Oct. 2 2018. [Online] Available: http://lteuniversity.com/cfs-file.ashx/_key/CommunityServer.Components.PostAttachments/00.00.00.00.99/3gppLTE.pdf
[37] C. Hoymann, W. Chen, J. Montojo, A. Golitschek, C. Koutsimanis, and X. Shen, "Relaying operation in 3GPP LTE: Challenges and solutions," *IEEE Commun. Mag.*, vol. 50, no. 2, pp. 156–162, Feb. 2012.
[38] J. Zyren and W. McCoy, "Overview of the 3GPP long term evolution physical layer," Freescale Semicond. Inc., Austin, TX, USA, White Paper 7, 2007, pp. 2–22.
[39] Anritsu. (2015). *LTE Resource Guide*. Accessed: May 7, 2018. [Online]. Available: http://www.cs.columbia.edu/6181/hw/anritsu.pdf
[40] R. F. Chisab and C. Shukla, "Performance evaluation Of 4G-LTE-SCFDMA scheme under SUI And ITU channel models," *Int. J. Eng. Technol.*, vol. 14, no. 1, pp. 58–69, 2014.
[41] M. Rinne and O. Tirkkonen, "LTE, the radio technology path towards 4G," *Comput. Commun.*, vol. 33, no. 16, pp. 1894–1906, 2010.
[42] A. de la Oliva, J. A. Hernandez, D. Larrabeiti, and A. Azcorra, "An overview of the CPRI specification and its application to C-RAN-based LTE scenarios," *IEEE Commun. Mag.*, vol. 54, no. 2, pp. 152–159, Feb. 2016.
[43] (Oct. 2015). *CPRI Specification V7.0*. [Online]. Available: http://www.cpri.info/downloads/CPRI_v_7_0_2015-10-09.pdf
[44] J. P. Castro, *The UMTS Network and Radio Access Technology*. Hoboken, NJ, USA: Wiley, 2001.
[45] H. Kaaranen, *UMTS Networks: Architecture, Mobility and Services*. Hoboken, NJ, USA: Wiley, 2005.
[46] T. Werthmann, H. Grob-Lipski, and M. Proebster, "Multiplexing gains achieved in pools of baseband computation units in 4G cellular networks," in *Proc. IEEE 24th Int. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)*, Sep. 2013, pp. 3328–3333.
[47] *IBM Completes Acquisition of ILOG*, IBM Corporation, Armonk, NY, USA, Jan. 2009.
[48] A. Q. Lawey, T. E. El-Gorashi, and J. M. Elmirghani, "Distributed energy efficient clouds over core networks," *J. Lightw. Technol.*, vol. 32, no. 7, pp. 1261–1281, Apr. 1, 2014.
[49] A. Q. Lawey, T. El-Gorashi, and J. M. Elmirghani, "Energy efficient cloud content delivery in core networks," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2013, pp. 420–426.
[50] X. Dong, T. El-Gorashi, and J. M. H. Elmirghani, "Green IP over WDM networks with data centers," *J. Lightw. Technol.*, vol. 29, no. 12, pp. 1861–1880, Jun. 15, 2011.
[51] N. I. Osman, T. El-Gorashi, L. Krug, and J. M. H. Elmirghani, "Energy-efficient future high-definition TV," *J. Lightw. Technol.*, vol. 32, no. 13, pp. 2364–2381, Jul. 1, 2014.
[52] H. M. M. Ali, T. E. El-Gorashi, A. Q. Lawey, and J. M. Elmirghani, "Future energy efficient data centers with disaggregated servers," *J. Lightw. Technol.*, vol. 35, no. 24, pp. 5361–5380, Dec. 15, 2017.
[53] A. M. Al-Salim, T. El-Gorashi, A. Lawey, and J. Elmirghani, "Energy efficient big data networks: Impact of volume and variety," *IEEE Trans. Netw. Service Manage.*, vol. 15, no. 1, pp. 458–474, Mar. 2018.
[54] I. WaÃ§mann, D. Versick, and D. Tavangarian, "Energy consumption estimation of virtual machines," in *Proc. 28th Annu. ACM Symp. Appl. Comput.*, 2013, pp. 1151–1156.
[55] *Cisco ASR 5000 Series Product Overview Release 12.0*, Cisco, San Jose, CA, USA, 2013.
[56] *Alcatel-Lucent 9926 Base Band Unit LR13.1.L*, Alcatel-Lucent, Paris, France, 2013.
[57] *Export Compliance Metrics for Intel Microprocessors*, Intel, Santa Clara, CA, USA, 2018.
[58] Sumitomo Electric. *FTE7502 EPON Optical Network Unit (10G ONU) Datasheet*. Accessed: Feb. 21, 2015. [Online]. Available: http://www.sumitomoelectric.com/onu-fte7502.html
[59] Cisco. *Cisco ME 4600 Series Optical Line Terminal Data Sheet*. Accessed: Feb. 19, 2015. [Online]. Available: http://www.cisco.com/c/en/us/products/collateral/switches/me-4600-series-multiservice-optical-access-platform/datasheet-c78-730445.html
[60] Alcatel-Lucent. *TRDU2x40-08 LTE 3GPP Band 20 LTE FDD Transmit Receive Duplexer Unit—800 MHz EDD Datasheet*. Accessed: Nov. 2, 2015. [Online]. Available: https://www.academia.edu/34356836/Alcatel-Lucent_9926_Hardware_Maintenance_and_Upgrade_Guide_for_FDD_LTE_Use_pursuant_to_applicable_agreements
[61] L. Nonde, T. E. H. El-Gorashi, and J. M. H. Elmirghani, "Energy efficient virtual network embedding for cloud networks," *J. Lightw. Technol.*, vol. 33, no. 9, pp. 1828–1849, May 1, 2015.
[62] A. Q. Lawey, T. E. H. El-Gorashi, and J. M. H. Elmirghani, "Renewable energy in distributed energy efficient content delivery clouds," presented at the IEEE Int. Conf. Commun. (ICC), Jun. 2015.
[63] X. Dong, T. E. H. El-Gorashi, and J. M. H. Elmirghani, "On the energy efficiency of physical topology design for IP over WDM networks," *J. Lightw. Technol.*, vol. 30, no. 11, pp. 1931–1942, Jun. 1, 2012.
[64] X. Dong, T. El-Gorashi, and J. M. H. Elmirghani, "IP over WDM networks employing renewable energy sources," *J. Lightw. Technol.*, vol. 29, no. 1, pp. 3–14, Jan. 1, 2011.
[65] GreenTouch. (2015). *GreenTouch Final Results from Green Meter Research Study*. Accessed: Jan. 7, 2016. [Online]. Available: http://www.greentouch.org/index.php?page=greentouch-green-meter-research-study
[66] J. Elmirghani *et al.*, "GreenTouch GreenMeter core network energy-efficiency improvement measures and optimization," *J. Opt. Commun. Netw.*, vol. 10, pp. A250–A269, Feb. 2018.
[67] D. P. Dobkin and S. P. Reiss, "The complexity of linear programming," *Theor. Comput. Sci.*, vol. 11, no. 1, pp. 1–18, May 1980.
[68] R. G. Gallager, "Distributed minimum hop algorithms," M.I.T. Lab Inf. Decis. Syst., Cambridge, MA, USA, Tech. Rep. LIDS-P1175, Jan. 1982.

AHMED N. AL-QUZWEENI received the B.Sc. and M.Sc. degrees in computer engineering from Nahrain University, Baghdad, Iraq, in 2001 and 2004, respectively, and the Ph.D. degree in communication networks from the University of Leeds, U.K., in 2019.

From 2005 to 2009, he was a Mobile Core Network Senior Engineer, where he was involved in short message system, intelligent network, PSTN, and billing systems. From 2009 to 2014, he was an Assistant Lecturer with the Department of Computer Communication, Al-Mansour University College, Baghdad. He was a Team Leader with ZTE Corporation for Telecommunication, Iraq. His current research interests include energy efficiency in optical and wireless networks, NFV, mobile networks, 5G networks, caching the contents, cloud computing, and the Internet of Things.

AHMED Q. LAWEY received the B.Sc. and M.Sc. degrees (Hons.) in computer engineering from Nahrain University, Iraq, in 2002 and 2005, respectively, and the Ph.D. degree in communication networks from the University of Leeds, U.K., in 2015. From 2005 to 2010, he was a Core Network Engineer with ZTE Corporation for Telecommunication, Iraq. He is currently a Lecturer in communication networks with the School of Electronic and Electrical Engineer, University of Leeds. His current research interests include energy efficiency in optical and wireless networks, big data, cloud computing, and the Internet of Things.

TAISIR E. H. ELGORASHI received the B.S. degree (Hons.) in electrical and electronic engineering from the University of Khartoum, Khartoum, Sudan, in 2004, the M.Sc. degree (Hons.) in photonic and communication systems from the University of Wales, Swansea, U.K, in 2005, and the Ph.D. degree in optical networking from the University of Leeds, Leeds, U.K., in 2010, where she is currently a Lecturer in optical networks with the School of Electrical and Electronic Engineering. Previously, she held a Postdoctoral research post at the University of Leeds, from 2010 to 2014, where she focused on the energy efficiency of optical networks investigating the use of renewable energy in core networks, green IP over WDM networks with datacenters, energy efficient physical topology design, the energy efficiency of content distribution networks, distributed cloud computing, network virtualization, and big data. She was a BT Research Fellow and developed energy efficient hybrid wireless optical broadband access networks and explored the dynamics of TV viewing behavior and program popularity, in 2012. The energy efficiency techniques developed during the Postdoctoral research contributed three out of eight carefully chosen core network energy efficiency improvement measures recommended by the GreenTouch consortium for every operator network worldwide. Her work led to several invited talks at GreenTouch, Bell Labs., the Optical Network Design and Modeling Conference, the Optical Fiber Communications Conference, the International Conference on Computer Communications, and the EU Future Internet Assembly in collaboration with Alcatel Lucent and Huawei.

JAAFAR M. H. ELMIRGHANI (M'92–SM'99) received the B.Sc. degree in electrical engineering (Hons.) from the University of Khartoum, in 1989, the Ph.D. degree in the synchronization of optical systems and optical receiver design from the University of Huddersfield, U.K., in 1994, and the D.Sc. degree in communication systems and networks from University of Leeds, U.K., in 2014. He is the Director of the School of Electronic and Electrical Engineering, Institute of Communication and Power Networks, University of Leeds, U.K., where he joined, in 2007, and prior to that as the Chair of optical communications with the University of Wales, from 2000 to 2007. He was with Swansea University, where he developed the Technium Digital (TD), a technology incubator/spin-off hub and the Director the Institute of Advanced Telecommunications. He has provided outstanding leadership in a number of large research projects at IAT and TD. He has coauthored *Photonic Switching Technology: Systems and Networks* (Wiley), and he has published over 450 papers. He has received in excess of 22 million in grants to date from EPSRC, the EU and industry, and he has held prestigious fellowships supported by the Royal Society and BT. His research interests include optical systems and networks. He is a Fellow of the IET and the Institute of Physics, and a Chartered Engineer. He was a member of the Royal Society International Joint Projects Panel and the Engineering and Physical Sciences Research Council (EPSRC) College. He was an IEEE Comsoc Distinguished Lecturer, from 2013 to 2016. He has received the IEEE Communications Society Hal Sobol Award, the IEEE Comsoc Chapter Achievement Award for excellence in chapter activities (both in international competition, in 2005), the University of Wales Swansea Outstanding Research Achievement Award, in 2006. He has received awards in international competition, including the IEEE Communications Society Signal Processing and Communication Electronics Outstanding Service Award, in 2009, the Best Paper Award from the IEEE ICC 2013. Related to green communications he has received awards, including the IEEE Comsoc Transmission Access and Optical Systems Outstanding Service Award, in 2015, in recognition of leadership and contributions in the areas of green communications, the GreenTouch 1000x Award, in 2015, for pioneering research contributions in the fields of energy efficiency in telecommunications, the IET 2016 Premium Award for the best paper in IET optoelectronics, and the shared the 2016 Edison Award in the collective disruption category with a team of six from GreenTouch for their joint work on the GreenMeter. He has received four prizes in the department for academic distinction with the University of Khartoum. He was the Co-Chair of the GreenTouch Wired, Core, and Access Networks Working Group and an Adviser of the Commonwealth Scholarship Commission. He was the Chair of the IEEE Comsoc Transmission Access and Optical Systems Technical Committee and of IEEE Comsoc Signal Processing and Communications Electronics Technical Committee. He was the Founding Chair of the Advanced Signal Processing for Communication Symposium, which was started at the IEEE GLOBECOM, in 1999, and he has continued since at every ICC and GLOBECOM. He was also the Founding Chair of the first IEEE ICC/GLOBECOM Optical Symposium at the GLOBECOM 2000, the Future Photonic Network Technologies, Architectures and Protocols Symposium. He has chaired this symposium, which continues to date under different names. He was the Founding Chair of the first Green Track Chair of ICC/GLOBECOM at the GLOBECOM 2011. He has been the Chair of the IEEE Green ICT initiative within the Future Directions Committee (FDC), IEEE Technical Activities Board (TAB), a pan of the IEEE societies initiative responsible for Green ICT activities across the IEEE, since 2012. He was on the technical program committee of 34 IEEE ICC/GLOBECOM conferences, from 1995 to 2016, including 15 times as the Symposium Chair. He has given over 55 invited and keynote talks over the past eight years. He was an Editor of the *IEEE Communications Magazine*, the IEEE Communications Surveys and Tutorials, and the IEEE Journal on Selected Areas in Communications series on green communications and networking. He is currently an Editor of the *Journal of Optical Communications* and IET Optoelectronics.