QoS Provisioning and Energy Saving Scheme for Distributed Cognitive Radio Networks Using Deep Learning

Mduduzi Comfort Hlophe and Bodhaswar T. Maharaj

Abstract: **One of the major challenges facing the realization of cognitive radios (CRs) in future mobile and wireless communications is the issue of high energy consumption. Since future network infrastructure will host real-time services requiring immediate satisfaction, the issue of high energy consumption will hinder the full realization of CRs. This means that to offer the required quality of service (QoS) in an energy-efficient manner, resource management strategies need to allow for effective trade-offs between QoS provisioning and energy saving. To address this issue, this paper focuses on single base station (BS) management, where resource consumption efficiency is obtained by solving a dynamic resource allocation (RA) problem using bipartite matching. A deep learning (DL) predictive control scheme is used to predict the traffic load for better energy saving using a stacked auto-encoder (SAE). Considered here was a base station (BS) processor with both processor sharing (PS) and first-come-first-served (FCFS) sharing disciplines under quite general assumptions about the arrival and service processes. The workload arrivals are defined by a Markovian arrival process while the service is general. The possible impatience of customers is taken into account in terms of the required delays. In this way, the BS processor is treated as a hybrid switching system that chooses a better packet scheduling scheme between mean slowdown (MS) FCFS and MS PS. The simulation results presented in this paper indicate that the proposed predictive control scheme achieves better energy saving as the traffic load increases, and that the processing of workload using MS PS achieves substantially superior energy saving compared to MS FCFS.**

Index Terms: **Bipartite matching, cognitive radio networks, deep learning, energy saving, mean slowdown, quality of service, resource allocation, resource percentage threshold, traffic prediction.**

I. INTRODUCTION AND BACKGROUND

BECAUSE of the increase in data rate requirements and the level of heterogeneity, the rate of change in network traffic expected in future mobile and wireless communication networks poses new challenges related to spectrum management and energy consumption [1]. The increase in multimedia services and other mission-critical applications that are hosted by network infrastructure, which demand immediate satisfaction in terms of quality of service (QoS) requirements, has led to even newer challenges in energy efficiency [2]. These have led to escalating concerns regarding energy efficiency in terms of network operational costs such that research on energy consumption and saving has gained huge attention. In mobile and wireless communication networks, base stations (BSs) consume much of the power and their power consumption varies over time [3]. The general BS operation mode can be described as a finite state machine, which can be explained in terms of a two-state Markov model, i.e., idle (OFF) and active (ON). Hence the energy consumption of a general BS depends on its mode of operation. When the BS is in its active mode, it has to process traffic streams from all its associated users, hence its computational components are faced with maximizing user satisfaction demands, while simultaneously minimizing their energy consumption. On the one hand, maximizing user satisfaction by upholding user QoS requirements through an increase in transmission power has a substantial effect on energy efficiency, while on the other hand reducing transmission power degrades QoS performance [4]. Therefore, quantifying the trade-off between energy consumption and the required QoS are key parameters in BS energy consumption that require balanced optimization.

Researchers in both industry and academia have proposed workable solutions to reduce network operational costs through the installation of energy-efficient hardware [5], but this approach does not seem to solve the problem completely. Reducing network operational costs by installing more energy-efficient hardware somehow requires a compromise between the energy cost and the user coverage drop. Thus, this approach somehow falls short in addressing the energy efficiency problem, since it might result in wireless access networks being almost invariably over-provisioned and under-provisioned with respect to the user traffic demands [6]. Even though resource over-provisioning may lead to better QoS provisioning in terms of negligible packet losses and transmission delays, it comes at non-negligible operational costs [7]. It is thus clear that resource over-provisioning needs to be replaced by optimal or even near-optimal energy-efficient solutions that allocate adequate network infrastructure based on current resource demand. Adequate provisioning, however, requires better understanding of the relationship between resource demand, available capacity and the transparency between real-time and best effort traffic streams. In light of the above, newer and more efficient resource management schemes, capable of controlling how many network resources are allocated at a certain time, can be extremely effective and provide quite large network operational cost reduc-

Manuscript received December 20, 2019; revised March, 2020; approved for publication by Namyoon Lee, Guest Editor, May 21, 2020.

This work was supported in part by the Association of Commonwealth Universities (ACU) under grant FE-2015-26 and in part by the Sentech Chair in Broadband Wireless Multimedia Communications (BWMC) at the University of Pretoria.

M. C. Hlophe and B. T. Maharaj are with the Department of Electrical, Electronics and Computer Engineering, University of Pretoria, Pretoria, South Africa, e-mail: u16250444@tuks.co.za and sunil.maharaj@up.ac.za

M. C. Hlophe is the corresponding author.

Digital Object Identifier: 10.1109/JCN.2020.000013

tions. As the energy efficiency of BSs is receiving more attention owing to several factors, there are also concerns about meeting the international mobile telecommunications (IMT)-2020 and beyond requirements, as well as challenges facing cognitive radio networks (CRNs).

In the overall energy efficiency network design and evaluation process, the primary objective is an adequate metric that is directly related to the optimized decisions across all the protocol layers. The most popularly used metric is bits-per-Joule, which is defined as the system throughput for unit energy consumption [8]. A great quantity of information-theoretic results for energy efficiency at the link level, based on this metric set the limitation on transmission power is set as a constraint, and it has been proven that the upper bound channel capacity per unit energy can only be achieved by utilizing an unlimited number of degrees of freedom per information bit [9]. Analyzing the "bits-per-Joule" capacity at network level proved that capacity increases with the number of nodes in the network, implying that large-scale energy-limited networks may only be suitable for delay-tolerant applications. For example, this metric has been widely used as the utility function in game-theoretic approaches for energy saving in wireless networks, where the energy consumption models only consider the transmission power associated with the data transmission rate [10]. However, the transmission power is only part of the overall energy budget and when the energy consumption of other parts (e.g., circuit power consumption of the transceiver) is taken into account, the energy-efficient schemes described in literature might not be appropriate to meet the IMT-2020 and beyond requirement specifications.

A. Meeting IMT-2020 and Beyond Requirement Specifications

Nowadays, global mobile data traffic is increasingly dominated by delay- and loss-intolerant traffic streams, which means that traffic congestion in the core network is an inevitable occurrence. This can quickly lead to overall network performance degradation resulting from the moderate-to-high traffic levels [11]. The consequence of this is the catastrophe of heavy burst losses, as the whole network might degenerate into chaos. The daily operations of network components that are pushing data and multimedia traffic into the internet require a significant increase in network energy consumption or no user will be able to use the network properly. However, despite the intense research efforts by both academia and standardization bodies in the quest to meet the requirements specified in the IMT-2020 [12], there is still considerable controversy concerning the definition of QoS and its direct influence on network resource provisioning. This is because huge research efforts have focused on the optimization of sum-rate to support high data transmission rates [13].

Apart from energy efficiency, a variety of objectives have been put forward in the preparation of the next generation of mobile and wireless communications. Other than achieving high data rates, these objectives are based on network metrics such as improved coverage with uniform user experience, higher reliability and lower latency, better energy efficiency, lower cost user devices and services and better scalability with the number of devices [14]. However, these objectives have to be realized simultaneously, and the challenge is that they are often coupled in a conflicting manner. Achieving improvements in one objective leads to degradation in the other. Consequently, the design of future mobile and wireless networks requires new optimization tools that are capable of properly handling both the existence of these objectives and intelligent trade-offs between them.

In CRNs, QoS provisioning requires a cross-layer design. Numerous studies have proposed allocating resources reasonably by considering QoS requirements for secondary users (SUs). For example, the authors in [15] considered the energy effect on QoS provisioning and proposed an energy-efficient channel hand-off strategy using partially observable Markov decision processes (POMDPs). In the development of the hand-off strategy using POMDPs, the authors considered imperfect channel sensing and residual energy of SUs, which considers beliefs about the operating and backup channels and the residual energy at the SU. An α-retry policy was proposed as a spectrum access strategy to enhance QoS for SUs in [16], where a preemptive priority queue was built as a two-dimensional discrete-time Markov chain (DTMC) and the blocking rate, the forced dropping rate, the throughput and the average delay of the SU packets were analyzed. From a spectrum management perspective [17] proposed a queuing theory-based analytical framework to analyze QoS for SUs. The spectrum resource management problem was considered for co-located SUs with both streaming and intermittent data and seamless end-to-end service was ensured by effectively identifying the number of backup channels. In another contribution, the authors in [18] proposed the use of priority queues as a resource allocation (RA) model for different traffic classes. They used packet priority to explore the relevance and implications of various heterogeneous classifications. The RA model incorporates the essential concepts of heterogeneity, which were developed with weight attached to differentiate between different traffic classes.

Meeting the IMT-2020 requirement and beyond specifications within the CRN space has presented itself as a huge challenge to researchers, since balancing the above approaches in an energy-efficient manner depends on several factors. Such factors are the degree of cooperation between primary users (PUs) and SUs and the reliability of spectrum sensing [19] exacerbating the problem of energy-efficient CRN operations. Because of the intermittent nature of SU connections in the primary networks, it is a difficult task to achieve energy-efficient CRN operations owing to the preemptive priority of PUs. Preemptive priority gives PUs absolute rights to use their licensed spectrum such that while SUs try to exploit transmission opportunities in free spectrum bands, PUs can resurface at any time and force SUs to terminate their transmissions. In that case, SUs are forced to leave the current spectrum band and handoff their activities to another channel that is deemed vacant through spectrum sensing. These transmission terminations lead to subsequent transmission delays and high energy consumption, which affect the latency and energy efficiency respectively, especially when the probability of forced terminations is high [20]. Thus, it seems as if there are apparently too many different requirements that need to be kept in mind when attempting to design lasting solutions. Unfortunately, there is no easy shortcut to achieving a long-lasting solution that balances energy-efficient RA and QoS in CRNs. This means that these two objectives cannot be treated separately because they are coupled - sometimes in a consistent fashion, but more often than not in conflicting ways, implying that improvements in one objective leads to deterioration in the other.

B. Technological Challenges Facing CRNs

The 5G era, which is driven by the Internet of things (IoT), introduces the requirements of high data rates, low latency, efficient use of spectrum resources and coexistence of different network technologies. Because of the existing spectrum quagmire, all future wireless devices need to be CR-capable, which makes CR technology the biggest enabler of 5G networks and beyond. Apart from the major concern of spectrum scarcity, there are several other issues and technological challenges that hinder the design, implementation and realization of a fully functional CRN. All future mobile and wireless network technologies will be deployed in a distributed fashion in order to exhibit significant gains in network capacity maximization. Wireless technologies deployed in this manner require significant coordination among network devices, which results in high communication overheads. Beyond the level of heterogeneity introduced by CRNs into the traditional wireless network, the other spectrum use cases that are supposed to benefit enormously from CR technology also have heterogeneous requirements. For example, the IoT and its other derivatives, such as the cognitive internet of people, services, data and things and several other variants of the IoT, with bandwidth-demanding services such as video streaming and video-on-demand, all pose great challenges to the issue of spectrum management.

Because of the challenges that come with the dynamics of channel availability, uncertainty of spectrum sensing and spectrum access, as well as PU activities, real-time SU transmissions might suffer. On the one hand, the network designer is faced with numerous design trade-offs, diversified network dynamics and limited resources, while on the other hand SUs with bandwidth-consuming services with stringent QoS constraints require resources to be allocated immediately. Thus, this necessitates a holistic cross-layer design approach to exploit the CR technology optimally, which calls for the development of intelligent and invisible paradigms for programmable and controllable networks to satisfy future requirements. Regarding the intelligent use of spectrum resources, the CR technology stands to benefit massively from the incorporation of artificial intelligence (AI) into its operation. From a cognitive radio (CR) perspective where the issues of QoS and energy efficiency have put enormous pressure on top of the existing spectrum shortage, achieving energy-efficient operations is a daunting task. With the interconnection of heterogeneous devices posing numerous challenges that may include high energy consumption, data rate requirements and intermittent connections because of SU mobility and PU activities, the incorporation of AI will ensure efficient decision-making. A perfect solution to this problem will require some kind of automated approach that achieves both QoS and energy efficiency, while also yielding a new set of network assurances such as reliability.

However, whatever automation network designers can come up with, it has to be within the perimeters of good energy consumption. One promising way is to use the massive amount of data that is generated by the great quantity of network equipment to develop predictive measures to deal with the energy efficiency problem. AI strategies such as deep architectures (i.e., deep learning (DL), deep reinforcement learning) can be used to analyze these data, extract the relevant patterns, make sense of the data and then prescribe energy-efficient actions to be taken by network equipment. This may lead to a realization of more effective and efficient energy-saving models for CRNs. Existing energy-saving schemes in traditional cellular networks are discussed in the following section.

II. EXISTING ENERGY EFFICIENCY SCHEMES

A. The EARTH Model

The first energy-saving technique was proposed in the energy aware radio and network technologies (EARTH) project, which is a concerted effort to achieve energy efficiency in wireless networks. The primary objective of the project was a reduction in energy consumption in mobile networks as environmental concerns such as global warming were gaining momentum [21]. To quantify the energy savings in a wireless network, the power consumption of the entire system needs to be captured and an appropriate energy efficiency evaluation framework (E3F) needs to be defined. The E3F is applied to provide an assessment of the BS energy efficiency of a third generation partnership project long term evolution (3GPP LTE) network deployed where BSs are switched ON/OFF based on traffic load [22]. This model is based on a finite-state machine model consisting of two operational states, P_{on}, which denotes the static/load-independent power consumption figure; P_{tx}, which denotes the dynamic/load-dependent power consumption figure whose consumption trajectory is scaled according to traffic load.

This technique offers energy-efficient communication in low traffic regimes and plays an important role in reducing overall network power consumption [23]. However, in terms of meeting the IMT-2020 and beyond requirement specification, one striking drawback is that when a BS is switched ON/OFF, a switching cost is incurred in terms of the energy and the time to transition from ON to OFF when there is little or no channel activity and vice versa, which is a significant amount that cannot be ignored. Except the evident and noticeable reductions of the operational costs faced by mobile operators, it also adapts the level of resource over-provisioning by re-associating traffic to moderately served BSs [24]. A considerable amount of energy and time is spent in turning on BS components, user data management and user re-association; which affects the QoS owing to server response times that also has a consequence in packet delays and losses. There is also limited support for newer systems and technologies and the (de)activation information is not defined.

B. Green Cellular Network Model

The second energy-saving technique is the green cellular network model, which was proposed to address the shortcomings of the EARTH model and reduce carbon dioxide (CO_2) emissions. This technique incorporates the use of both grid power supply and energy harvested from solar and wind sources. However, the basic structure of the EARTH model was not discarded, but was maintained and reused - meaning that the BS power consumption remained unaltered. Through this technique, BSs can selectively switch between grid and harvested energy in order to reduce network operation energy costs. In this technique, an energy-harvesting BS is co-located with a multi-access edge computing (MEC) server to reduce energy consumption further. The radio network and energy level information is communicated periodically to the MEC server through the radio network information services (RNIS) and the energy manager. The RNIS entity is responsible for selecting the appropriate energy source to fulfill the energy buffer and for monitoring the energy levels of the system.

The authors in [25] investigated an environment-aware framework for CRNs where PU and SU networks collaborate. The collaboration between the two networks is intended to maximize their profits and meet PU QoS and the total CO_2 emission. In terms of energy balancing, [26] proposed an energy-balancing strategy where the key technique is that each BS maintains two parameters. These parameters contain the trend of its previous energy consumption and then predicts its future quantity of energy, which is defined as the BS's potential energy capacity. Using this concept offers better solutions, but the simplifying assumptions made may often introduce inaccuracies when the switching is not optimized. In terms of green cellular networking, it is argued that energy saving can be achieved through the adoption of renewable sources of energy to make communication networks more energy-efficient. However, from a communication perspective, a more energy-efficient system is created where the power consumption of the BS is optimized rather than adding several more power sources.

C. The One-step-ahead Predictive Model

The final energy-saving strategy is the one-step-ahead predictive model, which has the potential of playing a decisive role in boosting energy efficiency in future mobile and wireless networks through the use of predictive analytics. This technique does not optimize the network operation only in terms of energy consumption, but also with respect to balancing the transmission rates and power consumption by predicting the future traffic load. In this way, the signaling overhead and circuit power in terms of proactive preparation of the required number of items of computational equipment to be allocated are taken into account a priori. The use of predictive analytics through one-step-ahead network traffic prediction puts the system ahead of the normal time by enabling it to take proactive decisions through the extraction and analysis of traffic patterns in network traffic trends and user behavior to predict future network behavior for better RA. Through this technique, systems can learn traffic profiles and automatically tune their computational parameters to accommodate future demands; thus data analytics promises to be a possible pathway towards achieving both QoS and energy efficiency objectives.

There are several prediction techniques that are suitable for training and testing CR systems, such as the bio-inspired metaheuristic system in [23], which has been inspired by swarm intelligence and has been seen to achieve better energy saving for 5G networks. The authors proposed the separation of the control and data planes and the use of particle swarm intelligence to handle the interaction and operation of users to achieve better energy efficiency and lower aggregate delays. However, the integration of real-time practical learning capabilities using such a soft computing technique is a challenge in CRNs [19]. Nevertheless, an investigation of applications from deep architectures has shown some significant benefits in enabling decision-making through prediction in the absence of real-time information. For example, artificial neural network (ANN) techniques studied in [27] and [28] proved to improve the accuracy and decrease the complexity of traffic prediction and RA, respectively. However, from both these works, it is clear that recent RA problems have become dynamic and require the enhancement of machine learning architectures in order to improve their inference capabilities. As just pointed out, [29] proposed an inference engine using fuzzy techniques to improve RA in CRNs using an improved channel allocation that considered signal strength as the decision variable for the channel access priority of SUs. A genetic algorithm (GA)-based RA technique was studied in [30]. GA was used to define the radio in the form of chromosomes and genes, where the users' QoS requirements were given as the input of the GA algorithm. The impact of the available spectrum resource size was analyzed in terms of both the population size and the number of defined chromosome genes in spectrum allocation efficiency.

In another contribution, an improved long short-term memory (LSTM) was used in [31] to obtain accurate and fast traffic flow forecasting in intelligent transportation systems. Moreover, a time series prediction for extracting useful information from historical records to determine their future values was studied in [32], where a random connectivity LSTM (RCLSTM) model was used to reduce the computational complexity associated with LSTM and was tested and verified for traffic prediction and user mobility in wireless networks. The RCLSTM was found to exhibit a certain level of sparsity, which appealingly reduces the computational complexity, making it suitable for latency-stringent applications. An online optimization algorithm called the energy aware and adaptive management (ENAAM), based on traffic prediction and foresighted control policies, was proposed in [33]. Here, the BSs and virtual machines (VMs) are dynamically switched ON/OFF to effect energy saving and QoS provisioning by exploiting short-term traffic load and harvested energy forecasts using LSTM. This contribution was inspired by the convergence of communication and computing has led to the emergence of mobile edge computing (MEC), where computing resources supported by VMs were distributed at the edge of the mobile network. BSs aiming at ensuring reliable and ultra-low latency services are equipped with an energy-harvesting system to reduce energy consumption.

One shortcoming of this traffic prediction technique is that it relies on the obvious seasonality of traffic that has been aggregated with the granularity of "hours of day". Longer time granularities, based on hours, can have significant short-comings in today's network traffic, since it does not exhibit the same seasonal behavior, but rather up-down linear trends. The traffic pattern in modern cellular networks has changed drastically since the emergence of smart phones owing to the many applications hosted by wireless networks. Applications for social networking (e.g., Facebook), for internet telephony (e.g., Skype), for micro-blogging (e.g., Twitter, Posterous, FriendFeed, etc.), for instant messaging (e.g., WhatsApp, Facebook Messenger, WeChat, Viber, etc.), consist mostly of real-time content and hence exhibit different traffic patterns compared to the traditional voice, text messaging, emails and web surfing. Nowadays, the behavior of user traffic is mostly application-specific and varies rapidly within short intervals of time. Thus, the "hours of day" granularity may present some anomalies with traffic trends that require a new traffic matrix at each time slot in order to be effective. Because of the change in traffic trends, there are significant changes in individual packet sizes, burst sizes, packet inter-arrival times, as well as the behavior of inactive periods. For this reason, a system capable of learning traffic and user behavior, predicting its future trends, and dynamically adjusting its computational resources to be assigned to traffic relations may turn early research into usable solutions.

D. Motivation and Contributions

The application of data analytics in wireless data promises to be a possible pathway towards achieving both QoS and energy efficiency objectives. This will help in obtaining tentative operating points for network equipment to achieve energy efficiency and network sustainability, which is an essential step in managing the high level of heterogeneity associated with future mobile and wireless networks. Then, depending on the applied network functions, the question regarding the balance between QoS provisioning and energy efficiency rests on the packet-service discipline to achieve better energy saving. To solve this problem, an interesting contribution is followed in [34]; however, the approach proposed in this paper differs from that, since it involves a separation principle that decouples the design of RA, energy consumption, and service QoS provisioning and makes the problem manageable. The major contributions of this paper are summarized as follows:

- Firstly, a distributed dynamic RA based on uplink (UL) power allocation and SU resource reservation protocol is proposed. Resource reservation is a transport layer protocol designed to reserve resources across a network for QoS using the integrated services model and is adopted in this paper to give SUs a high probability to complete their transmissions. The resource reservation problem is solved using geometric programming (GP) and a resource percentage threshold (RPT) serves as the portion of resources reserved for SUs. An optimal RA solution is obtained through a weighted bipartite matching from graph theory with a polynomial complexity of $\mathcal{O}(K^3)$ compared to the $\mathcal{O}(K!)$ of integer programming.
- Secondly, resource consumption efficiency obtained from the bipartite graph solution is then used to solve as weighted cost function in which power consumption is added together with different weights, reflecting their contribution to BS power consumption. This initiates a DL predictive control scheme with control actions derived to drive a stacked auto-encoder (SAE) in making future traffic predictions as well as providing computing performance measures.
- Finally, control actions are applied and using this formulation, relevant parameters of the operating environment such as workload arrival patterns, are estimated and used by the model to predict the future behavior over a finite horizon, T. Using the output of the SAE, which is simply a regression between the previous and current system states, appropriate packet-processing schemes are chosen between mean slowdown first-come-first-served (MS FCFS) and MS processor sharing (MS PS). The belief that MS is important in packet-by-packet processing as a measure of the system's energy efficiency is proven by the finding that different traffic flows consume significantly different amounts of resources and the choice of a processing scheme determines the overall energy efficiency of the system.

To achieve these objectives, the remainder of this paper is organized as follows: The system model of the resource-aware energy-efficient model is presented using both opportunistic spectrum access (extrinsic) and BS server opportunistic computing (intrinsic) in Section III. The mathematical description of the model is presented in both opportunistic access and traffic load power consumption and then an overall formulation is given in Section IV. The solutions for resource reservation and optimal RA are presented in Section V, and the solution of the computational-resource aware DL predictive scheme is given in Section VI. The experimental set-up for the DL predictive scheme is discussed in Section VII and the prioritization and energy consumption per service is discussed in Section VIII. The simulation parameters and the simulation results together with their discussions are given in Section IX. Finally, the validation of the main findings of this paper is given in the conclusion in Section X.

III. PROPOSED SYSTEM MODEL

Consider a single-cell spectrum-sharing scenario where a PU transmitter and a PU receiver coexist with a set $\mathcal{K} : k = 1, 2, \cdots, K$ SUs in an energy-constrained CRN. It is assumed that SUs are running real-time traffic and performing opportunistic transmission on the shared spectrum consisting of $\mathcal{J} : j = 1, 2, \cdots, J$ physical resource blocks (PRBs). Perfect channel state information (CSI) is assumed for all K SUs uniformly distributed in a cluster such that the distance between them and the BS is characterized by channel gains $g_{j,1}(t), \cdots, g_{j,K}(t)$ as illustrated in Fig. 1 below [35]. As shown in Fig. 1 above, resources are allocated using a mapping technique through an undirected bipartite graph. Since the CR technique is performed by SUs, and the transmission access and gateway are dealt with by the BS, a bipartite matching strategy in graph theory is applied to optimize the RA. The BS is assumed to utilize a hybrid access scheme where SUs can connect when there are free RBs and employs resource reservation for bandwidth estimation admission control. Therefore, the proposed system model is split into two parts, which are discussed in the following subsections.

A. Spectral Efficiency

Each SU k can transmit data through PRB j such that the transmission link is represented by $k \to j$, which is denoted as follows:

$$d_{jk} = \sqrt{(x_k - x_j)^2 + (y_k - y_j)^2}, \tag{1}$$

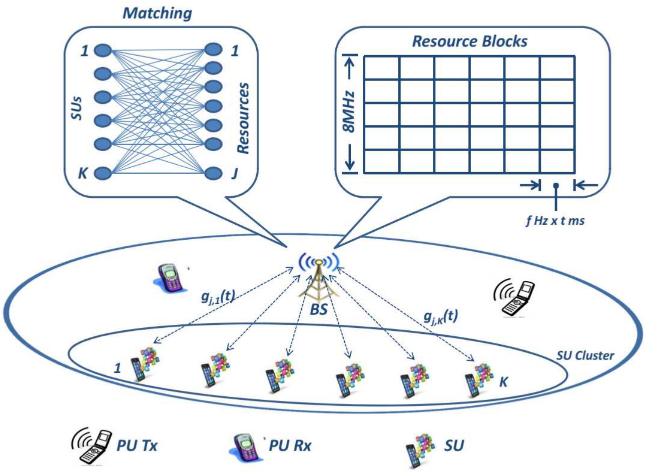

Fig. 1. RA characterized as a weighted bipartite matching problem.

which is a Euclidean distance measure where the k^{th} SU and the BS are located at (x_k, y_k) and (x_j, y_j), respectively. In order to increase the chances of SUs completing their transmissions, a resource reservation strategy is proposed where the BS reserves a fraction of the J PRBs for opportunistic SU transmissions, which the SUs can exploit with an optimal transmission power $P_{j,k}$ that is closely controlled by PUs' activities. Thus, the spectral efficiency of the k^{th} SU with access to the j^{th} PRB is formulated as follows:

$$R_{j,k} = \log_2(1+\gamma_{j,k}), \quad \text{where} \quad \gamma_{j,k} = \frac{P_{j,k} g_{j,k}}{\sigma^2 + \mathcal{I}_k}, \tag{2}$$

where $R_{j,k}$ is the Shannon bound of the spectral efficiency, and $\gamma_{j,k}$ is the signal-to-interference-plus-noise ratio (SINR) obtained through the designated transmission power $P_{j,k}$, the channel gain $g_{j,k}$, the noise spectral density σ^2, and the interference caused by simultaneous transmission $\mathcal{I}_k$ is described as follows:

$$\mathcal{I}_k = \sum_{j \in \mathcal{K} \setminus \{k\}} P_j |g_{j,k}|^2, \tag{3}$$

where P_j denotes the transmission power of the other PRBs, excluding the one assigned to the k^{th} SU. The channel gain $g_{j,k}$ is defined as a function of the distance measure in (1) and other terms such as the path-loss coefficient $n_{j,k}$, the channel fading coefficient $h_{j,k}$ and the path-loss exponent α_{pl} as given in [36]. In (3) is a measure of the multiple access interference originating from other SUs, which might be using the same access technique as the k^{th} SU, linear SINR prediction that employs constant first-order derivatives across adjacent transmissions is assumed. Therefore, according to the proposed resource reservation strategy, the UL rates achieved by SUs can be obtained as follows:

$$R_{j,k} = \frac{\xi \varsigma^{th}}{\lambda} \log_2\left(1+\gamma_{j,k}\right), \tag{4}$$

where $\xi\varsigma^{th}/\lambda$ is the fraction of resources reserved for SUs which represent the admission condition for each SU that requests a connection. The terms ξ and ς^{th} are computed as follows:

$$\xi(\gamma) = \log(1+\gamma) - \left(\frac{\gamma}{1+\gamma}\right)\log\gamma, \quad \varsigma^{th}(\gamma) = \frac{\gamma}{1+\gamma}, \tag{5}$$

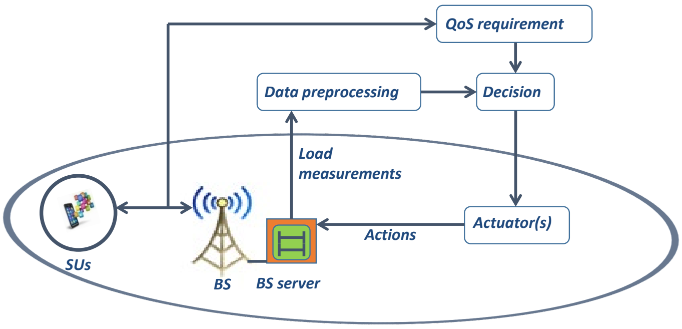

Fig. 2. High-level view of the decision-making life cycle for energy saving through BS traffic load prediction.

where the term γ is the most influential parameter as a factor selected to adapt the channel quality conditions to the QoS.

B. BS Server Opportunistic Computing

The BS is assumed to use the OpenFlow software defined networking (SDN) standard defined in [26], which is used to monitor network information and subsequently decide on best configurations to be applied in the entire CRN. The design objective is to maximize the CR system performance with respect to throughput while reducing infrastructure energy consumption. The high-level representation of traffic classification and decision-making life cycle is shown in Fig. 2 below. In Fig. 2 above, the circled SUs represent all the SUs in the cluster where the BS server operates using model-based reinforcement learning (RL), with an actuator that estimates the state of the processor using Markov decision processes (MDPs). The traffic streams from SUs consist of a variety of QoS requirements and the current load measurements are the inputs used by the actuator to make energy-efficient decisions and take subsequent transmission actions. This completes a model-based RL strategy whose process feeds the RL portion of the algorithm, which determines, in a single look-ahead, which possible scheduling scheme would provide the most effective energy saving, QoS, cost and response. For the balancing of QoS provisioning and energy saving, the BS server is assumed to employ MS on workload processing.

The optimization of the system performance and energy consumption involves performance specifications that are measured, such as the traffic load measurements and QoS requirements that form the system state $x \in X$. Therefore, as seen in Fig. 2 above, the BS has to make energy-saving decisions based on the following dynamic equation

$$\hat{x} = \phi\big(x(t), u(t)\big), \tag{6}$$

which describes a continuous-time non-linear input affine system, where $x(t)$ is the state of the system defined by the load-dependent power consumption $P(\upsilon, \rho, t)$, where υ denotes the BS switching mode, ρ is the system utilization, and the term $u(t)$ represents the control input to the system, which are the decisions made on which scheduling scheme to be used. In order to achieve the objective of the study, one must be cognizant that the main cause of high energy consumption in wireless networks is BS operation, whose process has proved difficult to optimize manually because of the complexity of the interactions of the equipment necessary for its operation.

IV. MATHEMATICAL PROBLEM FORMULATION

Since the rules and heuristics needed for every scenario that ensures everything from efficient RA to energy efficiency are difficult, particularly when interactions with the immediate environment are considered. In this paper, each of the parameters required to ensure energy-efficient RA will be formulated individually and will then be combined into a single formulation. Thus, the optimization problem for SU UL capacity maximization is formulated with the resource reservation technique whereby the BS dynamically reserves resources for SUs based on the number of PUs currently being served. The formulation of the optimization problem is as follows:

$$\mathbf{P1} = \arg\max_{\mathbf{P}} \sum_{j=1}^{J}\sum_{k=1}^{K} R_{j,k}, \quad \text{subject to} \tag{7}$$

$$\mathbf{C1} : \sum_{j=1}^{J}\sum_{k=1}^{K} P_{j,k} \leq P_{\max}, \quad \forall k \in K \tag{8}$$

$$\mathbf{C2} : Pr\left\{\sum_{k=1}^{K} P_{j,k}|g_{j,k}|^2 > I_{th}\right\} \leq \delta, \quad \forall k \in K \tag{9}$$

where $\mathbf{P}$ is the set of all individual UL transmission powers $P_{j,k}$, which for all $k \in K$, is limited by a power constraint $P_{\max}$ given in (8). This is the maximum allowed power, which is set individually for each SU owing to the power falloff of $d_{jk}^{-\alpha_{pl}}$ with distance d_{jk}. The constraint in (9) denotes that the probability that the interference threshold I_{th} is exceeded must not exceed δ.

A. Traffic Load and Power Consumption

Every BS activity has required power consumption implied in its energy consumption, thus it is assumed that the power consumed by the BS belongs to the following classes: $P_{\text{on}}(t)$, which represents the load-independent power consumption, $P_{\text{tx}}(t)$, which represents the load-dependent total transmission power, and $P_{\text{server}}(t)$, which is the load-dependent computational power consumption of the server. Therefore, the total BS power consumption is obtained as a combination of these classes as follows:

$$P(\upsilon, \rho, t) = \upsilon(t)P_{\text{on}}(t) + P_{\text{tx}}(t) + P_{\text{server}}(\rho, t), \tag{10}$$

where $\upsilon(t) \in \{\varepsilon, 1\}$, $\varepsilon \neq 0$ is the BS switching status indicator; 1 for active mode and ε for power-saving mode, $\rho(t)$ is the maximum server utilization factor at time slot t. In fact, ε is the operational, load-independent power consumption representing the normal BS power expenditure, which entails baseband processing, conversion, cooling, etc. The term $P_{\text{tx}}(t)$ represents the load-dependent total transmission power from the BS to the served SUs. The term $P_{\text{server}}(\rho, t)$ denotes the load-dependent computational energy consumption of the server, defined as follows:

$$P_{\text{server}}(\rho, t) = P_{\text{idle}}(t) + \rho(t)P_{\text{comp}}(t), \tag{11}$$

which describes the QoS part where the BS server dynamically adjusts its computational power based on the current traffic load and demand. The term $P_{\text{idle}}(t)$ is the server's load-independent operational component, and $P_{\text{comp}}(t)$ is the maximum power consumed by the server when operating at full power.

Assuming that the BS computational resources can be tuned, the term $P_{\text{comp}}(t)$ is linearly scaled with respect to $\rho(t)$. The term $\rho(t)$ denotes the slope of the trajectory that quantifies the load dependence. Therefore, $P(\upsilon, \rho, t)$ weighs the energy consumption due to the BS transmission and server computation. Since a single BS is considered, an optimization weight α is employed and the corresponding weighted cost function is defined as follows:

$$J(\upsilon, \rho, t) \triangleq \bar{\alpha}P\big(\upsilon(t), \rho(t), t\big) + \alpha(\varphi(t) - \rho(t)\big)^2, \tag{12}$$

where $\bar{\alpha} \triangleq 1 - \alpha$, $0 \leq \alpha \leq 1$, the quadratic term $(\varphi(t) - \rho(t))^2$ accounts for the QoS cost, where $\varphi(t) = \ell(t)/\ell_{\max}$ is the approximation of the normalized BS load at time slot t, as given in [39]. Hence, over the finite horizon $t = 1, \cdots, T$, the optimization problem is defined as follows:

$$\mathbf{P1}^* : \min_{\upsilon, \rho} J(\upsilon, \rho, t), \quad \forall t \in T \tag{13}$$

subject to

$$\begin{aligned} &\mathbf{C1}^* : 0 \leq \rho(t) \leq 1, \quad \mathbf{C2}^* : \upsilon(t) \in \{\varepsilon, 1\}, \\ &\mathbf{C3}^* : I_{\max} \geq I(t), \end{aligned} \tag{14}$$

then the vectors υ and ρ contain control actions for the considered time horizon T, i.e., $\upsilon = [\upsilon(1), \upsilon(2), \cdots, \upsilon(T)]$ and $\rho = [\rho(1), \rho(2), \cdots, \rho(T)]$. The constraint $\mathbf{C1}^*$ specifies the server utilization factor bounds, $\mathbf{C2}^*$ specifies the BS operation status, $\mathbf{C3}^*$ forces the required number of VMs, $I_{\max}$, always to be greater than or equal to a minimum number $I(t) \geq 1$.

B. Overall Optimization Problem

In order to achieve an optimal RA strategy, transmission collisions between PUs and SUs may result when the interference caused by SUs exceed δ, as shown in (9). In order to meet this condition, the RA problem has to be reduced to a bipartite matching problem. The advantage of this formulation is that globally optimal power allocations can be effectively computed for a variety of system-wide objectives and SU QoS constraints. Therefore, a single objective bipartite graph to realize the bipartite matching technique is constructed as illustrated in Fig. 3 below. In Fig. 3 above, on the left is the bipartite graph showing the whole matching process, while on the right is a graph edge showing individual allocation. In this illustration, the K SUs form a set $\mathcal{K} = \{1, \cdots, K\}$ and on the opposite side, the J PRBs form a set $\mathcal{J} = \{1, \cdots, J\}$; $\mathcal{K} \leq \mathcal{J}$. This constitutes a two-dimensional mapping problem where each SU may want a number of PRBs such that the inputs to the system are the number of SUs and the number of PRBs. Thus, during the RA process, it is paramount to consider certain parameters such as rate requirements and the weight a_{jk} of the link, which will be explained in detail in Section V.A. Therefore, since the spectrum resource is ξ, the formula for data rate R_{jk} can be formulated to

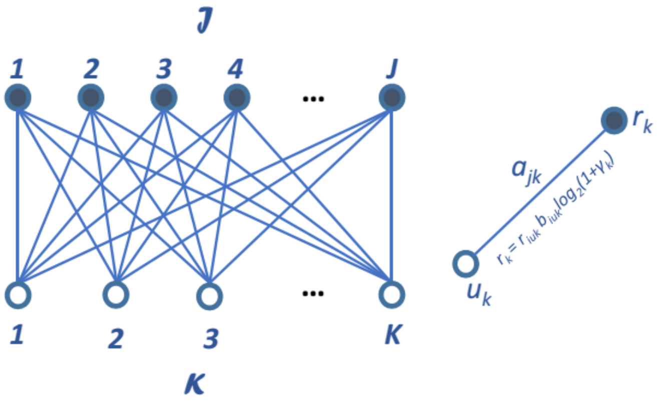

Fig. 3. Unconstrained weighted bipartite graph representation of resource allocation.

the system data rate as follows:

$$R_{sys} = \sum_{i=1}^{K} R_{j,k}. \tag{15}$$

In order to maximize (15), it is required that all the possible $K!$ combinations between SUs and PRBs are tried out. At this point, the weight of each connection link between the SU and the BS is calculated using a bipartite graph as follows:

$$w_{j,k} = \xi \left\{ J \times K \times \frac{R_{sys}}{\sum_{j=1}^{J}\sum_{k=1}^{K} R_{j,k}} \right\}, \tag{16}$$

where $J \times K$ is used to normalize the mean value of w_{jk}. The higher the value of w_{jk}, the higher the data rate attained, thus the optimal RA can be described by the following optimization problem:

$$\mathbf{P1}^{**} : \arg \max_{j \in \{1,2,\cdots,J\}} \left(\sum_{k=1}^{K} w_{j,k} \right) \tag{17}$$

subject to

$$\begin{aligned} &\mathbf{C1}^{**} : J \geq K, \quad \forall q_i, q_j \in \{1,2,\cdots,K\}, \\ &\mathbf{C2}^{**} : \{q_i, q_j\} \leq \{r_1, \cdots, r_J\}, \end{aligned} \tag{18}$$

where by assuming that the bipartite graph is perfectly symmetric, the original graph can be solved using a Hungarian matching algorithm.

V. PROPOSED SOLUTION FORMULATION

When an SU requests an UL connection, the BS checks that if by accepting the new SU connection its meets its admission control condition, availability of resources and interference constraints. Transmission resources are afforded to SUs only if these conditions are met, but mostly this RA is determined by the number of SUs already in the system. To obtain the new admission condition for a newly arriving SU, the number of SUs is increased by 1 and this new admission condition is compared with an admission bandwidth threshold, ξ^{su}, $\xi_{\text{new}} \geq \xi^{su}$, which is set to a minimum equi-spaced sub-channel per SU. This admission constraint is imposed to protect the integrity of PU transmissions; then the admission control probability ϕ for SUs within the BS coverage can be represented as follows:

$$\phi = \left\{ \begin{array}{ll} 1, & \xi_{\text{new}} \geq \xi^{su} \\ 0, & \text{Otherwise,} \end{array} \right\} \tag{19}$$

with $\xi_{\text{new}} = \xi\varsigma^{th}/(\lambda+1)$, where the denominator is the arrival rate λ in (4) increased by 1. Then, the total number of SUs attempting to connect to the BS can be given by

$$\lambda = \frac{\xi\varsigma^{th}}{\xi_{\text{new}}} - 1 = \xi\varsigma^{th}(\xi^{su})^{-1}. \tag{20}$$

So, if $0 \leq \frac{\xi\varsigma^{th}}{\xi_{\text{new}}} - 1 \leq \xi\varsigma^{th}(\xi^{su})^{-1}$, it means the BS is underloaded, thus the admission probability equals 1. If $\xi\varsigma^{th}/\xi_{\text{new}} - 1 > \xi\varsigma^{th}(\xi^{su})^{-1}$, the BS is overloaded and any new SU connection request will be rejected. This results in a new admission probability of $\xi\varsigma^{th}/(\lambda+1)$. In this case, the optimal transmission power and channel power gains per SU are obtained using GP.

A. Bipartite Matching Strategy for RA

This formulation constitutes a two-dimensional mapping problem where each SU may want a number of PRBs such that the inputs to the system are the number of SUs and the number of PRBs. Thus, during the allocation process, it is paramount to consider certain parameters, such as the weights of rate requirements, and a graph data structure is considered for this problem, as outlined below.

1) Denote the $\mathcal{K}$, $\mathcal{J}$ as bi-partition sets, $\mathcal{K} = \{1,2,\cdots,K\}$ and $\mathcal{J} = \{1,2,\cdots,J\}$ and consider them as two strictly independent sets, as discussed above.
2) Initiate labels u_j and v_j by $u_j = \max_{\forall i} w_{ji}$, which are supposed to support the SU data rates and label then according to $r_1, r_2, \cdots, r_J$, for the satisfaction of the k^{th} SU in terms of bandwidth demand b_i. Match rate requests $\{q_i, q_j\}$ to one or more of the available radio units, as illustrated in Fig. 4(a) below: This formulation is supported by Hall's theorem [40], where based on the SU traffic demands, RA will be performed by satisfying different design criteria, without loss of generality, q_i and q_j. In order to achieve RA for the the bandwidth demands q_i and q_j, an example of the solution is shown in Fig. 4 above, where the objectives are collected to form vectors of rate request, $\{q_i, q_j\}$ for SU traffic k, where the candidate PRBs must satisfy both criteria.
3) Construct a subgraph as in Fig. 4(b), which is a straightforward generalization of the bipartite graph matching in Fig. 3 to multi-dimensional matching, where Fig. 4(a) shows the RA using a bipartite graph and Fig. 4(b) is a solution of the bipartite matching problem by constructing a regenerative bipartite graph.
4) The weight $a_{kj} \in [0,1]$ is then added to the links connecting the SU to the corresponding PRBs, such that the number of PRBs consumed by each SU is represented as

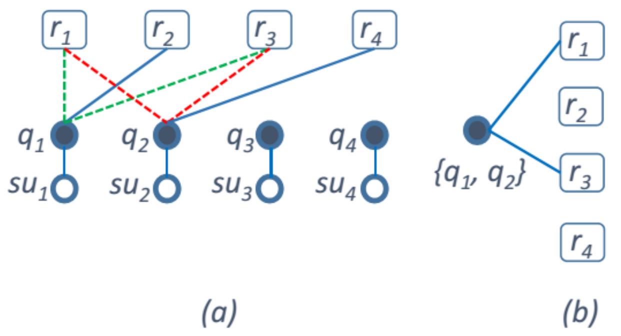

Fig. 4. Multi-objective oriented bipartite graphs and evolving of concepts to facilitate RA.

follows:

$$r_k = \sum_{j\in J} a_{kj} b_{iu_k}, \qquad b_{iu_k} = \frac{b_{\min}}{b_{\max}}, \tag{21}$$

where the term b_{iu_k} is the number of consumed PRBs, $b_{\min}$ and $b_{\max}$ are the minimum required long-term rate, which represents the rate QoS class requested by SU k and the maximum required rate, respectively.

5) The number of PRBs consumed is regarded as the load efficiency, which is the resource consumption efficiency. Using (16) and (21), this is given as

$$\rho(t) = \frac{R_{j,k}(t)}{r_k(t)}, \tag{22}$$

which is the long-term rate achieved by SU k.

B. *Efficiency of the Bipartite Matching Algorithm*

The bipartite matching algorithm, which is an efficient matching approach in graph theory, provides an optimal RA solution with a polynomial complexity of $\mathcal{O}(K^3)$ compared to the $\mathcal{O}(K!)$. Given a finite bipartite graph $\mathcal{G}$ with vertices that can be partitioned into two disjoint sets $\mathcal{K}$ and $\mathcal{J}$, all of its edges $\mathcal{E}$ connect a vertex in $\mathcal{K}$ to one in $\mathcal{J}$. A match is a subset of edges $K \subseteq \mathcal{E}$ no two members of which share a common vertex, where K is the cardinality of the node set of the bipartite graph.

For example, if data rates r_1, r_2, and r_3 satisfy demand q_i, and data rates r_1, r_3, and r_4 satisfy demand q_j, the request vector $\{q_i, q_j\}$ for SU_1 and SU_2 traffic only connects to r_1 and r_3. This is the intersection of sets formed by PRB matching both criteria for SU_1 and SU_2, as shown in the regenerative bipartite graph in Fig. 4(b). In this way, a regenerative bipartite graph is obtained for multi-objective RA, using a technique called preference requirement, which states that for any application that the SU is running, there is a preference requirement of data rate from the corresponding vertex in set $\mathcal{K}$ to set $\mathcal{J}$.

VI. MODELING THE PREDICTIVE CONTROL SCHEME USING DEEP LEARNING

Supposing that a naive method is applied at the current time t to predict the traffic load and then decide an appropriate trade-off between QoS provisioning and energy saving to be applied at time $t+1$. Referring back to Fig. 2, let the BS server be able to save the traffic profile for the previous and current time slots, $t-1$ and t, respectively, for it to be able to predict the traffic load for the time $t+1$ by curve-fitting and approximating its moving trend. Assuming that the BS's static energy and server utilization can be tuned to scale the server dynamic power consumption in proportion to the expected traffic load for the next time slot, it can make the necessary trade-off between QoS provisioning and energy saving.

Using the predicted value of the traffic load together with the values associated with the maximum BS capacity $\varphi(t)$, the energy consumption cost function $J(\upsilon, \rho, t)$ in (13) can be solved. To solve the cost function $J(\upsilon, \rho, t)$, an online management technique is employed whereby the cost function is treated as a Lyapunov candidate [45]. Therefore, using the Lyapunov technique, the BS processor is treated as a hybrid switching system where $J(\upsilon, \rho, t)$ is associated with a search of an optimum operating state, which when reached is maintained until parameters are updated. Then, a time series prediction technique can be applied to learn the function that maps a sequence of past observations as input to an output observation, but before that, the control actions that need to drive the time series prediction model have to be derived.

A. *Derivation of Control Actions*

In this subsection, using model-based RL and Lyapunov techniques, the control actions that will drive the time series prediction model are derived. Assuming that the control actions to drive the model in the next time slot $t{+}1$ are $\varrho(t) \triangleq (\upsilon(t), \rho(t))$, the system state vector, which contains the inputs to the time series model at time slot t, is denoted by $x(t) = (\varrho(t), \varphi(t))$. Here, the cost function $J(\zeta, \rho, t)$ is associated with reaching a certain state and maintaining it until the duration Δt elapses. This means that, in the sequel, the presentation $(\upsilon(t), \rho(t), t)$ is dropped to make way for $(x(t + \Delta T))$ in the sequel. Now, if the upcoming time slot is represented as $t{+}1 = t{+}\sum_{t=1}^{T} \Delta t$, which denotes the next transmission time interval where the predicted system state is $x(t+1)$, becomes the decision-making information state for the next time slot. Taking the optimal cost function $J(x(t))$ as a Lyapunov candidate and the input trajectory as $u(t)$, then $\Delta t = \sum_{t=1}^{t} \Delta t < T$, $1 \leq t \leq T$ is denoted as the time between two decision time steps, i.e., $d(t)$ and $d(t+1)$, as exemplified by the illustration in Fig. 5 below. With reference to the graphical representation of the auto-scaling in Fig. 5 above, the system faces serious concern when the load exceeds the 0.5 mark, which is indicated by the broken line. Assuming that the probability that the traffic flow information of the space-time points causes the future traffic flow, given the traffic load $\rho(t)$ at time step t, the decision-making step can be given as $\tau(t)$. The decision for the next transmission time is taken, the control $\varsigma(t)$ is applied at the beginning of the next time slot, whereas the offered load $\rho(t)$ is accumulated during the time slot and its value is only known at the end of each time slot. Then, it means that the decision on the next time slot $t+1$ is made at the end of time-slot t. The estimated system state for time-slot $t+1$ is

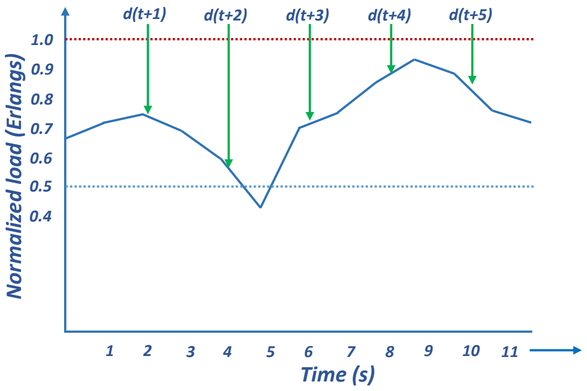

Fig. 5. Graphical representation of the time series prediction and decision steps.

given as follows (more details in Appendix A):

$$x(t+1)=\phi\big(x(t),\upsilon(t)\big), \tag{23}$$

where $\phi(t)$ is the behavior model that captures the relationship between $(x(t),\upsilon(t))$ and the next state $x(t+1)$.

VII. EXPERIMENTAL SET UP

The experimental set-up step, which is similar for the auto-regressive integrated moving average (ARIMA), the LSTM and the proposed SAE, are outlined in the following subsections.

A. Data Collection, Pre-processing and Normalization

In this section, a traffic flow data set was obtained using a traffic flow simulator in [42], aggregating it into data points separated by 1 second for 90 seconds (each for IN and OUT traffic data). The traffic streams for K SUs were represented in the form of a traffic flow matrix with a history of n time slots as follows:

$$\mathbf{S}^f=\begin{pmatrix}S_1\\S_2\\\vdots\\S_K\end{pmatrix}=\begin{pmatrix}S_1(t-n)&\cdots&S_1(t-1)\\S_2(t-n)&\cdots&S_2(t-1)\\\vdots&\vdots&\vdots\\S_K(t-n)&\cdots&S_K(t-1)\end{pmatrix}, \tag{24}$$

whose rows represent the historical traffic flow data during the previous n time slots. Thus, this traffic flow matrix, together with the system state and controls, is used as the input data for the predictive model to generate predictions for the time horizon T, i.e., $t, t+1,\cdots, t+T$.

B. Data Pre-processing

In this section, a multilayer perceptron with a nonlinear activation function, the logistic sigmoid, which has the squashing role in restricting from a node to (-1, 1), was used. This is represented using attention matrix $\mathbf{A}$, given as follows:

$$\mathbf{A}=\Phi\big(W(\mathbf{S}^s)+\zeta\big), \tag{25}$$

which can be interpreted as the probability that the traffic flow information of the space-time points causes the behaviour of future traffic flow. The term Φ is the sigmoid activation function between outputs and the hidden layers of the neural network, which is monotonic, continuous and differentiable, given as follows:

$$\Phi(b)=\Big(1+e^{\kappa(b_{\max}-b_{\min})}\Big)^{-1}, \tag{26}$$

where $\kappa\in[1,-1]$ is a constant that determines if the function is increasing or decreasing, and $b_{\min}$ is the QoS metric of a given service and determines the absissa shift of the function, whereby the absissa is the QoS metric $b_{\max}$, both given in (21). The term W is the weight that acts as a projection and vectorization of the speed matrix $\mathbf{S}^{\mathbf{s}}$ between the input and the hidden neurons, the speed matrix $\mathbf{S}^s$ has the same size as $\mathbf{S}^f$ in (24); and the term ζ is the bias variable without which a given layer will not be able to produce output in the next layer that differs from 0. Then, the traffic flow matrix $\mathbf{S}^f$ is point-wise multiplied by the attention matrix $\mathbf{A}$ to obtain a weighted traffic flow matrix $\mathbf{S}^{\mathbf{A}}$ for the application of the DL procedures. Then using feature mining, we obtain server instances of the applications. By server instances, we mean a set of features/attributes representing a specific occurrence of the problem.

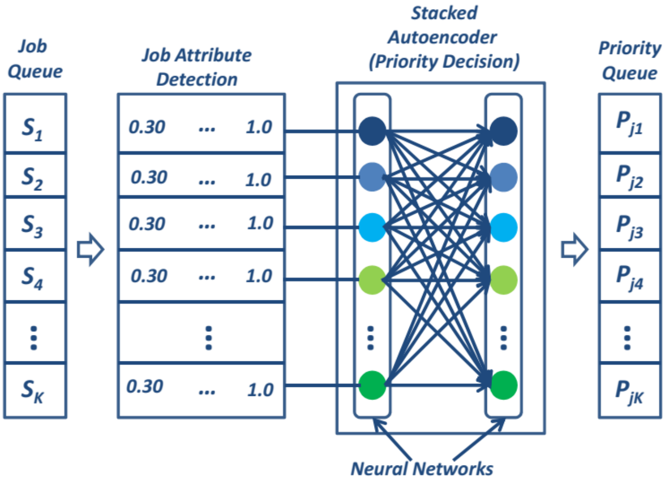

Fig. 6. SAE with job queue attribute detection and priority queue.

C. Training, Testing and Validation

During training and testing, the data set was split into 30% training and 70% testing data sets, respectively. The variations of the traffic load were aggregated in seconds on average of kilobits per second. Then, it was reshaped, services and jobs put into a job queue, extracted, processed and the service priorities made available in the priority queue and decisions for the next transmission slot were taken based on these priorities. The goal of the SAE was to use a speed matrix $\mathbf{S}^s$ of the space-time points of $\mathbf{S}^f$ to learn attention-weight matrix $\mathbf{A}$. Thus, given the traffic load $\rho(t)$ at time stamp t, the decision-making step given as $\tau(t)$ leads to a decision that should ensure that enough resources are allocated to serve the traffic until the next decision step. This is the number of virtual machine instances that need to be turned on to serve the traffic classes, such that the number of virtual machines allocated define the instance of a class. Here, a fully connected SAE consists of two encoders, each encoder consisting of a single hidden layer, as shown in Fig. 6 below. The service attributes are obtained from the job queue, i.e., resources (b_{iu_k}, $b_{\max}$), deadline, and processing time. The scheduler performs all sets of combinations of the application environment mentioned in the service attributes. All attributes per service per user are entered with calculated weights and assigned priorities. Assign the current value of $\rho(t)$ to a system state $x(t) = [\rho(t), q(t)]$ as the input vector that drives the system behavior at time t. Initialize the cost function $J(v, \rho, t)$ to zero; then begin a breadth-first search by building a tree of all future states up to the prediction depth T, i.e., $\hat{x}(t+1), \cdots, \hat{x}(t+T)$. Accumulate the cost as the search for future states travels through the tree, accounting for predictions and past outputs. Create a state-space $\mathcal{S}(t+n)$ using the set of states reached in every prediction depth $t+n$. For every prediction depth $t+n$, the search continues from the set of states $\mathcal{S}(t+n-1)$ reached at the previous step $t+n-1$, exploring all possibilities of obtaining the next system state. Update the accumulated cost as the result of the previous accumulated cost, plus the cost associated with the current time step $t+n$. When this exploration has finished, select the action at time $t+T$ that leads to the best final accumulated cost as the optimal operating value, as done in Appendix A.

D. Calculation of Performance Measures

Two ways of exploring the state space are used in this paper; one is the random technique, which uses the randomness of a random tree, while the other one uses the exploration technique from RL technique.

The **random technique** proceeds as follows: Let t be the current time, and $\rho(t+n-1)$ be the predicted traffic load in time slot $t+n-1, n = 2, \cdots, T$. It then performs a random prediction as follows: If the expected load difference $\rho(t+1) - \rho(t) > 0$ then the offered load in the next time slot is randomly selected in the range $[0.5, 1]$, otherwise it is selected evenly from the range $(0.0, 0.5)$. This is an exploration kind of behavior.

The **RL-based technique** substitutes the randomness of the exploration of the random tree with an exploitation technique from RL, which is a very naive technique that only cares when the traffic load exceeds 0.5. This means that it selects its prediction in the range $[0.5, 1]$, which saves a lot of time compared to the exploration technique.

VIII. PRIORITIZATION AND ENERGY CONSUMPTION PER SERVICE

Once the prediction horizon is fully explored, a unique sequence of reachable states $\hat{x}(t+1), \cdots, \hat{x}(t+T)$ with minimum cumulative cost is obtained. Two decisions have to be made when generating the class values; one prioritizes the QoS while the other prioritizes the energy saving. These are described as follows:

- **QoS prioritization:** This approach prioritizes QoS provision over energy saving, which means it allocates more resources to guarantee QoS. The allocation of more resources reduces energy efficiency, since more server instances will have to be launched at that given time. In this case, QoS requirements are guaranteed while energy saving suffers. In order to guarantee QoS, the decision $d(t)$ taken at time step t considers future traffic changes until time step $t+1$. The service class is generated as follows:

$$d(t) = \max(QoS(\lambda(t))), \forall t \in \{\tau(t), \tau(t+1)\}. \quad (27)$$

  The $QoS(\cdot)$ function takes the traffic load measured at time t as input and outputs $I(t)$, which is the number of virtual machines that are required to serve the measured traffic rates without violating the QoS.
- **Energy saving prioritization:** This approach assigns priority to energy saving by ignoring short-lived bursty traffic between steps t and $t+1$ to save on the energy consumption that comes with launching many virtual machines, thus accepting short-lived QoS degradation owing to the under-provision of virtual machines. The service class is generated as follows:

$$d(t) = \max\Big(QoS\big(\lambda(\tau(t))\big), QoS\big(\lambda(\tau(t))\big)\Big). \quad (28)$$

This gives the number of processing units by specifying the scheduling mechanism, as the idea is to concentrate the computational resources by choosing the appropriate scheduling mechanism for each service that will not drain too much of the systems' energy. Thus, for energy consumption per application, the following equation is used

$$EE(v, \rho) = \frac{v(t)P_{\text{on}}(t) + P_{\text{tx}}(t)}{\lambda} + \rho \cdot P_{\text{comp}}(\rho, t), \quad (29)$$

where λ is the packet arrival rate, $\rho = \lambda\varphi(t)\mathbb{E}[s]$ is the load factor, $\varphi(t)$ is the normalized traffic load, $\mathbb{E}[.]$ is the mean service rate of the server, and s is the service time per job. Equation (29) supports the principle of concentrating computation on a small number of processing units in order to minimize the server power consumption per application [52]. A simple analytic model that uses the combined energy-QoS cost function includes in its first part the well-known Pollaczek-Khintchine formula for $M/G/1$ queue [48] for the average response time, based on Poisson arrivals of jobs and general service time distributions, and in its second part the energy consumption per job, given as follows:

$$C(v, \rho) = a\mathbb{E}s\left[1 + \frac{\rho(1 + C_s^2)}{2(1-\rho)}\right] + b\left[\frac{v(t)P_{\text{on}}(t) + P_{\text{tx}}(t)}{\lambda} + m\mathbb{E}[s]\right]. \quad (30)$$

Here, $C_s^2 = \sigma_s^2/(1/\mu)^2$ is the squared coefficient of variation of service time, where σ_s^2 is the variance of the service time and μ is the mean service rate; constants a and b describe the relative importance placed on QoS and energy consumption, respectively. This allows us to compute the value of the arrival rate that minimizes $C(v, \rho)$. The result in (31) indicates the optimum setting of the load $\rho^* = \lambda^*\mathbb{E}[s]$ and its dependence on $v(t)P_{\text{on}}(t) + P_{\text{tx}}(t)$ and on the ratio b/a given as follows:

$$\rho^* = \frac{\sqrt{\frac{2b(v(t)P_{\text{on}}(t)+P_{\text{tx}}(t))}{a(1+C_s^2)}}}{1 + \sqrt{\frac{2b(v(t)P_{\text{on}}(t)+P_{\text{tx}}(t))}{a(1+C_s^2)}}}. \quad (31)$$

Algorithm 1: Algorithm to exhaustively evaluate operating states within the prediction horizon T

```
Input: Current state $x(t)$, Prediction horizon $T$
Output: $\varsigma(t)$, $EE(\upsilon,\rho)$, $C(\upsilon,\rho)$
Initialize all inputs: $\mathcal{S}_t = x(t)$
Assign $J(\hat{x}(t)) \leftarrow J(\hat{x}(t-1))$
For $t = 1 : T$ do
    Set next state $\mathcal{S}_{t+1} \neq \emptyset$
    For $x \in \mathcal{S}_t$ do
        Predict environment parameters for $t+1$
        Estimate the next state, $\hat{x} = \phi(x,u)$
        Set $\mathcal{S}_{t+1} = \mathcal{S}_{t+1} \cup \{\hat{x}\}$
        Find $x_{\min} \in \mathcal{S}_T$ with minimum cost $J(x)$
        (i.e., $\hat{u}(t) =$ input leading from $x(t)$ to $x_{\min}$
        If $J(\hat{x}) = J(x) - J(\hat{x})$ do
            Execute QoS priority using (27)
        Else
            Execute energy saving using (28)
        End If
        Compute service cost using (31)
    End For
    Predict the cost $J(\hat{x})$ using (34)
    Return $\varsigma$(t)
End For
```

Table 1. CRN traffic profiles considered in this paper.

CRN applications' traffic profiles			
Application/Job	**Des. delay**	$b_{\min}$ **bps**	$b_{\max}$ **bps**
Audio or VoIP service	180 ms	30 K	64 K [43]
Online gaming	150 ms	1 M	4 M [50]
Buffered video	2 ms	3 M	25 M [51]
Video conferencing	300 μs	256 K	20 M [49]

Equation (31) gives us a simple rule of thumb for selecting system load for optimum operation, depending on how we weigh the importance of energy consumption with respect to average response time or how fast we are getting the jobs done. We also see that ρ^* increases at the ratio $(b(\upsilon(t)P_{\text{on}}(t) + P_{\text{tx}}(t)))/a(1 + C_s^2)$. This tells us that the optimum load should increase with the expression $\upsilon(t)P_{\text{on}}(t) + P_{\text{tx}}(t)$ of the system, the relative importance that we place on energy, and with the squared coefficient of variation of service time. The algorithm for this process is shown in **Algorithm 1**:

IX. SIMULATION RESULTS AND DISCUSSION

To validate our main findings of this paper, a series of simulations were conducted using $MATLAB^{TM}$. The services considered are tabulated in Table 1 below. The features tabulated in Table 1 above are used to compute the resource consumption efficiency for each service run by the CR devices using (21). Simulation parameters are as shown in Table 2 below.

Table 2. Simulation parameters.

Simulation parameter	**Value**
Carrier frequency, f_c	2.1 GHz
System bandwidth, ξ	8 MHz
Maximum number of RBs, J	100
Noise power spectral density, σ	2×10^{-11} watts/Hz
BS operating power, P_{on}	40.25 dBm [10.6 W]
BS transmission power	46 dBm [40 W]
Dynamic maximum power, P_{serv}	56.7 dBm [472.3 W]
Energy consumption at P_{idle}	3 J
Path-loss model	$34.46 + 20\log_{10}(d_{jk})$
Symbol duration, T_s	500×10^3 symbols/sec
Maximum packet arrival rate, λ	60 packets/sec
Min. and Max. BS load, φ	$[5, 10]$ MB
Maximum decision interval, Δt	1 sec
Minimum number of VMs, $I(t)$	1
Maximum number of VMs, $I_{\max}$	30
Number of input layers	1
Activation function, hidden	Sigmoidal
Learning rate	0.3
Number of output layers	1

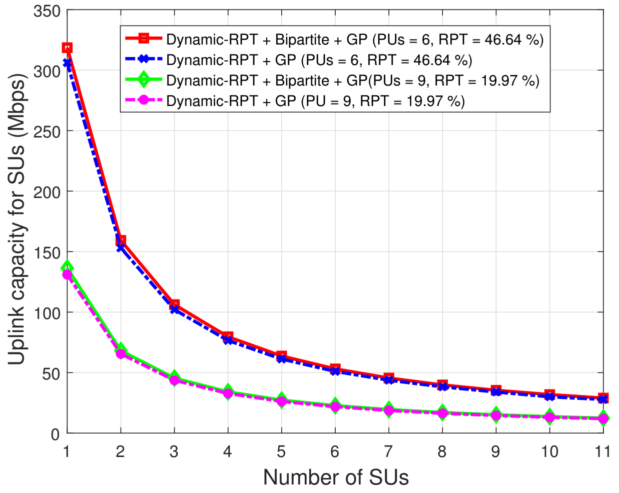

Fig. 7. UL achievable capacity per SU.

A. Resource Reservation and Allocation

In this subsection, it is considered that some resources are left for SUs using the RPT, which gives the amount of resources remaining for SUs when PUs are active in their channels. The bandwidth required by PUs is calculated and their required capacity is set to find the bandwidth percentage to be allocated to SUs, which gives the RPT. After the RPT has been calculated, the performance on the achievable capacity is evaluated using two variations of the algorithm; (i) Dynamic-RPT + Bipartite matching and GP and (ii) Dynamic-RPT + GP for six and nine PUs in the CRN. The achievable capacity and the percentage decrease in RPT as a function of an increasing SINR are illustrated in Fig. 7 and Fig. 8 below, respectively.

Fig. 7 above shows results for the RPT for SUs, given the number of PUs and resources allocated to them. It can be ob-

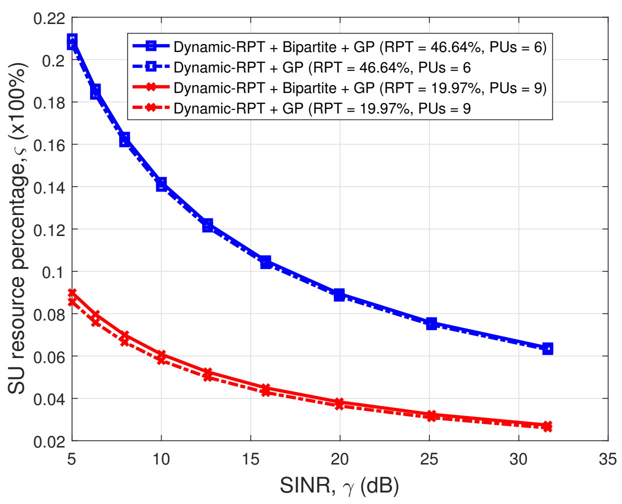

Fig. 8. Resource percentage threshold vs. SINR.

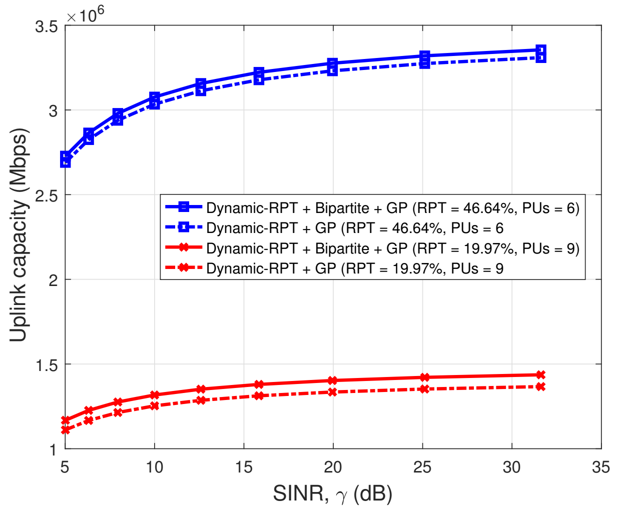

Fig. 9. UL achievable capacity per SU vs. SINR.

served that the achievable UL capacity for SUs decreases with an increase in the number of SUs being admitted, which suggests that their performance might be limited by interference from other users (PUs and SUs). However, by combining the dynamic resource percentage threshold with bipartite matching achieves better performance through the GP solution obtained in (7). It is shown that as the number of PUs is increased from six to nine, the RPT reduces from 46.64% to 19.97%, which is a 26.67% difference. With six PUs and only one SU admitted into the system, the achievable capacity is 318.56 Mbps and 306.16 Mbps at an RPT of 46.64%, which indicates that Dynamic-RPT and bipartite matching performs 3.54% better than Dynamic-RPT alone. As SUs are admitted into the system, the achievable capacity decreases such that at 11 SUs, the achievable capacity is 28.96 Mbps and 27.83 Mbps, which is a 0.3% difference between the two algorithms. When the number of PUs is increased to nine, the achievable capacity is 136.36 Mbps and 131.05 Mbps with one SU in the system. However, there is a difference of only 0.483 Mbps between the two algorithms at 11 SUs. With nine PUs, more asymptotic behavior is observed, as the number of SUs increases, which suggests that the RPT algorithm always allows for resources to be reserved.

Fig. 8 above shows the decline in the RPT as the SINR γ increases, and what can be observed is that the RPT decreases more rapidly when there are fewer PUs in the system than when there are more PUs. This observation can be explained as follows: with an increase in the number of available channels and good channel conditions, the probability of SUs' channel access increases, which also increases the possibility of SU collisions. It is a problem that still needs to be addressed. Furthermore, with a decrease in the number of available channels, even if the channel conditions improve, there are fewer transmission opportunities for SUs due to PU activities and an increase in the collision intensity among SUs. In the same fashion, the achievable UL capacity per SU as the SINR increases is evaluated in Fig. 9 below. As shown in Fig. 9, the performance of the UL capacity per SU as a function of an increasing SINR γ; however, there is a performance difference, as the number of PUs is increased from six to nine. As in the case of the performance shown in Fig. 7, it is shown here that the achievable UL capacity per SU decreases. However, the difference between the two algorithms is less when the number of PUs is kept constant. This shows that even though SUs might be expected to exploit much of the available capacity and achieve higher transmission throughput, this is affected by the number of available PUs. This can be explained using the response of the effective system bandwidth as a function of an increasing SINR, which is shown in Fig. 10 below. Fig. 10 above shows the response of the system bandwidth as a function of an increasing SINR γ for the two algorithms and changing number of SUs. The system bandwidth decreases drastically with an increase in PUs from six to nine. This result serves to confirm an old belief that SUs can have higher instantaneous throughput when there are more available channels, but the channel switching that takes place when PUs re-appear to claim some of their channels and the PU activities together with the contention among SUs create high liability, since there might not be available channels for some SUs.

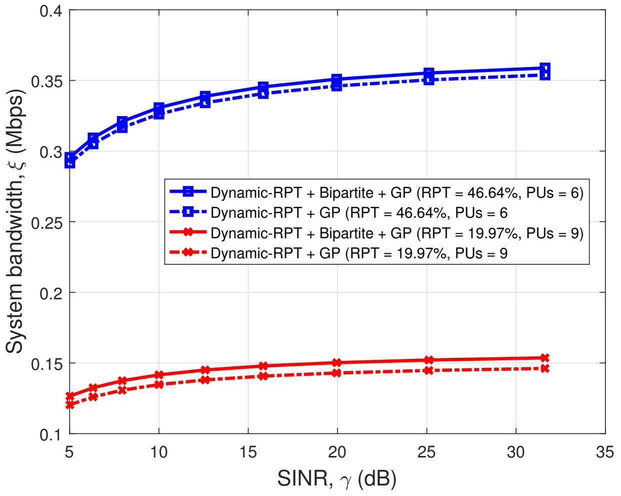

Fig. 10. System bandwidth vs. SINR.

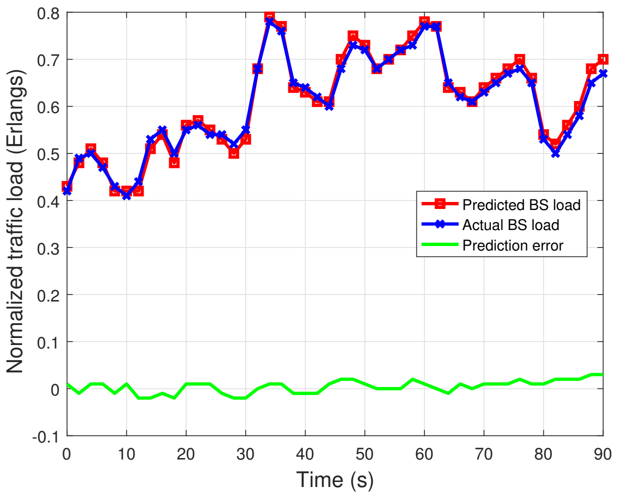

Fig. 11. BS load prediction using an SAE architecture.

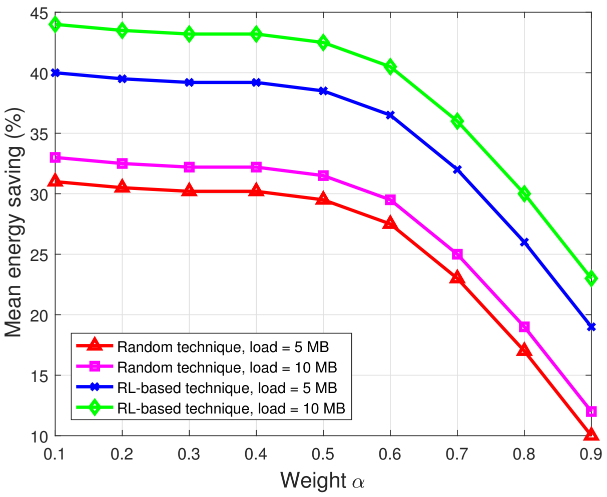

Fig. 12. Mean energy saving as a function of optimization weight α using SAE.

B. BS Traffic Load Prediction

Table 3, Table 4 and Table 5 below present a convergence comparison of the training and validation for three NN architectures used for forecasting/prediction purposes. All three prediction schemes are trained over 110 epochs, each epoch consisting of 2,000 individual training trials, and the RMSE is used as a performance measure of the models. The training and validation results of the ARIMA in performing the same prediction task are shown in Table 3. The training shown in Table 3 was performed over 110 epochs and the training convergence was observed at a loss of 0.11. The same performance evaluation was done for the LSTM, as shown in Table 4.

In Table 4 it can be observed that the LSTM begins with an average training error of 0.4036 (RMSE), which is better than that obtained using the ARIMA, i.e., 0.5094 (RMSE); it is a 10.58% improvement over ARIMA. However, at the end of the training only a 1.15% difference is observed. Compared to the ARIMA and LSTM, the SAE architecture performs well with an initial training error of 0.3115 (RMSE), which is 19.79% and 9.21% superiority over ARIMA and LSTM, respectively. Therefore, compared with the other two architectures, in the sequel, the SAE becomes the architecture of choice because of the lowest training error. The training of the SAE in traffic load prediction is performed where a careful choice of learning rate is made to be α = 0.3 to make the SAE training more reliable. Even though the training and subsequent optimization might take a little longer than it would with a higher learning rate, this is a better choice of the learning rate. A training rate higher than this one proved to be faster, while the convergence was poor because the weight updates became big, such that the optimizer overshot the minimum and made the training loss worse. The results for BS load prediction are illustrated in Fig. 11 below. As shown in Fig. 11, the BS load pattern prediction results were based on the SAE architecture for a time horizon of $T = 90$ seconds. The prediction error shows some stabilization as the training time increases, which indicates reliability of the training results to be obtained.

C. Energy Saving and QoS Provisioning

When the SAE is use in energy saving with respect to the optimization weight, α, the RL-based technique is compared with the random technique with the load varied from 5 MB and 10 MB. The performance results are shown in Fig. 12 below.

As shown in Fig. 12, the performance of energy saving is considered in terms of the traffic load with respect to the optimization parameter α. In the range $\alpha = [0.1, 0.4]$, the impact of the traffic load is clearly visible: as the load is increased (i.e., 10 MB), more energy is being saved compared to when there is less traffic load (i.e., 5 MB). This shows that as more load is allocated to each VM, energy saving increases owing to the reduction in the number of VMs required to be turned on. This means that concentrating the load in fewer VMs saves more energy than when the load is distributed over more VMs, i.e., when the load is 5 MB. Another notable observation is that there is more gain in energy saving with the RL-based technique compared to the random technique when the load is increased from 5 MB to 10 MB. However, in the range $\alpha = [0.4, 1]$, the energy saving drops for both algorithms, which indicates that as $\alpha \rightarrow 1$ the emphasis is placed on QoS than on energy saving and the system can allow for short-lived drops in energy saving.

D. Effect of Traffic Load on Server Response Times

In this subsection, the evaluation is based on the effect of the traffic arrival on server response times and server energy consumption, which in principle means concentrating on the server computational units' energy consumption. The results of the server energy consumption per job processed are illustrated in Fig. 13 below.

Fig. 13 above illustrates the server energy per packet vs. packet arrival rate λ as the traffic load ρ and service rate μ are varied. This evaluation is, in principle, concentrating on the server computational units' energy consumption. At first, the traffic load is varied, while the service rate is kept constant; then the traffic load is kept constant while the service rate is varied. In both cases the energy per packet decreases as the packet arrival rate increases. This observation can be explained by the

Table 3. Training and validation results for the ARIMA.

ARIMA											
Training epoch	10	20	30	40	50	60	70	80	90	100	110
Training loss	0.509	0.406	0.331	0.254	0.206	0.159	0.131	0.122	0.119	0.113	0.113
Validation loss	0.462	0.332	0.297	0.244	0.203	0.153	0.129	0.121	0.119	0.113	0.113

Table 4. Training and validation results for the LSTM.

LSTM											
Training epoch	10	20	30	40	50	60	70	80	90	100	110
Training loss	0.404	0.318	0.265	0.209	0.187	0.136	0.122	0.120	0.109	0.101	0.101
Validation loss	0.381	0.303	0.249	0.186	0.161	0.124	0.112	0.111	0.108	0.101	0.101

Table 5. Training and validation results for the SAE.

SAE											
Training epoch	10	20	30	40	50	60	70	80	90	100	110
Training loss	0.312	0.282	0.228	0.171	0.149	0.116	0.111	0.108	0.107	0.100	0.100
Validation loss	0.281	0.253	0.205	0.151	0.140	0.118	0.111	0.108	0.107	0.100	0.100

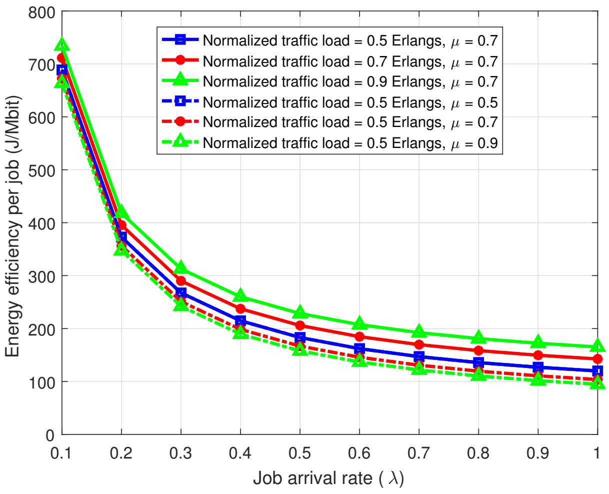

Fig. 13. Server energy consumption per processed job as a function of packet arrival rates.

fact that in the low packet arrival rate mode, the system expends more energy in switching ON the servers' computational units, leading to higher energy consumption. When the service rate is kept constant and the traffic load is increased, the energy per packet is more sensitive to the increase in packet arrival rate. However, when the service rate is varied and the traffic load is kept constant, the energy per packet decreases and is less sensitive to the increase in packet arrival rate. This shows that the server's energy consumption is more responsive to the traffic load than it is to the service rate. This observation is consistent with the results found in [52].

Table 6 and Table 7 show the energy consumption per packet vs. normalized traffic load as a function of the emphasis placed on QoS and energy saving, respectively. Table 6 shows the performance of the system when a, which is the importance placed on QoS, is varied and b, which is the importance placed on energy consumption, is kept at a minimum of 0.1; the energy consumption is low. This attests to the result obtained in (31), which shows that the optimum setting of the load $\rho^* = \lambda^* \mathbb{E}[s]$ will depend on $v(t)P_{\text{on}}(t) + P_{\text{tx}}(t)$ and on the ratio b/a.

Table 7 shows the performance of the system when the emphasis is placed on energy consumption b and the QoS priority is at its minimum. These results illustrate that when the emphasis is placed on QoS, the energy consumption will increase, since the system needs to guarantee QoS by allocating more resources, thus increasing energy consumption and reducing on energy saving. From these results, it can be noted that when the priority is the QoS, the energy saving decreases as the traffic load increases. This is because when the traffic load increases, the resource consumption efficiency increases as the computational resources have to stay ON for a long time, which increases energy consumption, thus reducing energy saving. However, when the priority is shifted to energy saving, energy consumption is low, as the system ignores the increase in traffic load and allows some minor QoS degradation. Therefore, fixing one parameter, either a and varying b or vice versa allows us to scale the system's response time and the energy consumption per application.

E. *Cost of Energy Consumption with Server Mean Slowdown*

Here, the variation in energy consumption cost as a function of the mean arrival rate λ for both FCFS and PS scheduling mechanisms with server MS is evaluated. The number of channels and the price per joule of energy are kept constant, while the mean service rate μ is varied. The traffic profiles in Table 1 are regarded as the QoS parameters, where $b_{\max}$ represents the size of the request, which is the maximum required bandwidth, and $b_{\min}$ represents the benchmark minimum required bandwidth as defined in (21).

Table 6. Energy consumption per service, varying a and keeping b at its minimum.

Energy consumption per packet (Joules/bit)									
Traffic load	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
a = 0.6, b = 0.1	63.5562	31.9717	21.4538	16.1872	12.9900	10.7812	9.0752	7.6060	6.2019
a = 0.7, b = 0.1	63.6072	32.0253	21.5100	16.2433	13.0400	10.8141	9.0746	7.5487	6.0569
a = 0.8, b = 0.1	63.6583	32.0789	21.5662	16.2995	13.0900	10.8471	9.0740	7.4914	5.9118
a = 0.9, b = 0.1	63.7093	32.1325	21.6224	16.3557	13.1400	10.8801	9.0735	7.4340	5.7668

Table 7. Energy consumption per packet, keeping a at its minimum and varying b.

Energy consumption per packet (Joules/bit)									
Traffic load	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
a = 0.1, b = 0.6	379.551	189.954	126.756	95.156	76.190	63.533	54.471	47.643	42.288
a = 0.1, b = 0.7	442.801	221.604	147.873	111.006	88.880	74.116	63.549	55.593	49.361
a = 0.1, b = 0.8	506.051	253.254	168.990	126.856	101.570	84.700	72.628	63.543	56.433
a = 0.1, b = 0.9	569.301	284.904	190.106	142.706	114.260	95.283	81.707	71.493	63.505

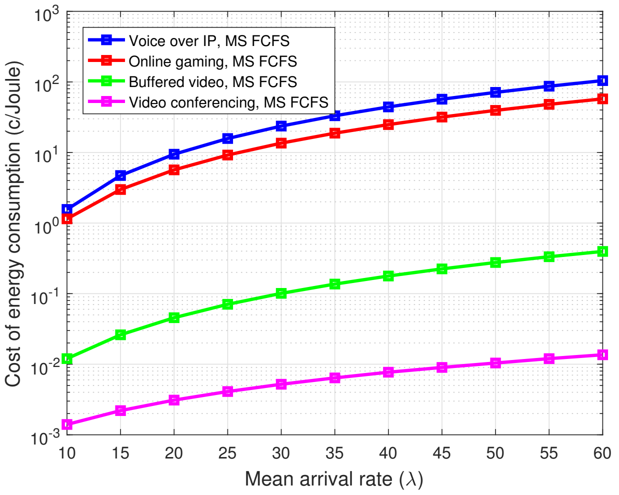

Fig. 14. Cost of energy consumption as a function of mean arrival rates with $\mu = 0.6$.

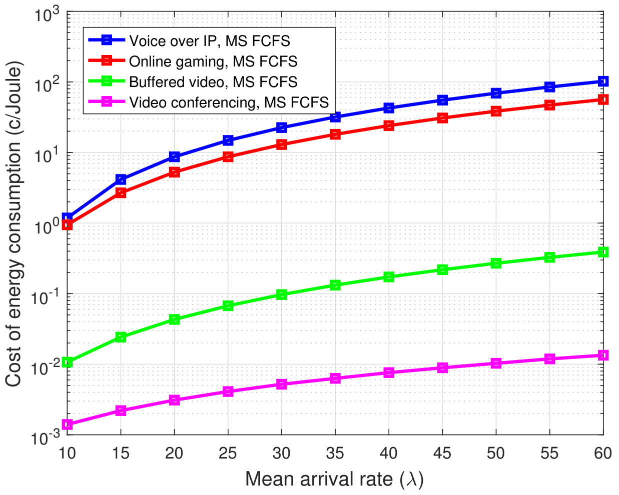

Fig. 15. Cost of energy consumption as a function of Mean arrival rates with $\mu = 0.7$.

E.1 Mean Slowdown First-come-first-served

The performance evaluation results for the MS FCFS scheduling mechanism for processing rates $\mu = 0.6$, $\mu = 0.7$, and $\mu = 0.8$ are shown in Figs. 14, 15, and 16, respectively.

In Figs. 14, 15, and 16 above, there is a noticeable increase in the energy consumption, since as the rate of job arrivals increase, more energy is expended in pushing them out of the system within their deadlines. The video conferencing application has the lowest cost on the system compared to the other applications, owing to the difference between the $b_{\min}$ and $b_{\max}$, as seen in Table 1. A greater difference between these values gives a large value of $b_{i_{u_k}}$ and the lower the value of the resource consumption efficiency becomes as can be seen in (22). However, the cost of energy consumption decrease by 0.43% when the service rate is increased from $\mu = 0.6$ to $\mu = 0.8$. Thus, for an FCFS scheduling scheme with server MS, the cost of energy consumption varies inversely with the service rate, which shows the variation in the cost of energy consumption as a function of arrival rate for the MS FCFS scheduling mechanism. The effect of increasing the arrival rate on energy consumption is investigated, keeping the number of VMs, service rate, and size of requests constant.

E.2 Mean Slowdown Processor Sharing

The performance evaluation results for the MS PS scheduling mechanism for processing rates $\mu = 0.6$, $\mu = 0.7$ and $\mu = 0.8$ are shown in Figs. 17, 18, and 19, respectively. When the scheduling scheme used is the MS PS, the cost of energy consumption is reduced, as shown in Fig. 17 below.

Figs. 17, 18, and 19 above show the variation in the cost of energy consumption as a function of an increasing arrival rate for the PS scheduling mechanism with MS. The energy consumption and its cost increase as the mean arrival rate increases, but it is not sensitive to a changing processing rate, hence the performance remains the same. It can be observed that there is substantial energy saving with the MS PS compared to the MS

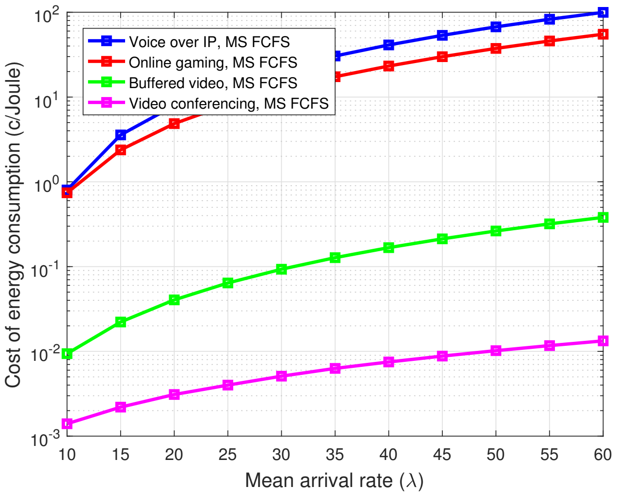

Fig. 16. Cost of energy consumption as a function of mean arrival rates with $\mu = 0.8$.

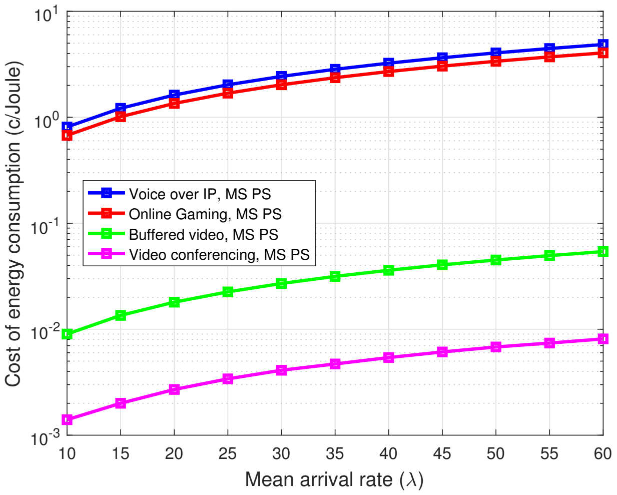

Fig. 17. Cost of energy consumption as a function of mean arrival rates with $\mu = 0.6$.

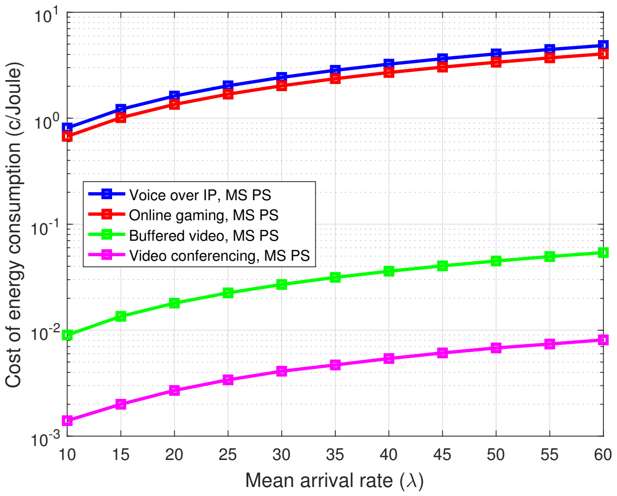

Fig. 18. Cost of energy consumption as a function of mean arrival rates with $\mu = 0.7$.

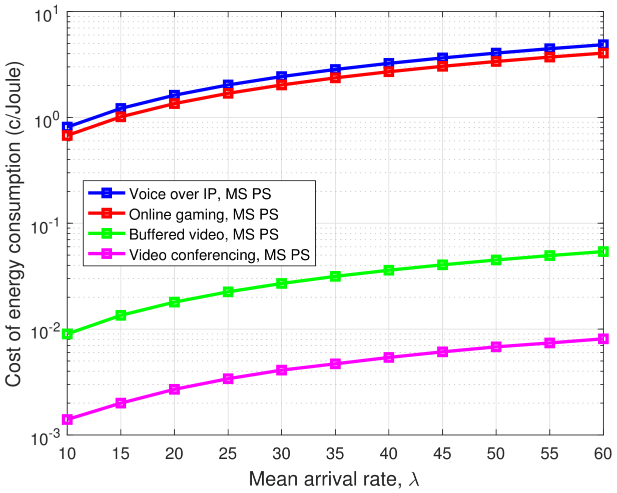

Fig. 19. Cost of energy consumption as a function of mean arrival rates with $\mu = 0.8$.

FCFS, with MS PS achieving higher energy saving without the cost of energy consumption increasing much as the mean arrival rate increases. This is because there is almost no queuing delay with the PS system, since all packets are served simultaneously such that the traffic load as the incoming traffic rate is never greater than its outgoing capacity. Except for the fact that MS is more energy efficient, which applies to both FCFS and PS, PS is an even more efficient scheduling scheme, since it concentrates job processing within a few computational units instead of distributing the workload across all processing units. With a few VMs commissioned, less energy is consumed and hence there is substantial energy saving when the MS PS scheduling mechanism is used compared to MS FCFS.

Detailed results for energy saving between the two workload scheduling schemes are tabulated in Table 8 and Table 9 below.

The results shown in Table 8 and Table 9 show that a decrease in the b_{iu_k} is also a decrease in the percentage energy saving. This can be explained based on the fact that as the difference between the required long-term resources $b_{\min}$ and the the maximum required bandwidth becomes large, more packet processing units are required. However, one striking observation is that as the packet arrival rates increase, the system energy efficiency also increases. This means that the energy consumption per packet decreases, as observed in Table 6 and Table 7. However, in this case it is the packet-service discipline that improves the performance. Regarding packet service discipline, the central issue that has been addressed is the notion of fairness, by realizing that different packets consume energy differently. In essence, the question is what packet-service discipline is deemed energy-fair. The answer to this question remains in the server itself, where traffic flows require different services and the scheduling properties that are required. The MS PS packet-service discipline offers significant energy savings that MS FCFS as justified by a study in [53].

Table 8. Percentage energy saving using MS FCFS scheduling mechanism.

FCFS with server slowdown												
λ Pkts/sec	10	15	20	25	30	35	40	45	50	55	60	
VoIP	0.118	0.414	0.869	1.482	2.253	3.182	4.269	5.515	6.919	8.481	10.201	%
Online gaming	0.095	0.268	0.527	0.870	1.296	1.807	2.403	3.083	3.848	4.696	5.630	%
Buffered video	0.001	0.002	0.004	0.007	0.010	0.013	0.017	0.022	0.027	0.033	0.039	%
Video conferencing	0.000	0.000	0.000	0.000	0.001	0.001	0.001	0.001	0.001	0.001	0.001	%

Table 9. Percentage energy saving using MS PS scheduling mechanism.

PS with server slowdown												
λ Pkts/sec	10	15	20	25	30	35	40	45	50	55	60	
VoIP	0.810	1.215	1.620	2.025	2.430	2.835	3.240	3.645	4.050	4.455	4.860	%
Online gaming	0.675	1.013	1.350	1.688	2.025	2.363	2.700	3.038	3.375	3.713	4.050	%
Buffered video	0.009	0.014	0.018	0.023	0.027	0.032	0.036	0.041	0.045	0.050	0.054	%
Video conferencing	0.001	0.002	0.003	0.003	0.004	0.005	0.005	0.006	0.007	0.007	0.008	%

X. CONCLUSION

In this paper, a QoS provisioning and energy saving scheme for single BS management where CRs run real-time services was proposed, with design expectations in line with the IMT-2020 and beyond requirement specifications. Since QoS constraints need to be satisfied, energy efficiency becomes a critical concern and energy management issues need to be addressed from a single BS perspective. Firstly, a distributed dynamic RA based on a resource reservation protocol was proposed to reserve resources to give SUs a high probability of completing their transmissions. The optimal RA solution was obtained through weighted bipartite matching with a polynomial complexity of $\mathcal{O}(K^3)$ compared to the $\mathcal{O}(K!)$ of integer programming. Resource consumption efficiency obtained through bipartite matching was then used to solve as weighted cost function in which power consumption was added together with different weights reflecting their contribution to BS power consumption. Using the derived control actions, workload arrival patterns were estimated and used by the DL model to predict the future behavior over a finite horizon, T. Using the output of the SAE, which is simply a regression between the previous and current system states, appropriate packet processing schemes were chosen between MS FCFS and MS PS. The simulation results obtained indicate that the predictive control achieves better energy saving as the traffic load increases, owing to a few VMs commissioned to serve the traffic load. This shows that concentrating the workload on a few computational resources saves energy owing to fewer VMs being turned on. Moreover, using the MS PS packet processing achieves 6.89924% more energy energy saving compared to MS FCFS. This shows that communication systems with PS represent adequate models for resource sharing, e.g., the bandwidth of communication systems, and can be adopted for balancing QoS provisioning and energy saving in future mobile and wireless network design.

APPENDIX A SOLUTION OF THE COST FUNCTION

Let $x(t+\Delta t)$ be the actual state when the sample-and-hold controllers $\{u(t),\cdots,u(t+\Delta t)\}$ are applied. Moreover, let $J(x(t+\Delta T))$ be the optimal cost obtained by solving (13) based on the new current state $x(t+\Delta T)$, provided the current cost function has decreased. Then, the condition that determines the next transmission time $t+1$ is obtained by checking if the optimal cost (i.e., the current cost function) regarded as a Lyapunov candidate is guaranteed to decrease, i.e.,

$$J\bigg(x(t+\Delta T)) - J(x(t)\bigg) < 0. \tag{32}$$

For more details in deriving this condition, considering **Lemma 3** in [46], the following result holds:

$$J^*\big(x^*(t+\Delta T)\big) - J^*\big(x(t)\big) \leq -\int_t^{t+\Delta T} \phi\big(x^*(s),u^*(s)\big)ds, \tag{33}$$

where $J^*(x^*(t+\Delta T))$ is the optimal cost obtained by solving (13) if the current state at $t+\Delta T$ is $x^*(t+\Delta T)$. This means that the optimal cost would be guaranteed to decrease if the actual state followed the optimal state trajectory $x(s)=x^*(s)$ for $s\in[t,t+\Delta T]$. From (33), we obtain

$$\begin{aligned} &J^*\big(x^*(t+\Delta T)\big) - J^*\big(x(t)\big) \leq J^*\big(x(t+\Delta T)\big) \\ &- J^*\big(x^*(t+\Delta T)\big) - \int_t^{t+\Delta T} \phi\big(x^*(s),u^*(s)\big)ds, \end{aligned} \tag{34}$$

where $\phi(x^*(s),u^*(s))$ (as in (6)) is known at t when the solution is obtained. Once solved, the control action that needs to be applied in the next time-slot, $t+1$, is $\varsigma(t)\triangleq(\upsilon(t),\rho(t))$. Because the next time-slot is taken as the next transmission time, it is also the next decision time. The decision on the trade-off between QoS and energy saving is taken based on the value of the optimal cost function $J^*(x^*(t+\Delta T))$ computed for the current state $x^*(t+\Delta T)$ at time $t+\Delta T$. Then the system state vector is denoted $x(t)=(I(t),E(t))$, which contains the number of available VMs and the energy levels. Therefore, $\varsigma(t)\triangleq(\upsilon(t),\rho(t))$

is the vector that determines the system behavior at time slot t such that the system evolution is described using the discrete-time state-space equation in (23).

REFERENCES

[1] E. J. Kitindi, S. Fu, Y. Jia, A. Kabir, and Y. Wang, "Wireless network virtualization with SDN and CRAN for 5G networks: Requirements, opportunities, and challenges," *IEEE Access*, vol. 5, pp. 19099–19155, Sept. 2017.

[2] R. Deng, J. Chen, C. Yuen, P. Cheng, and Y. Sun, "Energy-efficient cooperative spectrum sensing by optimal scheduling in sensor-aided cognitive radio networks," *IEEE Trans. Vehicular Tech.*, vol. 61, no. 2, pp. 716–725, Feb. 2012.

[3] G. Vallero, M. Deruyck, M. Meo, and W. Joseph, "Accounting for energy cost when designing energy-efficient wireless access networks," *Energies*, vol. 11, no. 3, pp. 1–21, Mar. 2018.

[4] Z. Liu, B. Liu, C. Chen, and C. W. Chen, "Energy-efficient resource allocation with QoS support in wireless body area networks," in *Proc. IEEE GLOBECOM*, 2015, pp. 1–6.

[5] S. Ahmed, Y. D. Lee, S. H. Hyun, and I. Koo, "A cognitive radio-based energy-efficient system for power transmission line monitoring in smart grids," *J. Sensors*, vol. 2017, pp. 1–12, Dec. 2017.

[6] P. Gandotra, R. K. Jha, and S. Jain, "Green communication in next generation cellular networks: A survey," *IEEE Access*, vol. 5, pp. 11727–11758, June 2017.

[7] M. K. Hanawal, F. Malik, and Y. Hayel, "Differential pricing of traffic in the internet," in *Proc. IEEE WiOpt*, May 2018, pp. 1–8.

[8] M. S. Mushtaq, S. Fowler, and A. Mellouk, "Power saving model for mobile device and virtual base station in the 5G era," in *Proc. IEEE ICC*, 2017, pp. 1–6.

[9] E. Agrell, "Conditions for a monotonic channel capacity," *IEEE Trans. Commun.*, vol. 63, no. 3, pp. 738–748, Dec. 2014.

[10] S. Kim, "Game theory for wireless ad hoc networks" in *Game Theory: Breakthroughs in Research and Practice*, Hershey, Pennsylvania, USA, IGI Global, pp. 353–368, 2018.

[11] B. Nleya and A. Mutsvangwa, "Enhanced congestion management for minimizing network performance degradation in OBS networks," *SAIEE Africa Research J.*, vol. 109, no. 1, pp. 48–57, Mar. 2018.

[12] S. Tombaz, A. Vastberg, and J. Zander, "Energy- and cost-efficient ultra high capacity wireless access," *IEEE Wireless Commun. Mag.*, vol. 18, no. 5, pp. 18–24, Oct. 2011.

[13] E. Bjornson, E. A. Jorswieck, M. Debbah, and B. Ottersten, "Multiobjective signal processing optimization: The way to balance conflicting metrics in 5G systems," *IEEE Signal Processing Magazine*, vol. 31, no. 6, pp. 14–23, Nov. 2014.

[14] N. N. Srinidhi, S. M. Dilip Kumar, and K. R. Venugopal, "Network optimizations in the internet of things: A review," *Eng. Sci. Tech., an Int. J.*, vol. 22, no. 1, pp. 1–21, Sept. 2018.

[15] U. Muhammad, S. K. Muhammad, V. Hiep, and I. Koo, "Energy-efficient channel hand-off for sensor network-assisted cognitive radio network," *Sensors*, vol. 15, no. 8, pp. 18012–18039, Aug. 2015.

[16] Y. Zhao, S. Jin, and W. Yue, "A novel spectrum access strategy with α-retry policy in cognitive radio networks: A queuing-based analysis," *J. Commun. Netw.*, vol. 16, no. 2, pp. 193–201, May 2014.

[17] R. Doost, M. Y. Naderi, and K. R. Chowdhury, "Spectrum allocation and QoS provisioning framework for cognitive radio with heterogeneous service classes," *IEEE Trans. Wireless Commun.*, vol. 13, no. 7, pp. 3938–3950, July 2014.

[18] B. Awoyemi, B. Maharaj, and A. Alfa, "Optimal resource allocation solutions for heterogeneous cognitive radio networks," *Digital Commun. Netw.*, vol. 3, no. 2, pp. 129–139, May 2017.

[19] M. J. Kaur, M. Uddin, and H. K. Verma, "Optimization of QoS parameters in cognitive radio using adaptive genetic algorithm," *International J. Next Generation Netw.*, vol. 4, no. 2, pp. 1–15, June 2012.

[20] L. Zhai, H. Wang, and C. Gao, "A spectrum access based on quality-of-service (QoS) in cognitive radio networks," *PLoS one*, vol. 11, no. 5, May 2016.

[21] M. Alsharif, J. Kim, and J. Kim, "Green and sustainable cellular base stations: An overview and future research directions," *Energies*, vol. 10, no. 5, pp. 1–27, Apr. 2017.

[22] G. Auer *et al.*, "How much energy is needed to run a wireless network?," *IEEE Wireless Commun. Mag.*, vol. 18, no. 5, pp. 40–49, Oct. 2011.

[23] M. W. Kang, and Y. W. Chung, "An efficient energy saving scheme for base stations in 5G networks with separated data and control planes using particle swarm optimization," *Energies*, vol. 10, pp. 1–28, Sept. 2017.

[24] D. Renga, H. A. H. Hassan, M. Meo, and L. Nuaymi, "Energy management and base station on/off switching in green mobile networks for offering ancilliary services," *IEEE Trans. Green Commun. Netw.*, vol. 2, no. 3, pp. 868–880, Sept. 2018.

[25] L. Sboui, H. Ghazzai, Z. Rekzi, and M.-S. Alouini, "On green cognitive radio cellular networks: Dynamic spectrum and operation management," *Special Section on Green Communications and Networking for 5G Wireless Networks, IEEE Access*, vol. 4, pp. 4046–4057, July 2016.

[26] J. Yang, B. Payne, M. Hitz, Z. Fei, L. Li, and T. Wei, "Location aided energy balancing strategy in green cellular networks," in *Proc. IEEE ICCCN*, 2014, pp. 1–6.

[27] G. Wang, C. Guo, S. Wang, and C. Feng, "A traffic prediction based sleeping mechanism with low complexity in femtocell network," in *Proc. IEEE ICC*, June 2013, pp. 560–565.

[28] X. Tan, H. Huang, and L. Ma, "Frequency allocation with artificial neural networks in cognitive radio system," in *Proc. IEEE TENCON Spring Conference*, 2013, pp. 366–370.

[29] Y. H. Wang and S. L. Liao, "Applying a fuzzy-based dynamic channel allocation mechanism to cognitive radio networks," in *Proc. IEEE WAINA*, 2017, pp. 564–569.

[30] Y. El Morabit, F. Mrabti, and E. H. Abarkan, "Spectrum allocation using genetic algorithm in cognitive radio networks," in *Proc. IEEE RAWSN*, 2015, pp. 90–93.

[31] B. Liu, J. Cheng, Q. Liu, X. Tang, "A long short-term traffic flow prediction method optimized by cluster computing," Preprints 2018, 2018080163, DOI: 10.20944/preprints201808.0163.v1

[32] Y. Hua *et al.*, "Deep learning with long short-term memory for time series prediction," *IEEE Commun. Mag.*, vol. 57, no. 6, pp. 114–119, June 2019.

[33] T. Dlamini, A. F. Gambin, D. Munaretto, and M. Rossi, "Online supervisory control and resource management for energy harvesting BS sites empowered with computation capabilities," *Wireless Commun. Mobile Comput.*, vol. 2019, 2019.

[34] T. Zhao, J. Wu, S. Zhou, and Z. Niu, "Energy-delay trade-offs of virtual base stations with a computational resource-aware energy consumption model," in *Proc. IEEE ICCS*, 2014, pp. 26–30.

[35] M. C. Hlophe and B. T. Maharaj, "Optimization and learning in energy efficient resource allocation for cognitive radio networks," in *Proc. IEEE VTC*, 2019, pp. 1–5.

[36] P. S. Aravind, J. Shah, and D. G. Kurup, "Bit error rate (BER) performance analysis of DASH7 protocol in Rayleigh fading channel," in *Proc. IEEE ICACCI*, 2018, pp. 695–698.

[37] F. Kelly, "Notes on effective bandwidths," *Stochastic Netw.: Theory and Applicat.*, vol. 4, pp. 141–168, 1996.

[38] E. Caushaj, I. Ivanov, H. Fu, I. Sethi, and Y. Zhu, "Evaluating throughput and delay in 3G and 4G mobile architectures," *J. Comput. Commun.*, vol. 2, no. 10 pp. 1–8, Jan. 2014.

[39] S. Ping, A. Aijaz, O. Holland, and A. H. Aghvami, "Green cellular access network operation through dynamic spectrum and traffic load management," in *Proc. IEEE PIMRC*, 2013, pp. 2791–2796.

[40] M. Leconte, M. Lelarge, and L. Massoulie, "Bipartite graph structures for efficient balancing of heterogeneous loads," *ACM Sigmetrics Performance Evaluation Review*, vol. 40, no. 1, pp. 41–52, June 2012.

[41] S. V. Weijs, R. Van Nooijen, N. Van De Giesen, "Kullback-Leibler divergence as a forecast skill score with classic reliability-resolution-uncertainty decomposition," "Monthly Weather Review", vol. 138, no. 9, pp. 3387–3399, Sept. 2010.

[42] JORG Database. Accessed September 10, 2018. [Online]. Available: https://www.comm.utoronto.ca/ jorg/teaching/ece466/labs/lab1/?C=N;O=D.

[43] ITU-T, "ITU-T Recommendation G.114 One-way transmission time," *ITU G.114*, 2003.

[44] X. Zhang, P. Wang, "Optimal trade-off between power saving and QoS provisioning for multicell cooperation networks," *IEEE Wireless Commun.*, vol. 20, no. 1, pp. 90–96, Feb. 2013.

[45] K. Hashimoto, S. Adachi, and D. V. Dimarogonas, "Self-triggered model predictive control for nonlinear input-affine dynamical systems via adaptive control samples selection," *IEEE Transactions on Automatic Control*, vol. 62, no. 1, pp. 177–189, Jan. 2017.

[46] H. Chen and F. Allgower, "A quasi-infinite horizon nonlinear model predictive control scheme with guaranteed stability," *Automatica*, vol. 34, no. 10, pp. 1205–1217, Oct. 1998.

[47] T. H. Nguyen, M. Forshaw, and N. Thomas, "Operating policies for energy efficient dynamic server allocation," *Electronic Notes in Theoretical Computer Science*, vol. 318, pp. 159–177, Nov. 2015.

[48] W. C. Chan, T. C. Lu, and R. J. Chen, "Pollaczek-Khinchin formula for the M/G/1 queue in discrete time with vacations," *IEE Proc. Comput. Digital Techniques*, vol. 144, no. 4, pp. 222–226, July 1997.

[49] M. Draoli, C. Gaibisso, M. Lancia, and E. A. Mastromartino, "Satisfying high quality requirements of videoconferencing on a packet switched network," in *Proc. INET*, 1997.

[50] J. Dilley and R. L. Armstrong, “How much speed you need for online gaming,” Accessed September 18, 2018, [Online]. Available: https://www.highspeedinternet.com/resources/how-much-speed-do-i-need-for-online-gaming/.

[51] M. Hosseini and V. Swaminathan, “Adaptive 360 VR video streaming: Divide and conquer,” in *Proc. IEEE ISM*, 2016, pp. 107–110.

[52] D. L. Nguyen, O. Berder, and O. Sentieys, “A low-latency and energy-efficient MAC protocol for cooperative wireless sensor networks,” in *Proc. IEEE GLOBECOM*, 2013, pp. 3826–3831.

[53] R. Jejurikar and R. K. Gupta, “Integrating processor slowdown and preemption threshold scheduling for energy efficiency in real time embedded systems,” in *Proc. IEEE RTCSA*, 2004.

Mduduzi Comfort Hlophe received his B. Eng. degree in Electronic Engineering from the University of Swaziland in 2012; an M. Eng. degree in Wireless Communication Networks with the Wireless Lab (YLab) at the University of Johannesburg in 2015. Currently pursuing a Ph.D. with the Broadband Wireless Multimedia Communications (BWMC) group in the Department of Electrical, Electronic and Computer Engineering at the University of Pretoria. His research interests include mathematical modeling of: Multivariate statistics, classification methods, knowledge discovery, reasoning with uncertainty and inference, predictive analytics and inference; with applications in mobile and wireless communications, robotics, finance and health.

Bodhaswar T. (Sunil) Maharaj received his Ph.D. in Engineering in the Area of Wireless communications from the University of Pretoria. Dr. Maharaj is a Full Professor and currently holds the research position of Sentech Chair in Broadband Wireless Multimedia Communications (BWMC) in the Department of Electrical, Electronic and Computer Engineering at the University of Pretoria. His research interests are in OFDM-MIMO systems, massive MIMO, cognitive radio resource allocation and 5G cognitive radio sensor networks.