Vakilinia *et al. EURASIP Journal on Wireless Communications and Networking* 2014, **2014**:201
http://jwcn.eurasipjournals.com/content/2014/1/201

EURASIP Journal on
Wireless Communications and Networking
a SpringerOpen Journal

RESEARCH **Open Access**

Optimal multi-dimensional dynamic resource allocation in mobile cloud computing

Shahin Vakilinia*, Dongyu Qiu and Mustafa Mehmet Ali

Abstract

In this paper, we propose a model for mobile application profiles, wireless interfaces, and cloud resources. First, an algorithm to allocate wireless interfaces and cloud resources has been introduced. The proposed model is based on the wireless network cloud (WNC) concept. Then, considering power consumption, application quality of service (QoS) profiles, and corresponding cost functions, a multi-objective optimization approach using an event-based finite state model and dynamic constraint programming method has been used to determine the appropriate transmission power, process power, cloud offloading and optimum QoS profiles. Numerical results show that the proposed algorithm saves the mobile battery life and guarantees both QoS and cost simultaneously. Moreover, it determines the best available cloud server resources and wireless interfaces for applications at the same time.

Keywords: Resource allocation; Mobile cloud computing; Wireless network cloud

1 Introduction

Popularity of smartphones and related applications in various fields are increasing in everyday life significantly. These devices have a wide range of features (e.g., high-speed processors and supporting multiple wireless interfaces). Furthermore, due to increasing complexity of applications, smartphones require a significant computational capability. In addition, they have become a primary computing platform for many users due to the well-developed applications in realms such as mobile commerce, mobile learning, mobile health care, mobile computing, mobile gaming, and etc. As applications become more and more complex, mobile users experience shorter battery lifetime. Most of the smartphone applications are QoS-sensitive and computation-intensive to perform on a mobile system. Mobile cloud computing is a new concept in which mobile users access the cloud virtual resources via the Internet. It is beneficial to QoS and battery saving by means of mobile data offloading. Mobile computation offloading technique shares application code between the cloud server and the mobile. Most of the time, mobile users need to maintain a low level of power consumption and thus computation must be performed in the cloud which comes with cost. Therefore, mobile users always face a trade-off between communication and computation [1].

On the other hand, wireless network cloud (WNC) [2] proposes an architecture to join wireless access systems to cloud computing and shift the processing of base stations with different technologies to a virtual cloud network. Therefore, all wireless technologies is converging and is suitable for next generation wireless networks. WNC and cloud radio access network (C-RAN) [3] using similar software-defined radio (SDR) concept tend to decrease wireless network operating cost while enhancing the total network performance. Accordingly, without doubt, the next generation of wireless networks (5G) movement toward wireless clouds is irresistible [4,5].

Despite flexibility and great potential applicability, resource allocation problem in heterogeneous wireless networks (HetNet) attributed with WNC and mobile cloud computing has received scarce attention as of today. Therefore, the prime contribution of the current research has been based on bridging HetNet with WNC and mobile cloud computing to better allocate resources to the end user. In addition, a multi-objective optimization problem considering cloud server power consumption, operating cost, and QoS followed by a detailed trade-off amongst user objectives have been studied.

*Correspondence: s_vakili@encs.conrcoria.ca
Department of Electrical and Computer Engineering, Concordia University, 1455 De Maisonneuve, Montreal, H4B 1R6, Canada

In this paper, we propose a model including the operators, clouds, applications, and mobile profile parameters. Due to the fact that a part of the algorithm has to be conducted in smartphones, complexity order of the problem becomes a vital parameter. Estimation and approximation techniques have been used to linearly approximate the parameters to decrease complexity order of our algorithm. Using dynamic constraint programming [6,7], event-based lexicographic multi-objective optimization method [8] and QoS-based resource allocation solutions [9,10] with consideration to the resources and applications constraints, network, and mobile resources have been allocated to applications simultaneously.

It is worthy of note that the main objective of this paper concerns performance metrics of mobile devices and users, regardless of cloud computing centers and wireless operators related challenges, [11-16] which have not been considered in this paper.

The rest of the paper is organized as follows: this study's related works is discussed in Section 2, in Section 3, the system model will be defined, followed by the optimization algorithm in Section 4. Within Section 5, numerical results reveal performance of the proposed multi-dimensional algorithm. Finally, Section 6 concludes the paper.

2 Related works

Rahimi et al., [17], Fernando et al., [18], and Dinh et al., [19] give an overview of the mobile cloud computing (MCC) presenting definition, architecture, applications, and approaches, then, on the corresponding challenges at the operational, user, and application levels have been discussed. They introduced MCC as the dominant computing model for mobile applications in the future.

Moreover, extensive research such as in [20-22] has been done over wireless local area network (WLAN)/ cellular interworking mechanisms, which combines WLANs and cellular data networks into integrated wireless data networks featured with QoS capabilities. Liu et al. [23] suggest a new dynamic load balance (DLB) scheme to improve communication performance focusing on underlying users. In their proposed scheme, joint session admission control is a basis for user mobility, cognition, and service arrival awareness in integrated 3G/WLAN networks. Gazis et al.and Luo et al. [24,25] recommend a standardization policy in the area of WLAN-cellular data network integration for different interworking architectures. Proposing the generic interworking architectures in the technical literature, [26] studies general aspects of integrated WLAN-cellular data networks. Access network discovery and selection function (ANDSF) suggests a function for selection of access network and control offloading amongst 3rd generation partnership project (3GPP) and other access networks. Such selections are based on the mobile battery saving, user preference, and operator policies. However, ANDSF does not consider application preferences, selection optimality, and simultaneous power allocation.

In general, international standards and standardization bodies such as WiMAX and 3GPP decide to move toward creating a seamless integrated wireless technology entitled HetNet [27]. HetNet by its nature includes a variety of wireless access technologies. Access networks are connected through a backbone which is a network core for all of them. Moreover, HetNet consists of both macro and micro cells as well as low power nodes which have distinct or overlapped coverage areas. When a multi-interface device moves within a HetNet environment, its default network for every connection can be determined based on a set of predetermined parameters of network nature such as QoS settings, signal strength, backbone utilization, speed preference, selected cost or service, and mobile node's remained battery life.

Furthermore, some researchers have studied power consumption in smartphones. Murmuria et al., [28] and Carroll and Heiser [29] measure, analyze, and model power usage of smartphones by characterizing their subsystems power usages. Balasubramanian et al. [30] consider wireless interface selection problem as a statistical decision problem and propose an algorithm to select the wireless network interface considering the context of the mobile applications in order to improve the battery lifetime. Hence, the features of wireless access interface selection also has fundamental impact on the performance of mobile computing applications and their power consumption.

There are some trade-offs amongst power consumption, QoS parameters, and costs. These objectives are dependent on network parameters, applications profiles, and cloud resources. Cuervo et al. [31] aim to optimize energy consumption of a mobile device by estimation and evaluating the trade-off between the energy consumed by local processing versus the transmission of code and data for cloud offloading. Decision process in [31] considers information and complex characteristics of the mobile environment. A framework for smartphones is introduced in [32]. It shifts smartphone application processing into the cloud centers. It is based on the concept of smartphone virtualization in the cloud and addresses lack of scalability by creating virtual machines of a complete smartphone system on the cloud. ThinkAir [32] provides on-demand resource allocation by dynamically managing VMs in the cloud via using an execution controller. The execution controller handles decision-making and communication with the cloud server. It considers execution time, energy, and cost to make decision in order to achieve optimum performance. With regard to the network profile parameters, device profile parameters, and program profile parameters of the smartphone, ThinkAir dynamically allocates the available cloud resources to the programs simultaneously. Kumar and Lu [33] suggest that cloud computing can potentially save energy through offloading of applications processing with limited reliability and quality of service requirements. This reflects the fact that for some applications such as delay-sensitive ones, migrated offloading to the clouds could not significantly offer energy savings to the smartphones while satisfying QoS parameters.

Trade-off between system throughput and energy consumption of mobile devices has been addressed in [34]. Based on the Lyapunov optimization approach, an online control algorithm is designed to balance energy and throughput. It maximizes a joint utility using stability-utility parameters while bounding the traffic queue length, via making instantaneous decisions to control the transmission pattern. The admission control algorithm diminishes the need for statistical estimation of traffic arrivals and link conditions.

In order to allocate resources amongst the cloud users efficiently, a communication framework amongst cloud users and service providers has been designed in [35]. There, authors propose a biding language in order to convert cloud user demands into the organized requests which helps cloud providers to support heterogeneous user demands while protecting the systems from selfish user behavior. Moreover, online compatible online cloud auction (COCA) mechanism is implemented to make users incentive to reveal their honest valuations. Finally, they have considered the sum of all the valuations of the allocated resources as the benchmark.

A QoS-aware resource-allocation multiple cooperative subtasks of jobs in cloud-based computing and data store services are investigated in [36]. Defining the objective function as a weighted sum of the expense and the job completion time and job execution time deadlines and budget constraints, game theory approach is used to solve the scheduling problem. First, considering users as their chosen strategy regardless of the others, a binary integer programming method is proposed to obtain the initial independent optimization solution. Then, an evolutionary strategy is designed to achieve the optimal solution.

Regarding the scalability advantage of public clouds and better QoS especially delay and power consumption of local clouds, MAPCloud is proposed in [37]. This provided a means to select local and public clouds for mobile applications in order to increase the performance and scalability of the applications. Interestingly, for a fixed price, MAPCloud decreases 32% of the delay and power consumption while providing scalability. Then, cloud resource allocation for mobile applications (CRAM) using heuristic methods has been developed as a resource allocation module for mobile applications achieving 84% of the optimal power saving solutions for large amount of users.

Rahimi et al. [38] focused on modeling the mobile applications as location-time workflows (LTW) of task. 2D location map is used to locate mobile hosts and cloud resources. Moreover, trajectory has been associated with mobile users. Defining QoS as a function of delay, power, and price, an efficient heuristic algorithm called MuSIC is proposed to maximize the mobile utilities while ensuring high-application QoS.

Applying the game theory approach, coalition of the cloud service providers is addressed in [39] where the uncertainty of internal users from each provider has been taken into account. First, with respect to randomness of demand, a stochastic linear programming game model to study the resource and revenue sharing for cloud providers is developed. Then, using the Markov chain to model coalitional arrangement, the coalitional game for forming the cooperation to share resource and revenue are investigated.

In this paper, we address performance modeling of mobile applications using MCC and WNC. A resource allocation algorithm is proposed to allocate resources and mobile transmission power and process power.

3 System model

In this section, we present a system model for optimum resource allocation. We assumed that mobile users access application clouds via WNC and the Internet. Figure 1 shows the presumed topology based on [2] and [19]. However, as of today, smartphones just support WLAN/cellular technologies simultaneously.

We assume that there are a number of active applications on a mobile phone that support both WLAN and cellular technologies. In order to achieve a better performance and improve power saving, a portion of the processing workload has to be offloaded to the clouds. As depicted in the Figure 2, each application must choose a proper wireless interface and a cloud network to offload the processes to. However, the said selection process depends on parameters such as battery lifetime and required processing load. In addition, a feasible QoS profile for the application needs to be determined.

In order to conceive this model presumption, we defined sets of variables according to application profiles, computing resource profiles, and network profiles. $A = \{1, 2, .., i, ..I\}$ states a set of mobile applications, $CR = \{1, .., j, , ..J\}$ states a set of available cloud computing resources, and $WN = \{WN_1, .., WN_k, , ..WN_K\}$ represents accessible wireless network interfaces collection. Note that$I$, J, and K are the number of active applications, available clouds, and wireless interfaces, respectively. Each collection element is a vector of characteristics which is related to the cost, power consumption, and

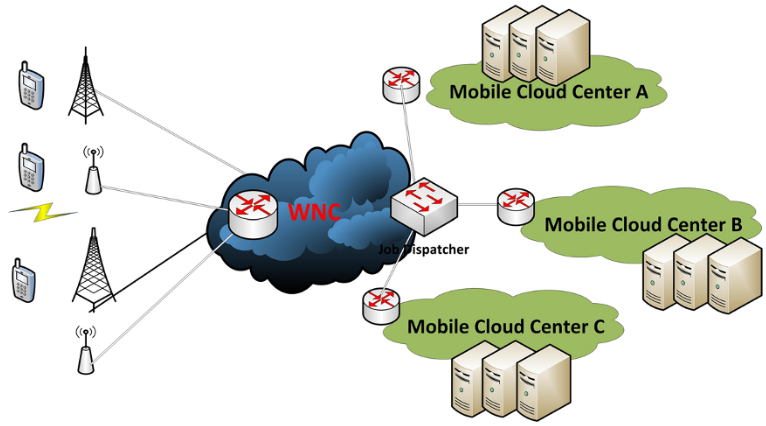

Figure 1 Network topology.

QoS. An operating point $\mathrm{OP_n} = (U_\mathrm{n}, P_\mathrm{n}, C_\mathrm{n})$ is defined to describe the mobile system behavior over the nth time slot. U_n shows the utilization associated with mobile user satisfaction and is strictly related to the QoS indexes of applications. P_n represents mobile phone power consumption, and C_n demonstrates mobile phone cost function. In this paper, appropriate $j \in CR$ and $k \in WN$ are assigned to the ith application in order for better formulation of controlling and optimizing the operating point. Therefore, operating point indicates important objectives of mobile user such as mobile power consumption at each time slot. All parameters of the model are detailed in Table 1.

Moreover, there are some limitations and restrictions on resources and user profiles which are strictly dependent on the mobile application QoS requirements and network parameters. In the following subsections, the objective functions and constraints will be investigated. Traffic rate of the ith application is defined discretely between λ^i_{min} and λ^i_{max} where i belongs to $\{1,2,\ldots,I\}$. Due to the bound limits, different functions of downlink traffics are linearly approximated using affine functions and the Taylor series. Such approximations decrease the complexity order in a dramatic way while errors remain small.

3.1 QoS utilization and constraints

As explained before, utilization function is related to application QoS characteristics. Objective utilization function will be as follows:

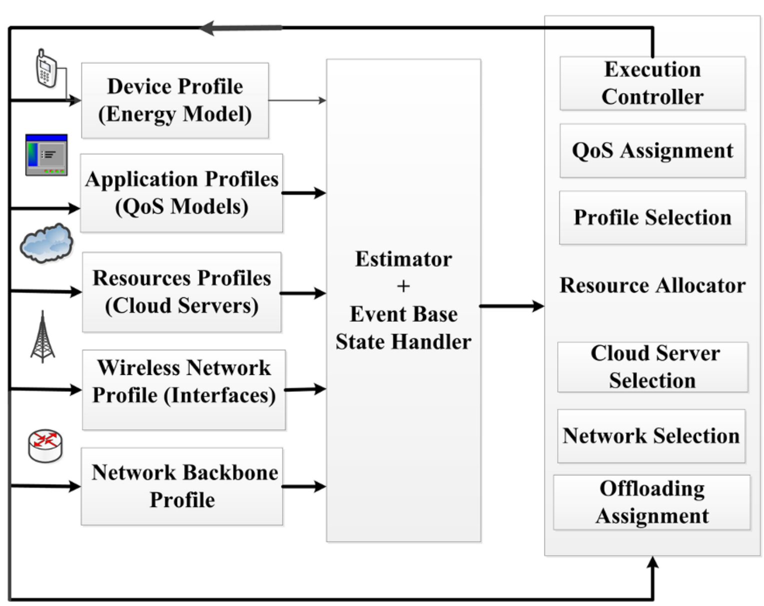

Figure 2 Algorithm architecture.

Table 1 Table of parameters

Parameters	Indicator
$\lambda^i_{\min}$	Minimum required incoming traffic rate of the *i*th application
$\lambda^i_{\max}$	Maximum required incoming traffic rate of the *i*th application
λ_i	Incoming traffic rate of the *i*th application
S_i	Effective processor speed (instructions per second) dedicated to the *i*th application in a mobile device
C_i	Instructions of *i*th application per time slot
$R^i_{\min}$	Minimum required transmission rate of the *i*th application
$R^i_{\max}$	Maximum required transmission rate of the *i*th application
R_i	Transmission rate of the *i*th application
γ_i	Instructions have to be processed in cloud servers *i*th application
χ_i^{dep}	Utilization factor of the *i*th application dependent on incoming traffic rate
χ_i^{indep}	Utilization factor of the *i*th application independent from incoming traffic rate
D_i^{th}	Delay threshold of *i*th application traffic
D_i	Delay of *i*th application traffic
η_i	Mobile data offloading to clouds instructions of *i*th application
TD_i	Uploading data of *i*th application
$R_k^{\max}$	Maximum achievable transmission rate using *k*th interface
H_k	Channel quality indicator of the *k*th interface
Dw_k^{th}	Achievable guaranteed downlink delay of *k*th interface
Dw_k	Downlink delay of *k*th interface
α_k^w	Cost coefficient of *k*th wireless download rate (per instruction)
θ_k	Downlink QoS exponent of *k*th interface
μ_k	Download rate of *k*th interface
$P_{\text{maint}}(k)$	Connection power consumption of *k*th interface
P_{Error}^{Th}	Error rate threshold
β_j^n	Delay between the wireless cloud and *j*th cloud server at the *n*th time slot
$\hat{\beta}_j^n$	Estimation of the β_j^n
S_j'	Effective processor speed of *j*th cloud server
α_j^{DL}	Cost coefficient of downlink traffic of *j*th cloud server
α_j^{UL}	Cost coefficient of uplink traffic of *j*th cloud server
α_j^{comp}	Cost coefficient cloud computation of *j*th cloud
P_{comp}	Process power consumption per processing speed unit
E^n	*n*th moment
$\epsilon(n)$	Mobile energy level at the *n*th time slot
$Bg(n)$	Mobile budget fee at the *n*th time slot

$$U = \sum_{i=1}^{I} (\chi_i^{\text{dep}}(\lambda_i, R_i) + \chi_i^{\text{indep}}) \tag{1}$$

where $\chi_i^{\text{dep}}(\lambda_i, R_i)$ depends on the upload and download rate. Conversely, χ_i^{indep} is independent from the upload and download rate in the *i*th application utilization function.

Delay process consists of two parts, namely processing delay and communication delay. Also, communication delay includes two parts: wireless link delay and internet network delay. Applications such as cloud mobile gaming (CMG) interact closely with cloud servers. Therefore, uplink delay is a significant parameter as well. In addition, effective capacity concept has been used to model downlink wireless link delay. Assuming that arrival rates and service rates of the wireless links are all stationary and independent, according to the Gartner-Ellis limits [40,41], wireless link delay violation probability of *k*th wireless interface is approximated by [42-44]:

$$pr(D_k > D_k^{Th}) \approx e^{-\theta \mu_k D_k^{Th}} \tag{2}$$

Internet delay (delay between the wireless network cloud and cloud servers) for interactive applications such as CMG denotes the round trip time delay, and for streaming applications denotes the one-way delay. Calculating the Internet delay requires a complicated procedure. However, assuming that the mobile cloud computing centers are near the wireless access network, Internet delay may be considered as a Gaussian random variable. Therefore, linear estimator [45] based on adaptive algorithm proposed in [46] is used to predict the Internet delay:

$$\hat{\beta}_j^n = \beta_j^{n-1} + \frac{r\sigma^n_{\beta_j}}{\sigma^n_{\hat{\beta}_j}}(\hat{\beta}_i^{n-1} - \beta_j^{n-1}) \tag{3}$$

where r is equal to

$$r = \sqrt{\frac{E^2(\beta_j^n - E(\beta_j^n))(\hat{\beta}_j^n - E(\hat{\beta}_j^n))}{(E(\beta_j^n - E(\beta_j^n))^2)(E(\hat{\beta}_j^n - E(\hat{\beta}_j^n))^2)}} \tag{4}$$

β_j^n represents the Internet delay of the *j*th cloud server with the application in the *n*th time slot. After cloud server selection process, a cloud server is mapped to the *i*th application. Hereafter, we assume that the *j*th cloud server and *k*th wireless interface are assigned to the *i*th application. Therefore, total delay of *i*th application could be written by

$$\begin{aligned} D_i = {} & \text{processdelay} + \text{wirelessuplinkdelay} + \text{internetdelay} \\ & + \text{cloudprocessdelay} + \text{wirelessdownlinkdelay} \end{aligned} \tag{5}$$

In Equation 5, the first part denotes the processing delay in the mobile phone related to the number of instructions per time slot executed in the mobile phone. Mobile process delay is approximated by $\frac{C_i - \eta_i}{S_i}$ where η_i represents the offloading instructions to the cloud and S_i denotes effective processor speed dedicated to *i*th application. The second part denotes the uploading delay of smartphone approximated by $\frac{TD_i}{R_i}$. The third part represents the network delay represented by β_j. The forth part denotes the processing delay of the cloud server considered as $\frac{\eta_i + \gamma_i}{S_j'}$. Finally, the last part implies the downlink delay of the *k*th interface represented by Dw_k. Dw_k is considered a random variable and its expected value is of great benefit to decision-making. The expected delay has to be less than the application delay threshold. QoS characteristics of the ith application could be written as follows:

$$Q_i = \{\lambda^i_{\min}, \lambda^i_{\max}, \lambda_i, R^i_{\min}, R^i_{\max}, R_i, D_i, D_i^{Th}\} \tag{6}$$

3.2 Power consumption process

Power consumption also consists of two main parts, namely transmission power consumption and processing power consumption. The formulation will be as follows:

$$P = \text{processing power} + \text{transmission power} \tag{7}$$

The first part indicates the processing power consumption, and the second part is the transmission power. Processing power may be approximated linearly as a function of effective processor speed dedicated to the applications.

$$\text{Processingpower} = \left(\sum_{i=1}^{I} P_{\text{comp}} S_i\right) \tag{8}$$

Transmission power itself consists of connection maintenance power consumption [47] and data transmission power consumption.

$$\text{Transmissionpower} = \left(\sum_{k=1}^{K} (P_k^{tr}(H_k, R_k) + P_{\text{maint}}(k))\right) \tag{9}$$

Transmission power depends on the channel state information (CSI) and transmission rate of the mobile phone. Without a doubt, OFDM is the dominant technology in the current and future transmission technologies. With respect to the CSI on the receiver side, 'Water filling' could be an optimum algorithm to allocate the transmission power to the sub-carriers. Considering a single antenna, it will be equal to

$$P_k^{tr}(H_k, R_k) = \sum_{m_k=1}^{M_k} \left(e^{ln2\left(\frac{R_k}{W_k} - \sum_{m_k=1}^{M_k} \log_2\left(\frac{h_{mk}}{\Gamma_k n_{mk}}\right)\right)} - \frac{h_{mk}}{\Gamma_k n_{mk}} \right) \tag{10}$$

where W_k represents the kth interface sub-channel bandwidth. h_{mk} is the mth sub-channel quality indicator of kth interface. Γ_k indicates coding gain of kth interface, n_{mk} states the mth sub-channel noise of the kth interface, and M_k represents the number of subcarriers of the kth interface.

In the rest of the paper, we use Equation (10) as the transmission power function. Connection maintenance power consumption has a linear relation with transmission time. According to central limit theorem, allocated processing power for applications is approximated by a Gaussian random variable. Then, mobile CPU process sharing feasibility is defined by $Pr\left(\sum_{i=1}^{I} S_i > S\right) \leq p$ therefore, $\left(\sum_{i=1}^{I} \mu_{S(i)} + \zeta \sum_{i=1}^{I} \sigma_{S(i)}\right) < S$ where $\zeta = \Phi^{-1}(1-p)$, Φ^{-1} is the inverse function of the CDF of normal distribution with $\mu_{S(i)}$ and $\sigma_{S(i)}$ as the first and the second moments, respectively. For the proof, see [48].

3.3 Cost function

The cost function consists of the following two parts:

$$\text{Cost} = \text{wirelessoperatorcosts} + \text{cloudservicecosts} \tag{11}$$

We assumed that each active application receives service from a specific cloud server and a wireless interface is selected for communication of each application. Based on the proposed cost model in [49], the mobile cloud service cost function could be written as follows:

$$\text{cloud service costs} = \sum_{j=1}^{J} \left(\alpha_{\text{comp}}(\gamma_j + \eta_j) + \alpha_j^{UL} R_j + \alpha_j^{DL} \lambda_j'\right) \tag{12}$$

where $\lambda_j^{'}$ denotes the sum of the incoming traffic rates of applications which the jth cloud server is assigned to them. We assumed that each application is linked to a cloud server. First part indicates the computation cost while the second and the third parts represent the cost associated with data upload and download to cloud servers, respectively. Wireless access network costs also could be approximated linearly by

$$\text{wirelessoperatorcosts} = \sum_{k=1}^{K} \alpha_k^w \lambda_k^{''} + \sigma_k^w R_k^{''} \tag{13}$$

$\lambda_k^{''}$ and $R_k^{''}$ denote the sum of incoming and outgoing traffic rates, respectively, of applications which the kth interface is assigned to them. Accordingly, the following characteristics for clouds and wireless network interfaces are proposed: $CR_j = \{\alpha_j^{UL}, \alpha_j^{DL}, \beta_j, S_j^{'}\}$ and $WN_k = \{\alpha_k^w, H_k, R_k^{\max}, P_{\text{maint(k)}}\}$ (See Table 1).

It is also possible to define objective functions and constraints with respect to application tasks instead of applications alone. Changing the scale from application to task increases resource allocation accuracy as well as complexity order of the algorithm.

4 Problem formulation and solution

4.1 Problem definition

Less power consumption, user satisfaction, and cost are of great interest to many mobile users. In this section, we propose a multi-objective dynamic resource allocation algorithm to optimize the aforementioned topics of interest in the form of objective function and processes with respect to the network resources and mobile and application constraints. Dynamic constraint programming and lexicographic-event-based optimization method [8] have been used to solve the multi-objective optimization problem. However, complexity order of the proposed algorithm also needs to be considered. Moreover, the previously highlighted measures of interest usually are not available in a closed form and are mostly obtained from numerical data. Henceforth, linear interpolation method for numerical data and the Taylor series for closed form function have been applied to approximate the input data or non-linear functions in a short interval, e.g., $\lambda_i \in [\lambda^i_{\min}, \lambda^i_{\max}]$. However, as it will be explained shortly, the proposed protocol architecture design does not depend on linear functions of the system model. In fact, application of non-linear functions will not impact the complexity order drastically.

Figure 3 shows the overall structure of the proposed algorithm. The algorithm takes the parameters of cloud profiles, wireless access networks, mobile devices, and mobile applications as its input. In addition, it linearly approximates inputs such as cloud server delay based on which the state and corresponding events are selected. In the next level, optimum wireless network interface, best available cloud servers, offloading coefficients of applications, processing power and transmission power of the mobile phone and optimum QoS profile will be selected. Obtaining an optimum offloading solution, the execution controller introduced in [32] manages the shared process between the mobile phone and cloud servers.

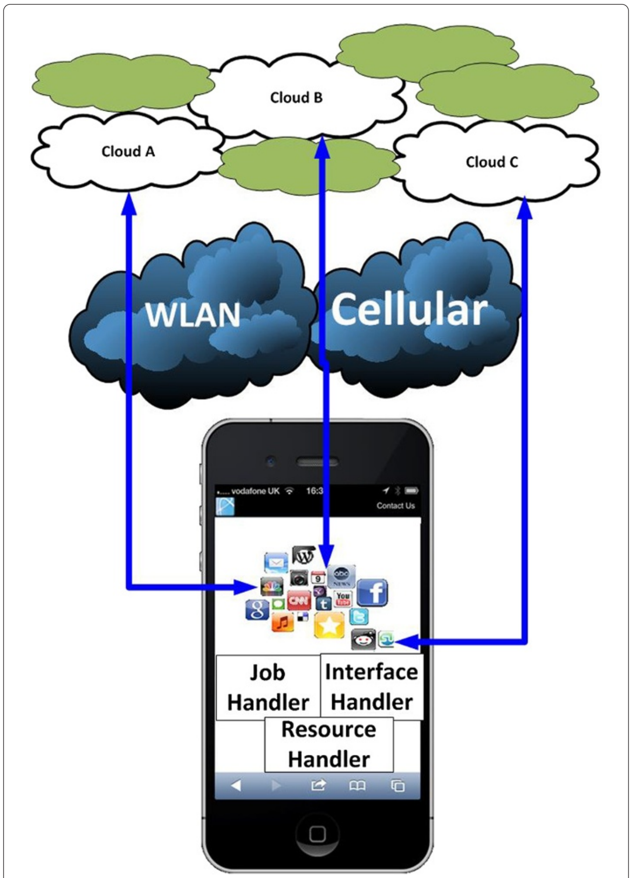

Figure 3 The proposed algorithm structure.

4.2 Problem formulation

Objective processes could be written as follows:

$$F^I_1(n) = -U(n) \tag{14}$$

$$F^I_2(n) = P(n) \tag{15}$$

$$F^I_3(n) = \mathrm{cost}(n) \tag{16}$$

Here, we used negative utilization factor to convert the maximization problem to a minimization problem. Mobile device and resources constraints are $\forall i \in A$

$$0 \leq \eta_i(n) \leq C_i(n) \tag{17}$$

$$\Sigma^I_{i=1}(\mu_{S_i}(n) + \zeta\sigma_{S_i}(n)) < S \tag{18}$$

$$\lambda^{''}_k(n) \leq \mu_k(n) \tag{19}$$

Application QoS profile constraints ($Q_{\min} < Q_i < Q_{\max}$) are considered as follows: $\forall i \in A$

$$E\{D_i(n)\} \leq D^{th}_i \tag{20}$$

$$\lambda^i_{\min} \leq \lambda^i(n) \leq \lambda^i_{\max} \tag{21}$$

$$R^i_{\min} \leq R^i(n) \leq R^i_{\max} \tag{22}$$

The current delay model approximates the real delay scenarios. To accommodate a more complicated type of

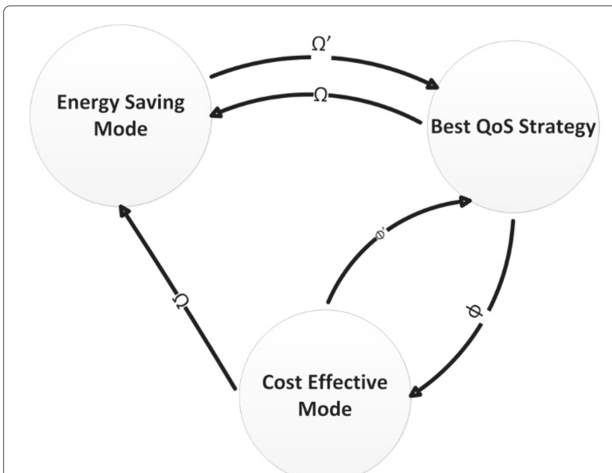

Figure 4 FSM structure.

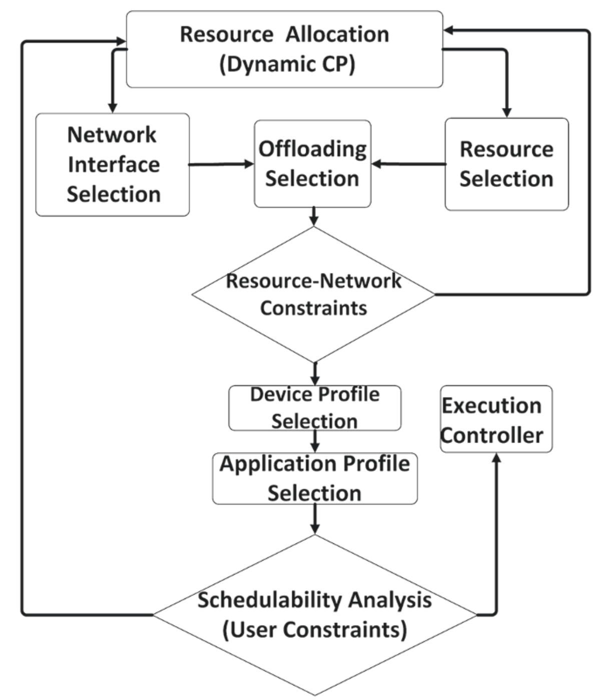

Figure 5 Algorithm architecture.

delay, Equation (20) can be replaced by arbitrary complex constraints.

We use lexicographic optimization method to solve the proposed multi-objective optimization problem. Objective functions are prioritized based on the state of the system. In fact, only one objective function is selected in each state and others are considered as constraints. The state of the system depends on the next occurring event. A finite state model (FSM) is proposed for optimal resource allocation while considering different events to transit amongst the states. Figure 4 shows the proposed FSM structure.

The states are as follows:

1. Best QoS strategy: in this state, we try to maximize user utilization while considering other objectives as constraint.
2. Cost-effective mode: in this state, considering the QoS and power consumption constraints, the proposed resource allocation algorithm attempts to minimize the cost of the system.
3. Energy-saving mode: in this state, the proposed algorithm minimizes the power consumption of the system considering QoS and cost constraints.

The transition events take place as detailed below:

Ω occurs when $\epsilon(n) \leq \epsilon(n^{th})$ where $\epsilon(n)$ and $\epsilon(n\text{th})$ denote the mobile energy in the nth time slot and its threshold, respectively. It shows that mobile energy is in a critical situation. $\acute{\Omega}$ occurs when: $\epsilon(n) \leq \epsilon(n^{th})$ and β happens when $\epsilon(n) \geq \epsilon(n-1)$ meaning that mobile device is being charged and is not in a critical situation energy wise. ϕ happens when $Bg(n) \leq Bg^{th}$ where Bg and Bgth denotes the mobile budget fee in the nth time slot and its threshold, respectively, reflecting the fact that mobile user budget is approaching levels lower than its threshold. Also, $\acute{\phi}$ occurs when $Bg(n) \geq Bg$th. QoS-sensitive state is considered as the initial state. According to the occurring events, dynamic constraint programming is applied to find the optimal solution. The state of the system is shown by x, where x belongs to the set of states; $X = \{1, 2, 3\}$. In each state, we solve the following optimization problem:

$$\begin{aligned}
&Argmin_{\;WN,CR,\eta,\lambda,R,S} F^{I}_{x|x} \\
&ST : F^{I}_{\acute{x}|x} \leq F^{th}_{\acute{x}} \quad \forall \acute{x} \in X, \acute{x} \neq x \\
&(17), (18), (19), (20), (21), (22)
\end{aligned}$$

where Equations (17), (18), (19) and Equations (20), (21), (22) are the resources and the mobile applications constraints, respectively, and

$$\begin{aligned}
\lambda &= \{\lambda_1, \ldots.\lambda_i, \ldots\lambda_I\} \\
\eta &= \{\eta_1, \ldots.\eta_i, \ldots\eta_I\} \\
R &= \{R_1, \ldots.R_i, \ldots R_I\} \\
S &= \{S_1, \ldots.S_i, \ldots S_I\}
\end{aligned}$$

Not only proper mobile cloud computing center and interface should be selected for the applications but also

Table 2 Numerical validation parameters

Parameters	Indicator
$\lambda^{i}_{\min}$	iid rv between 10 and 384 *kbps* with uniform distribution
$\lambda^{i}_{\max}$	iid rv between 100 and 2 *Mbps* with uniform distribution
S	800 MHz
C_i	iid rv between 100 and 10^7 with uniform distribution
$R_i^{\min}$	iid rv between 0 and 100 *kbps* with uniform distribution
$R_i^{\max}$	iid rv between 10 and 250 *kbps* with uniform distribution
χ_i^{dep}	iid rv between 0 and 1 with uniform distribution
χ_i^{indep}	iid rv between 0 and 1 with uniform distribution
D_i^{th}	iid rv between 50 *ms* and 20 *s* with uniform distribution
TD_i	iid rv between 0 and 100 *KB* with uniform distribution
$R_k^{\max}$	Maximum achievable transmission rate through using *k*th interface
Dw_k^{th}	50 *ms* interface
Dw_k	iid rv between 10 and 250 *kbps* with uniform distribution
α_k^{w}	iid rv between 10 and 250 *kbps* with uniform distribution
μ_k	k^{th} iid rv between 100 *kbps* and 2 *Mbps* with uniform distribution
β_j^{n}	iid rv between 20 *ms* and 5 *s* with uniform distribution
α_j^{DL}	Cost coefficient of downlink traffic of *j*th cloud server
α_j^{UL}	Cost coefficient of uplink traffic of *j*th cloud server
α_j^{comp}	Cost coefficient cloud computation of *j*th cloud
$P_{\text{maint}}(k)$	iid rv between 120 and 400 *mw* with uniform distribution for WiFi interfaces and iid rv between 500 and 800 *mW*
P_{comp}	$2.34 \times 10^{-10} W/Hertz$

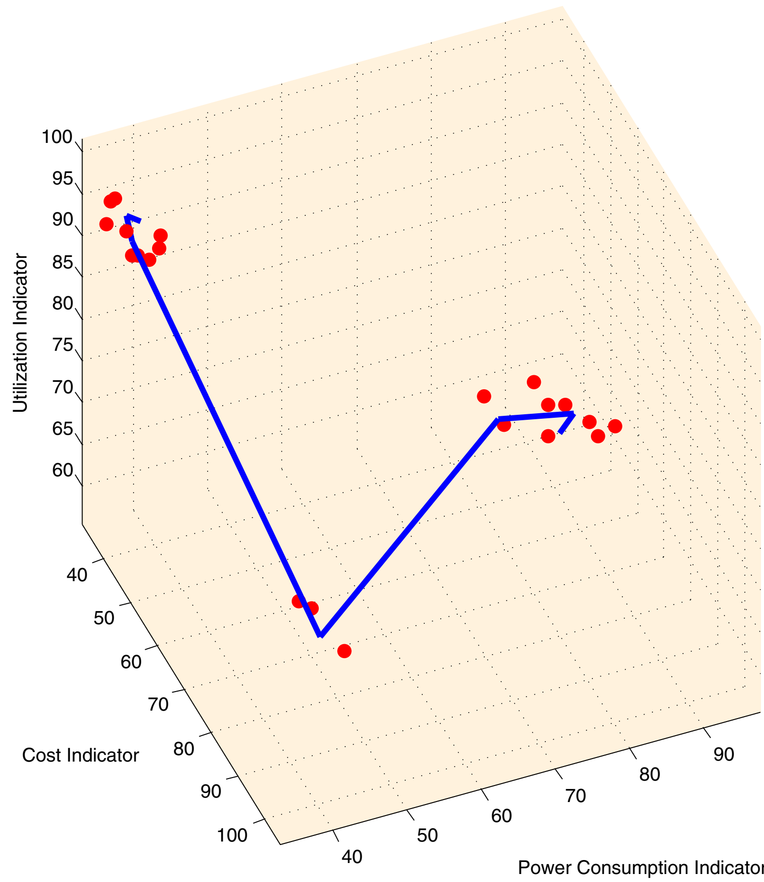

Figure 6 Operating point tracking on the time.

offloading, downloading, and uploading rates also should be determined in order to optimize the objective functions considering the constraints.

For instance, considering $x = 2$, the optimization problem will be as follows:

$$
\begin{aligned}
&Argmin_{\eta,\lambda,R,S}\left(\sum_{i=1}^{I} P_{\mathrm{comp}} S_i\right)+\left(\sum_{k=1}^{K}(Pt_k(H_k,R_k)+P_{\mathrm{maint}}(k))\right)\\
&ST: \sum_{i=1}^{I}\left(\chi_{i,\lambda}^{\mathrm{dep}}(\lambda_i)+\chi_{i,R}^{\mathrm{dep}}(R_i)+\chi_i^{\mathrm{indep}}\right)\le U_{th}\\
&\sum_{j=1}^{J}\left(\alpha_{\mathrm{comp}}(\gamma_j+\eta_j)+\alpha_j^{UL}R_j+\alpha_j^{DL}\lambda_j^{'}\right)\\
&\quad+\sum_{k=1}^{K}\left(\alpha_k^{w}\lambda_k^{"}+\sigma_k^{w}R_k^{"}\right)\le \mathrm{Cost}_{\mathrm{th}}\\
&E\{D_i(n)\}\le D_i^{th}\ \ \forall i\in A\\
&R_{\min}^{i}\le R^{i}(n)\le R_{\max}^{i}\ \ \forall i\in A\\
&0\le \eta_i(n)\le C_i(n)\ \ \forall i\in A\\
&\Sigma_{i=1}^{I}(\mu_{S_i}(n)+\zeta\sigma_{S_i}(n))<S\\
&\lambda_{\min}^{i}\le\lambda^{i}(n)\le\lambda_{\max}^{i}\ \ \forall i\in A\\
&\lambda_k(n)\le\mu_k(n)\ \ \forall k\in WN
\end{aligned}
$$

Here, first the best possible wireless network interface and cloud have to be selected for each active application. Next, process offloading, and variables such as download/upload rate and effective processor speed dedicated to the applications with the goal of minimizing the power consumption of the device will be calculated.

4.3 Problem solution

w_x^i is considered as the input of the proposed algorithm corresponding to the ith application such as WN, CR collections in the x state. u_x^i also is considered as the control variable vector related to the ith application such as offloading factor and ith application incoming traffic rate at the x state. u_x^i is selected from a predetermined set U. $U \subset R^I$ is restricted to Equations (17), (18), (19), and (21). We assumed that s_i is non-zero in all applications, because all applications need some process in a mobile phone. After multiplying Equation (5) by s_i, all constraints and objective functions will be linear in terms of η ,λ, and S. The only non-linear variable is power transmission rate. Bender decomposition method [7] is used to decompose the problem into functions linear in variables (i.e., η, λ, and S) and non-linear in variable R. Minimum amount of R_i could be easily found through Equations (5) and (20) as follows:

$$R_{\min}^{i}=\frac{D_i^{th}-(\hat{\beta}_j+\frac{C_i-\eta_i}{S_i}+\frac{\eta_i+\gamma_i}{S_j^{'}})-E\{Dw_i\}}{TD_i} \tag{23}$$

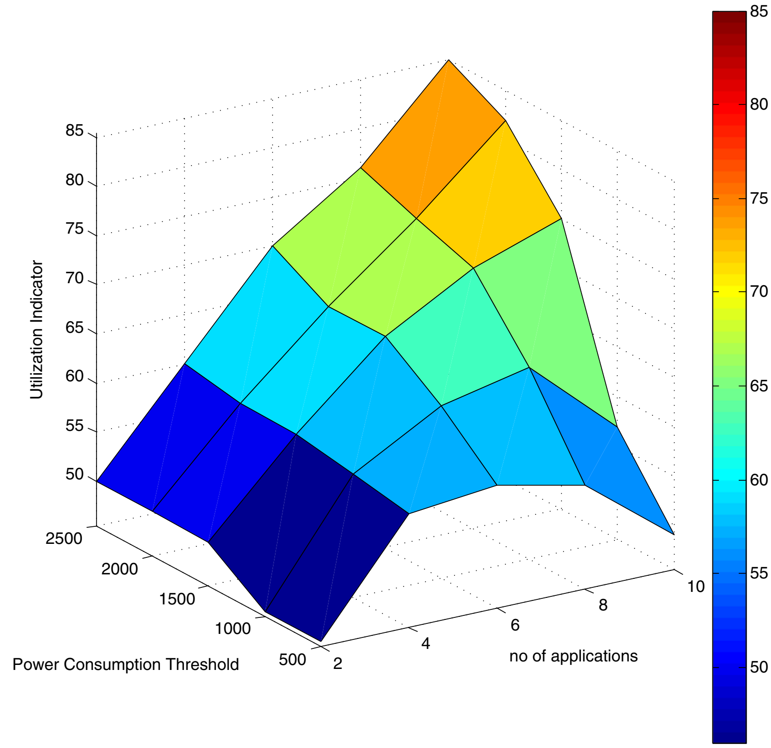

Figure 7 Utilization indicator for different power thresholds and different number of applications.

Then, R_i could be found by

$$R_i = R^i_{\text{min}} + \varepsilon_{R(i)} \tag{24}$$

where $\varepsilon_{R(i)}$ is dependent on the objective function and objective constraints. In order to find the $\varepsilon_{R(i)}$, simple incremental selection algorithm [50] is used. It ranges from 0 to $R^i_{\text{max}} - R^i_{\text{min}}$. However, if the transmission power consumption is linearly approximated in terms of transmission rate [30], then optimization problem will be simplified to a bilinear matrix inequality problem which could be solved with less complexity. $f^i_{\acute{x}}(u^i_x, w^i_x(j,k) \mid x)$ shows the $\acute{x}$th objective function derived from the ith application. Therefore, cost to go function could be written as below:

$$F^I_{x|x} = f^I_{x|x}(u^I_x, w^I_x(j,k) \mid x) + \sum_{i=1}^{I-1} f_{x|x}(u^i_x, w^i_x(j,k) \mid x) \tag{25}$$

Thus,

$$F^I_{x|x} = f^I_{x|x}(u^I_x, w^I_x(j,k) \mid x) + F^{I-1}_{x|x} \tag{26}$$

Also, the following constraint should be satisfied for the successive objective functions:

$$F^I_{\acute{x}|x} = f^I_{\acute{x}|x}(u^I_x, w^I_x(j,k) \mid x) + F^{I-1}_{\acute{x}|x} \leq F^{th}_{\acute{x}|x} \qquad \forall \acute{x} \in X, \acute{x} \neq x \tag{27}$$

Optimal solution to this problem could be found using dynamic programming (DP). It should be noted that applications are sorted according to their priorities and importance. Hence, the initial value of the objective is related to the most prioritized application. However, due to the constraints, computation complexity is much higher than a usual dynamic programming problem with brute-force search. A diagram based on [9] is proposed to find the optimal solution. Moreover, a method of learning from the mistakes [7] is used to restrict the feasible optimization region. Figure 5 shows the proposed algorithm structure. Using DP, the network and cloud resources are selected. In each DP step, linear programming output determines control variables of the system u^i_x. In order to decrease the complexity order, instead of brute-force search in resource and network selection, a policy is developed to assign the resources with the lowest objective values considering the application and resources constraints. The algorithm pseudo code is shown in Algorithm 1.

To improve modeling, TD_i and C_i could be approximated linearly by η_i and λ_i

$$TD_i = Y_{TD_i}\eta_i + Z_{TD_i}, C_i = Y_{C_i}\lambda_i + Z_{C_i} \tag{28}$$

Algorithm 1: Dynamic constraint programming

Data: CR,WN,$Q_{\min}$, $Q_{\max}$
Result: η , Q , F_x

events and states initialization;
while $\forall(u,w) \in \{\text{Valid}\} F_x^I(u^*,w^*) = \min F_x^I(u,w)$ **do**
  for $i = 1 \rightarrow I$ **do**
    for $j \in CR, \quad k \in WN$ **do**
      $w_x^i = Argmin_{j,k}(f_x^i \mid Q_{\min})$
    end for
    $u_x^i = LP_{\text{optimization}}(w_x^i \mid x)$ $//Argmin_{u_x^i} f(w_x^i, u_x^i \mid x)$
    if $Q(i) \notin (Q_{\min}(i), Q_{\max}(i))$ **then**
      update the valid region
      BREAK
    end if
    $f^*(w_x^i, u_x^i) = f(w_x^i, u_x^i)$
    $\forall x \in X, F_x^i(u^*,w^*) = F_x^{i-1}(u^*,w^*) + f_x^*(w_x^i, u_x^i)$
  end for
  if $\forall \acute{x} \in X, \acute{x} \neq x \quad F_x^I(u^*,w^*) \leq F_x^{th}$ **then**
    update the valid region
    BREAK
  end if
end while

where Y_{TD_i} and Z_{TD_i} are coefficients used in linear approximation of the uploading data of ith application in terms of offloading computation. Y_{C_i} is the incoming traffic dependent part of the ith application process while Z_{C_i} is its independent part.

For more precise resource allocation under some conditions, it is possible to approximate the functions by higher order of the Taylor series or other functions (e.g., exponential family). The proposed algorithm, using bender decomposition method, always breaks the optimization method into two different parts, namely linear and non-linear optimization part.

$$\begin{aligned} & Argmin_u g(u_{\text{nonlin}}) + p u_{\text{lin}} \\ & ST: z(u_{\text{nonlin}}) + q u_{\text{lin}} \leq G \\ & \qquad \text{resource constraints} \\ & \qquad \text{applications constraints} \end{aligned} \tag{29}$$

where $u_{\text{lin}}, u_{\text{nonlin}}$ represent linear and non-linear control variables, respectively. Therefore, master problem is divided into sub-problems. Using column generation techniques, two sub-optimization problems are minimized simultaneously. Integration of the two aforementioned linear and non-linear subproblems restricts the optimization feasible region and despite of the increasing complexity, it converges to an optimal solution.

5 Numerical results

In this section, some numerical results are presented to verify analysis of the previous section. Resources and application constraints are usually based on [28-30]. The resource characteristics considered in the numerical results are shown on Table 2. In the studied problem, we assumed that a main application such as CMG, video call, or media streaming is always present within the network, while considering presence of others as minor applications such as online social networks, health monitoring, or file and application download (I is considered a random variable between 2; 10 over the time). Fifty different cloud service providers with different characteristics have been considered in the network. The number of available WiFis is a random variable between 0 to 4. However, the number of available cellular networks varies from 1 to 5. If available resources are not enough for all applications, then the proposed algorithm allocates resources in order to maximize the objective function ignoring applications with less weight in the objective function. Connection maintenance power consumption includes elements such as receiving data power consumption. Receiving power consumption itself depends on several transmission parameters such as network contention [51].

Objective indicators are defined as follows: $u_{\text{indicator}}(n) = \frac{u(n)}{u_{\max}(n)} \times 100$, $P_{\text{indicator}}(n) = \frac{p_{\min}(n)}{p(n)} \times 100$ and $\text{cost}_{\text{indicator}}(n) = \frac{\text{cost}_{\min}(n)}{\text{cost}(n)} \times 100$ where $u_{\max}(n)$, $p_{\min}(n)$, and $\text{cost}_{\min}(n)$ represent the extremum achievable utilization, cost, and power consumption without considering the constraints of the mobile device at the nth time slot. Figure 6 shows the mobile device performance with respect to the time. As demonstrated, the mobile device had started with the best QoS strategy state and had a large utilization factor. After budget reduction state changed to cost effective mode, the algorithm minimized the cost of the mobile usage. Finally, when the remaining battery energy went lower than the threshold (10%) the state was changed to energy-saving mode which minimizes the device power consumption. Blue line shows the average tracking of the operating point through the time. Red points are the operating point samples during different time slots. Figure 7 shows power efficiency factor for different number of applications and different power consumption thresholds. As it is shown in Figure 7, generally with higher power consumption thresholds and more applications, utilization function tends to increase. However, for higher levels in number of applications with low power consumption threshold, utilization function is lower in comparison with applications

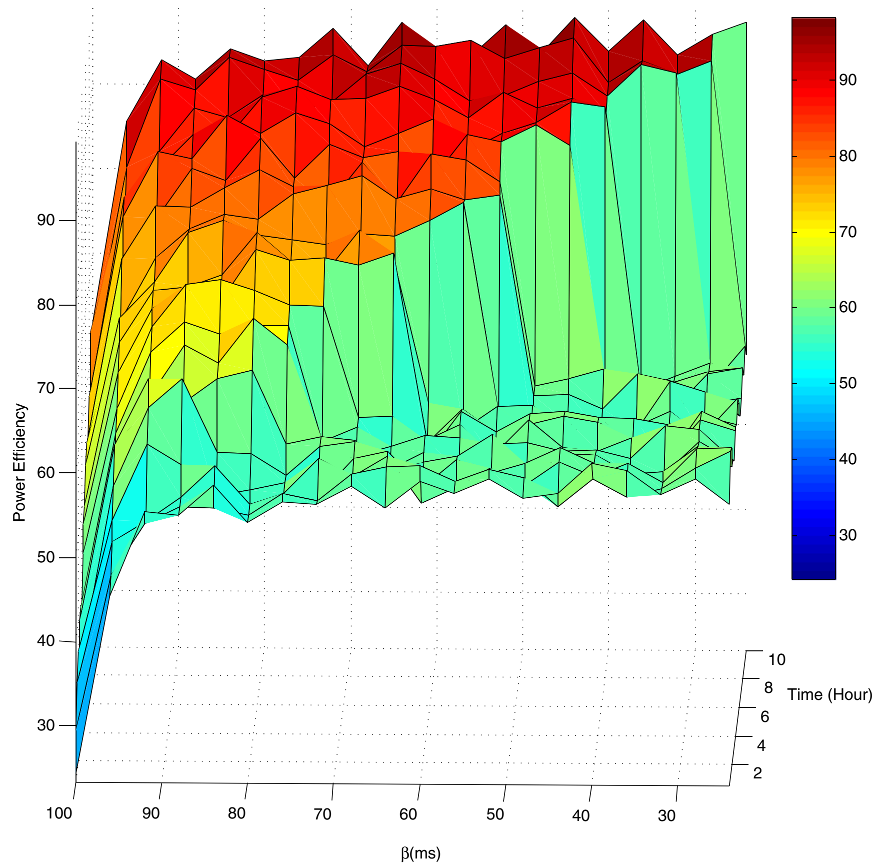

Figure 8 Power efficiency track over the time domain and different β parameters.

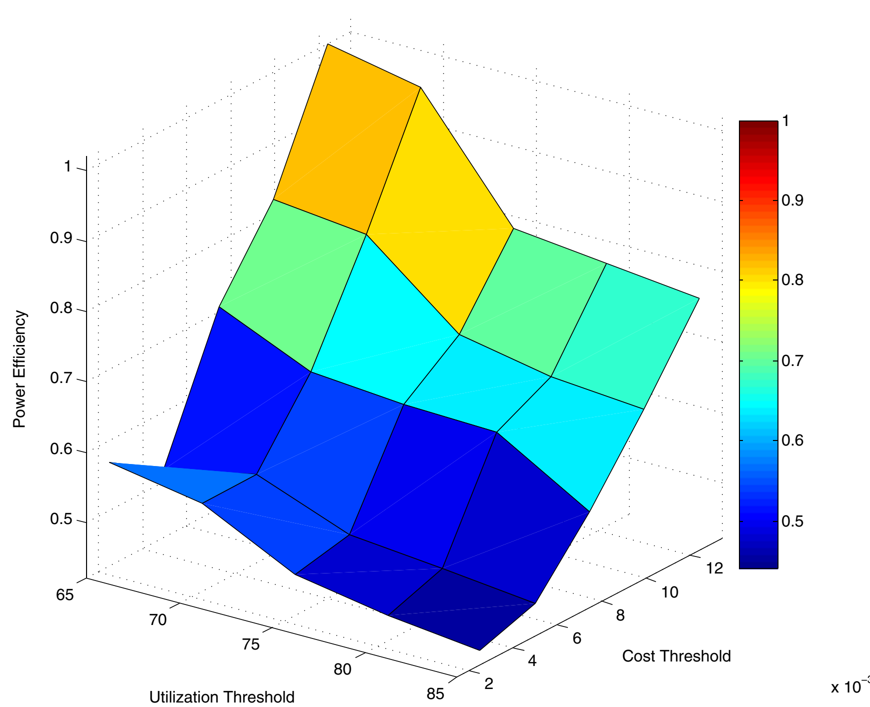

Figure 9 Power efficiency.

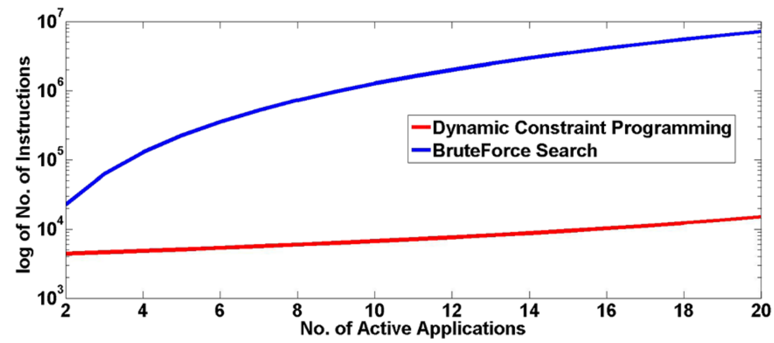

Figure 10 Complexity comparison.

1with the same power threshold limit due to larger number of constraints. Figure 8 depicts the system power efficiency over the time domain for different average network delays. Value of power efficiency function shows the state transition from the best QoS strategy or cost-effective mode to energy-effective mode. Figure 8 shows that state transition takes place earlier for larger amounts of $E(\beta)$. This indicates flexibility of the network in the proposed algorithm. However, for large values of $E(\beta)$, algorithm could not find a feasible optimum point satisfying all the constrains. In addition, power efficiency decreased in a dramatic way. Also, Figure 9 shows the power efficiency in terms of different utilization thresholds and cost thresholds. It is obvious that with decrease in average utilization threshold, while average cost threshold increases, power efficiency increases as well.

The complexity order of the proposed optimization algorithm is less than the complexity of brute-force search method. As depicted in Figure 10, the programming effort required which is defined by logarithm of the syntax lines to code the algorithm for the proposed dynamic constant programming is much less than the brute-force search.

6 Conclusions

In this paper, based on WNC concept, a system model for next generation of mobile communication has been considered. Cost, QoS, and power consumption functions are defined based on the system model. Next, a multi-dimensional optimization algorithm is proposed to optimize the objectives of a mobile user. The proposed multi-dimensional optimization algorithm takes network parameters, mobile device, and application constraints as input to optimally select the network resources and applications QoS profiles with optimum offloading coefficients. The proposed algorithm is established on event-based lexicographic optimization method and dynamic constraint programming. Numerical results for different environmental variables revealed that the proposed algorithm could be dynamically adaptive to environmental parameters variation. We have solved the optimization problem assuming particular linear approximations which may not be always valid. The next step could be extending the current work to the case of nonlinear functions and processes. In addition to the objectives of mobile users, performance metrics of cloud computing data centers and wireless operators can be considered as well.

Competing interests
The authors declare that they have no competing interests.

Received: 16 July 2014 Accepted: 7 October 2014
Published: 28 November 2014

References

1. X Hu, G Xing, JYT Leung, Exploring the interplay between computation and communication in distributed real-time scheduling. Compu. IEEE Trans. **60**(12), 1759–1771 (2011)
2. Y Lin, L Shao, Z Zhu, Q Wang, R Sabhikhi, Wireless network cloud: architecture and system requirements. IBM J. Res. Dev. **54**(1), 4–12 (2010)
3. S Jimaa, KK Chai, Y Chen, Y Alfadhl, in *Wireless and Mobile Computing,(WiMob), IEEE 7th International Conference On Networking and Communications, Shanghai, China*. LTE-A: an overview and future research areas, (2011), pp. 395–399
4. M Milosavljevic, S Sofianos, P Kourtessis, JM Senior, in *2013 IEEE International Conference On Communications Workshops (ICC), Budapest, Hungary*. Self-organized cooperative 5G RANS with intelligent optical backhauls for mobile cloud computing, (2013), pp. 900–904
5. P Demestichas, A Georgakopoulos, D Karvounas, K Tsagkaris, V Stavroulaki, J Lu, C Xiong, J Yao, 5G on the horizon: key challenges for the radio-access network. IEEE Vehicular Technol. Mag. **8**(3), 47–53 (2013)
6. PE Hladik, H Cambazard, AM Déplanche, N Jussien, Solving a real-time allocation problem with constraint programming. J. Syst. Softw. **81**(1), 132–149 (2008)
7. PE Hladik, H Cambazard, AM Déplanche, N Jussien, Dynamic constraint programming for solving hard real-time allocation problems. Network. **4**(m1), 2 (2005)
8. RT Marler, JS Arora, Survey of multi-objective optimization methods for engineering. Struct. Multidisciplinary Optimization. **26**(6), 369–395 (2004)
9. C Lee, J Lehoczky, D Siewiorek, R Rajkumar, J Hansen, in *20th IEEE Symposium on Real-Time Systems, Phoenix, Arizona, USA*. A scalable solution to the multi-resource qos problem, (1999), pp. 315–326
10. R Rajkumar, C Lee, JP Lehoczky, DP Siewiorek, in *19th IEEE Symposium on Real-Time Systems, Philadelphia, USA*. Practical solutions for QoS-based resource allocation problems, (1998), pp. 296–306
11. Q Duan, Y Yan, AV Vasilakos, A survey on service-oriented network virtualization toward convergence of networking and cloud computing. Netw. Service Manag. IEEE Trans. **9**(4), 373–392 (2012)

12. F Xu, F Liu, H Jin, AV Vasilakos, Managing performance overhead of virtual machines in cloud computing: a survey, state of the art, and future directions. Proc. IEEE. **102**(1), 11–31 (2014)
13. L Wang, F Zhang, JA Aroca, AV Vasilakos, K Zheng, C Hou, D Li, Z Liu, Greendcn: a general framework for achieving energy efficiency in data center networks. Selected Areas Commun. IEEE J. **32**(1), 4–15 (2014)
14. L Wang, F Zhang, AV Vasilakos, C Hou, Z Liu, Joint virtual machine assignment and traffic engineering for green data center networks. ACM SIGMETRICS Perform. Eval. Rev. **41**(3), 107–112 (2014)
15. D López-Pérez, X Chu, AV Vasilakos, H Claussen, Power minimization based resource allocation for interference mitigation in OFDMA femtocell networks. Selected Areas Commun. IEEE J. **32**(2), 333–344 (2014)
16. D López-Pérez, X Chu, AV Vasilakos, H Claussen, On distributed and coordinated resource allocation for interference mitigation in self-organizing LTE networks. IEEE/ACM Trans. Netw. (TON). **21**(4), 1145–1158 (2013)
17. MR Rahimi, J Ren, CH Liu, AV Vasilakos, N Venkatasubramanian, Mobile cloud computing: a survey, state of art and future directions. Mobile Netw. Appl. **19**(2), 133–143 (2014)
18. N Fernando, SW Loke, W Rahayu, Mobile cloud computing: a survey. Future Generation Comput. Syst. **29**(1), 84–106 (2013)
19. HT Dinh, C Lee, D Niyato, P Wang, A survey of mobile cloud computing: architecture, applications, and approaches. Wireless Commun. Mobile Comput. **13**(18), 1587–1611 (2013)
20. A Balasubramanian, R Mahajan, A Venkataramani, in *Proceedings of the 8th ACM International Conference on Mobile Systems, Applications, and Services*. Augmenting mobile 3G using WiFi (San Francisco, CA, USA, 2010), pp. 209–222
21. K Lee, I Rhee, J Lee, S Chong, Y Yi, *Mobile data offloading: how much can WiFi deliver?* vol. 21, (2013), pp. 536–550
22. B Han, P Hui, A Srinivasan, Mobile data offloading in metropolitan area networks. ACM SIGMOBILE Mobile Comput. Commun. Rev. **14**(4), 28–30 (2011)
23. Q Liu, J Yuan, X Shan, Y Wang, W Su, in *Global Mobile Congress (GMC), Shanghai, China*. Dynamic load balance scheme based on mobility and service awareness in integrated 3G/WLAN networks, (2010), pp. 1–6
24. V Gazis, N Alonistioti, L Merakos, Toward a generic 'always best connected' capability in integrated WLAN/UMTS cellular mobile networks (and beyond). Wireless Commun. IEEE. **12**(3), 20–29 (2005)
25. J Luo, R Mukerjee, M Dillinger, E Mohyeldin, E Schulz, Investigation of radio resource scheduling in WLANs coupled with 3G cellular network. Commun. Mag. IEEE. **41**(6), 108–115 (2003)
26. AK Salkintzis, C Fors, R Pazhyannur, WLAN-GPRS integration for next-generation mobile data networks. Wireless Commun. IEEE. **9**(5), 112–124 (2002)
27. C Ng, E Paik, T Ernst, M Bagnulo. Analysis of multihoming in network mobility support, RFC 4980 Internet, Draft draft IETF nemo multihoming issues 05 (IETF USA, 2007)
28. R Murmuria, J Medsger, A Stavrou, JM Voas, in *2012 IEEE Sixth International Conference On Software Security and Reliability (SERE), Gaithersburg, MD, USA*. Mobile application and device power usage measurements, (2012), pp. 147–156
29. A Carroll, G Heiser, in *Proceedings of the 2010 USENIX Annual Technical conference, Boston, MA, USA*. An analysis of power consumption in a smartphone (USENIX Association, 2010), pp. 271–285
30. N Balasubramanian author=Balasubramanian, A, pp. 280–293
31. E Cuervo, A Balasubramanian, D Cho, A Wolman, S Saroiu, R Chandra, P Bahl, in *Proceedings of the 8th International ACM Conference on Mobile Systems, Applications, and Services, San Francisco, CA, USA*. Maui: making smartphones last longer with code offload (ACM, 2010), pp. 49–62
32. S Kosta, A Aucinas, P Hui, R Mortier, X Zhang, in *Proceedings of IEEE INFOCOM, Orlando, FL, USA*. ThinkAir: dynamic resource allocation and parallel execution in the cloud for mobile code offloading (IEEE, 2012), pp. 945–953
33. K Kumar, YH Lu, Cloud computing for mobile users: can offloading computation save energy? Computer. **43**(4), 51–56 (2010)
34. W Fang, Y Li, H Zhang, N Xiong, J Lai, AV Vasilakos, On the throughput-energy tradeoff for data transmission between cloud and mobile devices. Inf. Sci. **283**, 79–93 (2014)
35. H Zhang, B Li, H Jiang, F Liu, AV Vasilakos, J Liu, in *Proceedings IEEE INFOCOM 2013, Turin, Italy*. A framework for truthful online auctions in cloud computing with heterogeneous user demands, (2013), pp. 1510–1518
36. G Wei, AV Vasilakos, Y Zheng, N Xiong, A game-theoretic method of fair resource allocation for cloud computing services. J. Supercomputing. **54**(2), 252–269 (2010)
37. MR Rahimi, N Venkatasubramanian, S Mehrotra, AV Vasilakos, in *Proceedings of the 5th IEEE/ACM Fifth International Conference on Utility and Cloud Computing, Chicago, USA*. MAPCloud: mobile applications on an elastic and scalable 2-tier cloud architecture (IEEE Computer Society, 2012), pp. 83–90
38. MR Rahimi, N Venkatasubramanian, AV Vasilakos, in *IEEE Sixth International Conference On Cloud Computing,Santa Clara, CA, USA*. MuSIC: mobility-aware optimal service allocation in mobile cloud computing (IEEE, 2013), pp. 75–82
39. D Niyato, AV Vasilakos, Z Kun, in *Proceedings of the 2011 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, Newport Beach, CA, USA*. Resource and revenue sharing with coalition formation of cloud providers: Game theoretic approach (IEEE Computer Society, 2011), pp. 215–224
40. CS Chang, *Performance Guarantees in Communication Networks*. (Springer-Verlag, New York, USA, 2000), pp. 102–245
41. D Wu, R Negi, Effective capacity: a wireless link model for support of quality of service. Wireless Commun. IEEE Trans. **2**(4), 630–643 (2003)
42. J Tang, X Zhang, Quality-of-service driven power and rate adaptation over wireless links. Wireless Commun. IEEE Trans. **6**(8), 3058–3068 (2007)
43. J Tang, X Zhang, Cross-layer modeling for quality of service guarantees over wireless links. Wireless Commun. IEEE Trans. **6**(12), 4504–4512 (2007)
44. ZL Zhang, End-to-end support for statistical quality of service guarantees in multimedia networks. PhD Thesis Dissertation
45. R Deutsch, *Estimation Theory*. (Prentice-Hall, New Jersey, USA, 1965)
46. A Kansal, A Karandikar, in *IEEE Global Telecommunications Conference, San Antonio, TX, USA, vol. 4*. Adaptive delay estimation for low jitter audio over INTERNET (IEEE, 2001), pp. 2591–2595
47. A Rahmati, L Zhong, in *Proceedings of the 5th International Conference on Mobile Systems, Applications and Services, San Juan, Puerto Rico, USA*. Context-for-wireless: context-sensitive energy-efficient wireless data transfer (ACM, 2007), pp. 165–178
48. M Wang, X Meng, L Zhang, in *Proceedings of IEEE INFOCOM 2011, Shanghai, China*. Consolidating virtual machines with dynamic bandwidth demand in data centers, (2011), pp. 71–75
49. F Chen, K Guo, J Lin, T La Porta, in *Proceedings of IEEE INFOCOM, Orlando, FL, USA*. Intra-cloud lightning: building CDNs in the cloud (IEEE, 2012), pp. 433–441
50. Q Du, X Zhang, QoS-aware base-station selections for distributed MIMO links in broadband wireless networks. Selected Areas Commun. IEEE J. **29**(6), 1123–1138 (2011)
51. J Manweiler, R Roy Choudhury, in *Proceedings of the 9th International Conference on Mobile Systems, Applications, and Services, Washington DC, USA*. Avoiding the rush hours: WiFi energy management via traffic isolation (ACM, 2011), pp. 253–266

doi:10.1186/1687-1499-2014-201
Cite this article as: Vakilinia *et al.*: **Optimal multi-dimensional dynamic resource allocation in mobile cloud computing.** *EURASIP Journal on Wireless Communications and Networking* 2014 **2014**:201.