Information Sciences 283 (2014) 79–93

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

On the throughput-energy tradeoff for data transmission between cloud and mobile devices

Weiwei Fang [a,*], Yangchun Li [a], Huijing Zhang [a], Naixue Xiong [b], Junyu Lai [c], Athanasios V. Vasilakos [d]

[a] School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China
[b] School of Computer Science, Colorado Technical University, Colorado Springs, Colorado 80907, USA
[c] School of Aeronautics and Astronautics, University of Electronic Science and Technology of China, Chengdu 611731, China
[d] Department of Computer Science, Kuwait University, Kuwait

ARTICLE INFO

Article history:
Received 29 October 2013
Received in revised form 26 May 2014
Accepted 15 June 2014
Available online 26 June 2014

Keywords:
Mobile cloud computing
Throughput-energy tradeoff
Lyapunov optimization
Dynamic scheduling
Queue stability

ABSTRACT

Mobile cloud computing has recently emerged as a new computing paradigm promising to improve the capabilities of resource-constrained mobile devices. As the data processing and storage are moved from mobile devices to powerful cloud platforms, data transmission has become an important issue affecting user experiences of mobile applications. One of the challenges is how to optimize the tradeoff between system throughput and energy consumption, which are potentially conflicting objectives. Inspired by the feasibility of transmission scheduling for prefetching-friendly or delay-tolerant applications, we mathematically formulate this problem as a stochastic optimization problem, and design an online control algorithm to balance such an energy-performance tradeoff based on the Lyapunov optimization framework. Our algorithm is able to independently and simultaneously make control decisions on admission and transmission to maximize a joint utility of the average application throughput and energy cost, without requiring any statistical information of traffic arrivals and link bandwidth. Rigorous analysis and extensive simulations have demonstrated both the system stability and the utility optimality achieved by our control algorithm.

1. Introduction

Mobile devices, such as smartphone, PDA, and tablet-PC, are gradually becoming an important part of human life as the most convenient communication tools unbounded by time and space. However, mobile devices are facing severe challenges in resources (e.g., computation, storage and energy) and communications (e.g., bandwidth and mobility), which greatly impede the improvement of user experiences. In recent years, the paradigm of mobile cloud computing has been introduced to extend capabilities of mobile devices, by taking advantage of high-speed wireless communications and high-performance cloud platforms [7,21] to help gather, store and process data for the mobile devices [22,14]. In this new paradigm, the cloud-based mobile applications usually require frequent data exchanges between mobile devices and cloud platforms [44]. With the surging popularity of mobile cloud computing, it has been found that traffic from mobile devices grows three times faster

* Corresponding author. Tel.: +86 1051683617.
E-mail addresses: wwfang@bjtu.edu.cn (W. Fang), phantom_lyc@sina.com (Y. Li), huijingzhangpau@gmail.com (H. Zhang), nxiong@coloradotech.edu (N. Xiong), laijy@uestc.edu.cn (J. Lai), vasilako@ath.forthnet.gr (A.V. Vasilakos).

http://dx.doi.org/10.1016/j.ins.2014.06.022
0020-0255/

than that from wired devices, and will exceed the latter by year 2016 [8]. Unlike wired devices, current mobile devices and their daily use are seriously constrained by the energy capacity of batteries [23,17]. Hence, the data transmission strategy for mobile devices has to take into account not only providing sufficient system throughput to satisfy application requirements, but also conserving precious battery energy to prolong operational lifetime.

This paper explores such a throughput-energy trade-off for prefetching-friendly or delay-tolerant applications on mobile devices having multiple types of network interfaces. To support pervasive Internet access, current mobile devices are increasingly being equipped with more than one wireless interfaces [9], such as WiFi, 3G-HSDPA, and 4G-LTE [3,13], that have substantially different characteristics. Firstly, the availability of these types of networks can vary significantly at different places [32,35]. At least as of now, the coverage and penetration capabilities of 3G and 4G are much higher than those of WiFi. Secondly, the achievable data rates of these interfaces can vary considerably over time, and are sometimes far less than the nominal values specified by the technical standards [13,35]. For example, it is reported in [13] that the downlink bandwidth measured in U.S. range from 0.35Mbps to 19.27 Mbps for 802.11g WiFi (up to 54 Mbps), and from 2.08 Mbps to 30.80 Mbps for LTE (up to 100 Mbps). Thirdly, the energy costs for transmitting a given amount of data over these wireless interfaces could differ by an order of magnitude [13,30]. It has been revealed by empirical studies in [13] that the radio power level has a linear relationship with link transmission bandwidth, while the power coefficients are distinct for different types of interfaces. Based on above-mentioned reasons, the selection of available wireless interfaces for data transmission has direct impacts on system throughput and energy consumption of mobile devices.

Fortunately, many of the mobile applications are naturally prefetching-friendly or delay-tolerant, to different degrees, so that it is possible to switch among multiple radio interfaces to achieve energy-efficient data transfer. For example, mobile users may rely on the online map service when visiting new places. When they use the mobile client to request online maps, the cloud could pre-push some potentially useful maps (e.g., historic sites) to the device during periods of good connectivity of a lower-energy link. On the other hand, delay-tolerant applications such as online social networking can be set to fetch the content update at specific intervals. This create opportunities for energy saving by deferring the transmission of the latest data until a satisfactory link connection becomes available. However, one-sided pursuit of energy saving may degrade system throughput and application performance.

To address the challenges above, we propose a new online control algorithm, MOTET, for Mobile device Optimization on the Throughput-Energy Tradeoff using the Lyapunov optimization framework [27]. MOTET maximizes a joint utility of the sum throughput of applications and the energy costs of device, by independently and simultaneously making online decisions to control admission and transmission behaviors [11,20]. It is associated with two control parameters, i.e., throughput-energy parameter and stability-utility parameter, which can be appropriately tuned to provide a desired performance tradeoff among application throughput, energy cost and service delay. Specifically, a non-negative parameter θ is used to normalize the values of throughput and energy to make them comparable in the utility function. Meanwhile, MONET can obtain a time-average utility within a deviation of $\mathcal{O}(1/V)$ from optimality, while bounding the traffic queue length and the traffic service delay by $\mathcal{O}(V)$, where V is a non-negative control parameter representing a design knob of the stability-utility tradeoff (i.e., how much we emphasize utility maximization compared to system stability). MOTET operates without requiring any statistical knowledge of traffic arrivals and link conditions, and is computationally efficient for implementing on resource-constrained mobile devices. We thoroughly analyze the performance of our new proposed online control algorithm with rigorous theoretical analysis. To complement the analysis, we conduct a simulation study to evaluate MOTET using datasets from real-world measurements on wireless link bandwidth and transmission energy consumption of mobile devices [13]. Experimental results demonstrate that MOTET can approach a time-average utility that is arbitrarily close to the optimum, while still maintaining strong stability and low congestion. Furthermore, with an appropriate throughput-energy parameter, MOTET can significantly reduce the energy expenditure while only incurring a marginal sacrifice in the system throughput. To our knowledge, prior work has not explored such a tradeoff issue on mobile devices, and our use of the Lyapunov optimization framework for solving this issue is also novel.

The rest of this paper is organized as follows: Section 2 reviews some related work. Section 3 describes the theoretical model for throughput-energy tradeoff, and also formulates the objective problem. Section 4 presents the optimal control algorithm, and Section 5 provides an analysis on performance bounds of our algorithm. Section 6 shows the performance evaluation results. Finally, Section 7 concludes the paper.

2. Related work

The contribution of our work lies in the intersection of the following two important cutting-edge research topics.

2.1. Efficient transmission techniques for mobile devices

Advances in the portability and capability of mobile devices, together with widespread 3G/4G networks and WiFi access, have brought rich mobile application experiences to end users. Some existing approaches focus on improving system throughput for real-time applications, such as streaming and downloading. A feedback-based control technique is proposed in [28] to automatically adapt streaming parameters to bandwidth fluctuations so as to improve throughput and reduce delay. COMBINE [2] uses both the WiFi and the 3G/4G links in combination for collaborative downloading. An application-layer

active wireless network switching system is proposed in [40] to automatically select the best wireless connection for data transmission on the smartphone. A recent work, Adapp [6], is capable to optimally select the network service that suits an application best in terms of user-desired quality of experience (QoE). However, these methods focus predominantly on throughput enhancement rather than energy saving.

Current mobile devices are severely constrained by limited battery capacity. It has long been revealed that the power drained by network interfaces constitutes a large fraction of the total power used by a mobile device [36]. Such a situation becomes even worse with the rising popularity of mobile cloud computing [18]. Hence, recent efforts have been made on transmission scheduling of radio interfaces to improve energy efficiency of mobile devices. Some of them are proposed for mobile devices with a single wireless interface. Neely [25] developed a joint strategy (EECA) for power allocation and admission control that satisfies the constraints on transmit power while maximizing the throughput metric. Based on empirical studies on tail energy (i.e., energy consumed in high-power state after data transfer is completed), TailEnder [5] aims to aggregate small transmissions into large ones through prefetching and delayed transfer, so that the occurrence of tails (and thus energy consumption) can be reduced. Similar schemes include TailTheft [19] and Traffic-Backfilling [15]. On the other hand, some prior studies try to achieve energy efficiency by scheduling data transfer among several available interfaces on mobile devices. CoolSpots [30] aims to reduce the energy consumption by intelligently deciding whether and when to use WiFi and Bluetooth based on an application's bandwidth requirement. Context-for-Wireless [33] employs the statistical information of historical context to decide whether and when to power-on the WiFi interface to improve the energy efficiency of data transfer in cellular networks. Both SALSA [32] and eTime [35] apply Lyapunov optimization techniques to make online transmission decisions to balance the tradeoff between energy consumption and transmission delay. However, these four schemes are interested in determining the lowest energy link among a set of available links at a given instant, without taking any application requirement on network throughput into consideration.

2.2. Lyapunov optimization techniques for stochastic systems

Lyapunov optimization [27] is a recently developed technique for solving problems of joint system stability and performance optimization on stochastic networks, especially communication and queueing systems. To achieve this goal, network algorithms are designed to make control actions that greedily minimize a bound on the following drift-plus-penalty expression in each time slot t:

$$\Delta(t) + Vc(t)$$

where $\Delta(t)$ (Lyapunov drift) represents the congestion state of queue backlog, $c(t)$ denotes the objective function to be optimized, and V is a non-negative weight that is chosen as desired to affect a performance tradeoff between backlog reduction and penalty minimization. Unlike Markov Decision Process [38] and Dynamic Programming [1], Lyapunov optimization does not require knowledge of the statistics of related stochastic models, but instead the queue backlog information, to make online control decisions. These two traditional techniques suffer from the so-called "curse of dimensionality" problem [27], and result in hard-to-implement systems where significant re-computation might be needed when statistics change [37]. In contrast, Lyapunov optimization algorithms commonly have a better computational complexity, and are easy to be implemented in practical systems [32,35]. By now, this new technique has been applied in solving many stochastic network optimization problems, including workload/resource scheduling among data centers [39,31,43], power management in smart grid [37,10,41], and energy/throughput optimization for wireless systems [26,32,35]. Among them, the work most relevant to ours is that in [43], which makes online decisions on request admission control, routing and virtual machine scheduling to balance the tradeoff between the throughput performance and power consumption in SaaS (i.e., Software-as-a-Service) clouds. This work inspires us to exploit a similar throughput-energy tradeoff issue in mobile cloud computing scenarios.

3. Basic throughput-energy tradeoff model

As shown in Fig. 1, we consider a mobile device user who has M heterogeneous applications $m \in \{0, 1, 2, \ldots, M\}$ running on a cloud platform [34]. The whole system operates in discrete time with unit time slots $t \in \{0, 1, 2, \ldots\}$. The data

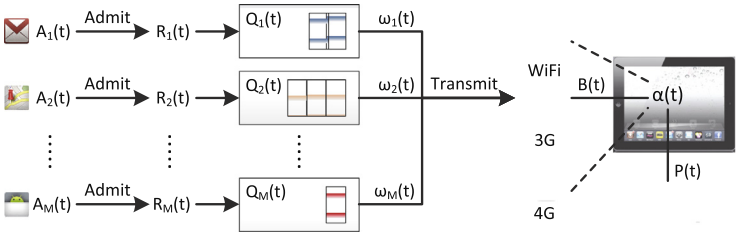


Fig. 1. The system model for MOTET.

(i.e., content) generated by each application is processed in a corresponding queue, denoted as $\mathbf{Q}(t) \triangleq (Q_1(t), \ldots, Q_M(t))$, in which $Q_m(t)$ represents the queue backlog of application m's data to be transmitted from the cloud to the mobile device at the beginning of time slot t. In every time slot t, we denote the amount of newly generated data arriving at queue Q_m by $A_m(t)$, where $\mathbf{A}(t) = (A_1(t), \ldots, A_M(t))$ denotes the arrival vector. We assume that $\mathbf{A}(t)$ are independent and identically distributed (i.i.d.) over time slots [32,27], with the expectations of $\mathbb{E}\{\mathbf{A}(t)\} = \lambda \triangleq (\lambda_1, \ldots, \lambda_M)$. We also assume that there exists a finite maximum A_m^{max} such that $0 \leqslant A_m(t) \leqslant A_m^{max}$ for all t and all $m \in \{1, \ldots, M\}$. However, we do not assume any priori knowledge of the statistics of $A_m(t)$. For example, $A_m(t)$ could be a Markov-modulated process with time-varying instantaneous rates where the transition probabilities between different states are unknown. This models a scenario with unpredictable and time-varying traffic arrivals.

3.1. Control decisions and queue dynamics

Under the data arrival model above, we focus on two important control decisions to be made. In each time slot t, the first control decision of a mobile device is to determine the amount of application data out of $A_m(t)$, denoted by $R_m(t)$, that can be admitted by Q_m into the system. Any new data that is not admitted is treated as declined. This can easily be generalized to the case where arrivals that are not immediately accepted are stored in a buffer for future admission decision. Thus, for all m and t, we have $0 \leqslant R_m(t) \leqslant A_m(t)$.

While the amount of admitted data $R_m(t)$ are waiting in the corresponding queue Q_m, the other important control decision in time slot t is to determine whether to use any of the available links (e.g., WiFi, 3G or 4G) to transfer data, and if so, which one would be used. Let $\alpha(t)$ denotes this transmission decision, and the vector of data service rates [26] $\omega(t) = (\omega_1(t), \ldots, \omega_M(t))$ is jointly determined by $\alpha(t)$ and the current link condition $L(t)$. We specify $L(t)$ as the achievable link (i.e., downlink [35]) bandwidth of wireless interfaces in time slot t, as bandwidth is the most critical factor in both the system throughput and the energy consumption of data transmission [13]. Specifically, the network controller observes the current $L(t)$ and selects $\alpha(t)$ within some abstract set $\mathcal{A}$ that specifies the decision options. Then, the service rates for slot t can be given by functions $\hat{\omega}_m(\alpha, L)$ as $\omega_m(t) = \hat{\omega}_m(\alpha(t), L(t))$ for each $m \in \{1, \ldots, M\}$. We assume a maximum transmission rate ω_m^{max}, regardless of $\alpha(t)$ and $L(t)$, so that $0 \leqslant \hat{\omega}_m(\alpha(t), L(t)) \leqslant \omega_m^{max}$.

Finally, we can capture the following queueing dynamics over time for applications $m \in \{1, \ldots, M\}$ of a mobile device user:

$$Q_m(t+1) = \max[Q_m(t) - \omega_m(t), 0] + R_m(t) \tag{1}$$

Accordingly, we can formally define the stability constraint on the queues, which ensures that the average queue length is finite. The queue stability of this system can be defined as:

$$\overline{Q} \triangleq \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{m=1}^{M} \mathbb{E}\{Q_m(t)\} < \infty \tag{2}$$

With the data transmission scheduling above, it is intuitive that the wireless link currently providing higher bandwidth may be more preferable to be used for satisfying throughput requirements. However, this will lead to a corresponding rise in battery energy consumption for the mobile applications. We shall characterize such a throughput-energy tradeoff in the following subsection.

3.2. Characterizing and optimizing the energy-performance tradeoff

3.2.1. System throughput

For the cloud-based mobile applications, one of the most important performance metrics is the overall application throughput in terms of the generated data that can be admitted and transmitted. Specifically, for each application $m \in \{1, \ldots, M\}$, we define the time-average throughput r_m of a mobile application as:

$$r_m \triangleq \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\{R_m(t)\} \tag{3}$$

Then, $\sum_{m=1}^{M} r_m$ is the throughput objective that is expected to be maximized, subject to the following two constraints: (1) $0 \leqslant r_m \leqslant \lambda_m$, because the time-average throughput r_m cannot exceed the time-average arrival rate λ_m for any application $m \in \{1, \ldots, M\}$; (2) $0 \leqslant r_m \leqslant \omega_m^{max}$, because the time-average throughput r_m cannot exceed the maximal service rate of Q_m for any application $m \in \{1, \ldots, M\}$.

3.2.2. Energy consumption

It has been revealed by recent studies [13,35] that, the amount of energy consumed for data transfer by a mobile device is primarily associated with the wireless interface used and its current link bandwidth, as formally characterized as follows:

$$P(t) = [\alpha B(t) + \beta]\tau \tag{4}$$

where α and β denote the empirical coefficients in the power model, and different types of interfaces have distinct power coefficients [13]. Besides, τ denotes the time span of one time slot, and $B(t) = \widehat{B}(\alpha(t), L(t))$ is the bandwidth of current selected link in time slot t. Since the system bandwidth is shared by all the M applications, we know that $B(t)\tau = \sum_{m=1}^{M} \omega_m(t)$.

Based on the power model above, the time-average energy consumption can be defined as:

$$p \triangleq \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\{P(t)\} \tag{5}$$

Then, p is the energy objective that is expected to be minimized.

3.2.3. A unified objective from an economic perspective

In Sections 3.2.1 and 3.2.2, we have derived both the performance metric and the energy metric. However, the primary challenge is how to optimize the tradeoff between these two potentially conflicting objectives in a unified manner. To address this issue, we construct a unified utility objective to couple both sides in an economic way.

Specifically, we price the throughput of each application $m \in \{1, \ldots, M\}$ as a logarithmic function $g(r_m) = log(1 + r_m)$, according to the law of diminishing marginal utility in microeconomics [43]. This law states that the marginal utility gained by consuming equal successive units of a good will decline as the amount consumed increases. Such a rule has already been adopted by wireless network algorithms (e.g., [12]) and ISP bandwidth pricing schemes (e.g., [4]).

Given the revenue brought by the system throughput and the incurred energy consumption, our objective is to maximize the time-average utility as follows:

$$\begin{aligned} \max \quad & \sum_{m=1}^{M} g(r_m) - \theta p \\ s.t. \quad & 0 \leqslant r_m \leqslant \lambda_m, \forall m \in \{1, \ldots, M\} \\ & 0 \leqslant r_m \leqslant \omega_m^{max}, \forall m \in \{1, \ldots, M\} \\ & \overline{Q} < \infty \\ & \alpha(t) \in \mathcal{A}, \forall t \\ & 0 \leqslant R_m(t) \leqslant A_m(t) \leqslant A_m^{max}, \forall m \in \{1, \ldots, M\}, \forall t \end{aligned} \tag{6}$$

where $\theta \geqslant 0$ is a scaler used to normalize the values of throughput and energy to make them comparable in the utility function. In special cases, when $\theta = 0$, the controller does not consider the energy consumption, whereas when $\theta \to \infty$, the controller does not consider the system throughput. Note that such an affine combination of two performance metrics is a common approach in multi-objective optimization.

Our objective is to design a flexible and efficient online control algorithm that can solve this long-term optimization problem. As stated in Section 2, traditional techniques require substantial statistics of system dynamics (e.g., traffic arrivals and link conditions), and suffer from excessive computational complexity. By comparison, the recently developed Lyapunov optimization framework has shown its efficacy and efficiency in designing online control algorithms for such constrained optimization of time-varying systems, without requiring any prior knowledge or prediction on system dynamics. In particular, the model above well fits the framework for optimizing functions of time averages [27].

4. Online control algorithm design

4.1. Problem transformation

To solve problem (6), we need to transform it to a solvable form. The Lyapunov optimization framework has provided the "auxiliary variables" to transform the problem into a new one involves only time averages rather than functions of time averages, and hence can be solved by the drift-plus-penalty framework [27]. Since the utility function $g(r_m)$ is concave and nonlinear, we introduce auxiliary variables γ_m for each admitted traffic stream $R_m(t), \forall m \in \{1, \ldots, M\}$, as follows:

$$\begin{aligned} \max \quad & \sum_{m=1}^{M} g(\gamma_m) - \theta p \\ s.t. \quad & \gamma_m \leqslant r_m, \in \{1, \ldots, M\} \\ & 0 \leqslant r_m \leqslant \lambda_m, \forall m \in \{1, \ldots, M\} \\ & 0 \leqslant r_m \leqslant \omega_m^{max}, \forall m \in \{1, \ldots, M\} \\ & \overline{Q} < \infty \\ & \alpha(t) \in \mathcal{A}, \forall t \\ & 0 \leqslant R_m(t) \leqslant A_m(t) \leqslant A_m^{max}, \forall m \in \{1, \ldots, M\}, \forall t \\ & 0 \leqslant \gamma_m(t) \leqslant A_m^{max}, \forall m \in \{1, \ldots, M\}, \forall t \end{aligned} \tag{7}$$

It is obvious that problem (7) is equivalent to that of the original problem (6), since the function $g(x) = log(1+x)$ is concave, continuous and non-decreasing. To solve problem (7), the inequality constraint $\gamma_m \leqslant r_m$ is transformed into a queue stability model [27]. Specifically, virtual queues H_m for each $R_m(t)$ is introduced. Define $H_m(0) = 0$, and define the virtual queue $H_m(t)$ for $t \in \{0, 1, 2, \ldots\}$, according to the following update:

$$H_m(t+1) = \max[H_m(t) - R_m(t) + \gamma_m(t), 0] \tag{8}$$

where $\gamma_m(t)$ denotes a process of non-negative auxiliary variables that is determined in every time slot t to satisfy the constraint $\gamma_m \leqslant r_m$ and $\gamma_m \triangleq \lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\{\gamma_m(t)\}$.

Lemma 1. *For all time slots t and a given $m \in \{1, \ldots, M\}$, $H_m(t)$ is stable if and only if $\gamma_m \leqslant r_m$.*

Proof. The proof of this lemma is similar to that of Theorem 2.4 of [27], and is omitted here for brevity. □

Besides, we develop a novel virtual queue, called an ϵ-persistent service queue, that can ensure bounded worst case delay guarantee on any buffered traffic arrivals in Q_m. To this end, for any $m \in \{1, \ldots, M\}$ define a virtual queue Z_m with initial backlog $Z_m(0) = 0$, and with queue update:

$$Z_m(t+1) = \max[Z_m(t) - \omega_m(t) + \epsilon_m 1_{\{Q_m(t)>0\}}, 0] \tag{9}$$

where $\epsilon_m > 0$ are pre-specified constants, and $1_{\{Q_m(t)>0\}}$ is an indicator function that is 1 if $Q_m(t) > 0$, and 0 else. The intuition is that Z_m has the same service process as Q_m, but has an arrival process that adds ϵ_m whenever the actual queue backlog is non-empty, which ensures that Z_m grows if there is data traffic in the Q_m queue that has not been serviced for a long time.

4.2. Lyapunov optimization

Let $\mathbf{\Theta}(t) \triangleq (\mathbf{Q}(t), \mathbf{H}(t), \mathbf{Z}(t))$ be a concatenated vector of all $Q_m(t), H_m(t)$, and $Z_m(t)$ queues. As a scalar measure of all the queue lengths, we define a quadratic Lyapunov function [24] as follows:

$$L(\mathbf{\Theta}(t)) \triangleq \frac{1}{2}\sum_{m=1}^{M}[Q_m(t)^2 + H_m(t)^2 + Z_m(t)^2] \tag{10}$$

Then, the one-slot conditional Lyapunov drift $\Delta(\mathbf{\Theta}(t))$ is defined as:

$$\Delta(\mathbf{\Theta}(t)) \triangleq \mathbb{E}[L(\mathbf{\Theta}(t+1)) - L(\mathbf{\Theta}(t))|\mathbf{\Theta}(t)] \tag{11}$$

Following the drift-plus-penalty framework in Lyapunov optimization, our control algorithm is designed to make decisions on $R_m(t)$, $\alpha(t)$, and $\gamma_m(t)$ to minimize an upper bound on the following drift-plus-penalty term in each time slot:

$$\Delta(\mathbf{\Theta}(t)) - V\mathbb{E}\left\{\sum_{m=1}^{M} g(\gamma_m(t)) - \theta P(t)|\mathbf{\Theta}(t)\right\} \tag{12}$$

where V is a non-negative parameter set by system operators to control the tradeoff between utility maximization (i.e., problem (7)) and system stability. Note that the sign of the penalty expression in (12) is negative, since we need to transform the maximization problem (6) into an equivalent one of penalty minimization in Lyapunov optimization [27,43].

Theorem 1. *(Drift-plus-Penalty Bound) Under any control algorithm, the drift-plus-penalty expression has the following upper bound for all t, all possible values of $\mathbf{\Theta}(t)$, and all parameters $V \geqslant 0$:*

$$\begin{aligned}\Delta(\mathbf{\Theta}(t)) - V\mathbb{E}\left\{\sum_{m=1}^{M} g(\gamma_m(t)) - \theta P(t)|\mathbf{\Theta}(t)\right\} \leqslant B + \mathbb{E}\left\{\sum_{m=1}^{M} Z_m(t)\epsilon_m|\mathbf{\Theta}(t)\right\} - \mathbb{E}\left\{\sum_{m=1}^{M}[Vg(\gamma_m(t)) - H_m(t)\gamma_m(t)]|\mathbf{\Theta}(t)\right\} \\ + \mathbb{E}\left\{\sum_{m=1}^{M} R_m(t)[Q_m(t) - H_m(t)]|\mathbf{\Theta}(t)\right\} \\ - \mathbb{E}\left\{\sum_{m=1}^{M}[(Q_m(t) + Z_m(t))\omega_m(t)] - V\theta P(t)|\mathbf{\Theta}(t)\right\}\end{aligned} \tag{13}$$

where $B = \sum_{m=1}^{M}(A_m^{max})^2 + \frac{1}{2}\sum_{m=1}^{M}(\omega_m^{max})^2 + \frac{1}{2}\sum_{m=1}^{M}\max\left[\epsilon_m^2, (\omega_m^{max})^2\right]$.

Proof. Squaring both sides of the queueing dynamic (1), and using the fact that for any $Q \geqslant 0, b \geqslant 0, A \geqslant 0, (max[Q-b,0]+A)^2 \leqslant Q^2 + A^2 + b^2 + 2Q(A-b)$, we have:

$$[Q_m(t+1)]^2 \leqslant [Q_m(t)]^2 + [R_m(t)]^2 + [\omega_m(t)]^2 + 2Q_m(t)[R_m(t) - \omega_m(t)]$$

Summing the above over $m = 1, \ldots, M$ and using the fact that $R_m(t) \leqslant A_m^{max}$ and $\omega_m(t) \leqslant \omega_m^{max}$, we have:

$$\sum_{m=1}^{M}([Q_m(t+1)]^2-[Q_m(t)]^2)\leqslant\sum_{m=1}^{M}\left(A_m^{max}\right)^2+\sum_{m=1}^{M}\left(\omega_m^{max}\right)^2+2\sum_{m=1}^{M}Q_m(t)[R_m(t)-\omega_m(t)]$$

Repeating the above steps for the queue $H_m(t)$ and $Z_m(t)$, by using the fact that $R_m(t)\leqslant A_m^{max},\gamma_m(t)\leqslant A_m^{max}$, and for any $Q\geqslant 0,b\geqslant 0,A\geqslant 0,(max[Q-b+A,0])^2\leqslant Q^2+max(A^2,b^2)+2Q(A-b)$, we have:

$$\sum_{m=1}^{M}([H_m(t+1)]^2-[H_m(t)]^2)\leqslant\sum_{m=1}^{M}\left(A_m^{max}\right)^2+2\sum_{m=1}^{M}H_m(t)[\gamma_m(t)-R_m(t)]$$

$$\sum_{m=1}^{M}([Z_m(t+1)]^2-[Z_m(t)]^2)\leqslant\sum_{m=1}^{M}\max\left[\epsilon_m^2,\left(\omega_m^{max}\right)^2\right]+2\sum_{m=1}^{M}Z_m(t)[\epsilon_m-\omega_m(t)]$$

Combining these three bounds together, and taking the expectation with respect to $\mathbf{\Theta}(t)$ on both sides, we arrive at the following one-slot conditional Lyapunov drift $\Delta(\mathbf{\Theta}(t))$:

$$\begin{aligned}\Delta(\mathbf{\Theta}(t))\leqslant B&+\mathbb{E}\left\{\sum_{m=1}^{M}Z_m(t)\epsilon_m|\mathbf{\Theta}(t)\right\}+\mathbb{E}\left\{\sum_{m=1}^{M}[H_m(t)\gamma_m(t)]|\mathbf{\Theta}(t)\right\}+\mathbb{E}\left\{\sum_{m=1}^{M}R_m(t)[Q_m(t)-H_m(t)]|\mathbf{\Theta}(t)\right\}\\&-\mathbb{E}\left\{\sum_{m=1}^{M}[(Q_m(t)+Z_m(t))\omega_m(t)]|\mathbf{\Theta}(t)\right\}\end{aligned}$$

where $B=\sum_{m=1}^{M}\left(A_m^{max}\right)^2+\frac{1}{2}\sum_{m=1}^{M}\left(\omega_m^{max}\right)^2+\frac{1}{2}\sum_{m=1}^{M}\max\left[\epsilon_m^2,\left(\omega_m^{max}\right)^2\right]$.

Now adding to both sides the penalty expression, i.e., the term $-V\mathbb{E}\left\{\sum_{m=1}^{M}g(\gamma_m(t))-\theta P(t)|\mathbf{\Theta}(t)\right\}$, we prove the theorem. □

Hence, rather than directly minimize the drift-plus-penalty expression in every time slot, we can actually seek to minimize the bound given in the right-hand-side of (13).

4.3. MOTET algorithm

The minimization of the right-hand-side of (13) can be decoupled to a series of independent sub-problems, which can be computed independently and simultaneously in a decentralized fashion.

Specifically, in each time slot t, based on online observations on the current queue states $\mathbf{\Theta}(t)=(\mathbf{Q}(t),\mathbf{H}(t),\mathbf{Z}(t))$ and link condition $L(t)$, MOTET performs the following four phases of control operations, including: (1) auxiliary variables selection, (2) admission decision control, (3) transmission decision control, and (4) all queue updates.

4.3.1. Auxiliary variables

For each application $m\in\{1,\ldots,M\},\gamma_m(t)$ is determined by minimizing the term $-\mathbb{E}\{\sum_{m=1}^{M}[Vg(\gamma_m(t))-H_m(t)\gamma_m(t)]|\mathbf{\Theta}(t)\}$. That is, application m observes $H_m(t)$ and chooses $\gamma_m(t)$ as the solution to:

$$\begin{aligned}&\max_{\gamma_m(t)}\ Vg(\gamma_m(t))-H_m(t)\gamma_m(t)\\&s.t.\quad 0\leqslant\gamma_m(t)\leqslant A_m^{max},\forall m\in\{1,\ldots,M\}\end{aligned}\tag{14}$$

Differentiating the objective function in (14) with respect to $\gamma_m(t)$ can yield the peak value of the function when $\gamma_m(t)=\frac{V}{H_m(t)}-1$. By taking the constraints in (14) into consideration, we obtain the optimal solution to problem (14):

$$\gamma_m(t)=\begin{cases}0, & H_m(t)>V\\ \frac{V}{H_m(t)}-1, & \frac{V}{A_m^{max}}\leqslant H_m(t)\leqslant V\\ A_m^{max}, & H_m(t)<\frac{V}{A_m^{max}}\end{cases}\tag{15}$$

It is obvious that the selection of auxiliary variables can be separately performed for each application $m\in\{1,\ldots,M\}$. The time complexity for this step is $\mathcal{O}(M)$.

4.3.2. Admission control

For each application $m\in\{1,\ldots,M\},R_m(t)$ is determined by minimizing the term $\mathbb{E}\{\sum_{m=1}^{M}R_m(t)[Q_m(t)-H_m(t)]|\mathbf{\Theta}(t)\}$. That is, application m observes $H_m(t)$ and $Q_m(t)$, and chooses $R_m(t)$ from $[0,A_m(t)]$ by:

$$R_m(t)=\begin{cases}A_m(t), & \text{if } Q_m(t)\leqslant H_m(t)\\ 0, & \text{otherwise}\end{cases}\tag{16}$$

This is a simple threshold-based admission control strategy. On the one hand, when $Q_m(t)\leqslant H_m(t)$, the newly arrived data are all admitted into the system. Actually, this not only reduces the value of $H_m(t)$ so that γ_m will be more closer to r_m, but also increases the throughput r_m so that the obtained utility will be improved. On the other hand, when $Q_m(t)>H_m(t)$, then all the arrived data will be denied to ensure the stability of queue Q_m. The time complexity for this step is $\mathcal{O}(M)$.

4.3.3. Transmission control

For each application $m \in \{1,\ldots,M\}$, $\alpha(t)$ is determined by minimizing the term $-\mathbb{E}\left\{\sum_{m=1}^{M}[(Q_m(t)+Z_m(t))\omega_m(t)]-V\theta P(t)|\mathbf{\Theta}(t)\right\}$. That is, the system observes $\mathbf{Q}(t),\mathbf{Z}(t)$ and $L(t)$, and chooses $\alpha(t)$ as the solution to:

$$\begin{aligned}&\max_{\alpha(t)}\sum_{m=1}^{M}[(Q_m(t)+Z_m(t))\hat{\omega}_m(\alpha(t),L(t))]-V\theta P(\widehat{B}(\alpha(t),L(t)))\\&s.t.\quad \alpha(t)\in\mathcal{A}\end{aligned}\tag{17}$$

From the objective in (17), it is obvious that the interface providing higher link rate (i.e., $\omega_m(t)$) while consuming less battery energy (i.e., $P(t)$) will be more preferable to be used for data transfer. The time complexity for this step is $\mathcal{O}(M*|\mathcal{A}|)$, where $|\mathcal{A}|$ denotes the cardinality of $\mathcal{A}$ and represents the total amount of available transmission options.

4.3.4. Queue updates

The queues $Q_m(t),H_m(t)$ and $Z_m(t)$ can be updated according to (1), (8) and (9), by using the optimal values of $\gamma_m(t),R_m(t)$ and $\alpha(t)$ determined in the phases above. The time complexity for this step is $\mathcal{O}(M)$.

5. Algorithm analysis

In this section, we analyze the performance bound of our MOTET algorithm.

Theorem 2. *(Algorithm Performance) Implementing the MOTET algorithm in every time slot for any fixed control parameter $V \geqslant 0$, yields the following performance bounds:(1) The worst case queue backlog for each queue Q_m is upper bounded by a finite constant Q_m^{max} for all t:*

$$Q_m(t) \leqslant Q_m^{max} \triangleq V + 2A_m^{max}\tag{18}$$

(2) Assuming FIFO service, and $Q_m(t) \leqslant Q_m^{max}, Z_m(t) \leqslant Z_m^{max}$ for all $t \in \{0,1,2,\ldots\}$. Then, the worst-case delay for data in Q_m is upper bounded by the constant D_m^{max} defined below:

$$D_m^{max} \triangleq \lceil (Q_m^{max}+Z_m^{max})/\epsilon_m \rceil\tag{19}$$

(3) The time average utility achieved by the MOTET algorithm is within B/V of the optimal utility c^*:

$$\liminf_{t\to\infty}\left\{\sum_{m=1}^{M}g(r_m)-\theta p\right\} \geqslant c^* - \frac{B}{V}\tag{20}$$

Proof. (1) We first prove that $H_m(t) \leqslant H_m^{max} \triangleq V + A_m^{max}$ for all t. Suppose this is true for a particular slot t. We show it also holds for $t+1$. If $H_m(t) < V$, then $H_m(t+1) \leqslant V + A_m^{max} \triangleq H_m^{max}$, because it can increase by at most A_m^{max} in one slot, as $\gamma_m \leqslant A_m^{max}$. If $H_m(t) > V$, then the auxiliary variable decision rule chooses $\gamma_m(t) = 0$ according to (15). Thus, the queue H_m cannot increase in the next slot, and we have $H_m(t+1) \leqslant H_m(t) \leqslant H_m^{max}$. This proves the H_m^{max} bound.

We now prove that $Q_m(t) \leqslant Q_m^{max}$ for all t. Assume this is true for a particular slot t. We prove it also holds for slot $(t+1)$. If $Q_m(t) \leqslant H_m^{max}$, then $Q_m(t+1) \leqslant H_m^{max} + A_m^{max} \triangleq Q_m^{max}$, because it can increase by at most A_m^{max} in one slot, as $R_m(t) \leqslant A_m(t) \leqslant A_m^{max}$. If $Q_m(t) > H_m^{max}$, then the admission decision will choose $R_m(t) = 0$ according to (16). Thus, the queue Q_m cannot increase in the next slot, and we have $Q_m(t+1) \leqslant Q_m(t) \leqslant Q_m^{max}$. This proves the Q_m^{max} bound.

(2) Fix any slot $t \geqslant 0$, and let $A_m(t)$ represents the data that arrives on this slot. From (1), the earliest time it can depart the queue is slot $t+1$. We show that all of this data departs on or before time $t + D_m^{max}$. Suppose that this assumption is not true, we shall reach a contradiction. It must be $Q_m(\varphi) > 0$ for all $\varphi \in \{t+1,\ldots,t+D_m^{max}\}$ (else, we would clear the data by $t + D_m^{max}$). From (9), for all $\varphi \in \{t+1,\ldots,t+D_m^{max}\}$:

$$Z_m(\varphi+1) = \max[Z_m(\varphi) - \omega_m(\varphi) + \epsilon_m, 0]$$

and hence for all $\varphi \in \{t+1,\ldots,t+D_m^{max}\}$:

$$Z_m(\varphi+1) \geqslant Z_m(\varphi) - \omega_m(\varphi) + \epsilon_m$$

Summing the above over $\varphi \in \{t+1,\ldots,t+D_m^{max}\}$ yields:

$$Z_m\left(t+D_m^{max}+1\right) - Z_m(t+1) \geqslant D_m^{max}\epsilon_m - \sum_{\varphi=t+1}^{t+D_m^{max}}\omega_m(\varphi)$$

Rearranging terms in the above inequality and using the fact that $Z_m(t+D_m^{max}+1) \leqslant Z_m^{max}$ and $Z_m(t+1) \geqslant 0$ yields:

$$D_m^{max}\epsilon_m - Z_m^{max} \leqslant \sum_{\varphi=t+1}^{t+D_m^{max}}\omega_m(\varphi)\tag{21}$$

Because service is FIFO, the data $A_m(t)$ that arrives in slot t is placed at the end of the queue in slot $t+1$ according to the queue dynamics, and this data is fully served only when all of the backlog $Q_m(t+1)$ has departed. That is, the last of the $A_m(t)$ data departs in the slot $t+T$, where $T>0$ is the smallest integer for which $\sum_{\varphi=t+1}^{t+D_m^{max}}\omega_m(\varphi) > Q_m(t+1)$. Because we have assumed that not all of the $A_m(t)$ data departs by time $t+D_m^{max}$, we must have:

$$\sum_{\varphi=t+1}^{t+D_m^{max}}\omega_m(\varphi) < Q_m(t+1) \leqslant Q_m^{max} \tag{22}$$

Combining (21) and (22) yields:

$$D_m^{max}\epsilon_m - Z_m^{max} < Q_m^{max}$$

Therefore,

$$D_m^{max} < (Q_m^{max} + Z_m^{max})/\epsilon_m$$

This contradicts the original definition of D_m^{max} given in the theorem.

(3) According to Caratheodory's theorem [39], we can easily prove that there exists a randomized stationary control policy π that chooses feasible control decisions $\gamma_m(t), R_m(t)$ and $\alpha(t)$, independent of the current queue backlogs, and achieves the following guarantees:

$$\mathbb{E}\{c^{\pi}(t)\} = c^* \tag{23}$$
$$\mathbb{E}\{R_m^{\pi}(t)\} = \mathbb{E}\{\omega_m^{\pi}(t)\} \tag{24}$$
$$\mathbb{E}\{\gamma_m^{\pi}(t)\} = \mathbb{E}\{R_m^{\pi}(t)\} \tag{25}$$
$$\epsilon_m = \mathbb{E}\{\omega_m^{\pi}(t)\} \tag{26}$$

Because, in every time slot t, our implementation seeks to minimize the right-hand-side of the drift-plus-penalty expression in (13):

$$\begin{aligned}\Delta(\mathbf{\Theta}(t)) - V\mathbb{E}\left\{\sum_{m=1}^{M} g(\gamma_m(t)) - \theta P(t)|\mathbf{\Theta}(t)\right\} \leqslant B &+ \mathbb{E}\left\{\sum_{m=1}^{M} Q_m(t)[R_m^{\dagger}(t) - \hat{\omega}_m(\alpha^{\dagger}(t), L(t))]|\mathbf{\Theta}(t)\right\} \\ &+ \mathbb{E}\left\{\sum_{m=1}^{M} H_m(t)\left[\gamma_m^{\dagger}(t) - R_m^{\dagger}(t)\right]|\mathbf{\Theta}(t)\right\} \\ &+ \mathbb{E}\left\{\sum_{m=1}^{M} Z_m(t)[\epsilon_m - \hat{\omega}_m(\alpha^{\dagger}(t), L(t))]|\mathbf{\Theta}(t)\right\} - V\mathbb{E}\{c^{\dagger}(t)|\mathbf{\Theta}(t)\}\end{aligned} \tag{27}$$

where $\gamma_m^{\dagger}(t), R_m^{\dagger}(t)$, $\alpha^{\dagger}(t)$ and $c^{\dagger}(t)$ are the resulting decisions and attributed values under any alternative (possibly randomized) policy (denoted by †).

Let $c(t) = \sum_{m=1}^{M} g(\gamma_m(t)) - \theta P(t)$. Now fix $\delta > 0$. Since the resulting values of $\gamma_m^{\dagger}(t)$, $R_m^{\dagger}(t), \alpha^{\dagger}(t)$ are independent of the current queue backlogs $\mathbf{\Theta}(t)$, we know from (23)–(26) that the policy † can achieve the following:

$$\sum_{m=1}^{M}\mathbb{E}\{R_m^{\dagger}(t)|\mathbf{\Theta}(t)\} = \sum_{m=1}^{M}\mathbb{E}\{R_m^{\dagger}(t)\} \leqslant \sum_{m=1}^{M}\mathbb{E}\{\hat{\omega}_m(\alpha^{\dagger}(t), L(t))\} + \delta \tag{28}$$

$$\sum_{m=1}^{M}\mathbb{E}\{\gamma_m^{\dagger}(t)|\mathbf{\Theta}(t)\} = \sum_{m=1}^{M}\mathbb{E}\{\gamma_m^{\dagger}(t)\} \leqslant \sum_{m=1}^{M}\mathbb{E}\{R_m^{\dagger}(t)\} + \delta \tag{29}$$

$$\epsilon_m \leqslant \sum_{m=1}^{M}\mathbb{E}\{\hat{\omega}_m(\alpha^{\dagger}(t), L(t))\} + \delta \tag{30}$$

$$-\mathbb{E}\{c^{\dagger}(t)|\mathbf{\Theta}(t)\} = -\mathbb{E}\{c^{\dagger}(t)\} \leqslant -c^* + \delta \tag{31}$$

Plugging the above (28)–(31) into the right-hand-side of (27) and taking $\delta \to 0$ yields:

$$\Delta(\mathbf{\Theta}(t)) - V\mathbb{E}\{c(t)|\mathbf{\Theta}(t)\} \leqslant B - Vc^* \tag{32}$$

This is in the exact form for application of the Lyapunov Optimization Theorem (Theorem 4.2 in [27]). Hence, all queues can be proved as mean rate stable. By the Lyapunov Optimization Theorem, we take expectations of both sides of the above, and then have:

$$\mathbb{E}\{L(\mathbf{\Theta}(t+1))\} - \mathbb{E}\{L(\mathbf{\Theta}(t))\} - V\mathbb{E}\{c(t)\} \leqslant B - Vc^*$$

By summing the above inequality over time slots $t \in \{0, 1, \ldots, T-1\}$, dividing the result by t, and using the fact that $L(\mathbf{\Theta}(t)) \geqslant 0$ and $L(\mathbf{\Theta}(0)) = 0$, we have:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\{c(t)\} \geqslant c^* - \frac{B}{V} \tag{33}$$

Since the function $g(x)$ is concave, we can have the following relationship according to Jensen's inequality [27]:

$$g\left(\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\{\gamma_m(t)\}\right) \geqslant \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\{g(\gamma_m(t))\} \tag{34}$$

Plugging (34) into (33) and taking a lim inf of both sides yields:

$$\liminf_{t\to\infty}\left\{\sum_{m=1}^{M} g(\gamma_m) - \theta p\right\} \geqslant c^* - \frac{B}{V} \tag{35}$$

Because $H_m(t)$ is stable and $r_m \geqslant \gamma_m$ can be satisfied according to Lemma 1. Then for all $m \in \{1,\ldots,M\}, g(r_m) \geqslant g(\gamma_m)$. Plugging this inequality into (35) yields (20). □

6. Performance evaluation

In this section, we evaluate the proposed MOTET algorithm using datasets from real-world measurements on wireless link bandwidth and transmission energy consumption of mobile devices.

6.1. Simulation setup

We consider a mobile device equipped with three state-of-the-art wireless interfaces, i.e., WiFi, 3G-HSDPA, and 4G-LTE. Then, $\mathcal{A}$ = {"WiFi", "3G-HSDPA", "4G-LTE", "Idle"}. The link bandwidth (i.e., L) traces we use are the UMICH measurement datasets from 4GTest Project [13]. According to observed bandwidth statistics in the traces, we set τ as a moderate value (60 s) [35]. A portion of two-hour bandwidth variation for different interfaces is shown in Fig. 2. In the simulations, we assume that the bandwidth of selected link is used and shared by all M applications on the basis of the current traffic arrival rate, which reflects the urgency of communication demands [6]. Meanwhile, we use the power model derived in [13]. Specifically, the parameters for power consumption of wireless interfaces are listed in Table 1.

To simulate applications with different volumes of data traffic, we generate synthetic traffic load according to the traffic patterns of current mobile applications. We consider totally $M = 10$ applications running in the cloud environment, with data arrivals following the Poisson Process [35,16] (note that our algorithm does not have any special requirement on this traffic pattern). The applications are classified into 3 groups, i.e., high-rate group (4 applications with $\lambda_m = 5MB$ and $A_m^{max} = 10MB$), middle-rate group (4 applications with $\lambda_m = 10MB$ and $A_m^{max} = 20MB$) and high-rate group (2 applications with $\lambda_m = 20MB$ and $A_m^{max} = 40MB$).

To fully investigate the MOTET performance, we compare it with an online algorithm "Fastest", which always chooses the interface currently providing the highest bandwidth among the three. According to Theorem 2.4 in [27], we use a simple admission policy to enforce that $R_m(t) = 0$ when r_m is larger than the time average of $\omega_m(t)$, so as to guarantee the rate stability of queue Q_m.

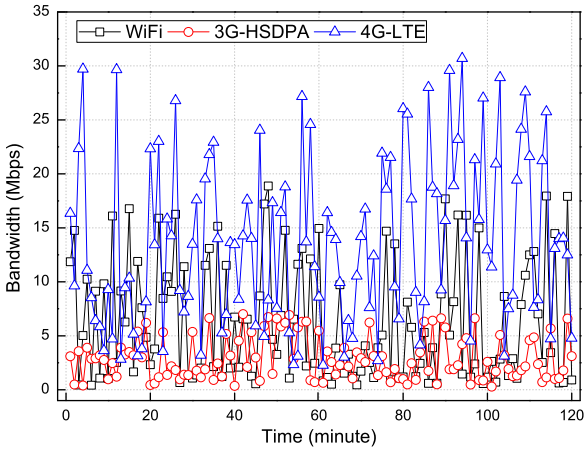


Fig. 2. Two-hour trace of communication bandwidth for different wireless interfaces.

Table 1
Power parameters for different wireless interfaces.

	α (mW/Mbps)	β (mW)
WiFi	137.01	132.86
3G-HSPDA	122.12	817.88
4G-LTE	51.97	1288.04

6.2. Results and analysis

We conduct the following analysis on critical factors to characterize their impacts on MOTET performance. Noted that the average delay metric in the results is calculated by weighted averaging the queueing delay for transmitted data, and the weight coefficient is the serviced amount.

6.2.1. Impact of the stability-utility parameter V

We first fix $T = 10,000$ time slots and $\theta = 0.01$, and then run experiments with different V values. The simulation results are shown in Fig. 3. We have several observations concerning these results.

Firstly, as the value of V increases, the time-average utility achieved improves significantly and converges to the maximum level for larger values of V (Fig. 3(a)), while both the average queue backlog (Fig. 3(d)) and the average service delay (Fig. 3(e)) almost increase linearly. This quantitatively confirms the $[\mathcal{O}(1/V), \mathcal{O}(V)]$ utility-stability tradeoff in Theorem 2. Meanwhile, the Fastest algorithm cannot achieve the optimal maximum utility, because its objective is to maximize the throughput but not the throughput-energy tradeoff.

Secondly, the average system throughput is monotonically increasing with V, while the average energy consumption is monotonically decreasing with V, as shown in Fig. 3(b) and (c). These two observations conform to our expectations on the optimization objective in (6). We are interested in analyzing two special cases with boundary values of V (i.e., $V = 1$ and $V = 4000$). When $V = 1$, MOTET scarcely considers the optimization objective in (6). According to (14)–(16), MOTET always chooses 0 or $[\frac{V}{H_m(t)} - 1]$ for $\gamma_m(t)$, making $H_m(t)$ smaller than $Q_m(t)$ and 0 chosen for $R_m(t)$. This results in the low system throughput of MOTET, as can be observed in Fig. 3(b). Meanwhile, the energy factor $P(t)$ has far less impact on the optimization of (17) than the bandwidth factor $\omega_m(t)$, which will lead to the selection of higher-bandwidth (and most probably higher-energy) link. This results in the high energy consumption of MOTET, as can be observed in Fig. 3(c). On the other hand, when V grows to 4000, MOTET has quite a lot of opportunities to choose relatively larger values for $\gamma_m(t)$ and $R_m(t)$, thus making it have almost the same throughput performance (i.e., γ_m) as Fastest. Meanwhile, MOTET can choose a suitable slot with a link providing higher rate while consuming less energy for data transfer based on the rule in (17), thus consuming much less energy than Fastest (as shown in Fig. 3(c)). However, such a performance improvement starts to diminish with excessive increases of V, which can adversely aggravate the congestion of system queues (as can be predicted from Fig. 3(c) and (d)). The Lyapunov optimization framework itself does not give any guidance on parameter selection, and how to identify a good V value is left as future work.

Thirdly, when the V value is relatively small, the optimization to the objective function may be rather useless and unsatisfactory. For example, when $V < 1000$, MOTET is inferior to Fastest, as shown in Fig. 3(a). The reason is similar to that for $V = 1$ in last paragraph. This observation implies that when applying the Lyapunov optimization technique for multiple objectives, especially those potentially conflicting ones (e.g., throughput and energy), the algorithm may not exhibit its superiority over others immediately when $V > 0$, and some attempts and adjustments may be necessary. This is very different from the algorithms on single objective optimization [35,39].

Moreover, we compare the theoretical and the experimental upper bound of queue lengths. The results are shown in Table 2. In our experiments, the queue lengths are smaller than the mathematical bounds, especially when V is relatively larger.

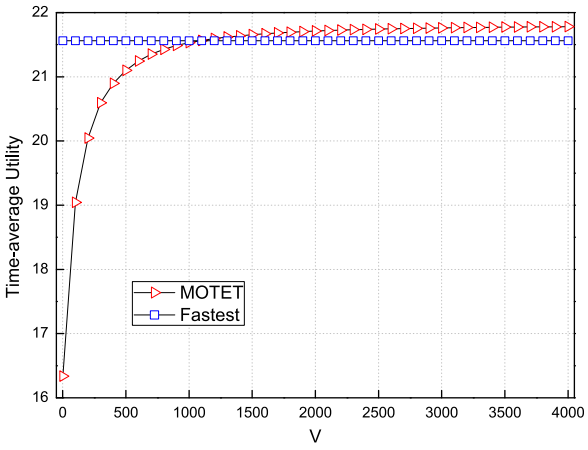
6.2.2. Impact of the long-term time slot T

We fix $V = 3000$ and $\theta = 0.01$, and then vary T from 500 time slots to 10,000 time slots, which is a sufficient range for exploring the characteristics of MOTET. Corresponding results of the two algorithms are shown in Fig. 4.

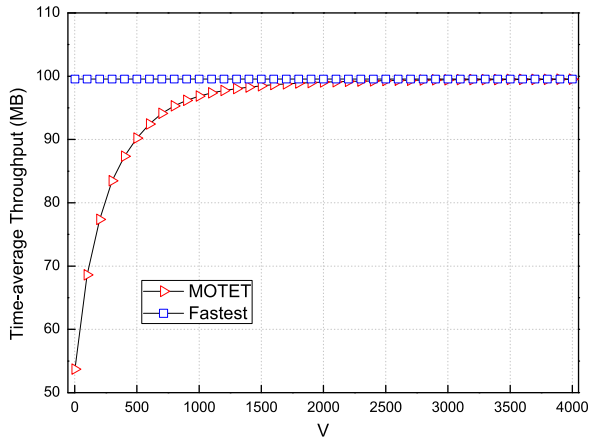
We can observe that MOTET achieves a higher utility than Fastest, which results in higher average queue backlogs and higher average service delays. These results are consistent with those in Fig. 3(a), (d) and (e) (when $V = 3000$). Besides, it is obvious that changing T has relatively small impact on the system stability. The fluctuations on the achieved utility, queue backlog, and queue delay are $[-0.06\%, +0.08\%]$, $[-4.53\%, +12.30\%]$, and $[-4.35\%, +11.08\%]$, respectively, for MOTET, and are $[-0.06\%, +0.13\%]$, $[-5.99\%, +11.72\%]$, and $[-6.73\%, +13.18\%]$, respectively, for Fastest. All these observations confirm the usability and effectiveness of the stability guarantee mechanisms employed by MOTET and Fastest.

6.2.3. Impact of the throughput-energy parameter θ

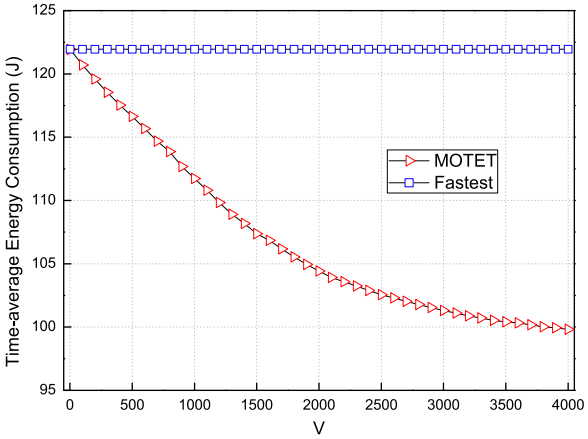
We fix $V = 3000$ and $T = 10,000$ time slots, and then run experiments with different θ values. Fig. 5 plots the simulation results under the two control algorithms with various values of parameter θ. From Fig. 5(a) we can see that as θ goes from 0



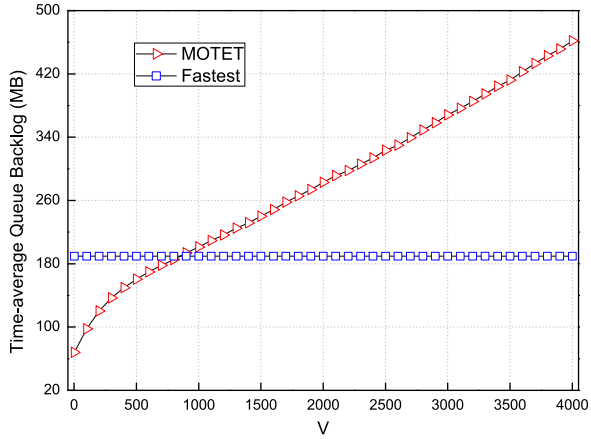
(a) Achieved utility



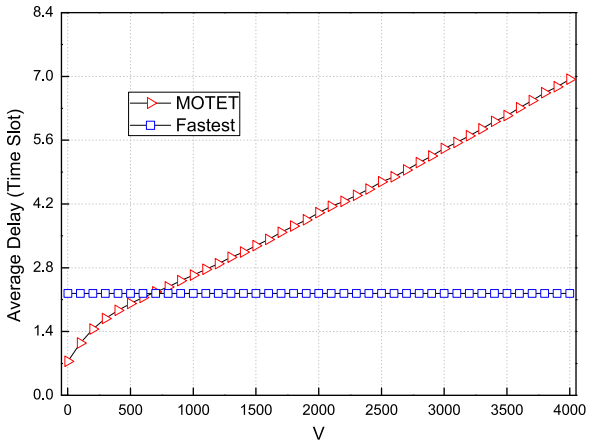
(b) System throughput



(c) Energy consumption



(d) Queue backlog



(e) Queue delay

Fig. 3. MOTET performance under different V value.

to 0.1, the value of objective utility decreases significantly for both of the algorithms. For Fastest, the increase of θ only has an impact on the final calculation of utility function in (6). According to (17), the increase of θ will make MOTET more inclined to reduce energy consumption on data transmission by using the lower-energy link for data transfer. Consequently, we can

Table 2
Comparison on theoretical bound (T) and experimental bound (E) of queue length.

		Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8	Q_9	Q_{10}
	T	20	20	20	20	40	40	40	40	80	80
$V = 0$	E	10	10	10	10	20	20	20	20	38	38
	T	21	21	21	21	41	41	41	41	81	81
$V = 1$	E	12.24	13.92	13.56	15.29	23.58	24.48	24.34	23.75	40.49	40.44
	T	120	120	120	120	140	140	140	140	180	180
$V = 100$	E	22.75	22.87	24.31	22.81	27.64	26.43	27.47	26.04	38.91	40.44
	T	1020	1020	1020	1020	1040	1040	1040	1040	1080	1080
$V = 1000$	E	47.4	43.10	63.10	44.90	79.29	96.12	81.86	89.97	75.54	81.11

notice the decreasing trend in both system throughput and energy consumption from the results in Fig. 5(b) and (c). In particular, when $\theta > 0.01$, the system throughput begins to fall drastically, indicating the energy factor becomes dominating in the utility optimization. However, we can also find from the results that, with an appropriate parameter θ, MOTET can significantly reduce the energy consumption while only incurring a marginal decrease in system throughput. For example, when $\theta = 0.01$, the time-average energy consumption is reduced by 16.93% at the cost of only 0.13% decrease in the time-average system throughput.

Nevertheless, the value of θ should be properly selected for guaranteeing the basic requirements on system throughput. An extreme case is when θ is considerably large, MOTET will decline to admit any data arrival and use any wireless interface for the energy saving purpose, which deviates the original design intention of this algorithm.

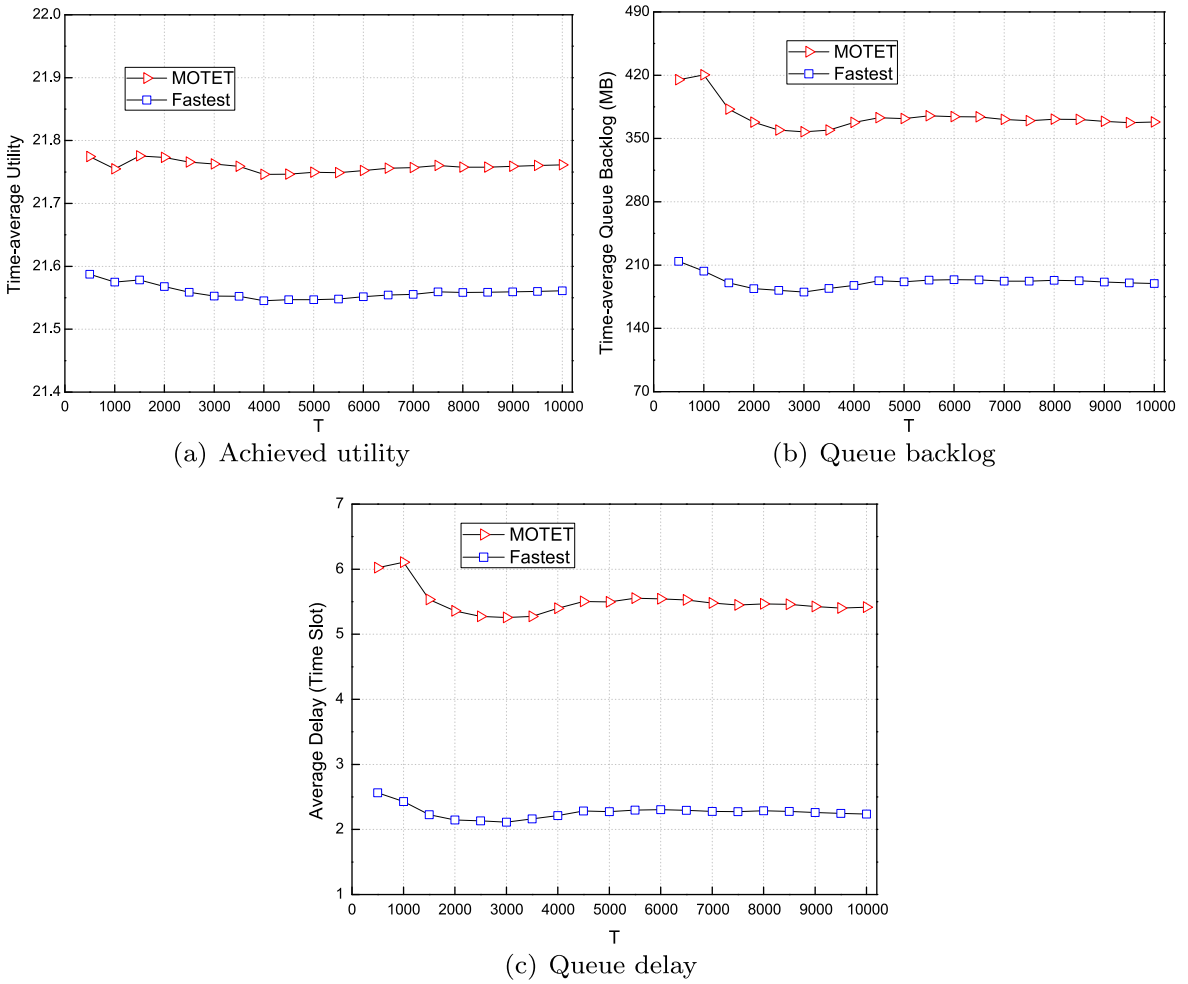
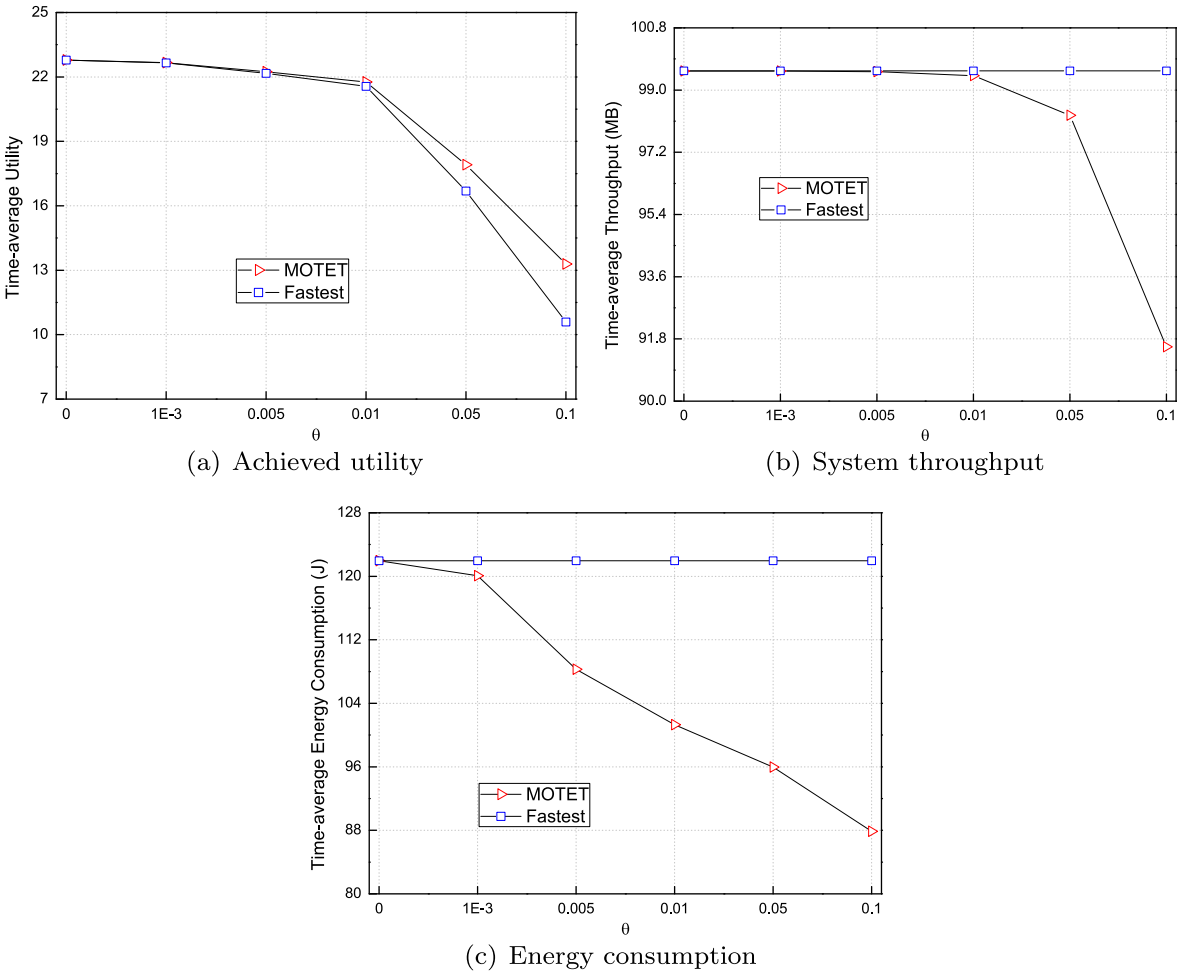


Fig. 4. MOTET performance under different T value.



(a) Achieved utility (b) System throughput

(c) Energy consumption

Fig. 5. MOTET performance under different θ value.

7. Conclusion

To reduce energy consumption for cloud-based mobile applications and prolong operational lifetime for energy-constrained mobile devices, we design and analyze an optimal online control algorithm, MOTET, to balance the tradeoff between the system throughput and the energy consumption in mobile cloud scenarios. By applying rigorous Lyapunov optimization framework, MOTET is able to independently and simultaneously make the control decisions on traffic admission and data transmission for the mobile device. Through in-depth theoretical analysis and extensive simulation experiments, we demonstrate that MOTET can achieve a desired performance tradeoff among application throughput, energy cost and service delay. In particular, it can approach the optimality within a diminishing gap of $\mathcal{O}(1/V)$, while guaranteeing system stability and bounding the queue length and the service delay by $\mathcal{O}(V)$, where V is a tunable control parameter. As future work, we will address the problem of how to select good parameter values for V and θ in MOTET [29]. Another important work is to further evaluate MOTET using a prototype implementation on the modern smartphone platform [13,32,42].

Acknowledgment

We would like to thank Dr. Junxian Huang (University of Michigan) for providing us the UMICH dataset from 4GTest measurement [13]. We are also grateful to Dr. Yunlu Liu (China Mobile Research Institute) for useful discussions. This work was supported by the National Natural Science Foundation of China under Grants 61202430 and 61303245, the State Key Lab of Astronautical Dynamics of China under Grant 2014ADL-DW0401, and the Science and Technology Foundation of Beijing Jiaotong University under Grant 2012RC040.

References

[1] T. Amin, I. Chikalov, M. Moshkov, B. Zielosko, Dynamic programming approach to optimization of approximate decision rules, Inform. Sci. 221 (0) (2013) 403–418. <http://www.sciencedirect.com/science/article/pii/S0020025512006111>.
[2] G. Ananthanarayanan, V.N. Padmanabhan, L. Ravindranath, C.A. Thekkath, Combine: leveraging the power of wireless peers through collaborative downloading, in: Proceedings of the 5th International Conference on Mobile Systems, Applications and Services, New York, NY, USA, 2007.
[3] Apple, Apple – ipad – view all the technical specifications, Apple Inc., 2013. <http://www.apple.com/ipad/specs/>.
[4] AT&T, AT&T high speed internet access, AT&T Inc., 2013. <http://www.att.com/shop/internet.html>.
[5] N. Balasubramanian, A. Balasubramanian, A. Venkataramani, Energy consumption in mobile phones: a measurement study and implications for network applications, in: Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference, New York, NY, USA, 2009.
[6] A. Chakraborty, S. Das, Adapp: an adaptive network selection framework for smartphone applications, in: Proceeding of the 2013 Workshop on Cellular Networks: Operations, Challenges, and Future Design, New York, NY, USA, 2013.
[7] P.-C. Chao, H.-M. Sun, Multi-agent-based cloud utilization for the it office-aid asset distribution chain: an empirical case study, Inform. Sci. 245 (0) (2013) 255–275.
[8] Cisco, Cisco visual networking index: forecast and methodology, 2012–2017, Cisco Systems Inc., 2013. <http://www.cisco.com/>.
[9] A. Gani, G.M. Nayeem, M. Shiraz, M. Sookhak, M. Whaiduzzaman, S. Khan, A review on interworking and mobility techniques for seamless connectivity in mobile cloud computing, J. Netw. Comput. Appl. 43 (0) (2014) 84–102. <http://www.sciencedirect.com/science/article/pii/S1084804514000927>.
[10] Y. Guo, M. Pan, Y. Fang, Optimal power management of residential customers in the smart grid, IEEE Trans. Parallel Distrib. Syst. 23 (9) (2012) 1593–1606.
[11] F. Gustavo, L. Fidel, J.O. Fajardo, Qos-oriented admission control in hsdpa networks, Netw. Protocols Algor. 1 (1) (2009) 52–61.
[12] I.-H. Hou, P.R. Kumar, Utility-optimal scheduling in time-varying wireless networks with delay constraints, in: Proceedings of the 7th ACM International Symposium on Mobile Ad Hoc Networking and Computing, New York, NY, USA, 2010.
[13] J. Huang, F. Qian, A. Gerber, Z.M. Mao, S. Sen, O. Spatscheck, A close examination of performance and power characteristics of 4g lte networks, in: Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services, New York, NY, USA, 2012.
[14] K. Kumar, J. Liu, Y.-H. Lu, B. Bhargava, A survey of computation offloading for mobile systems, Mob. Netw. Appl. 18 (1) (2013) 129–140.
[15] H.A. Lagar-Cavilla, K. Joshi, A. Varshavsky, J. Bickford, D. Parra, Traffic backfilling: subsidizing lunch for delay-tolerant applications in umts networks, in: Proceedings of the 3rd ACM SOSP Workshop on Networking, Systems, and Applications on Mobile Handhelds, New York, NY, USA, 2011.
[16] K. Lee, J. Lee, Y. Yi, I. Rhee, S. Chong, Mobile data offloading: how much can wifi deliver?, IEEE/ACM Trans Netw. 21 (2) (2013) 536–550.
[17] S.-H. Lim, S.W. Lee, M. Sohn, B.-H. Lee, Energy-aware optimal cache consistency level for mobile devices, Inform. Sci. 230 (0) (2013) 94–105. mobile and Internet Services in Ubiquitous and Pervasive Computing Environments. <http://www.sciencedirect.com/science/article/pii/S0020025512006287>.
[18] F. Liu, P. Shu, H. Jin, L. Ding, J. Yu, D. Niu, B. Li, Gearing resource-poor mobile devices with powerful clouds: architectures, challenges, and applications, IEEE Wireless Commun. 20 (3) (2013) 14–22.
[19] H. Liu, Y. Zhang, Y. Zhou, Tailtheft: leveraging the wasted time for saving energy in cellular communications, in: Proceedings of the Sixth International Workshop on MobiArch, New York, NY, USA, 2011.
[20] F. LIuís, J. Teodor, A review of the architecture of admission control schemes in the internet, Netw. Protocols Algor. 5 (3) (2013) 1–32.
[21] J. Lloret, M. Garcia, J. Tomas, J.J. Rodrigues, Architecture and protocol for intercloud communication, Inform. Sci. 258 (0) (2014) 434–451. <http://www.sciencedirect.com/science/article/pii/S0020025513003691>.
[22] R.M. Reza, J. Ren, C. Harold. Liu, A.V. Vasilakos, V. Nalini, Mobile cloud computing: a survey, state of art and future directions, ACM/Springer Mobile Appl. Netw. (MONET) 19 (2) (2014) 133–143.
[23] A.P. Miettinen, J.K. Nurminen, Energy efficiency of mobile clients in cloud computing, in: Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing, Berkeley, CA, USA, 2010.
[24] L.A. Mozelli, R.M. Palhares, G.S. Avellar, A systematic approach to improve multiple lyapunov function stability and stabilization conditions for fuzzy systems, Inform. Sci. 179 (8) (2009) 1149–1162.
[25] M. Neely, Energy optimal control for time-varying wireless networks, IEEE Trans. Inform. Theory 52 (7) (2006) 2915–2934.
[26] M. Neely, Opportunistic scheduling with worst case delay guarantees in single and multi-hop networks, in: Proceedings IEEE INFOCOM 2011, Shanghai, China, 2011.
[27] M.J. Neely, Stochasitic Network Optimization with Application to Communication and Queueing Systems, Morgan & Claypool, 2010.
[28] G. Paravati, C. Celozzi, A. Sanna, F. Lamberti, A feedback-based control technique for interactive live streaming systems to mobile devices, IEEE Trans. Consumer Electron. 56 (1) (2010) 190–197.
[29] W. Pedrycz, W. Homenda, Building the fundamentals of granular computing: a principle of justifiable granularity, Appl. Soft Comput. 13 (10) (2013) 4209–4218. <http://www.sciencedirect.com/science/article/pii/S1568494613002068>.
[30] T. Pering, Y. Agarwal, R. Gupta, R. Want, Coolspots: reducing the power consumption of wireless mobile devices with multiple radio interfaces, in: Proceedings of the 4th International Conference on Mobile Systems, Applications and Services, New York, NY, USA, 2006.
[31] X. Qiu, H. Li, C. Wu, Z. Li, F. Lau, Cost-minimizing dynamic migration of content distribution services into hybrid clouds, in: Proceedings IEEE INFOCOM 2012, Orlando, FL, USA, 2012.
[32] M.-R. Ra, J. Paek, A.B. Sharma, R. Govindan, M.H. Krieger, M.J. Neely, Energy-delay tradeoffs in smartphone applications, in: Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services, New York, NY, USA, 2010.
[33] A. Rahmati, L. Zhong, Context-for-wireless: context-sensitive energy-efficient wireless data transfer, in: Proceedings of the 5th International Conference on Mobile Systems, Applications and Services, New York, NY, USA, 2007.
[34] M. Shiraz, A. Gani, R. Khokhar, R. Buyya, A review on distributed application processing frameworks in smart mobile devices for mobile cloud computing, IEE Commun. Surv. Tutor. 15 (3) (2013) 1294–1313.
[35] P. Shu, F. Liu, H. Jin, M. Chen, F. Wen, Y. Qu, B. Li, etime: Energy-efficient transmission between cloud and mobile devices, in: Proceedings IEEE INFOCOM 2013, Turin, Italy, 2013.
[36] M. Stemm, R.H. Katz, Measuring and reducing energy consumption of network interfaces in hand-held devices, IEICE Trans. Commun. E80-B (8) (1997) 1125–1131.
[37] R. Urgaonkar, B. Urgaonkar, M.J. Neely, A. Sivasubramaniam, Optimal power cost management using stored energy in data centers, in: Proceedings of the ACM SIGMETRICS Joint International Conference on Measurement and Modeling of Computer Systems, New York, NY, USA, 2011.
[38] X. Xu, L. Zuo, Z. Huang, Reinforcement learning algorithms with function approximation: recent advances and applications, to be published in Inform. Sci. (0). <http://www.sciencedirect.com/science/article/pii/S0020025513005975>.
[39] Y. Yao, L. Huang, A. Sharma, L. Golubchik, M. Neely, Data centers power reduction: A two time scale approach for delay tolerant workloads, in: Proceedings IEEE INFOCOM 2012, Orlando, FL, USA, 2012.
[40] A. Yutaka, M. Hiroshi, Y. Yusuke, T. Shigeaki, F. Akira, Application-layer active wireless network switching on a smartphone, in: Proceedings of the 2nd Workshop on Smart Mobile Applications, Nwcastle, UK, 2012.
[41] Y. Zhang, Y. Wang, X. Wang, Testore: exploiting thermal and energy storage to cut the electricity bill for datacenter cooling, in: Proceedings of the 8th International Conference on Network and Service Management, Laxenburg, Austria, 2013.
[42] Y. Zhao, J. Wu, The design and evaluation of an information sharing system for human networks, IEEE Trans. Parallel Distrib. Syst. 25 (3) (2014) 796–805.
[43] Z. Zhou, F. Liu, H. Jin, B. Li, B. Li, H. Jiang, On arbitrating the power-performance tradeoff in saas clouds, in: Proceedings IEEE INFOCOM 2013, Turin, Italy, 2013.
[44] Y. Zhuang, N. Jiang, Z. Wu, Q. Li, D.K. Chiu, H. Hu, Efficient and robust large medical image retrieval in mobile cloud computing environment, Inform. Sci. 263 (0) (2014) 60–86. <http://www.sciencedirect.com/science/article/pii/S002002551300738X>.