Energy-Efficient Resource Allocation of Wireless Systems with Statistical QoS Requirement

Changyang She, Chenyang Yang
School of Electronics and Information Engineering,
Beihang University, Beijing, China
Email: {cyshe,cyyang}@buaa.edu.cn

Lingjia Liu
Department of Electrical Engineering and Computer Science,
University of Kansas, USA
Email: lingjialiu@gmail.com

***Abstract*—This paper strives to improve the *energy efficiency* (EE) of wireless systems by resource allocation for delay sensitive traffic with the statistical *quality of service* (QoS) requirement, where a delay bound and a small violation probability is ensured. To avoid wasting energy when serving random traffic, we turn the QoS exponent, a key parameter to characterize statistical QoS guarantee under the framework of effective bandwidth and effective capacity, into a QoS function dependent on the queue length. A method to design queue length based resource allocation strategy is proposed, which maximizes the EE under the constraint on the QoS function. By taking an OFDM system serving Poisson source as an example, a closed form subcarrier allocation strategy with constant transmit power on each subcarrier is derived. Simulation and numerical results show that the EE achieved by the proposed strategy is much higher than the relevant strategies in the literature considering statistical QoS provision, and is very close to the EE upper bound achieved when the delay bound is infinite.**

***Keywords:* energy efficiency, delay sensitive traffic, statistical QoS requirement, queue state information, resource allocation.**

I. Introduction

Energy efficiency (EE) has become an important design goal for wireless systems [1]. Since maximizing the EE of a system should not sacrifice the quality of service (QoS) required by each user, the feature of the traffic need to be taken into account for the EE-oriented design.

In future wireless communications, a significant portion of traffic is delay sensitive, e.g., video/audio and interactive data transmission. The delay performance metrics studied in the literature can generally be categorized as deterministic delay bound, average delay requirement and statistical QoS requirement (see [2] and references therein). For wireless systems, the hard delay bound guarantee is usually either impossible or too expensive to achieve in terms of transmit power. For real time multimedia applications, the average delay guarantee does not ensure the delay performance required by the services. The statistical QoS requirement, defined as a delay bound and a delay violation probability, is more relevant for wireless multimedia transmission. For example, in the fourth generation (4G) LTE/LTE-Advanced systems, the upper bound on the delay-violation probability for VoIP is 2% while the delay threshold is 50 ms for radio access networks [3].

Effective bandwidth [4] and effective capacity [5] are useful tools to study resource allocation with statistical QoS requirement. The notion of effective capacity stems from information theory, which characterize the maximal service rate provided by a wireless system for delay sensitive traffic. Recently, energy efficient resource allocation considering the statistical QoS provision has drawn growing attention [2, 6, 7]. In [2], the transmit power was minimized with effective capacity constraint for different systems. In [6], the EE was maximized for a single carrier system, which is defined as the ratio of effective capacity to the overall transmit and circuit power consumption. In [7], the EE was maximized for an orthogonal frequency division multiple access (OFDMA) system under the constraint on effective capacity for each user, where the EE is defined as the ratio of sum effective capacity to the overall power consumption. In [2, 6, 7], the investigated resource allocation strategies only depend on the channel state information of the underlying systems, and the arrival rate of sources are assumed to be constant. However, in practice, a lot of traffic has random arrival rate. When both sources and channels are random, the resource allocation should depend on the queue length, as implied in the pioneering work on analyzing minimal transmit power with average delay constraint [8]. Yet it remains unclear how to incorporate the queue state information into the framework of effective capacity.

In this paper, we design resource allocation to maximize the EE of a system supporting the traffic with statistical QoS requirement, by using effective bandwidth and effective capacity as tools. The EE is defined as the ratio of accumulate data transmitted to the overall energy consumed during a certain period, which is an *energy per bit* metric [9]. To simplify the analysis, we take an OFDM system as an example, where the transmit power is constant on each subcarrier. To accommodate random arrival process, we derive queue length based subcarrier allocation strategy. Simulation results show that the proposed strategy achieves much higher EE than existing strategies considering statistical QoS requirement, especially when the delay bound is tight or the traffic is bursty.

II. System Model and Problem Formulation

A. System Model and Statistical QoS Requirement

Consider a queueing system, where the source arrives in the buffer of the transmitter at rate $a(t)$, departs from the buffer at rate $b(t)$, and the system provides the service with rate $s(t)$

at time t, which are related as

$$b(t) = \begin{cases} \min\{a(t), s(t)\}, Q(t) = 0 \\ s(t), \qquad\qquad\qquad Q(t) \neq 0 \end{cases}, \quad (1)$$

where $Q(t) = A(t) - B(t)$ is the queue length at time t, $A(t) \triangleq \int_0^t a(\tau) d\tau$ and $B(t) \triangleq \int_0^t b(\tau) d\tau$ are the total number of bits arrived and departed in the interval of $[0, t]$. We assume that the queue is in steady state during time interval $[0, t]$, and denote Q_∞ as the steady state queue length.

The statistical QoS requirement is defined as $(D_{\max}, \varepsilon_D)$, where $D_{\max}$ is the delay bound and ε_D is the maximal delay violation probability allowed by the traffic. In this paper, we restrict on the queuing delay and ignore the coding delay.

The effective capacity and effective bandwidth are useful tools for designing resource allocation satisfying statistical QoS requirement. The effective bandwidth of $a(t)$ is [4]

$$EB(\theta^c) = \lim_{t\to\infty} \frac{1}{\theta^c t} \ln \mathbb{E}\left[e^{\theta^c A(t)}\right], \quad (2)$$

where θ^c is a nonnegative constant, called *QoS exponent*.

The effective capacity of $s(t)$ can be expressed as [5]

$$EC(\theta^c) = -\lim_{t\to\infty} \frac{1}{\theta^c t} \ln \mathbb{E}\left[e^{-\theta^c S(t)}\right], \quad (3)$$

where $S(t) \triangleq \int_0^t s(\tau) d\tau$ is the accumulated data that can be transmitted in the interval of $[0, t]$.

If the arrival and the service processes $a(t)$ and $s(t)$ are independent from each other and

$$EB(\theta^c) = EC(\theta^c), \quad (4)$$

the probability of the queue length exceeding $Q_{\max}$ satisfies

$$\Pr(Q_\infty > Q_{\max}) \approx \eta e^{-\theta^c Q_{\max}} \triangleq \varepsilon_Q, \quad (5)$$

where $\eta \triangleq \Pr(Q_\infty > 0)$ is the non-empty probability of the buffer [5], and θ^c is the unique solution of (4) when $\mathbb{E}[a(t)] < \mathbb{E}[s(t)]$ [10]. The approximation in (5) is accurate when $Q_{\max}$ is sufficiently large according to large deviation principle [4]. In fact, it is accurate even for smaller values of $Q_{\max}$, as shown in [5].

To use the notions of effective bandwidth and effective capacity to describe the statistical QoS requirement, we need to find the connection of the delay bound with the QoS exponent. When (4) is satisfied, the probability that the queue delay exceeds $D_{\max}$ can be approximated as [10]

$$\Pr(Delay > D_{\max}) \approx e^{-\theta^c EB(\theta^c) D_{\max}} \triangleq \varepsilon_D. \quad (6)$$

Comparing (5) and (6), we can find that when $Q_{\max} = EB(\theta^c) D_{\max}$ and $\varepsilon_Q = \eta \varepsilon_D$, $(D_{\max}, \varepsilon_D)$ and $(Q_{\max}, \varepsilon_Q)$ are equivalent in the sense that they can be satisfied with the same QoS exponent θ^c. In the sequel we use the maximum queue length $Q_{\max}$ and its violation probability ε_Q to characterize the statistical QoS requirement.

B. Queue Length Based Statistical QoS Requirement

The resource allocation derived from (4) is independent of the arrival process, which is the best solution for constant source [4]. In practice, many sources are random. For a random arrival process $a(t)$, the resource allocation (or equivalently the service process $s(t)$), should depend on the queue information in order not to waste the physical resources. To this end, we set the QoS exponent as a function of the queue length. Such a idea was first proposed in [4] for designing dynamic bandwidth allocation of wired line communications.

To introduce the *QoS function* in wireless link, we divide the original queue with length $Q_{\max}$ into two parts, $0 = Q_0 < Q_1 < Q_2 = Q_{\max}$. Assume that the transmission resource allocation is constant when Q_∞ stays in $(Q_0, Q_1]$ or $(Q_1, Q_2]$, but may be different when Q_∞ changes from one interval to the other. Denote $\theta(Q_j)$ as the QoS exponent when $Q_\infty \in (Q_{j-1}, Q_j], j = 1, 2$. For a given resource allocation strategy, we can obtain a unique $\theta(Q_j)$ from

$$EC[\theta(Q_j)] = EB[\theta(Q_j)], j = 1, 2. \quad (7)$$

Then, the event $Q_\infty > Q_{\max}$ can be divided into two events: Q_∞ exceeds Q_1 with the resource allocation strategy in $(Q_0, Q_1]$ and Q_∞ increases from Q_1 to $Q_{\max}$ with the resource allocation strategy in $(Q_1, Q_2]$. The tail probability of the queue length in (5) now satisfies

$$\begin{aligned} &\Pr(Q_\infty > Q_{\max}) \\ &= \Pr(Q_\infty > Q_1) \Pr(Q_\infty > Q_{\max} | Q_\infty > Q_1) \\ &\approx \eta e^{-\sum_{j=1}^{2} \theta(Q_j)(Q_j - Q_{j-1})} \end{aligned} \quad (8)$$

By dividing the original queue into infinite number of small queues whose sizes are infinitesimal, (8) becomes (See [4] and the reference therein)

$$\Pr(Q_\infty > Q_{\max}) \approx \eta \exp\left\{-\int_0^{Q_{\max}} \theta(q) dq\right\}, \quad (9)$$

where $\theta(q)$ is called *QoS function*, which can be found from

$$EC[\theta(q)] = EB[\theta(q)]. \quad (10)$$

To avoid serving empty buffer, we let $s(t) = 0$ when $q = 0$. In this case, $\mathbb{E}[s(t)] = 0 < \mathbb{E}[a(t)]$, and therefore (10) has no nonnegative solution for $\theta(q = 0)$ [10]. In the following, we only define the *QoS function* for $q > 0$.

(9) suggests that the behavior of the queue can be controlled by the *QoS function*. To guarantee the statistical QoS requirement $(Q_{\max}, \varepsilon_Q)$, we let the approximation of $\Pr(Q_\infty > Q_{\max})$ equal to ε_Q. Then, from (9) we can obtain

$$\int_0^{Q_{\max}} \theta(q) dq = \ln\left(\frac{\eta}{\varepsilon_Q}\right) \triangleq \bar{\theta} Q_{\max}. \quad (11)$$

Replacing θ^c in (6) with $\bar{\theta}$ [4], the required average *QoS function* $\bar{\theta}$ to satisfy $(D_{\max}, \varepsilon_D)$ can be obtained from

$$\bar{\theta} EB(\bar{\theta}) D_{\max} \approx -\ln \varepsilon_D. \quad (12)$$

Comparing (12) and (11), we can find that when $Q_{\max} = EB(\bar{\theta}) D_{\max}$ and $\varepsilon_Q = \eta \varepsilon_D$, $(D_{\max}, \varepsilon_D)$ and $(Q_{\max}, \varepsilon_Q)$ can be satisfied with the same value of $\bar{\theta}$.

In this paper, we strive to design energy efficient resource allocation, and use (11) as the QoS constraint, which is called *queue length based resource allocation* (QRA).

In existing works [2, 6, 7], the resource allocation is independent of the queue length, which is called *channel based resource allocation* (CRA).

C. Problem Formulation

EE is defined as the ratio of accumulated data transmitted to the accumulated energy consumed [9], which is

$$EE \triangleq \lim_{T\to\infty} \frac{B\left(T\right)}{\int_0^T P_{tot}\left(t\right)dt}, \tag{13}$$

where $P_{tot}(t)$ is the total power consumption at time t, including both transmit power and circuit power consumed for transmission and by operating the system.

It is easy to show that when the tail probability of the queue length decays exponentially, the arrival and departure processes satisfy

$$\lim_{T\to\infty} \Pr\left\{\left|A\left(T\right)/T - B\left(T\right)/T\right| > \varepsilon\right\} = 0. \tag{14}$$

It means that $\frac{1}{T}B\left(T\right)$ converges to $\frac{1}{T}A\left(T\right)$ in probability. Then, for a stationary and ergodic source $a(t)$, the EE of the system can be redefined as

$$EE = \frac{\mathbb{E}\left[a\left(t\right)\right]}{\lim\limits_{T\to\infty} \frac{1}{T}\int_0^T P_{tot}\left(t\right)dt}. \tag{15}$$

For a source with given average arrival rate, maximizing the EE is equivalent to minimizing the denominator of (15). For the QRA strategy, the resource allocation depends on the queue length, and thus the total power consumption determined by the resource allocation is a function of Q_∞. Since we have assumed that the queue has already in steady state at the initial time $t = 0$, the total power consumption is stationary. Assume that it is ergodic, then $\lim\limits_{T\to\infty} \frac{1}{T}\int_0^T P_{tot}\left(t\right)dt$ converges to $\mathbb{E}_{Q_\infty}\left[P_{tot}\left(Q_\infty\right)\right]$ in probability of 1. To design an optimal QRA, which maximizes the EE under the statistical QoS constraint, we can first find the optimal *QoS function* from the following optimization problem,

$$\min_{\theta(q)} \quad \mathbb{E}_{Q_\infty}\left[P_{tot}\left(Q_\infty\right)\right] = \int_0^{Q_{\max}} P_{tot}\left(q\right)d\Pr\left(Q_\infty \le q\right) \tag{16}$$

$$\text{s.t} \quad (11),$$

where the relation of $P_{tot}(q)$ with $\theta(q)$ depends on the specific system and source as well as the resource allocation strategy.

III. Queue Length Based Resource Allocation

In this section, we first find a numerical solution of problem (16), which is applicable to arbitrary wireless system and arbitrary traffic model. To illustrate how to design a QRA strategy for a particular system, we then take an OFDM system serving Poisson source as an example. Yet the idea of problem-solving can also be applied to other systems such as multi-antenna OFDMA systems serving the sources modeled in [11].

A. Solution of Problem (16)

In order to derive $\mathbb{E}_{Q_\infty}\left[P_{tot}\left(Q_\infty\right)\right]$, we need to derive the *cumulative distribution function* (CDF) of Q_∞. Similar to deriving (9), we can derive $\Pr(Q_\infty \ge q) \approx \eta e^{\int_0^q \theta(x)dx}$, from which the CDF of Q_∞ can be approximated as

$$\Pr\left(Q_\infty \le q\right) \approx 1 - \eta\left\{\exp\left[-\int_0^q \theta\left(x\right)dx\right]\right\}. \tag{17}$$

Substituting (17) into the objective function of (16), we have

$$\mathbb{E}_{Q_\infty}\left[P_{tot}\left(Q_\infty\right)\right] \approx \eta\int_0^{Q_{\max}} \theta\left(q\right)P_{tot}(q)e^{\left[-\int_0^q \theta(x)dx\right]}dq, \tag{18}$$

which is a function of $\theta(q)$. To save energy, the QRA avoids to serve empty buffer. This implies that $\eta \approx 1$ when the delay bound is larger than the inter-arrival time of the packets.

Denote $y\left(q\right) \triangleq \int_0^q \theta\left(x\right)dx$, and then its first-order derivative $\dot{y}(q) = \theta\left(q\right)$. From (18), the optimal problem in (16) that minimizes $\mathbb{E}_{Q_\infty}\left[P_{tot}\left(Q_\infty\right)\right]$ under the statistical QoS constraint can be re-formulated as

$$\min_{y(q)} \quad \int_0^{Q_{\max}} \dot{y}\left(q\right)P_y[\dot{y}(q)]\exp\left[-y\left(q\right)\right]dq \tag{19}$$

$$\text{s.t.} \quad y\left(0\right) = 0, y\left(Q_{\max}\right) = \bar{\theta}Q_{\max}, \tag{20}$$

where $P_y[\dot{y}(q)] = P_{tot}(q)$, which is a function of $\theta(q)$ depending on system and source, and constraint (20) comes from the definition of $y(q)$ and by substituting $y(q)$ into (11).

Problem (19) is a functional extremum problem, which can be solved through solving Euler-Lagrange equation, which is a necessary condition that the global optimal solution of problem (19) satisfies [12]. Denote $L\left[y(q), \dot{y}(q), q\right] \triangleq \dot{y}\left(q\right)P_y[\dot{y}(q)]\exp\left[-y\left(q\right)\right]$, the Euler-Lagrange equation can be expressed as

$$\frac{\partial L}{\partial y(q)} - \frac{d}{dq}\frac{\partial L}{\partial \dot{y}(q)} = 0. \tag{21}$$

If $P_y[\dot{y}(q)]$ is two-order differentiable with respect to $\dot{y}(q)$, upon substituting $L\left[y(q), \dot{y}(q), q\right]$, (21) can be re-expressed as

$$\begin{aligned}&\left\{2\frac{\partial P_y\left[\dot{y}(q)\right]}{\partial \dot{y}(q)} + \dot{y}(q)\frac{\partial^2 P_y\left[\dot{y}(q)\right]}{\partial [\dot{y}(q)]^2}\right\}\ddot{y}(q)\\ &= \dot{y}^2(q)\frac{\partial P_y\left[\dot{y}(q)\right]}{\partial \dot{y}(q)}.\end{aligned} \tag{22}$$

(22) is a second order *ordinary differential equation*(ODE), and (20) is the boundary conditions of the ODE. The solution of this ODE, i.e., $y(q)$, can be numerically obtained by, e.g., *shooting method* [13]. Then, the optimal *QoS function* $\theta(q) = \dot{y}(q)$ can be obtained for any given $P_y[\dot{y}(q)]$.

B. QRA for OFDM System with Poisson Source

Consider an OFDM system, where the maximum number of subcarriers is $N_{\max}$ and the separation between adjacent subcarriers is B. To maximize EE under the statistical QoS constraint, we can jointly optimize the overall transmit power and the number of active subcarriers $N_s(q)$ for transmission.

To simplify the analysis, we only optimize $N_s(q)$ while the transmit power at each active subcarrier is set as equal to a given value. The method for solving the joint optimization problem is tedious but similar.

Assume that the wireless channel is subject to frequency-selective block fading, which is constant within duration T_f but varies independently from one duration to another. In the kth duration, the channel gain on the subcarriers, $g_i(k), i = 1, 2, ..., N_{\max}$, are independent and identically distributed (i.i.d.). Assume that channel distributed information is known at the transmitter, and instantaneous channel information is only available at the receiver. Then, it is optimal for the the active subcarriers having equal transmit power. Suppose that the Shannon capacity achieving coding is used [7]. Then, the accumulate amount of data that can be transmitted within the kth duration of T_f can be expressed as

$$\int_{(k-1)T_f}^{kT_f} s(\tau) d\tau = T_f B \sum_{i=1}^{N_s(q)} \log_2 \left[1 + \frac{g_i(k) P_T}{\sigma_0^2}\right], \quad (23)$$

where P_T is the transmit power on each subcarrier, and σ_0^2 is the noise variance. The accumulated data that the system can serve during interval $[0, nT_f], n \in \mathbb{N}$, can be expressed as $S(nT_f) = \sum_{k=1}^{n} \int_{(k-1)T_f}^{kT_f} s(\tau) d\tau$. When $t = nT_f$, substituting $S(nT_f)$ and (23) into (3), the effective capacity of the OFDM system can be obtained as

$$EC(\theta, N_s) = -\frac{N_s(q)}{T_f \theta(q)} \ln \mathbb{E}_{\gamma_i(k)} \left\{ [1 + \gamma_i(k)]^{-\frac{\theta(q) B T_f}{\ln 2}} \right\}, \quad (24)$$

where $\gamma_i(k) \triangleq \frac{g_i(k) P_T}{\sigma_0^2}$ is the received SNR on the ith subcarrier in the kth duration. Since the channel gains $g_i(k), i = 1, 2, ..., N_s^c(q)$ are i.i.d., the statistics of the SNR does not depend on i and k. Therefore, we omit the indices i and k in the sequel for simplicity.

Consider a Poisson source $a(t)$. Denote $N_p(t)$ as the accumulated number of packets arrived in time interval $[0, t]$, which follows Poisson distribution with parameter λ,

$$\Pr[N_p(t) = n] = \frac{(\lambda t)^n}{n!} e^{-\lambda t}. \quad (25)$$

Denote a_0 as the size of each packet. Then, the accumulated number of bits arrived in time interval $[0, t]$ $A(t) = N_p(t) a_0$. From (2), the effective bandwidth of the Poisson source can be obtained as

$$\begin{aligned} EB(\theta) &= \lim_{t \to \infty} \frac{1}{t\theta(q)} \ln \left\{ \sum_{n=0}^{\infty} \left[e^{\theta(q) n a_0} \frac{(\lambda t)^n}{n!} e^{-\lambda t} \right] \right\} \\ &= \frac{\lambda}{\theta(q)} \left[e^{\theta(q) a_0} - 1 \right]. \quad (26) \end{aligned}$$

In the following, we find the expression of $P_{tot}(q)$ as the function of $\theta(q)$ for the OFDM system serving Poisson source.

The *QoS function* can be obtained by equating the effective capacity and effective bandwidth, i.e., from (10). Upon substituting (24) and (26) into (10), we can obtain $\theta(q)$ as a function of $N_s(q)$, from which we can further obtain the queue length based subcarrier allocation strategy as follows,

$$N_s(q) = -\frac{\lambda T_f \left[e^{\theta(q) a_0} - 1\right]}{\ln \mathbb{E}_\gamma \left[(1+\gamma)^{-\frac{B T_f \theta(q)}{\ln 2}} \right]}. \quad (27)$$

Then, the total power consumption as a function of $\theta(q)$ is

$$\begin{aligned} P_{tot}(q) &= (P_T + P_C) N_s(q) + P_0, \quad (28) \\ &= -\frac{\lambda T_f \left[e^{\theta(q) a_0} - 1\right] (P_T + P_C)}{\ln \mathbb{E}_\gamma \left[(1+\gamma)^{-\frac{B T_f \theta(q)}{\ln 2}} \right]} + P_0, \quad (29) \end{aligned}$$

where P_T and P_C are the transmit power and the circuit power on each subcarrier, respectively, and P_0 is the fixed circuit power consumed at the transmitter when no subcarrier is used.

The relationship between $P_{tot}(q)$ and $\theta(q)$ in (29) is complex. To gain useful insight, we use the following function to fit the expression (29) as in [14],

$$P_{tot}(q) = c_1 \theta^{c_2}(q) + c_3, \quad (30)$$

where c_1, c_2 and c_3 are constants.

Substituting (30) into (22), the Euler-Lagrange equation becomes

$$\ddot{y}(q) - \frac{1}{c_2 + 1} \dot{y}^2(q) = 0. \quad (31)$$

By solving this ODE under the boundary conditions in (20), we can obtain a solution for $\dot{y}(q) = \theta(q)$ as follows,

$$\theta^*(q) = (c_2 + 1) \left[\frac{Q_{\max}}{1 - \exp\left(-\frac{\bar{\theta} Q_{\max}}{c_2 + 1}\right)} - q \right]^{-1}, \quad (32)$$

which is the only solution of (31) and hence is the global optimal solution of problem (19). $\theta^*(q)$ is an increasing function of q, which reflects the fact that the service becomes more urgent when the queue is longer.

The optimal QRA strategy can be obtained by substituting (32) into (27). In low SNR scenario, for Rayleigh fading channel, it is not hard to derive that the closed-form QRA strategy can be approximated as

$$N_s^*(q) \approx \frac{T_f \lambda \left[e^{\theta^*(q) a_0} - 1\right]}{\ln \left[1 + \frac{T_f B \bar{\gamma} \theta^*(q)}{\ln 2}\right]}, \quad (33)$$

which increases with $\theta^*(q)$ and hence grows with q, where $\bar{\gamma}$ is the average receive SNR at each active subcarrier in a duration.

C. Discussion

The proposed QRA differs from relevant resource allocation methods in literature both in the objective function and in the QoS constraint. All existing methods using effective capacity as a design tool [2, 6, 7] are essentially the CRA strategy, because they are all independent of the queue length.

For the considered OFDM system where only the number of active subcarriers is optimized, the objective function in [6] can be expressed as

$$EE\left(N_s^c\right) \triangleq \frac{EC_S(\theta^c, N_s^c)}{P_{tot}^c} = \frac{N_s^c \ln \mathbb{E}_\gamma \left[(1+\gamma)^{-\frac{\theta^c B T_f}{\ln 2}} \right]}{T_f \theta^c [\left(P_T^S + P_C^S\right) N_s^c + P_0]}, \tag{34}$$

which is a monotonically increasing function of N_s^c. As a result, the maximum EE is achieved when $N_s^c = N_{\max}$. Such a strategy is referred to as *CRA 1*.

When the objective function defined in (15) is used for the considered optimization under the following constraint

$$EC(\theta^c) \geq EB(\theta^c), \tag{35}$$

the optimal value of N_s^c is the minimum number of active subcarrier that satisfies (35). This can be regarded as an extended method from [2], where the circuit power was not considered. This strategy is referred to as *CRA 2*.

IV. Simulation and Numerical Results

In this section, we evaluate the EE achieved by the OFDM system serving delay sensitive traffic with random arrival, when either the proposed QRA strategy or the CRA strategy is applied.

In the simulation, the packet arrival process is Poisson process with packet size a_0 and average packet arrival rate λ [15]. The channel is constant within each duration T_f and subject to i.i.d. Rayleigh fading among different durations.

The number of active subcarriers is decided at the beginning of each transmit time interval (TTI) with duration ΔT [16]. At initial time 0, the buffer is empty. In the end of each TTI the queue length, which is used to design the QRA strategy for the following TTI, is computed.

The number of active subcarriers are obtained from different strategies as summarized in the sequel. For the *QRA strategy*, when $q = 0$, no subcarrier is used. When $q > 0$, $N_s(q)$ is computed from (27), where $\theta(q)$ is obtained from (32). If $N_s(q) \geq N_{\max}$ in simulation, we decrease the fitting parameter c_2. For the CRA strategies, θ^c is obtained from (6). Two CRA strategies only differ in the objective functions, which are reduced from [6] and [2] without optimizing transmit power. With *CRA 1* , $N_s^c = N_{\max}$. With *CRA 2*, the number of active subcarriers is obtained from solving $EC\left(\theta^c, N_s^c\right) = EB\left(\theta^c\right)$. The total power consumed in each TTI is computed with (28). The EE is then computed as the ratio of the accumulated data transmitted to the total energy consumed during all TTIs in the interval $[0, T]$ [9].

We use circuit power consumption parameters of the micro BS [9], and the results are similar for other kinds of BSs. The statistical QoS requirement comes from [17]. All the simulation parameters are listed in Table I.

To show the impact of the approximations employed in deriving the QRA strategy for OFDM system, we also provide the numerical result of EE achieved when $D_{\max} = \infty$, in which case the approximations in (5) and (6) are accurate and no other approximations are introduced. The number of active subcarriers when $D_{\max} = \infty$ is obtained by solving problem (16) with $\bar{\theta} = 0$. When $\int_0^1 \theta\left(q\right) dq = 0$, $\theta(q) \equiv 0$ and constraint (11) degenerates into $\mathbb{E}[s(t)] = \mathbb{E}[a(t)]$. Therefore, N_s^c is independent of q, which can be obtained from $BN_s^c \mathbb{E}_\gamma \log_2\left(1+\gamma\right) = \lambda a_0$. With such a loosened QoS constraint, the achieved EE is an upper bound for the EE of the systems with arbitrary delay bound requirement.

TABLE I
List of Simulation Parameters

Average receive SNR $\bar{\gamma}$	10 dB
Delay upper bound D_{max}	$2 \sim 100$ ms
Delay bound violation probability ε	0.01
Packet size of the source a_0	$2 \sim 10$ kbits
Average packet arrival rate λ	200 packets/s
Transmit time interval ΔT	1 ms
Channel coherence time T_f	2 ms
Total number of subcarriers $N_{\max}$	128
Separation among adjacent subcarrier B	15 kHz
Transmit power per subcarrier P_T	$22.1/N_{\max}$ W
Circuit power per subcarrier P_C	$13.6/N_{\max}$ W
Fixed circuit power P_0	11.3 W

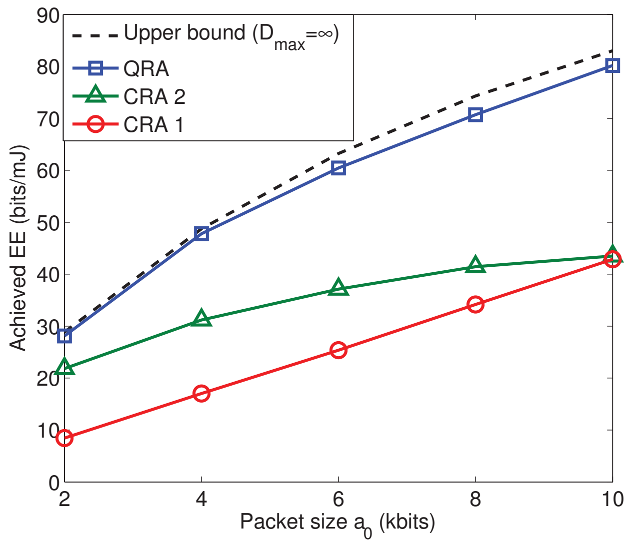

Fig. 1. EE vs. packet size a_0, $D_{\max} = 5$ ms, $\varepsilon_D = 0.01$, and $\lambda = 200$ packets/s.

Figure 1 shows the achieved EE versus packet size a_0, where the average packet arrival rate λ is fixed, and hence the average arrival rate $\mathbb{E}[a(t)]$ increases with a_0. According to the definition in [11], the burstiness of a counting process can be expressed as $I_a \triangleq \lim_{t\to\infty} \frac{\mathrm{Var}\{A(t)\}}{\mathbb{E}\{A(t)\}}$. Note that the definition of burstiness is only for counting process, for the Poisson source, we do not count the number of packets, but the number of bits. Thus, the burstiness of Poisson process can be derived as $I_a = a_0$, which depends on the number of bits per packet. The results in this figure show that the *QRA strategy* can achieve much higher EE than the CRA strategies, especially when the

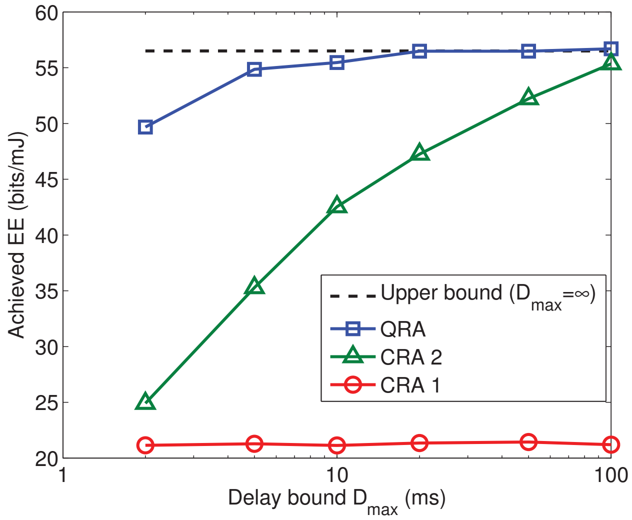

Fig. 2. EE vs. $D_{\max}$, $a_0 = 5$ kbits and $\varepsilon_D = 0.01$.

source becomes more bursty. Moreover, the achieved EE of the *QRA strategy* is close to the upper bound of EE, which implies that the approximations used to design the *QRA strategy* are accurate.

Figure 2 shows the EE versus the required delay bound. Since the TTI is 1 ms in the simulation, the minimum value of $D_{\max}$ is 2 ms (when $\Delta T \to 0$, $D_{\max}$ can be arbitrarily small). When $D_{\max}$ increases, both the EE achieved by the *QRA strategy* and *CRA 2* approach to the upper bound of EE. The EE of the *QRA strategy* is close to the upper bound when $D_{\max} \geq 5$ ms. To ensure the same statistical QoS requirement, *CRA 1* achieves much lower EE for a large range of delay bound requirement. The EE achieved by *CRA 1* is low due to the implicit assumption of full buffer. It is worthy to note that with optimized transmit power the resource allocation method in [6] is able to provide higher EE than the result in this figure, which however will not change the conclusion.

V. Conclusion

In this paper, we have designed energy efficient resource allocation strategy for delay sensitive traffic with statistical QoS requirement. In order not to waste the transmission resources for random arrival process, we incorporated the queue state information into the framework of effective capacity and effective bandwidth by introducing the *QoS function* that depends on the queue length. After translating the statistical QoS requirement into a constraint on the *QoS function*, we provided a method for finding the *QoS function* that maximizes the EE under the QoS constraint for general systems and sources. Then, by taking an OFDM system that serves Poisson source as an example, the optimal subcarrier allocation strategy with closed form was provided. Simulation results showed that the queue length based subcarrier allocation strategy achieves much higher EE than existing strategies especially when the arrival process is bursty or the delay bound is tight.

References

[1] G. Y. Li, Z. Xu, C. Xiong, C. Yang, S. Zhang, Y. Chen, and S. Xu, "Energy-efficient wireless communications: tutorial, survey, and open issues," *IEEE Wireless Commun. Mag.*, vol. 18, no. 6, pp. 28–35, Dec. 2011.

[2] X. Zhang and J. Tang, "Power-delay tradeoff over wireless networks," *IEEE Trans. Commun.*, vol. 61, no. 9, pp. 3673–3684, Sep. 2013.

[3] 3GPP, *Further Advancements for E-UTRA Physical Layer Aspects*. TSG RAN TR 36.814 v9.0.0, Mar. 2010.

[4] C.-S. Chang and J. A. Thomas, "Effective bandwidth in high-speed digital networks," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 6, pp. 1091–1100, Aug. 1995.

[5] D. Wu and R. Negi, "Effective capacity: A wireless link model for support of quality of service," *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 630–643, Jul. 2003.

[6] L. Liu, "Energy-efficient power allocation for delay-sensitive traffic over wireless systems," in *Proc. IEEE ICC*, Jun. 2012.

[7] C. Xiong, G. Y. Li, Y. Liu, Y. Chen, and S. Xu, "Energy-efficient design for downlink OFDMA with delay-sensitive traffic," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 3085–3095, Jun. 2013.

[8] R. A. Berry and R. G. Gallager, "Communication over fading channels with delay constraints," *IEEE Trans. Inf. Theory*, vol. 48, no. 5, pp. 1135–1149, May 2002.

[9] G. Auer, O. Blume, V. Giannini, I. Gódor, *et al.*, "D 2.3: Energy efficiency analysis of the reference systems, areas of improvements and target breakdown," *EARTH*, Nov. 2010. [Online]. Available: https://www.ict-earth.eu/publications/deliverables/deliverables.html

[10] J. Tang and X. Zhang, "Cross-layer-model based adaptive resource allocation for statistical QoS guarantees in mobile wireless networks," *IEEE Trans. Wireless Commun.*, vol. 7, no. 6, pp. 2318–2328, Jun. 2008.

[11] G. L. Choudhury, D. M. Lucantoni, and W. Whitt, "Squeezing the most out of ATM," *IEEE Trans. Commun.*, vol. 44, no. 2, pp. 203–217, Feb. 1996.

[12] V. I. Arnold, *Mathematical methods of classical mechanics*. Springer, 1989.

[13] U. M. Ascher, R. M. M. Mattheij, and R. D. Russell, *Numerical solution of boundary value problems for ordinary differential equations*. SIAM, 1995.

[14] M. A. Zafer and E. Modiano, "Delay-constrained energy efficient data transmission over a wireless fading channel," in *Proc. IEEE Inf. Theory and Appl. Workshop*, Feb. 2007.

[15] H. Chen, H. C. B. Chan, C.-K. Chan, and V. C. M. Leung, "QoS-based cross-layer scheduling for wireless multimedia transmissions with adaptive modulation and coding," *IEEE Trans. Commun.*, vol. 61, no. 11, pp. 4526 – 4538, Nov. 2013.

[16] F. Capozzi, G. Piro, L. Grieco, G. Boggia, and P. Camarda, "Downlink packet scheduling in LTE cellular networks: Key design issues and a survey," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 2, pp. 678 – 700, 2013.

[17] C.-F. Tsai, C.-J. Chang, F.-C. Ren, and C.-M. Yen, "Adaptive radio resource allocation for downlink OFDMA/SDMA systems with multimedia traffic," *IEEE Trans. Wireless Commun.*, vol. 7, no. 5, pp. 1734–1743, May. 2008.