Energy-Efficient Resource Allocation for MIMO-OFDM Systems Serving Random Sources with Statistical QoS Requirement

Changyang She, Chenyang Yang, and Lingjia Liu

***Abstract*—This paper optimizes resource allocation that maximizes the energy efficiency (EE) of wireless systems with statistical quality of service (QoS) requirement, where a delay bound and its violation probability need to be guaranteed. To avoid wasting energy when serving random sources over wireless channels, we convert the QoS exponent, a key parameter to characterize statistical QoS guarantee under the framework of effective bandwidth and effective capacity, into multi-state QoS exponents dependent on the queue length. To illustrate how to optimize resource allocation, we consider multi-input-multi-output orthogonal frequency division multiplexing (MIMO-OFDM) systems. A general method to optimize the *queue length based bandwidth and power allocation (QRA)* policy is proposed, which maximizes the EE under the statistical QoS constraint. A closed-form optimal QRA policy is derived for massive MIMO-OFDM system with infinite antennas serving the first order autoregressive source. The EE limit obtained from infinite delay bound and the achieved EEs of different policies under finite delay bounds are analyzed. Simulation and numerical results show that the EE achieved by the QRA policy approaches the EE limit when the delay bound is large, and is much higher than those achieved by existing policies considering statistical QoS provision when the delay bound is stringent.**

***Index Terms*—Energy efficiency, statistical QoS requirement, queue length based resource allocation, effective capacity, effective bandwidth**

I. Introduction

Energy efficiency (EE) is an important design goal for wireless systems [1]. Different from another critical goal, spectral efficiency (SE), the fundamental problem tackled by EE-optimal design is shifted from maximizing the data rate that *can be* transmitted into conveying the data that *need to be* transmitted. This suggests that providing higher rate than that required to ensure the quality of service (QoS) of each user is a waste of physical resources. This also indicates that the full buffer assumption behind designing SE-optimal transmission strategies needs a careful rethinking, and the EE-optimal design should also adapt to traffic variations rather than only adapt to channel variations [2]. As a consequence, the features of undergoing traffic in a system should be taken into account in order to maximize EE.

In future wireless communications, a significant portion of traffic is delay sensitive, e.g., video/audio and interactive data transmission, which requires low end-to-end delay. The delay performance metrics studied in the literature can generally be categorized into deterministic delay bound [3], average delay [4], and statistical QoS requirement [5]. In wireless systems, the deterministic delay bound is usually either impossible or too expensive to guarantee in terms of transmit power. For multimedia applications, the average delay guarantee does not necessarily ensure the delay performance required by the services where a data packet becomes useless once its delay requirement is violated. The *statistical QoS requirement*, defined as a delay bound $D_{\max}$ and a delay violation probability ε_D, is more relevant in this context. For example, in the fourth generation (4G) LTE/LTE-Advanced systems, the upper bound on ε_D for VoIP is 2% while $D_{\max}$ is 50 ms for radio access networks [6].

Effective bandwidth [7] and effective capacity [8] are powerful tools to study resource allocation with statistical QoS requirement. The notion of effective capacity stems from information theory, which characterizes the maximal constant arrival rate a wireless system can support under the statistical QoS constraint. Differing from Shannon capacity, using effective capacity as a metric allows one to analyze delay-sensitive traffic with various delay requirements in a unified manner [9–12]. Recently, energy efficient resource allocation policies subject to statistical QoS requirement were studied for single user [13–15] and multi-user wireless systems [16]. Though not explicitly optimizing towards EE, the resource allocation policies in [5] that minimizes the transmit power of orthogonal frequency division multiplexing (OFDM) systems under the effective capacity constraint can be easily extended to an EE-optimal design by taking into account the circuit power consumed for operating the system. To reflect the impact of such a QoS provision, a concept of *delay sensitive EE* was introduced [13], which is defined as the ratio of effective capacity to the total transmit and circuit power consumption and employed as the objective function to optimize the transmit power. In a parallel line of works in [14,15], the same objective function was used for optimizing power allocation respectively over frequency-selective and Nakagami fading channels. The

Manuscript received 4 October 2014; revised 4 April 2015 and 17 July 2015; accepted 11 September 2015.

Changyang She and Chenyang Yang are with the School of Electronics and Information Engineering, Beihang University, Beijing 100191, China (email: {cyshe,cyyang}@buaa.edu.cn).

C. She and C. Yang's work is supported in part by National Natural Science Foundation of China (NSFC) under Grant 61120106002, and National Basic Research Program of China, 973 Program 2012CB316003.

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

L. Liu's work is supported in part by NSF under grants ECCS-1228071, CCF-1422241, and ECCS-1509514.

optimal solutions were obtained, and the impacts of the delay requirement indicated by a QoS exponent, as well as the circuit power, transmit power constraint, and channel feature were investigated. In [16], the delay sensitive EE was maximized for a multi-user multi-carrier system under the constraint on effective capacity imposed by each user, where interesting results for the tradeoff between EE and delay as well as the connection between spectral efficient and energy efficient designs were obtained. By maximizing the delay sensitive EE, these works are essentially targeted to analyze the maximal achievable EE of the system subject to the statistical QoS requirement [13–16].

While important results have been obtained to optimize both spectral efficient and energy efficient resource allocation by using effective bandwidth and effective capacity, a common assumption behind all existing studies in the literature [5, 13–16] is that the data arrives at the buffer at a constant rate. In real-world systems, data traffic hardly has a constant arrival rate. In fact, the traffic may vary in large scale (i.e., the average arrival rate changes in different hours in a day) and in small scale (i.e., the instantaneous arrival rate is random). As indicated by a seminal work on analyzing minimal transmit power with average delay constraint [4], when *either* source *or* channel is stochastic, the resource allocation should depend on the queue state information (QSI), i.e., queue length for the considered memoryless source. Otherwise, the average power cannot be minimized for a target average delay. Nonetheless, how to incorporate the QSI into the framework of effective capacity remains an open problem.

In this paper, we revisit the problem of EE-optimal resource allocation with statistical QoS requirement and go further step by considering the more general, practical, yet more challenging scenario where *both* source *and* channel are random. Moreover, we consider the traffic with different average arrival rates instead of studying the maximal achievable EE by implicitly assuming high traffic load. We strive to establish a framework to optimize energy efficient resource allocation with statistical QoS provision for these practical scenarios. The problems needing to tackle are how to introduce proper QoS constraints with effective bandwidth and effective capacity and define a proper objective function such that EE can be truly maximized. To illustrate how the framework can be applied, we consider multi-input-multi-output (MIMO)-OFDM systems. To accommodate the randomness in both the source and the channel in the EE-optimal problem with statistical QoS requirement, we propose *queue length based resource allocation* (QRA). Under the framework of effective bandwidth and effective capacity, QoS exponent is a key parameter to characterize statistical QoS. Inspired by a pioneering work in designing wired networks [7], we convert the QoS exponent into multi-state QoS exponents dependent on the queue length. Then, the statistical QoS requirement can be transformed into a constraint on the multi-state QoS exponents. Since effective capacity reflects the maximal constant arrival rate a system can support, using the objective function in [13–16] for optimization assumes full buffer implicitly. As a result, the average data rate of the system achieved by the policies in [13–16] may be much higher than the average arrival rate of a source, which determines the actual data rate required by the user. Given that the basic principle behind the EE-optimal design is not to waste physical resources, providing maximal service capability is inefficient when the traffic load is low. Therefore, the EE we optimized is the ratio of the throughput to the overall power consumption, which is an *energy per bit metric* [17]. The major contributions of this paper are summarized as follows:

- *Framework:* By introducing the multi-state QoS exponents into effective capacity and effective bandwidth, we build a general framework to optimize energy efficient resource allocation for random sources with statistical QoS provision over random channels under different traffic load.
- *Solution:* A method to find QRA policy for MIMO-OFDM system is provided, where the number of active subcarriers and transmit power are joint optimized to maximize the EE under the statistical QoS constraint. By taking first order autoregressive source served by massive MIMO-OFDM system as an example, we derive a closed-form optimal solution for the QRA policy, which gives insight on understanding the behavior of energy efficient resource allocation policy for the system with random arrivals.
- *Performance:* We provide the ultimate EE limit, which is achievable when $D_{\max} \to \infty$. We show that if a resource allocation policy serves empty buffer, e.g., the policy only adapts to channel variations, then the policy cannot achieve the EE limit. Simulation results demonstrate that the QRA policy achieves higher EE than relevant policies in the literature for delay sensitive traffic with random arrival rate even under high traffic load. With less stringent delay bounds, the EE achieved by the QRA policy approaches the EE limit rapidly.

II. System Model and Problem Formulation

A. System Model

We employ cross-layer resource allocation to improve EE. The overall system architecture can be found from many existing works, e.g., [12, 15], where the data frames from the upper layer of the protocol stack are first divided into multiple packets at data-link layer, generating the data source. The packets arrive at the buffer of the BS, and then are split into bit-streams at physical layer, where the transmission resources are allocated. To improve the performance of the system and ensure the QoS of the users, the resources need to be allocated according to the QSI and/or channel state information (CSI) in each transmit time interval (TTI) [18].

Consider a MIMO-OFDM system, where a BS equipped with N_T antennas serves a user with a single antenna. The maximal transmit power of the BS is $P_T^{\max}$, and the system has overall $N_S^{\max}$ subcarriers with subcarrier separation B. To save energy with ensured QoS, the service ability of the system can be adjusted by allocating transmit power and bandwidth

in each TTI. P_T^c and N_S^c respectively denote the allocated transmit power and the number of active subcarriers (that is equivalent to the used bandwidth).

Assume frequency-selective block fading channels, which are constant within duration T_c and vary independently from one duration to another. Denote the spatial channel vector on subcarrier j as $\mathbf{h}_j \in \mathbb{C}^{1\times N_T}$, whose elements are assumed as independent Gaussian distributed with zero mean and variance μ. Each subcarrier is assumed to experience independent identically distributed (i.i.d.) channel fading contaminated by white Gaussian noise with variance σ_0^2.[1]

Assume that instantaneous channel vectors $\mathbf{h}_j, j = 1, \cdots N_S^c$ (i.e., CSI) are perfectly known at the user but only channel distribution information is known at the BS.[2] Then, channel capacity can be achieved by equally allocating transmit power among subcarriers and antennas. When the Shannon capacity achieving coding is used, the maximum instantaneous transmission rate within the duration of the kth fading block is [16]

$$s\left(t\right) = B\sum_{j=1}^{N_S^c}\log_2\left(1+\frac{P_T^c}{\sigma_0^2 N_T N_S^c}\mathbf{h}_j{\mathbf{h}_j}^H\right) \quad \text{(bits/s)}, \quad (1)$$

where $t \in (kT_c, (k+1)T_c]$, $k \in \mathbb{Z}$. As discussed in [16], when the rateless codes such as Luby transform and Raptor codes are used, the BS can adjust transmission rate to channel realizations and achieve $s\left(t\right)$ without the knowledge of CSI. The transmission rate is also called service rate in the sequel, since it reflects the service ability of a system with given transmission resources. Due to the block channel fading, the service process $\{s(t), t \geq 0\}$ is a random process.

Throughout the analysis of this paper, we consider a fluid queueing model as in [11, 13]. Clearly, such a model is an approximation of the actual system. However, the model is accurate to characterize prevalent cellular systems where the inter-arrival time between the packets and the TTI are much shorter than the required delay bound [6, 18, 19]. Assume that the buffer size is infinite, which is reasonable for modern BSs, since large memory is cheap nowadays. At time t, the data traffic from upper layer in the protocol stack enters a first-in-first-out buffer at the BS with instantaneous arrival rate $a(t)$ and is transmitted to the user with instantaneous departure rate $b(t)$. The dynamics of the queue length can be expressed as $\frac{dQ(t)}{dt} = a(t) - b(t)$, where $Q(t)$ is the queue length at time t. Moreover, as shown in [20],

$$\frac{dQ\left(t\right)}{dt} = \begin{cases} \max\left\{a\left(t\right)-s\left(t\right),0\right\}, Q\left(t\right)=0, \\ a\left(t\right)-s\left(t\right), \qquad\qquad Q\left(t\right)>0. \end{cases} \quad (2)$$

Therefore, the throughput of the system, $b(t)$, is related with $a(t)$ and $s(t)$ as

$$b\left(t\right) = \begin{cases} \min\{a\left(t\right),s(t)\}, Q\left(t\right)=0, \\ s\left(t\right), \qquad\qquad Q\left(t\right)>0. \end{cases} \quad (3)$$

B. Statistical QoS Requirement

The statistical QoS requirement of delay sensitive traffic is defined as $(D_{\max}, \varepsilon_D)$, where $D_{\max}$ is the delay bound and ε_D is the maximal delay violation probability. In this paper, we consider the queuing delay at the BS and ignore the coding and transmission delay in other layers of the protocol stack as in the literature in the context [5, 13–16].

Effective bandwidth and effective capacity are powerful tools to design resource allocation policy ensuring statistical QoS requirement [7,8]. When the assumptions for the Gärtner-Ellis theorem are satisfied, the effective bandwidth of random arrival process $\{a(t), t \geq 0\}$ can be expressed as [7]

$$E_B\left(\theta\right) = \lim_{t\to\infty}\frac{1}{\theta t}\ln\mathbb{E}\left[e^{\theta\int_0^t a(\tau)d\tau}\right], \quad (4)$$

where $\theta > 0$ is the *QoS exponent*. A large value of θ indicates a stringent delay bound, and *vice versa*. The effective capacity of random service process $\{s(t), t \geq 0\}$ can be expressed as [8]

$$E_C\left(\theta\right) = -\lim_{t\to\infty}\frac{1}{\theta t}\ln\mathbb{E}\left[e^{-\theta\int_0^t s(\tau)d\tau}\right]. \quad (5)$$

According to the asymptotic analysis in [21], for stationary arrival and service processes with the average arrival rate $\mathbb{E}[a(t)]$ less than the average service rate $\mathbb{E}[s(t)]$, the probability that the queue length exceeds a given threshold $Q_{\max}$ decays exponentially fast as the threshold increases. Assume that the queueing system is in steady state, and denote the steady state queue length as Q_∞. A *necessary and sufficient condition* for $\Pr\{Q_\infty > Q_{\max}\} \leq e^{-\theta Q_{\max}}$ to hold is $E_C\left(\theta\right) \geq E_B\left(\theta\right)$ [11]. The value of QoS exponent satisfying the QoS requirement, $(Q_{\max}, \varepsilon_Q)$, can be determined from [8, 12]

$$\varepsilon_Q = \Pr\left\{Q_\infty > Q_{\max}\right\} \approx \eta e^{-\theta^c Q_{\max}}, \quad (6)$$

where $\eta \triangleq \Pr\left\{Q_\infty > 0\right\}$ is the buffer non-empty probability. The approximation is accurate when the maximum queue length $Q_{\max}$ is long according to the large deviation principle [7].

The value of θ^c can also be determined from the following approximation [9]

$$\varepsilon_D = \Pr\left\{D_\infty > D_{\max}\right\} \approx \eta e^{-\theta^c E_B(\theta^c) D_{\max}}, \quad (7)$$

where D_∞ is the queueing delay of the data in steady state. When $Q_{\max} = E_B(\theta^c)D_{\max}$ and $\varepsilon_Q = \varepsilon_D$, the requirements $(D_{\max}, \varepsilon_D)$ and $(Q_{\max}, \varepsilon_Q)$ are equivalent in the sense that they are satisfied with the same QoS exponent θ^c.

[1] These assumptions for channel statistics in time-, frequency- and spatial-domain are only for better tractability and easy exposition, which are frequently made in the literature, e.g., [10,16]. In practical systems, these assumptions may not hold, yet the proposed framework is still applicable. For example, when the channel gains among subcarriers are not i.i.d., the QRA policy will provide higher EE gain over existing policies as shown in the simulation later.

[2] Such an assumption is only made for mathematical tractability. When the CSI is known at the BS, there is no closed-form expression for the effective capacity [10], which will lead to rather involved optimization. On the other hand, this assumption is reasonable when the user moves very fast and the channel coherent time is much shorter. In this scenario, the training overhead to obtain the CSI at the BS will be too heavy.

In existing works considering statistical QoS requirement [5, 10, 13–16], the resource allocation policies only depend on channel information but are independent of queue length. We refer to them as the *channel based resource allocation* (CRA) policies. It is worth noting that these CRA policies in literature can be divided into two kinds. The first kind of CRA policies allocate resources based on the channel distribution information [13,16]. The second kind of CRA policies allocate resources based on the CSI [5, 10, 14, 15]. In this paper, we allocate physical layer resources adaptive to Q_∞ and statistical information of both source and channel by devising a QRA policy. Unless otherwise specified, the CRA policies to be compared in the analysis and simulation later is the first kind policies, which is referred to as the CRA policies for brevity.

C. Queue Length Based QoS Exponent

As observed in [7] and [4], the CRA policies are the best solutions in terms of minimizing power or maximizing capacity for the traffic with constant arrival rate. Yet practical data packets in real-world applications hardly arrive at the buffer with constant rates. For a stochastic arrival process $\{a(t), t \geq 0\}$, the resource allocation policy (and therefore the service process $\{s(t), t \geq 0\}$) should also depend on the QSI except the CSI in order not to waste the physical resources. To design a QSI-based policy, we introduce a notion of *multi-state QoS exponents*, which depend on the queue length. Such an idea was first proposed in [7] for designing dynamic bandwidth allocation of wired line communications, where the channels are deterministic.

To introduce the multi-state QoS exponents to wireless systems where the channels are random, we use the same procedure as in [7] to emphasize the difference. We start by dividing the original queue with length $Q_{\max}$ into two segments. Consider a piece-wise resource allocation policy, which does not change when Q_∞ lies in $(0, \frac{Q_{\max}}{2}]$ or $(\frac{Q_{\max}}{2}, Q_{\max}]$, but updates when Q_∞ varies from one segment to another. Denote the QoS exponents and the resource allocation policies when $Q_\infty \in \left(\frac{i-1}{2}Q_{\max}, \frac{i}{2}Q_{\max}\right]$ as θ_i and ϕ_i, respectively, $i = 1, 2$. As shown in Appendix II in [11], for any given value of θ_i, the policy ϕ_i in wireless systems should be designed such that

$$E_C(\theta_i, \phi_i) \geq E_B(\theta_i), i = 1, 2. \tag{8}$$

Different from [7], the left hand side of (8) is no longer a deterministic service rate but the effective capacity of a stochastic service process. With the piece-wise resource allocation policy, the event $Q_\infty > Q_{\max}$ consists of two events: Q_∞ exceeds $\frac{Q_{\max}}{2}$ with policy ϕ_1 in $(0, \frac{Q_{\max}}{2}]$, and Q_∞ increases from $\frac{Q_{\max}}{2}$ to $Q_{\max}$ with policy ϕ_2 in $(\frac{Q_{\max}}{2}, Q_{\max}]$. Then, the probability of a queue length exceeding $Q_{\max}$ in (6) becomes

$$\begin{aligned}
&\Pr\{Q_\infty > Q_{\max}\} \\
&= \Pr\left\{Q_\infty > \frac{Q_{\max}}{2}\right\} \Pr\left\{Q_\infty > Q_{\max} | Q_\infty > \frac{Q_{\max}}{2}\right\} \\
&\approx \eta \exp\left[-\left(\frac{Q_{\max}}{2}\theta_1 + \frac{Q_{\max}}{2}\theta_2\right)\right].
\end{aligned} \tag{9}$$

If $\frac{Q_{\max}}{2}$ is large enough (which is true when it is much larger than the accumulated data arrived in each TTI of ΔT, $A_\Delta(t) \triangleq \int_t^{t+\Delta T} a(\tau)\, d\tau$ [22]), then the approximation in (9) will be accurate (see [7] and the references therein). In prevalent cellular systems, $\Delta T \ll D_{\max}$, which implies that $A_\Delta(t) \ll Q_{\max}$. This suggests that we have $\frac{Q_{\max}}{2} \gg A_\Delta(t)$, i.e., the approximation is accurate. Further considering that the buffer non-empty probability $\eta \leq 1$, we can obtain an upper bound of the probability as

$$\Pr\{Q_\infty > Q_{\max}\} \leq \exp\left[-(\theta_1 + \theta_2)\frac{Q_{\max}}{2}\right]. \tag{10}$$

As in [7], we generalize the results by dividing $Q_{\max}$ into N_q segments, each with length $l = \frac{Q_{\max}}{N_q}$. We refer to $Q_\infty \in ((i-1)l, il]$ as the ith state of the queue. For a given N_q, when $Q_{\max} \gg A_\Delta(t)$, we have $l \gg A_\Delta(t)$, and thus the effective bandwidth and effective capacity can be applied in each segment. The QoS exponent for the queue length lying in the ith segment is a constant. Then, the N_q-state QoS exponent, θ_i, is a function of $\frac{i}{N_q}$, $i = 1, 2, ..., N_q$. With a multi-state resource allocation policy $\{\phi_1, ..., \phi_{N_q}\}$ associated with $\{\theta_1, ..., \theta_{N_q}\}$, when (8) is satisfied in each state, the upper bound in (10) becomes

$$\Pr\{Q_\infty > Q_{\max}\} \leq \exp\left(-l\sum_{i=1}^{N_q}\theta_i\right), \tag{11}$$

The QoS requirement $(Q_{\max}, \varepsilon_Q)$ can be guaranteed by letting the upper bound of $\Pr\{Q_\infty > Q_{\max}\}$ equal to ε_Q. This suggests that the multi-state QoS exponents $\theta_1, ..., \theta_{N_q}$ should satisfy

$$\frac{1}{N_q}\sum_{i=1}^{N_q}\theta_i = \frac{\ln(1/\varepsilon_Q)}{Q_{\max}} \triangleq \bar{\theta}. \tag{12}$$

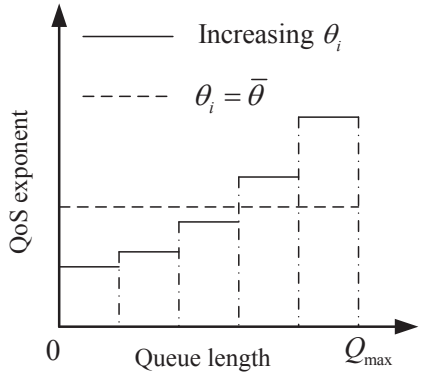
Remark 1: To satisfy the requirement of $(D_{\max}, \varepsilon_D)$, the multi-state QoS exponents, $\theta_1, ..., \theta_{N_q}$, should increase with queue length. That is to say, if $i \leq j$, then $\theta_i \leq \theta_j$.

We use a simple example to explain this. Consider a random source with QoS requirement $(D_{\max}, \varepsilon_D)$, which is served by a system whose received signal is contaminated by additive white Gaussian noise (AWGN). Since the channel is deterministic, the effective capacity of the system is equal to the service rate, which is not random.

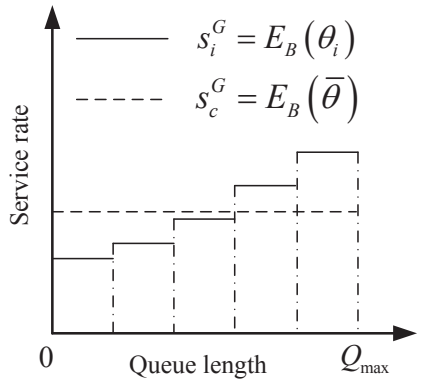
If $\theta_1, ..., \theta_{N_q}$ do not depend on queue state, then the service rate will be a constant. In order to guarantee the QoS, the service rate should satisfy $s_c^G = E_B(\bar{\theta})$, where $\bar{\theta}$ is defined in (12) [7]. With such a constant service rate s_c^G, the queueing delay of the data arriving at the buffer at time t can be expressed as $\frac{Q(t)}{s_c^G}$, which increases with $Q(t)$.

If θ_i increases with queue length Q_∞ as illustrated in Fig. 1(a), then to ensure the QoS the service rate should satisfy $s_i^G = E_B(\theta_i)$, $i = 1, ..., N_q$. Because effective bandwidth increases with the QoS exponent, s_i^G increases with Q_∞, as illustrated in 1(b). When $Q(t)$ stays in the ith state, the queueing delay is $D(t) = \frac{Q(t)}{s_i^G}$. If $Q(t)$ is long such that $s_i^G > s_c^G$, then the queuing delay $D(t)$ will be shorter than

$\frac{Q(t)}{s_c^G}$. Otherwise, the delay will be longer than $\frac{Q(t)}{s_c^G}$, which however does not matter, because in this case $Q(t)$ is short and the queuing delay is much less than the delay bound $D_{\max}$. It indicates that the delay violation probability when served with s_i^G is less than that when served with s_c^G. Given that $(D_{\max}, \varepsilon_D)$ can be guaranteed when served with s_c^G, $(D_{\max}, \varepsilon_D)$ can also be satisfied when the multi-state QoS exponents increase with Q_∞.



(a) Queue length dependent QoS exponent.



(b) Queue length dependent service rate.

Fig. 1. An example in Remark 1, where $N_q = 5$.

D. Problem Formulation for Energy-Efficient QRA Policy

We optimize queue length based bandwidth and transmit power allocation policy that maximizes the EE of MIMO-OFDM systems.[3] Denote the number of subcarriers and transmit power used for supporting the required QoS exponent θ_i when $(i-1)l < Q_\infty \leq il$ as $N_S(\theta_i)$ and $P_T(\theta_i)$, and denote the QRA policy as $\phi_i = \{N_S(\theta_i),\ P_T(\theta_i)\}$, where $N_S(\theta_i) \leq N_S^{\max}$ and $P_T(\theta_i) \leq P_T^{\max}$, $i = 1,..., N_q$.

The total power consumption includes the powers consumed for transmitting data and for operating the system. For the considered problem at hand, we are only interested in the impact of number of active subcarriers on the circuit power. With the QRA policy, both the transmit power and circuit power depend on the multi-state QoS exponents. The total power consumption of the BS in the ith state can be modeled as

$$P_{tot}(\theta_i) = \frac{1}{\rho} P_T(\theta_i) + P_{cs} N_S(\theta_i) + P_c, \tag{13}$$

where ρ is the power amplifier efficiency, P_{cs} is the circuit power equivalently consumed for each subcarrier, and P_c is the fixed circuit power independent from $N_S(\theta_i)$, all can be obtained from the flexible model in [23]. Note that in practical systems ρ is not a constant. For simplicity and a fair comparison with existing policies [13, 14, 16], we assume that the power amplifier works in linear region, where ρ does not change with the transmit power. When more realistic power consumption model is taken into account, the proposed framework will not change.

The EE is defined as the ratio of average throughput $\mathbb{E}[b(t)]$ to the average power consumption at the BS [17]. For ergodic arrival process and service process, it is not hard to show that $\mathbb{E}[a(t)] = \mathbb{E}[b(t)]$ when (6) is satisfied. Then, the EE can be expressed as

$$EE \triangleq \frac{\mathbb{E}\left[b\left(t\right)\right]}{\mathbb{E}\left[P_{tot}(\theta_i)\right]} = \frac{\mathbb{E}\left[a\left(t\right)\right]}{\mathbb{E}\left[P_{tot}(\theta_i)\right]}, \tag{14}$$

where the expectation of the power consumption is taken over Q_∞.

For the source with given average arrival rate, maximizing the EE is equivalent to minimizing the denominator of (14). Then, the problem of optimizing QRA policy to maximize the EE of MIMO-OFDM system under the statistical QoS constraint can be formulated as

$$\min_{\phi_i, i=1,\ldots,N_q} \mathbb{E}[P_{tot}(\theta_i)] \tag{15}$$

$$\text{s.t} \quad \frac{1}{N_q}\sum_{i=1}^{N_q} \theta_i = \bar{\theta}, \tag{15a}$$

$$E_C\left(\theta_i, \phi_i\right) \geq E_B\left(\theta_i\right), (i-1)l < Q_\infty \leq il, \tag{15b}$$

$$0 \leq N_S(\theta_i) \leq N_S^{\max}, 0 \leq P_T(\theta_i) \leq P_T^{\max}, i = 1, ..., N_q, \tag{15c}$$

where $\bar{\theta} = \frac{\ln(1/\varepsilon_Q)}{Q_{\max}}$. In order to satisfy the requirement $(Q_{\max}, \varepsilon_Q)$, the multi-state QoS exponents should meet the constraint in (15a), and the resource allocation policy should satisfy the constraint in (15b). Given that serving an empty buffer will waste the service capability, the QRA policy will let $s(t) = 0$ when $Q(t) = 0$.

To find the solution of problem (15), we first need to find the relationship of $P_{tot}(\theta_i)$ with ϕ_i and then with θ_i. Note that for any given value of θ_i, multiple resource allocation policies can satisfy the sufficient and necessary condition for the QoS requirement to hold (i.e., constraint (15b)), which leads to different values of power consumption. To find the relation between $P_{tot}(\theta_i)$ and θ_i that can finally maximize EE, we select the policy that minimizes $P_{tot}\left(\theta_i\right)$ for any given θ_i under constraints (15b) and (15c). We refer to such an optimization problem as subproblem I. Denote the optimal policy and the corresponding minimum total power

[3] We consider such a multi-antenna multi-carrier system just for illustrating how QRA policy can be optimized. We can also consider single antenna OFDM or single carrier MIMO systems. In the problem at hand, the difference between these systems lies in the power consumption model, which however will not change the optimization framework we proposed.

as $\phi_i^* = \{N_S^*(\theta_i), P_T^*(\theta_i)\}$ and $P_{tot}^*(\theta_i), \forall \theta_i \in (0, \theta_{\max}]$, respectively, where $\theta_{\max}$ is the QoS exponent that the system can support with $N_S^{\max}$ and $P_T^{\max}$. Then, the relation is $P_{tot}^*(\theta_i)$.

The second step of solving problem (15) is to find the multi-state QoS exponents. Note that the required values of θ_1,..., θ_{N_q} cannot be uniquely determined from (15a) as in determining the required QoS exponent in CRA policies from (6). Besides, the values of θ_i will affect the distribution of the queue length, and hence affect the average total power consumption. Therefore, we find the optimal multi-state QoS exponents that minimize $\mathbb{E}[P_{tot}^*(\theta_i)]$ under constraints (15a) and $\theta_i \leq \theta_{\max}$. We refer to such an optimization problem as subproblem II. Denote the optimal solution as $\{\theta_i^*, i = 1,..., N_q\}$. Then, the final optimized QRA policy can be obtained as $\{N_S^*(\theta_i^*), P_T^*(\theta_i^*), i = 1,..., N_q\}$.

III. Solution of Queue length based resource allocation policy

In this section, we first present a general approach to find the solution of problem (15) without specifying the source. Specifically, we first solve subproblem I, and then solve subproblem II. Then, to provide a closed-form policy and gain useful insight, we proceed to derive the optimal QRA policy in closed-form for massive MIMO system serving an AR(1) source.

A. Optimizing Resource Allocation Given Multi-state QoS Exponents

To establish the relation between the minimum total power consumption and QoS exponent (i.e., $P_{tot}^*(\theta_i)$), we find the resource allocation policy that minimizes the instantaneous total power consumption under constraints (15b) and (15c) from subproblem I as follows,

$$\min_{\phi_i=\{N_S(\theta_i),P_T(\theta_i)\}} P_{tot}(\theta_i) = \frac{P_T(\theta_i)}{\rho} + P_{cs}N_S(\theta_i) + P_c \tag{16}$$

$$\text{s.t} \quad E_C(\theta_i, \phi_i) \geq E_B(\theta_i), i = 1, ..., N_q, \tag{16a}$$

$$0 \leq N_S(\theta_i) \leq N_S^{\max}, 0 \leq P_T(\theta_i) \leq P_T^{\max}, \tag{16b}$$

where we have relaxed the number of subcarriers as continuous variables for simplicity.[4]

From (1) and (5), the effective capacity in (16a) can be obtained as

$$E_C(\theta_i, \phi_i) = -\frac{N_S(\theta_i)}{T_c\theta_i} \ln \mathbb{E}_{\mathbf{h}} \left\{ \left[1 + \frac{P_T(\theta_i)}{\sigma_0^2 N_T N_S(\theta_i)} \mathbf{h}\mathbf{h}^H \right]^{-\frac{T_c B \theta_i}{\ln 2}} \right\}. \tag{17}$$

[4]Later, we will use simulation to show that the relaxation has little impact on the EE performance of QRA policy.

The objective function and the constraints in (16b) are linear. Since $E_C(\theta_i, \phi_i)$ in (17) is joint concave in $N_S(\theta_i)$ and $P_T(\theta_i)$ (as proved in Appendix A), the problem is convex. Therefore, the optimal solution of problem (16), $\phi_i^* = \{N_S^*(\theta_i), P_T^*(\theta_i)\}$, can be found by standard tools such as the interior-point method [24]. Substituting $N_S^*(\theta_i)$ and $P_T^*(\theta_i)$ into the objective function (16), $P_{tot}^*(\theta_i)$ can be obtained. Solving problem (16) with different values of $\theta_i \in (0, \theta_{\max}]$, we can numerically obtain the relation between $P_{tot}(\theta_i)$ and θ_i as $P_{tot}^*(\theta_i)$, $\forall \theta_i \in (0, \theta_{\max}]$[5]. Then, the minimum average power consumption with given QoS exponents can be expressed as

$$\mathbb{E}[P_{tot}^*(\theta_i)] = \sum_{i=1}^{N_q} P_{tot}^*(\theta_i) \Pr\{(i-1)l < Q_\infty \leq il\}, \tag{18}$$

from which we can further find the optimal values of θ_1,..., θ_{N_q}.

B. Optimizing Multi-state QoS Exponents

To find the optimal QoS exponents θ_1^*,..., $\theta_{N_q}^*$ from subproblem II, i.e., to minimize $\mathbb{E}[P_{tot}^*(\theta_i)]$ in (18) under constraints (15a) and $\theta_i \leq \theta_{\max}$, we need the probabilities that Q_∞ stays in the N_q states. However, the probability $\Pr\{(i-1)l < Q_\infty \leq il\}$, $i = 1,..., N_q$, is very hard to obtain. To cope with this difficulty, we turn to minimize an upper bound of $\mathbb{E}[P_{tot}^*(\theta_i)]$.

The minimal average total power consumption in (18) can be expressed as follows,

$$\begin{aligned} &\mathbb{E}[P_{tot}^*(\theta_i)] \\ &= \sum_{i=1}^{N_q} P_{tot}^*(\theta_i) \{\Pr\{Q_\infty \geq (i-1)l\} - \Pr\{Q_\infty \geq il\}\} \end{aligned} \tag{19}$$

$$\begin{aligned} &= P_{tot}^*(\theta_1) \Pr\{Q_\infty \geq 0\} \\ &\quad + \sum_{i=1}^{N_q-1} [P_{tot}^*(\theta_{i+1}) - P_{tot}^*(\theta_i)] \Pr\{Q_\infty \geq il\} \\ &\quad - P_{tot}^*(\theta_{N_q}) \Pr\{Q_\infty \geq Q_{\max}\}, \end{aligned}$$

where $\Pr\{Q_\infty > il\}$ is the complementary cumulative distributed function (CCDF) of Q_∞ at the boundaries of the N_q states.

Since the queue length violation probability decays exponentially as the queue length increases, $\Pr\{Q_\infty > Q_{\max}\} \ll \Pr\{Q_\infty \geq il\}$, $i = 1,..., N_q - 1$. Hence, we have

$$\begin{aligned} \mathbb{E}[P_{tot}^*(\theta_i)] \approx\ & P_{tot}^*(\theta_1) \Pr\{Q_\infty \geq 0\} + \\ & \sum_{i=1}^{N_q-1} [P_{tot}^*(\theta_{i+1}) - P_{tot}^*(\theta_i)] \Pr\{Q_\infty \geq il\}. \end{aligned} \tag{20}$$

[5]Since effective bandwidth increases with QoS exponent θ and effective capacity decreases with θ [7, 8], $EC(\theta, \phi_{\max}) - EB(\theta)$ decreases with θ. This suggests that we can use linear searching method to find the value of $\theta_{\max}$ [24].

As discussed in existing works such as [5, 11], the transmit power (and thus the total power consumption) increases with the QoS exponent in order to satisfy the statistical QoS requirement. Moreover, as discussed in Remark 1, $\theta_{i+1} \geq \theta_i$. Then, $P^*_{tot}(\theta_{i+1}) - P^*_{tot}(\theta_i) > 0$. Because each weighting coefficient of $\Pr\{Q_\infty \geq il\}$, $i = 0, 1, ..., N_q - 1$ in (20) is positive, the upper bound of $\Pr\{Q_\infty > il\}$ can be applied in deriving the upper bound of $\mathbb{E}[P^*_{tot}(\theta_i)]$.

Similar to deriving (11), we can derive the upper bound as follows,

$$\Pr\{Q_\infty > il\} \leq \exp\left(-l\sum_{n=1}^{i}\theta_n\right), i = 1, ..., N_q. \quad (21)$$

where $l \gg A_\Delta(t)$.

By substituting (21) into (19), the upper bound of $\mathbb{E}[P^*_{tot}(\theta_i)]$ can be expressed as follows,

$$\mathbb{E}[P^*_{tot}(\theta_i)] \leq \sum_{i=1}^{N_q} P^*_{tot}(\theta_i)\left[\exp\left(-l\sum_{n=1}^{i-1}\theta_n\right) - \exp\left(-l\sum_{n=1}^{i}\theta_n\right)\right]. \quad (22)$$

To optimize the multi-state QoS exponents, we introduce an auxiliary function, $\theta(q) \geq 0, q \in (0, Q_{\max}]$, with which the QoS exponents can be obtained as

$$\theta_i = \frac{1}{l}\int_{(i-1)l}^{il}\theta(q)\,dq, i = 1, ..., N_q. \quad (23)$$

Then, $l\sum_{i=1}^{N_q}\theta_i = \int_0^{Q_{\max}}\theta(q)\,dq$. To guarantee that θ_i increases with queue length, we find a $\theta(q)$ that increases with q. Substituting (23) into (22), we have

$$\mathbb{E}[P^*_{tot}(\theta_i)] \leq \sum_{i=1}^{N_q} P^*_{tot}(\theta_i)\times \left\{\exp\left[-\int_0^{(i-1)l}\theta(q)\,dq\right] - \exp\left[-\int_0^{il}\theta(q)\,dq\right]\right\}. \quad (24)$$

When $l \ll Q_{\max}$, the sum in (24) can be accurately approximated by an integration, and thus the upper bound of $\mathbb{E}[P^*_{tot}(\theta_i)]$ can be re-expressed as follows,

$$\mathbb{E}[P^*_{tot}(\theta_i)] \leq \int_0^{Q^{\max}} P^*_{tot}[\theta(q)]\,d\exp\left[-\int_0^q\theta(x)\right] = \int_0^{Q^{\max}}\theta(q)P^*_{tot}[\theta(q)]\,e^{\left[-\int_0^q\theta(x)dx\right]}dq, \quad (25)$$

where $P^*_{tot}[\theta(q)]$ can be obtained from solving a new problem, which is obtained from problem (16) by replacing $\theta(q)$ with θ_i.

Upon substituting (23) into (15a), the constraints becomes,

$$\frac{1}{Q_{\max}}\int_0^{Q_{\max}}\theta(q)\,dq = \bar{\theta}. \quad (26)$$

Then, subproblem II can be transformed into another problem of optimizing $\theta(q) \in (0, \theta_{\max}]$ to minimize the upper bound of the average power consumption in (25) under constraints (26). From the solution of this problem, we can obtain the multi-state QoS exponents from (23).

For notational simplicity in the following derivations, define $y(q) \triangleq \int_0^q\theta(x)dx, q \in (0, Q_{\max}]$, and then its first-order derivative $\dot{y}(q) = \theta(q), \forall q \in (0, Q_{\max}]$. The problem to find the optimal auxiliary function can be expressed as

$$\min_{y(q)} \int_0^{Q_{\max}}\dot{y}(q)P^*_{tot}[\dot{y}(q)]\exp[-y(q)]dq \quad (27)$$

$$\text{s.t.} \quad y(0) = 0, y(Q_{\max}) = \bar{\theta}Q_{\max}, \quad (27a)$$

$$\dot{y}(q) \leq \theta_{\max}, \forall q \in (0, Q_{\max}], \quad (27b)$$

where constraint $(27a)$ comes from the definition of $y(q)$ and by substituting $y(q)$ into (26).

Problem (27) is a functional extremum problem with two boundary values and one inequality constraint. To solve the problem, we introduce a function $\alpha(q)$ to turn the inequality constraint (27b) into the following equality constraint

$$g[y(q), \dot{y}(q), q] \triangleq \dot{y}(q) - \theta_{\max} + \alpha^2(q) = 0. \quad (28)$$

Then, the solution of problem (27) can be found by solving Euler-Lagrange equations, the necessary condition that the global optimal solution should satisfy [25]. Denote the *Lagrangian* as $L[y(q), \dot{y}(q), q] \triangleq \dot{y}(q)P^*_{tot}[\dot{y}(q)]\exp[-y(q)]$. Then, the *augmented Lagrangian* can be expressed as $L'[y(q), \dot{y}(q), q] \triangleq L[y(q), \dot{y}(q), q] + v(q)g[y(q), \dot{y}(q), q]$ [26], where $v(q)$ is the *Lagrange multiplier function*, and the necessary conditions are

$$\frac{\partial L'[y(q), \dot{y}(q), q]}{\partial\alpha(q)} - \frac{d}{dq}\frac{\partial L'[y(q), \dot{y}(q), q]}{\partial\dot{\alpha}(q)} = 0,$$
$$\frac{\partial L'[y(q), \dot{y}(q), q]}{\partial y(q)} - \frac{d}{dq}\frac{\partial L'[y(q), \dot{y}(q), q]}{\partial\dot{y}(q)} = 0,$$

(27a) and (28),

where the first two equations are the Euler-Lagrange equations, which can be derived as

$$\alpha(q)v(q) = 0, \quad (29)$$

$$\frac{\partial L[y(q), \dot{y}(q), q]}{\partial y(q)} - \frac{d}{dq}\frac{\partial L[y(q), \dot{y}(q), q]}{\partial\dot{y}(q)} - \frac{dv(q)}{dq} = 0. \quad (30)$$

If $P^*_{tot}[\dot{y}(q)]$ is two-order differentiable with respect to $\dot{y}(q)$, by substituting $L[y(q), \dot{y}(q), q]$ into (30), the necessary conditions can be re-expressed as

$$\left\{2\frac{\partial P^*_{tot}[\dot{y}(q)]}{\partial\dot{y}(q)} + \dot{y}(q)\frac{\partial^2 P^*_{tot}[\dot{y}(q)]}{\partial[\dot{y}(q)]^2}\right\}\ddot{y}(q) - \frac{dv(q)}{dq} = \dot{y}^2(q)\frac{\partial P^*_{tot}[\dot{y}(q)]}{\partial\dot{y}(q)}, \quad (31)$$

(27a), (28) and (29).

Depending on whether the constraint (27b) (i.e., (28)) is active or not, we find the solution of the auxiliary function in three cases as follows.

Case 1: Constraint (27b) *is always inactive (i.e.,* $\theta(q) <$

$\theta_{\max}, \forall q \in (0, Q_{\max}]$*)*

In this case, $\dot{y}(q) < \theta_{\max}, \forall q \in (0, Q_{\max}]$, from (28) we have $\alpha(q) \neq 0$. Then, from (29) we know that $v(q) = 0$ and $\frac{dv(q)}{dq} = 0$, and (31) becomes a second order *ordinary differential equation* (ODE), where $(27a)$ is the boundary conditions of the ODE. The solutions of this ODE, $y(q)$, can be numerically obtained by, e.g., *shooting method* [27], from which we select the one that yields minimal average power consumption, which is the global optimal solution of problem (27). Then, $\theta^*(q) = \dot{y}(q)$ is the optimal auxiliary function.

To guarantee the statistical QoS, the auxiliary function should be an increasing function of q, as we discussed in Remark 1 from intuition. In the sequel, we confirm that $\theta^*(q)$ indeed increases with q. To do this, we prove that $\ddot{y}(q) = \frac{d\theta^*(q)}{dq}$ is positive.

Specifically, from (31) we have $\ddot{y}(q) = \left\{\dot{y}^2(q)\frac{\partial P^*_{tot}[\dot{y}(q)]}{\partial \dot{y}(q)}\right\} \Big/ \left\{2\frac{\partial P^*_{tot}[\dot{y}(q)]}{\partial \dot{y}(q)} + \dot{y}(q)\frac{\partial^2 P^*_{tot}[\dot{y}(q)]}{\partial[\dot{y}(q)]^2}\right\}$. In [11] (see page 1173 and Fig. 5), it was shown that $\frac{\partial P^*_{tot}[\dot{y}(q)]}{\partial \dot{y}(q)}$ and $\frac{\partial^2 P^*_{tot}[\dot{y}(q)]}{\partial[\dot{y}(q)]^2}$ are positive for a system with constant arrival rate. When the data arrives with random rate, the effective bandwidth increases with $\dot{y}(q)$, and hence the required total power increases faster than that serving with constant arrival rate, i.e., $\frac{\partial P^*_{tot}[\dot{y}(q)]}{\partial \dot{y}(q)} > 0$ and $\frac{\partial^2 P^*_{tot}[\dot{y}(q)]}{\partial[\dot{y}(q)]^2} > 0$. Therefore, $\ddot{y}(q) > 0, \forall q \in (0, Q_{\max}]$.

Case 2: Constraint (27b) *is always active (i.e.,* $\theta(q) = \theta_{\max}, \forall q \in (0, Q_{\max}]$*)*

In this case, we obtain a unique solution $\theta^*(q) = \theta_{\max}, \forall q \in (0, Q_{\max}]$. In other words, the optimal multi-state QoS exponents satisfy $\theta^*_i = \theta_{\max}$, $i = 1,..., N_q$.

Case 3: Constraint (27b) *is neither always active nor always inactive.*

As discussed in previous cases, when (27b) is inactive, $\dot{y}(q) < \theta_{\max}$ and increases with q, when (27b) is active, $\dot{y}(q) = \theta_{\max}$. This suggests that $\dot{y}(q)$ is a piecewise function in case 3. Further considering that $\dot{y}(q)$ should be an increasing function of q, its optimal value can be expressed as follows,[6]

$$\theta^*(q) = \begin{cases} \text{the same form as in case 1}, q \leq q_{th}, \\ \theta_{\max}, q > q_{th}, \end{cases}$$

where constraint (27b) is inactive when $q \leq q_{th}$ and active when $q > q_{th}$, and q_{th} can be found from problem (27) in $(0, Q_{\max}]$.

In general, the global optimal solution of the functional extreme problem is very hard to find, as analyzed in [28]. Fortunately, for some cases where $P^*_{tot}[\theta(q)] = c_1\theta^{c_2}(q) + c_3$ as addressed in [29], we can show that the local optimal solution of $\theta(q)$ is unique and hence is global optimal. Given the optimized auxiliary function $\theta^*(q), q \in (0, Q_{\max}]$, the multi-state QoS exponents can be obtained from (23), $\theta^*_i = \frac{1}{l}\int_{(i-1)l}^{il} \theta^*(q)\, dq$, $i = 1,..., N_q$. The corresponding solution of problem (16), $\{N^*_S(\theta^*_i), P^*_T(\theta^*_i), i = 1,..., N_q\}$, is the finally optimized QRA policy.

The following proposition proved in Appendix B addresses the global optimality of the original QRA optimization problem.

Proposition 1. *If* $\{N^*_S(\theta_i)\ P^*_T(\theta_i), i = 1,..., N_q\}$ *is the global optimal solution of subproblem I, and* $\{\theta^*_i, i = 1,..., N_q\}$ *is the global optimal solution of subproblem II, then the optimized QRA policy is the global optimal solution of problem* (15)*.*

Remark 2: Since an upper bound of average power consumption is minimized and approximations are introduced, the obtained θ^*_i is not the global optimal solution of subproblem II. Hence, the proposed QRA policy is not the global optimal solution of the original problem in general. In the next section, we will show that the QRA policy obtained from sequentially solving the two subproblems is the global optimal solution of problem (15) when $D_{\max} \to \infty$.

C. Closed-Form QRA Policy for a Special Case

To gain useful insight, we provide a closed-form QRA policy. To this end, we take a massive MIMO-OFDM system serving a first order autoregressive (i.e. AR(1)) source as an example, where the number of subcarriers and transmit power constraints in (15c) are not considered.

As discussed in [30], AR(1) process can be used to model several kinds of video sources. For an AR(1) source, the amount of data in each frame of duration T_a satisfies [31]

$$\int_{nT_a}^{(n+1)T_a} a(\tau)\, d\tau = \rho_A \int_{(n-1)T_a}^{nT_a} a(\tau)\, d\tau + \beta w(n),$$

where $\rho_A \in [0, 1]$ and $\beta > 0$ are two constants, and $w(n)$ is Gaussian white noise of mean m_w and variance 1. Assume that the arrival rate $a(t)$ is constant within each frame but changes according to the AR(1) model. From [32], the effective bandwidth of a continuous AR(1) source can be obtained as

$$E_B(\theta_i) = \frac{1}{T_a}\left(m_a + \frac{r_a^2}{2}\theta_i\right) \quad \text{(bits/s)}, \tag{32}$$

where $r_a = \frac{\beta}{1-\rho_A}$ and $m_a = r_a m_w$.

When $N_T \to \infty$, the maximal achievable rate of massive MIMO-OFDM system without CSI at the BS can be expressed as [33]

$$s^M_i = BN_S(\theta_i)\log_2\left[1 + \frac{\mu P_T(\theta_i)}{N_S(\theta_i)\sigma_0^2}\right] \quad \text{(bits/s)}, \tag{33}$$

which is deterministic, and hence the effective capacity is $E_C(\theta_i, \phi_i) = s^M_i$.

As proved in Appendix C, the solution of problem (16) is optimal when the constraint in (16a) holds with equality (i.e., $E_C(\theta_i, \phi_i) = E_B(\theta_i)$). Considering the effective bandwidth of AR(1) source in (32) and the effective capacity of massive MIMO-OFDM system in (33), the Karush-Kuhn-Tucker (KKT) conditions of problem (16) without constraints on $N_S^{\max}$

[6]When $q \leq q_{th}$, $\dot{y}(q)$ can be obtained from solving (31) with $\frac{dv(q)}{dq} = 0$ and boundary conditions, which are $y(0) = 0, y(q_{th}) = y(Q_{\max}) - \theta_{\max}(Q_{\max} - q_{th})$.

and $P_T^{\max}$ can be derived as,

$$P_{cs} - \frac{v_0 B}{\ln 2}\left\{\ln\left[1+\frac{\mu P_T(\theta_i)}{N_S(\theta_i)\sigma_0^2}\right] - \frac{\frac{\mu P_T(\theta_i)}{N_S(\theta_i)\sigma_0^2}}{1+\frac{\mu P_T(\theta_i)}{N_S(\theta_i)\sigma_0^2}}\right\} = 0, \tag{34a}$$

$$\frac{1}{\rho} - \frac{v_0 B}{\ln 2}\frac{\frac{\mu}{\sigma_0^2}}{1+\frac{\mu P_T(\theta_i)}{N_S(\theta_i)\sigma_0^2}} = 0, \tag{34b}$$

$$E_B(\theta_i) - BN_S(\theta_i)\log_2\left[1+\frac{\mu P_T(\theta_i)}{N_S(\theta_i)\sigma_0^2}\right] = 0, \tag{34c}$$

where v_0 is the Lagrange multiplier for constraint (16a), and (34c) comes from $E_C(\theta_i,\phi_i) = E_B(\theta_i)$.

Finding v_0 from (34b) and then substituting to (34a), we have

$$\left\{1+\frac{\mu P_T(\theta_i)}{N_S(\theta_i)\sigma_0^2}\right\}\ln\left[1+\frac{\mu P_T(\theta_i)}{N_S(\theta_i)\sigma_0^2}\right] - \frac{\mu P_T(\theta_i)}{N_S(\theta_i)\sigma_0^2} = \frac{\mu P_{cs}\rho}{\sigma_0^2}. \tag{35}$$

The left hand side of (35) only depends on the transmit power on each subcarrier, which is denoted as $P_{TS_i} \triangleq \frac{P_T(\theta_i)}{N_S(\theta_i)}$. To find P_{TS_i} from (35), define $f_q(P_{TS_i}) \triangleq \left(1+\frac{\mu}{\sigma_0^2}P_{TS_i}\right)\ln\left(1+\frac{\mu}{\sigma_0^2}P_{TS_i}\right) - \frac{\mu}{\sigma_0^2}P_{TS_i}$. It is easy to derive that $f_q(0) = 0$ (which is less than $\frac{\mu P_{cs}\rho}{\sigma_0^2}$), $f_q(\infty) \to \infty$ (which is larger than $\frac{\mu P_{cs}\rho}{\sigma_0^2}$), and $f_q'(P_{TS}) > 0$. Thus, there is a unique solution of P_{TS_i} that satisfies $f_q\left(P_{TS_i}^*\right) = \frac{\mu P_{cs}\rho}{\sigma_0^2}$, which is independent of θ_i and $E_B(\theta_i)$. Hence, we can ignore index i of $P_{TS_i}^*$ for simplicity. With P_{TS}^*, the achievable rate on each subcarrier $r_s \triangleq B\log_2\left(1+\frac{\mu}{\sigma_0^2}P_{TS}^*\right)$. After substituting (32) into (34c), and using $E_C(\theta_i,\phi_i) = N_S^*(\theta_i)\, r_s$ and $P_T^*(\theta_i) = P_{TS}^* N_S^*(\theta_i)$, we can derive the optimal resource allocation policy with any given QoS exponents as follows,

$$N_S^*(\theta_i) = \frac{1}{T_a r_s}\left(m_a + \frac{r_a^2}{2}\theta_i\right), P_T^*(\theta_i) = \frac{P_{TS}^*}{T_a r_s}\left(m_a + \frac{r_a^2}{2}\theta_i\right) \tag{36}$$

Then, from (13) we obtain the total power consumption of the BS as

$$P_{tot}^*(\theta_i) = \left(\frac{P_{TS}^*}{T_a r_s \rho} + \frac{P_{cs}}{T_a r_s}\right)\left(m_a + \frac{r_a^2}{2}\theta_i\right) + P_c, \tag{37}$$

which is a linear function of θ_i.

Upon substituting (37) into (31), the ODE[7] is simplified to $\ddot{y}(q) = \frac{1}{2}\dot{y}^2(q)$, where θ_i in (37) is replaced with $\theta(q) = \dot{y}(q)$. By solving this ODE under the boundary conditions in (27a), we can obtain

$$\theta^*(q) = \frac{2}{\left[\frac{Q_{\max}}{1-\exp\left(-\frac{\bar{\theta}Q_{\max}}{2}\right)} - q\right]}, q \in (0, Q_{\max}], \tag{38}$$

which is the unique solution of (31) and hence is the global optimal solution of problem (27). Substituting (38) into (23), we have

$$\theta_i^* = \frac{2N_q}{Q_{\max}}\ln\left[\frac{N_q - (i-1)\left(1-\sqrt{\varepsilon_Q}\right)}{N_q - i\left(1-\sqrt{\varepsilon_Q}\right)}\right], i = 1, ..., N_q, \tag{39}$$

where $\bar{\theta} = \frac{\ln(1/\varepsilon_Q)}{Q_{\max}}$ and $l = \frac{Q_{\max}}{N_q}$ have been applied. Substituting (39) into (36), the closed-form QRA policy, $\{N_S^*(\theta_i^*), P_T^*(\theta_i^*), i = 1,..., N_q\}$, is obtained.

Remark 3: Even for the special case of $N_T \to \infty$ where the service process is not random due to channel hardening, the optimized resource allocation policy still depends on the queue length. This confirms that CRA policies are not optimal for random arrival process [4].

IV. Discussions

To show the impact of the approximations and bounds introduced in the design on the final performance of the proposed solutions, in this section we first provide the *EE limit* of the system by finding the maximal EE for $D_{\max} \to \infty$. Then, we compare the EE achieved by the massive MIMO-OFDM system with the proposed closed-form QRA policy for the traffic with arbitrary values of $D_{\max}$ with the *EE limit*, which implicitly shows that the policy is the global optimal solution of the original problem (15) in the special case. Finally, we explain why the QRA policy achieves higher EE than the CRA policies.

A. EE Limit: The Maximal EE Achieved by Traffic With $D_{\max} \to \infty$

When $D_{\max} \to \infty$, $\theta_i \equiv 0$, $i = 1,..., N_q$, the constraint in (15b) degenerates into $\mathbb{E}_{\mathbf{h}}[s(t)] \geq \mathbb{E}[a(t)]$. With the relaxed QoS constraint, the achieved EE of the system approaches a limit. In other words, the EE achieved by a system serving traffic with infinite delay bound is the ultimate EE upper bound of the system serving traffic with finite delay bounds. For the delay tolerant traffic with $D_{\max} \to \infty$, the resource allocation is independent of the queue length, and therefore the corresponding policy is a CRA policy. Further recalling that the average of the total power consumption in (15) is taken over the queue length, the *EE limit* can be achieved by the resource allocation policy $\phi^{\lim}$ found from the following problem,

$$\begin{aligned}\min_{\phi^{\lim}=\{N_S^{\lim},P_T^{\lim}\}} \quad & P_{tot}^{\lim} = P_T^{\lim}/\rho + P_{cs}N_S^{\lim} + P_c, \\ \text{s.t} \quad & \mathbb{E}_{\mathbf{h}}[s(t)] \geq \mathbb{E}[a(t)], \\ & 0 \leq N_S^{\lim} \leq N_S^{\max}, 0 \leq P_T^{\lim} \leq P_T^{\max},\end{aligned} \tag{40}$$

which is the degenerated form of problem (15) and also problem (16). According to Appendix A, the problem is convex, which can be solved by using the interior-point method.

Then, the EE limit can be obtained by substituting $\mathbb{E}[a(t)]$ and the minimized value of $P_{tot}^{\lim}$ into (14). Since the power consumption and resource allocation do not depend on QSI, the EE limit is accurate, which is obtained without using

[7]Since constraints in (15c) is not considered, $v(q) = 0$. Thus, $\frac{dv(q)}{dq} = 0$, and (31) is an ODE.

the upper bound of the CCDF of Q_∞ in (21) and the approximations in (6) and (7).

A *necessary condition for a transmit policy to achieve the EE limit* is

$$\mathbb{E}_{\mathbf{h},Q_\infty}[s(t)] = \mathbb{E}[a(t)]. \tag{41}$$

This is because if $\mathbb{E}_{\mathbf{h},Q_\infty}[s(t)] - \mathbb{E}[a(t)] > 0$, the system will serve empty buffer in some TTIs, the EE can always be improved if the system provides no service when $Q(t) = 0$.

B. EE Achieved by the Closed-form QRA Policy in the Special Case

For the massive MIMO-OFDM system with $N_T \to \infty$ serving the AR(1) source with arbitrary delay bound $D_{\max}$, when the number of subcarriers and transmit power are jointly optimized, both the transmit power P_{TS}^* and service rate r_s on each subcarrier are constant, as discussed in section III.C. Therefore, the optimal total power consumption in (37) can be re-written as

$$P_{tot}^*(\theta_i^*) = \left(\frac{P_{TS}^*}{\rho r_s} + \frac{P_{cs}}{r_s}\right) s_i^M + P_c, \tag{42}$$

where $s_i^M = E_B(\theta_i^*)$ when $(i-1)l < Q_\infty \le il$, $i = 1,..., N_q$. We can see that the total power consumption linearly increases with the service rate s_i^M.

On the one hand, as indicated in Appendix C, the optimal solution of problem (40), which is degenerated from problem (16), is obtained when $\mathbb{E}_{\mathbf{h}}[s(t)] = \mathbb{E}[a(t)]$. In massive MIMO-OFDM system, the EE limit with $D_{\max} \to \infty$ can be achieved with $s_i^M = \mathbb{E}[a(t)]$, $i = 1,..., N_q$. With the corresponding policy $\phi^{\lim} = \{N_S^{\lim}, P_T^{\lim}\}$, the EE limit of this system is

$$EE^{\lim} = \frac{\mathbb{E}[a(t)]}{P_{tot}^{\lim}} = \left\{\left(\frac{P_{TS}^*}{\rho r_s} + \frac{P_{cs}}{r_s}\right) + \frac{P_c}{\mathbb{E}[a(t)]}\right\}^{-1}. \tag{43}$$

On the other hand, when the QRA policy in section III.C is used to serve the traffic with any finite values of $D_{\max}$, $\mathbb{E}_{Q_\infty}\left(s_i^M\right) = \mathbb{E}[a(t)]$. By taking the average over Q_∞, from (42) we have

$$\mathbb{E}_{Q_\infty}[P_{tot}^*(\theta_i^*)] = \left(\frac{P_{TS}^*}{\rho r_s} + \frac{P_{cs}}{r_s}\right)\mathbb{E}_{Q_\infty}\left(s_i^M\right) + P_c. \tag{44}$$

By substituting (44) into $EE = \frac{\mathbb{E}[a(t)]}{\mathbb{E}[P_{tot}^*(\theta_i^*)]}$, we see that the achieved EE is the same as (43).

This implies that the closed-form QRA policy is the global optimal solution of problem (15). Although this is not true for general cases, the EE achieved by the QRA policy for the traffic with finite $D_{\max}$ is close to the EE limit, as will be shown via simulations in next section.

C. Comparison Between QRA and CRA Policies

Existing energy efficient resource allocation methods using effective capacity as a design tool are essentially CRA policies [5, 13–16], denoted as $\phi^c = \{N_S^c, P_T^c\}$. The proposed QRA policy differs from these relevant methods both in the objective function and in the QoS constraint.

For MIMO-OFDM system, the objective function in [13–16] can be expressed as

$$\begin{aligned} EE^c &\triangleq \frac{E_C(\theta^c, \phi^c)}{P_{tot}^c} \\ &= \frac{-\frac{N_S^c}{T_c\theta^c}\ln \mathbb{E}_{\mathbf{h}}\left[\left(1 + \frac{P_T^c}{\sigma_0^2 N_T N_S^c}\mathbf{h}\mathbf{h}^H\right)^{-\frac{T_c B\theta^c}{\ln 2}}\right]}{P_T^c/\rho + P_{cc}}, \end{aligned} \tag{45}$$

where P_{cc} is the total circuit power consumed by the BS, which is modeled as a constant in existing literatures [13–16].

In [13–15], no QoS constraint is imposed. In [16], the statistical QoS requirement is ensured by using the following constraint

$$E_C(\theta^c, \phi^c) \ge E_B(\theta^c). \tag{46}$$

Recalling that (46) is the *necessary and sufficient condition* for satisfying the QoS requirement [11], the policies in [13–15] do not always guarantee the QoS. In the sequel, we refer to the bandwidth and power allocation policy that maximizes (45) under constraint (46) as *CRA1* policy, which is a single user version of that proposed in [16]. We refer to the power allocation policy that minimizes the transmit power under constraint (46) as *CRA2* policy, which is a modified version of that proposed in [5] with a single user and only channel distribution information at the BS.

Example: To understand the inherent difference between the CRA policies intuitively, we assume constant arrival rate as in [5,13–16] (i.e., $a(t) = a$), and consider a simple scenario where the transmit power is allocated and the number of active subcarriers is fixed.

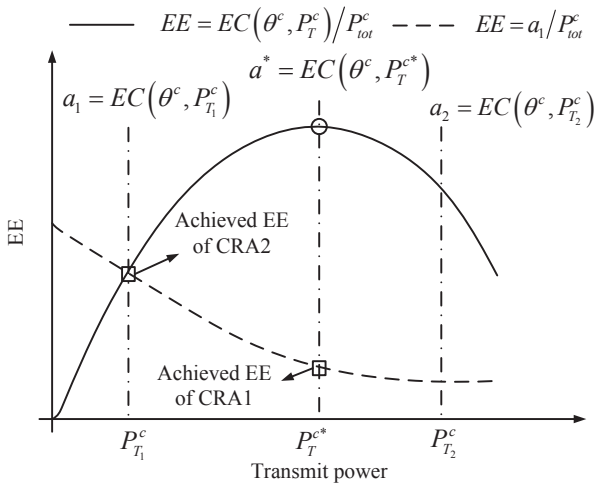


Fig. 2. Illustration of the difference between the CRA policies.

The EEs respectively defined in (45) and (14) versus transmit power are shown in Fig. 2. As proved in [16], the EE^c defined in (45) first increases and then decreases with P_T^c. Without constraint (46), the EE^c-maximal transmit power is P_T^{c*}. When constraint (46) is imposed, the solution of the CRA1 policy can be obtained as follows depending on the traffic load. If the arrival rate $a < E_C(\theta^c, P_T^{c*})$,

the optimal transmit power is still P_T^{c*}. If the arrival rate $a \geq E_C(\theta^c, P_T^{c*})$, e.g., $a = a_2$, and the minimum transmit power that can guarantee the QoS requirement is $P_{T_2}^c$, then the corresponding maximal value of EE^c satisfying the QoS will be less than $E_C(\theta^c, P_T^{c*})/P_{tot}$. On the other hand, the EE defined in (14) is a decreasing function of P_T^c, and thus the optimal transmit power of the CRA2 policy is always achieved when $a = E_C(\theta^c, P_T^c)$. Comparing the two CRA policies, it is easy to see that they are the same if $a \geq E_C(\theta^c, P_T^{c*})$. To differentiate the two policies, we consider the scenario where $a = a_1 < E_C(\theta^c, P_T^{c*})$. For a given source, the arrival rate is fixed, and the average departure rate $\mathbb{E}[b(t)]$ will not exceed the arrival rate a_1. Further considering that when (6) is satisfied, $\mathbb{E}[a(t)] = \mathbb{E}[b(t)]$ [21]. As a consequence, the really achieved EEs of the CRA1 and CRA2 policies are respectively $\frac{a_1}{P_T^{c*}/\rho + P_{cc}}$ and $\frac{a_1}{P_{T_1}^c/\rho + P_{cc}}$, where the EE achieved by the CRA2 policy is higher than that achieved by the CRA1 policy (see the marks on the dash line). This is because by using the objective function EE^c, the CRA1 policy implicitly assumes that there is always enough data to be transmitted. When the arrival rate is low in the system, such a *full buffer assumption* causes the EE loss.

Now we come back to the general cases under random sources and random channels.

To ensure the statistical QoS requirement, the CRA policies should satisfy $E_C(\theta^c, \phi^c) \geq E_B(\theta^c)$. In fading channels, $\mathbb{E}_{\mathbf{h}}[s(t)] > E_C(\theta^c, \phi^c)$ [8]. Similarly, for random sources, $E_B(\theta^c) > \mathbb{E}[a(t)]$ [7]. Therefore, $\mathbb{E}_{\mathbf{h}}[s(t)] > \mathbb{E}[a(t)]$ is true when CRA policies are applied to serve delay sensitive traffic with random sources *or* random channels. For example, when a AR(1) source is served by the massive MIMO-OFDM system using the CRA policies, we have

$$\begin{aligned} &\mathbb{E}_{\mathbf{h}}[s(t)] - \mathbb{E}[a(t)] = s_c^M - \mathbb{E}[a(t)] \\ &\geq E_B(\theta^c) - \mathbb{E}[a(t)] = \frac{r_a^2}{2T_a}\theta^c > 0, \end{aligned} \tag{47}$$

where s_c^M is the service rate of massive MIMO-OFDM system with CRA policies, and (32) and $\mathbb{E}[a(t)] = \frac{m_a}{T_a}$ are applied. (47) implies that the CRA policies will serve empty buffer for delay sensitive traffic with random arrival rate and deterministic service rate. Similar conclusion can be obtained when the arrival rate is deterministic and the service rate is random. Recalling the necessary condition for a transmit policy to achieve the EE limit discussed in section IV.A, the CRA policies serving traffic with finite delay bound cannot achieve the EE limit.

By contrast, $\mathbb{E}_{\mathbf{h},Q_\infty}[s(t)] = \mathbb{E}[a(t)]$ for the QRA policy since it provides no service when $Q(t) = 0$. Therefore, the QRA policy is possible to achieve the EE limit owing to satisfying the necessary condition in (41).

Remark 4: Both the QRA and CRA policies we addressed only assume channel distribution information known at the BS. If CSI is available at the BS, the optimization variables, objective function and QoS constraint to derive a QRA policy will change. Specifically, the power allocation among subcarriers and antennas should also be optimized. The expectation of the objective function in problem (15) should also be taken over CSI, and $E_C(\theta_i, \phi_i)$ in (15a) should be the effective capacity with known CSI, which has no closed-form expression as shown in [10].[8]

Remark 5: The study in [4] indicates that under average delay requirement, when *either* the source *or* the channel is random, the minimal average transmit power of a QRA policy is lower than that of a CRA policy. For the single antenna narrow band system considered in [4], the randomness of channel leads to the randomness of service rate. However, for massive MIMO system, although the channel is random, the service rate is deterministic, then the gap between transmit power of a QRA policy and that of a CRA policy only comes from the randomness of the arrival rate under the average delay requirement. By taking the massive MIMO-OFDM system as an example, we can extend the above result in [4] to statistical QoS requirement. Specifically, as shown in (43) and (44), the EE achieved by the QRA policy equals to the EE limit. On the other hand, (47) indicates that with CRA policies, $\mathbb{E}_{\mathbf{h}}[s(t)] > \mathbb{E}[a(t)]$ if the source is random. Since the necessary condition to achieve the EE limit in (41) is not satisfied, the EE limit cannot be achieved by CRA policies. Therefore, the QRA policy can achieve higher EE than CRA policies in the massive MIMO-OFDM system for random sources.

V. Simulation and Numerical Results

In this section, we first show the relationship between the two QoS requirements $(D_{\max}, \varepsilon_D)$ and $(Q_{\max}, \varepsilon_Q)$. Then, we validate the approximations and bounds used in the derivations. Finally, we evaluate the EE achieved by MIMO-OFDM systems using the QRA policy.

In the simulation, the channel is constant within each duration T_c and subject to i.i.d. Nakagami-m distribution among different durations. $m = 1$ corresponds to Rayleigh fading. As m increases, the fluctuation of the channel decreases. At initial time, the buffer is empty. The resources are allocated at the beginning of each TTI [18]. The queue length used for the QRA policy in next TTI is computed in the end of each TTI. The micro BS parameters in [17] are applied. For other kinds of BS, the simulation results are similar. When the maximum bandwidth is used, the power consumed for baseband processinge is 13.6 W, which linearly increases with the number of subcarriers as indicated in [23]. Therefore, the circuit power for one subcarrier is $13.6/N_S^{\max}$ W. Simulation parameters in the sequel are listed in Table I, unless otherwise specified.

We compare the QRA policy with three kinds of CRA policies modified from [5,16], whose θ^c is obtained from (7). For the CRA1 policy, we solve P_T^c and N_S^c that maximizes (45) under constraint (46). For the CRA2 policy, we solve the optimization problem that minimize P_T^c under constraint (46).

[8]In massive MIMO systems, it is easy to extend the QRA policy into the scenario with CSI at the BS. Simulation results for comparing the QRA and CRA policies with CSI at BS are similar to those for the policies without CSI at BS, which are not provided for conciseness.

To solve the above problems, exhaustive searching is applied. The third policy is a directly extension of the CRA2 policy by not serving empty buffer (i.e., $P_T = 0$ when $Q(t) = 0$), which is referred to as the extended CRA2 policy. Noting that the policy in [5] was derived assuming CSI available at BS, the CRA2 policies are modified from [5] by assuming channel distribution information at the BS in order to compare all the policies fairly. The optimization results show that the optimal number of subcarriers for CRA1 policy is $N_S^{\max}$. Besides, to minimize P_T^c under constraint (46), CRA2 and the extended CRA2 will also use all the subcarriers. Therefore, when evaluate the achieved EE of the CRA policies, we set the circuit power $P_{cc} = P_{cs}N_S^{\max} + P_c$. The total power consumed in each TTI is computed with (13), and the EE is the ratio of the accumulated data transmitted to the overall energy consumed in all TTIs during the simulation time of 1000 s.

TABLE I
LIST OF SIMULATION PARAMETERS [17, 18, 34]

Delay bound $D_{\max}$ and violation rate ε_D	5 ms and 0.01
Average packet arrival rate λ and packet size u of Poisson source	200 $\sim$ 2000 packets/s and 2 $\sim$ 4 kbits
Frame duration of AR(1) source T_a	1 ms
Parameters of AR(1) source ρ_A and m_w	0.5 and 1
Channel coherence time T_c and transmit time interval ΔT	2 ms and 1 ms
Total number of subcarriers $N_S^{\max}$ and subcarrier separation B	128 and 15 kHz
Number of transmit antennas N_T	4
Maximum transmit power $P_T^{\max}$ and power amplifier efficiency ρ	8.0 W and 28.5 %
Fixed circuit power consumption P_c and that for one subcarrer P_{cs}	14.3 W and $13.6/N_S^{\max}$ W
Circuit power consumption for CRA policies with $N_S^c = N_S^{\max}$, P_{cc}	$14.3 + 13.6$ W
Average receive SNR when $P_T = P_T^{\max}$ and $N_S = N_S^{\max}$	10 dB

A. Relation of Two QoS Requirements

When the source is constant and the resource allocation is independent of queue length, it was shown in [13] that if $(Q_{\max}, \varepsilon_Q)$ is satisfied then $(D_{\max}, \varepsilon_D)$ will be ensured. When both source and channel are stochastic meanwhile the resource allocation policy depends on queue length, it is very hard to quantitatively characterize the relationship between these two QoS requirements.

To circumvent this difficulty, we consider the equivalent effective bandwidth $E_B(\bar{\theta})$ as in [7], which is the equivalent minimal constant service rate that can guarantee $(Q_{\max}, \varepsilon_Q)$. The relation between $(Q_{\max}, \varepsilon_Q)$ and $(D_{\max}, \varepsilon_D)$ with constant service rate is discussed in Appendix D. In general wireless systems $s(t)$ is random, which differs from the wired systems considered in [7]. Nonetheless, in modern high throughput systems such as MIMO-OFDM system the service rate is close to constant.[9] To see this, we use $I_s \triangleq \text{Var}[s(t)]/\{\mathbb{E}[s(t)]\}^2$ to reflect the dispersion of a random process $s(t)$ [22]. The channels are frequency-selective with different coherent bandwidths, which are modeled as "frequency domain block fading", where several adjacent subcarriers are with identical channel gain and compose an independent subchannel, and the channel gains among different subchannels are i.i.d.. Denote the mean and the variance of the data rate of each subchannel as m_s and σ_s^2, respectively. As the number of i.i.d. subchannels N_{iid} becomes large, according to the central limit theorem, the service rate of the system $s(t)$ can be approximated as a Gaussian random variable with mean $m_s N_{\text{iid}}$ and variance $\sigma_s^2 N_{\text{iid}}$, and then $I_s = \frac{\sigma_s^2}{m_s^2 N_{\text{iid}}}$ becomes small. Moreover, as the number of antennas increases, σ_s decreases due to the channel hardening. In Table II, we show the simulation results of I_s for different values of N_{iid} and N_T with average receive SNR, $\bar{\gamma} = \mathbb{E}_{\mathbf{h}}\left(\frac{\mu P_T^{\max}}{\sigma_0^2 N_T N_S^{\max}}\mathbf{h}\mathbf{h}^H\right) = \frac{\mu P_T^{\max}}{\sigma_0^2 N_S^{\max}} = 10$ dB. The results are obtained when all the $N_S^{\max}$ subcarriers are active. Thus, N_{iid} is determined by the coherent bandwidth (e.g., when $N_{\text{iid}} = 128$, the coherent bandwidth equals to the subcarrier separation B, and the channel gains on different subcarriers are i.i.d.). We can see that even when N_{iid} and N_T are not large, say $N_{\text{iid}} = 16$ and $N_T = 2$, the instantaneous service rate is nearly deterministic.[10]

As a result, the service rate of MIMO-OFDM system with policy ϕ_i that satisfies $E_B(\theta_i) = E_C(\theta_i, \phi_i)$, denoted as $s_0(\phi_i)$, is approximately equal to $E_B(\theta_i)$, which is a deterministic and increasing function of θ_i [7]. Since θ_i increases with queue lendth, $s_0(\phi_i)$ increases with queue length as well. Noting that $E_B(\bar{\theta})$ is equivalent to $E_B(\theta_i)$, i = 1,..., N_q in the sense that they guarantee the same $(Q_{\max}, \varepsilon_Q)$ [7] and further considering *Remark 1*, $(D_{\max}, \varepsilon_D)$ can be satisfied with $s_0(\phi_i)$, i = 1,..., N_q. In other words, if a QRA policy satisfies $(Q_{\max}, \varepsilon_Q)$, it will also ensure $(D_{\max}, \varepsilon_D)$.

TABLE II
I_s OF MIMO-OFDM SYSTEMS, $m = 1$ (RAYLEIGH FADING)

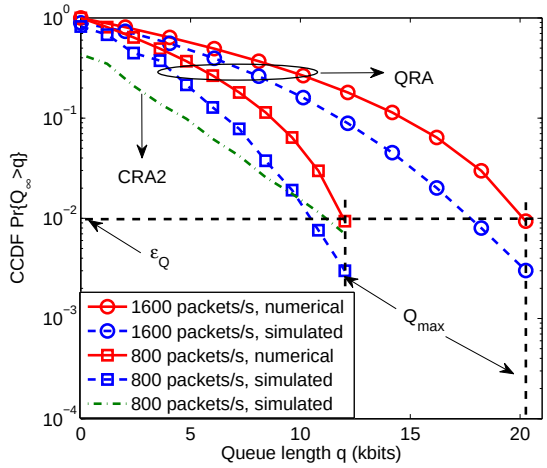
	$N_{\text{iid}} = 16$	$N_{\text{iid}} = 32$	$N_{\text{iid}} = 64$	$N_{\text{iid}} = 128$
$N_T = 2$	0.0686%	0.0343%	0.0172%	0.0086%
$N_T = 4$	0.0661%	0.0331%	0.0166%	0.0083%
$N_T = 8$	0.0640%	0.0321%	0.0160%	0.0080%

B. Validation of the Analysis

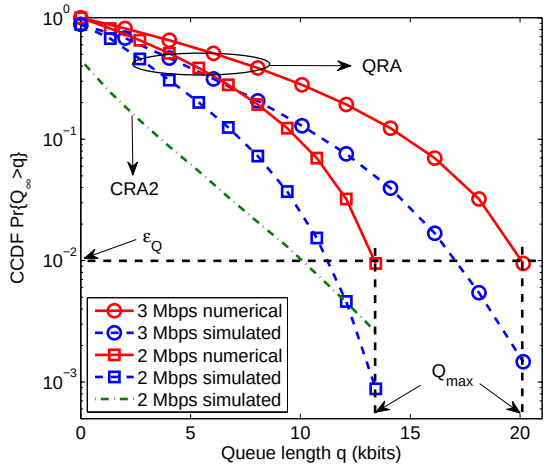
To show that the proposed policy is not restricted to AR(1) source, we also consider discrete Poisson source [35] with average packet arrival rate λ and packet size u, where the distribution of the accumulated data arrived in one TTI is $\Pr\left[\int_t^{t+\Delta T} a(\tau)\, d\tau = nu\right] = \frac{(\lambda\Delta T)^n}{n!}e^{-\lambda\Delta T}u$.

[9] It is worth to note that when considering multicell networks with inter-cell interference, such a conclusion is no longer true. With random service rate, an upper bound of delay violation probability can be characterized by the queue length violation probability as in [13].

[10] Recall that for massive MIMO system with $N_t \to \infty$ the instantaneous service rate *is* deterministic, as shown in (33). Further simulations show that the channel correlation among different antennas has little impact on the results, which are not provided for conciseness.



(a) Poisson sources with $u = 2$ kbits

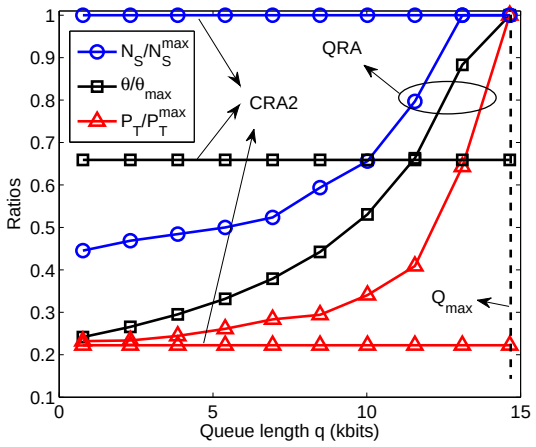


(b) AR(1) sources with different $\mathbb{E}[a(t)]$

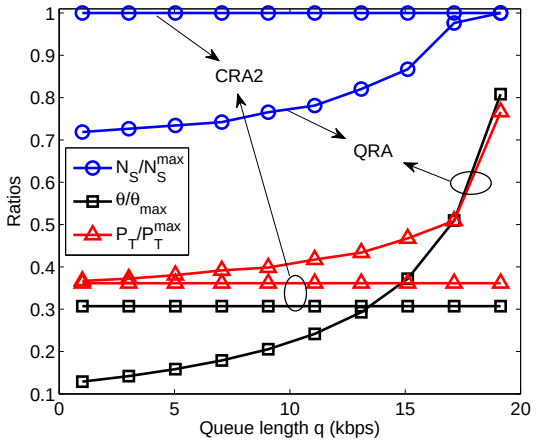
Fig. 3. CCDFs of the queue length, where $N_{\text{iid}} = 128$, $m = 1$ (Rayleigh fading).

Figure 3 shows the upper bound of the CCDFs of queue length for the MIMO-OFDM systems using the QRA or CRA2 policy to serve the discrete Poisson or AR(1) source, where the numerical results are obtained from (21). We can observe that the simulated CCDFs are lower than the numerical results for different sources with different values of λ or $\mathbb{E}[a(t)]$. Besides, when $q = Q_{\max}$, $\Pr\{Q_\infty > Q_{\max}\} \leq \varepsilon_Q$, which implies that (11) is valid and $(Q_{\max}, \varepsilon_Q)$ can be satisfied with the QRA policy. Furthermore, the queuing delay and violation probability achieved by the QRA policy for all the sources in the simulation are equal to (4.0 ms, 0.01). This indicates that with the QRA policy, the statistical QoS requirement $(D_{\max}, \varepsilon_D)$ can be guaranteed. It is shown that the CCDF with CRA2 policy is lower than that with QRA policy in a wide region. However, once the QoS requirement, $(Q_{\max}, \varepsilon_Q)$, is satisfied, a lower CCDF is not helpful for improving the experience of the user. On the other hand, the EE depends on the CCDF in all range of queue length. As a result, with the CRA2 policy, extra energy is used to provide a lower CCDF that is unnecessary. By controlling the CCDF of queue length, the QRA policy can save energy and hence improve the EE.

Figure 4 shows the transmit power and the number of active subcarriers for the systems using the QRA policy or the CRA2 policy to support discrete Poisson or AR(1) source. The corresponding multi-state QoS exponents of the QRA policy and the QoS exponent of the CRA2 policy are also shown. To plot the results in the same figure, the values are normalized by their corresponding maximum. The curves are not very smooth, because the values of $N_S(\theta_i)$ are integer. We can see that the multi-state QoS exponents as well as the amount of resources allocated by the QRA policy increase with the queue length. In contrast, the allocated resources and the QoS exponent of CRA2 are independent of queue length.



(a) Poisson source, where $u = 4$ kbits and $\lambda = 400$ packets/s



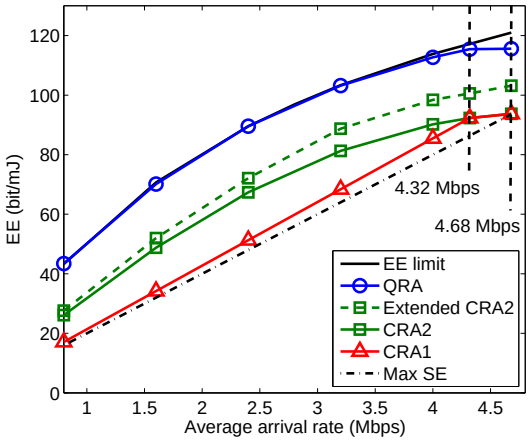
(b) AR(1) source, $r_a = m_a = 3$ kbits

Fig. 4. Allocated resources and normalized multi-state QoS exponents versus queue length, $N_{\text{iid}} = 128$, $m = 1$ and $N_q = 10$.

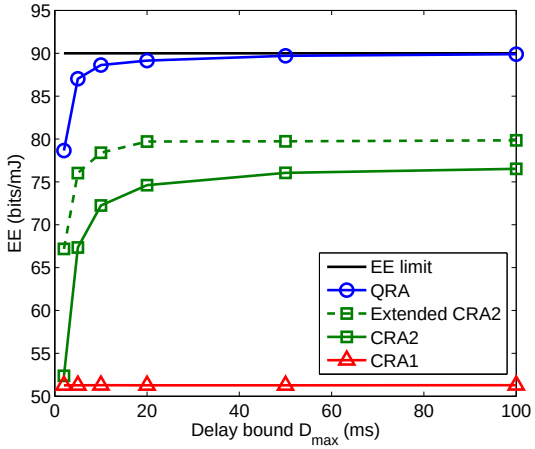
C. Comparison of Achieved EE with the EE Limit

Considering that the CRA policies are proposed for the sources with constant arrival rate, in the sequel we consider

both discrete Poisson source and constant rate source. The simulation results with AR(1) source are similar to that with Poisson source, which are not provided for conciseness.



(a) EE versus $\mathbb{E}[a(t)] = \lambda u$, $D_{\max} = 5$ ms and $\varepsilon_D = 0.01$.



(b) EE versus $D_{\max}$, $\lambda = 600$ packets/s.

Fig. 5. EE achieved by different policies, Poisson source, $u = 4$ kbits, $N_{\text{iid}} = 128$, $m = 1$ (Rayleigh fading).

Figure 5(a) shows the simulation results of the EE achieved by different policies. To see the difference in maximizing EE and SE, the policy maximizing the effective capacity under constraint (46) proposed in [16] is simulated (with legend "max SE', which always employs $N_S^{\max}$ and $P_T^{\max}$). To observe the impact of not serving empty buffer, we compare with the extended CRA2 policy. We can see an EE gain of the QRA policy over the CRA2 policy, which comes from the queue length based policy and the minimized total power consumption. The extended CRA2 policy achieves higher EE than the CRA2 policy, but the gain is not significant. The EE loss from the CRA2 to the CRA1 comes from the full buffer assumption. When $\mathbb{E}[a(t)] \geq 4.32$ Mbps, with the CRA1 policy the maximum $EE^c = E_C(\theta^c, \phi^c)/P_{tot}^c$ is achieved when $E_C(\theta^c, \phi^c) = E_B(\theta^c)$,[11] and the achieved EE is the same as the CRA2 policy. When $\mathbb{E}[a(t)] < 4.32$ Mbps, where full buffer assumption no longer holds, the maximum EE^c is attained when $E_C(\theta^c, \phi^c) > E_B(\theta^c)$, where the EE of the CRA1 policy is less than that of the CRA2 policy as discussed in Fig. 2. As expected, the "max SE" policy achieves the same EE as the CRA1 policy only when $\mathbb{E}[a(t)]$ reaches the maximal value that can be support by the system with $N_S^{\max}$ and $P_T^{\max}$ without violating $(D_{\max}, \varepsilon_D)$ (which is 4.68 Mbps here). To support this high traffic load, all the policies will use $N_S^{\max}$ and $P_T^{\max}$ when $Q_\infty > 0$. It is shown that the EE achieved by the QRA policy is closed to the EE limit except for the case when $\mathbb{E}[a(t)]$ reaches the maximal value. This implies that the relaxation in problem (16) has little influence on the EE performance of QRA policy. Besides, when $\mathbb{E}[a(t)]$ reaches the maximal value, all the resources are allocated to guarantee the statistical QoS, and thus the QRA policy cannot further adjust the transmit power and the number of subcarriers to save energy.

In Fig. 5(b), since $\lambda = 600$ packets/s, the average inter-arrival time between packets is around 2 ms. When the delay bound is tight, e.g., $D_{\max} = 2$ ms, the fluid model is inaccurate, yet the QRA policy is still applicable and approaches the EE limit surprisingly fast. It is worthy to note that although $s(t)$ is nearly deterministic in this simulation according to Table II, there is an EE-delay trade-off when $D_{\max}$ is not large. This is inconsistent with the results in [10], because the source arrives in random rate here. When $D_{\max}$ increases, the achieved EE of the CRA2 policy increases, but cannot approach the EE limit because it only minimizes the transmit power. The achieved EE of the CRA1 policy is low. This is because when $D_{\max} \geq 2$ ms the maximum EE^c is achieved when $E_C(\theta^c, \phi^c) > E_B(\theta^c)$ (i.e., full buffer assumption does not hold).

For single antenna narrow band systems, the service/transmit rate in fading channels is random. Therefore, random channels implies random service rates in early works in the literature [4]. Yet prevalent and future wireless systems are wideband and with multiple antennas, where the service rates are nearly deterministic even under fading channels. This indicates that it is the randomness of the service rate rather than the channel is more revelent in the context. To show how random service rate $s(t)$ affects the achieved EE of different policies, in Fig. 6 we simulate the source with constant arrival rate, and consider two kind of channels. The first kind channels are frequency selective channels with different coherent bandwidth as that to obtain Table II. When the number of independent subchannels grows, $s(t)$ approaches the average data rate of the system in fading channels. The second kind channels are Nakagami-m channels with different values of m. When the value of m grows, $s(t)$ approaches the data rate in AWGN channel. To capture the essence of

[11] This is the full buffer condition for EE-maximization problem. The condition indicates that for a traffic with given delay bound, the traffic load is so high that the effective bandwidth of the source is equal to the effective capacity of the system.

the problem, we consider single antenna system and set the number of active subcarriers as $N_S^{\max}$. Then, only transmit power allocation is optimized for all the policies, and the EE limit is obtained by solving problem (40) with $N_S^{\lim} = N_S^{\max}$.

The results show that the difference among the EEs achieved by the QRA and CRA2 policies and the EE limit reduce with the number of independent subchannels and the value of m. In other words, when $s(t)$ is random, the QRA policy can achieve higher EE than the CRA policies. This means that the conclusion drawn in [4] for average delay (i.e., the resource allocation policy should depend on QSI if the arrival rate and service rate are not both deterministic) is also true for statistical QoS provision. When $a(t)$ is constant and $s(t)$ is nearly deterministic, both the QRA and CRA2 policies can approach the EE limit when delay bound is large, which is consistent with the conclusion in [4] and the results in [10]. Since the objective function of the CRA1 policy $\frac{E_C(\theta^c, P_T^c)}{P_{tot}^c}$ increases with the number of independent subchannels and m, when serving the constant source, more independent subchannels or large values of m lead to a more deviation from the full buffer assumption. As a result, the achieved EE of the CRA1 policy reduces when the service rate becomes less random.

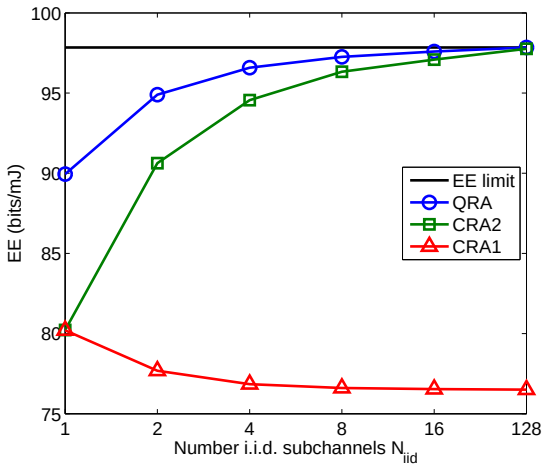
VI. Conclusion

In this paper, we proposed energy efficient bandwidth and transmit power allocation policy for MIMO-OFDM systems with statistical QoS requirement under various traffic loads. To accommodate the randomness either in sources or in channels, the policy depends on the queue length, which is optimized by introducing multi-state QoS exponents into the framework of effective bandwidth and effective capacity. A general solution was provided to optimize the QRA policy that maximizes the EE under the statistical QoS constraints, and a closed-form optimal solution was obtained for a massive MIMO-OFDM system serving AR(1) source. The EE limit of a system with various delay bound requirements was derived. Our analysis confirmed that the QRA is superior to the policies only adaptive to channel condition for delay sensitive traffic with random arrival rate. Simulation results show that the QRA policy provides substantial EE gain over relevant policies for randomly arrived data especially when the delay bound is stringent, which comes from adapting to the queue length and removing the full buffer assumption. Moreover, the achieved EE of the QRA policy approaches the EE limit at a surprisingly rapid rate.

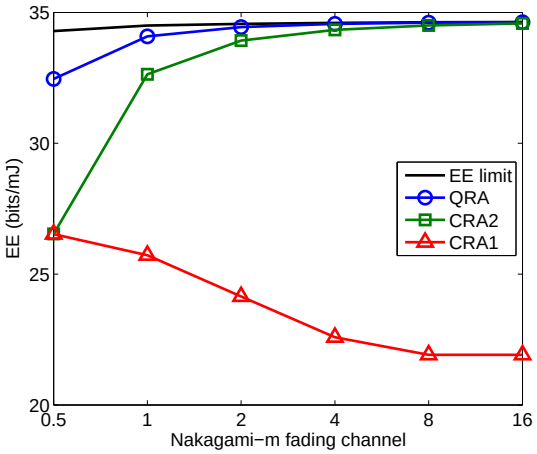
Appendix A
Proof of Concavity of Effective Capacity in (16a)

Proof: Denote $\xi = \mathbf{h}\mathbf{h}^H$, and define $f_p(x) \triangleq \left(1 + \frac{x}{\sigma_0^2 N_t}\mathbf{h}\mathbf{h}^H\right)^{-\frac{T_c B \theta_i}{\ln 2}} = (1+\kappa_1 \xi x)^{-\kappa_2}$, and $g_p(x) \triangleq -\frac{1}{T_c\theta_i}\ln x$, where $\kappa_1 = \frac{1}{\sigma_0^2 N_t}$ and $\kappa_2 = \frac{T_c B\theta_i}{\ln 2}$. Then, the effective capacity in (16a) (expressed in (17)) can be expressed as

$$E_C(\theta_i, \phi_i) = N_S(\theta_i)\, g_p\left\{\mathbb{E}_\xi\left\{f_p\left[\frac{P_T(\theta_i)}{N_S(\theta_i)}\right]\right\}\right\}. \quad \text{(A.1)}$$



(a) $a(t) = a_0 = 3.5$ Mbps, $m = 1$



(b) $a(t) = a_0 = 1$ Mbps, $N_{\text{iid}} = 1$.

Fig. 6. EE achieved by power allocation serving a source with constant rate, $N_T = 1$. The values of a_0 are selected to ensure the full buffer assumption satisfied at one kind of channel in each figure.

The first and second-order derivatives of $f_p(x)$ are $\frac{df_p(x)}{dx} = -\kappa_1\kappa_2\xi(1+\kappa_1\xi x)^{-\kappa_2-1} < 0$ for all $x \geq 0$ and $\frac{d^2 f_p(x)}{dx^2} = \kappa_2(\kappa_2+1)\kappa_1{}^2\xi^2(1+\kappa_1\xi x)^{-\kappa_2-2} > 0$ for all $x \geq 0$. Therefore, $f_p(x)$ is a decreasing and convex function. Because $g_p(x)$ is also a convex function, and taking the expectation over ξ does not change the monotonicity and convexity of $f_p(x)$, $g_p\{\mathbb{E}_\xi\{f_p[P_T(\theta_i)]\}\}$ is a concave function of $P_T(\theta_i)$ [24]. Since the effective capacity in (A.1) is a perspective function of $g_p\{\mathbb{E}_\xi\{f_p[P_T(\theta_i)]\}\}$, it is jointly concave in $N_S(\theta_i)$ and $P_T(\theta_i)$ [24]. ■

Appendix B
Proof of Proposition 1

Proof: To prove that the optimized QRA policy, $\{N_S^a(\theta_i^a),\ P_T^a(\theta_i^a),\ i = 1,...,N_q\}$, is the global optimal solution of problem (15), we show that the average total power of arbitrary feasible solution of problem (15) is higher than that achieved by the optimized QRA policy.

Denote an arbitrary feasible solution of problem (15) as $\{N_S^a(\theta_i^a),\ P_T^a(\theta_i^a),\ i = 1,...,N_q\}$, where $\theta_1^a,...,\ \theta_{N_q}^a$ are the related multi-state QoS exponents. Then, the achieved average total power consumption can be expressed as

$$\mathbb{E}\left[P_{tot}^a(\theta_i^a)\right] = \sum_{i=1}^{N_q} P_{tot}^a(\theta_i^a)\Pr\left\{(i-1)l < Q_\infty \le il\right\}.$$

For any given θ_i^a, $i = 1,...,\ N_q$, the optimal solution of subproblem I can be expressed as $\{N_S^*(\theta_i^a),\ P_T^*(\theta_i^a), i = 1,...,\ N_q\}$, which minimizes the instantaneous total power consumption in each of the N_q states. Denote the corresponding minimal total power as $P_{tot}^*(\theta_i^a)$, then $P_{tot}^a(\theta_i^a) \ge P_{tot}^*(\theta_i^a)$, $i = 1,...,\ N_q$. Moreover, as shown in (21), the probability $\Pr\left\{(i-1)l < Q_\infty \le il\right\}$, $i = 1,...,\ N_q$, is determined by the multi-state QoS exponents. This means that when $\theta_1^a,...,\ \theta_{N_q}^a$ are given, the probabilities are constant. Hence we have

$$\mathbb{E}\left[P_{tot}^a(\theta_i^a)\right] \ge \mathbb{E}\left[P_{tot}^*(\theta_i^a)\right]. \tag{B.2}$$

Furthermore, the global optimal solution of subproblem II, $\{\theta_i^*,\ i = 1,...,\ N_q\}$, can minimize the average total power consumption, $\mathbb{E}\left[P_{tot}^*(\theta_i)\right]$, $\forall\theta_i \in (0,\theta_{\max}]$, $i = 1,...,\ N_q$. Thus,

$$\mathbb{E}\left[P_{tot}^*(\theta_i^a)\right] \ge \mathbb{E}\left[P_{tot}^*(\theta_i^*)\right]. \tag{B.3}$$

From (B.2) and (B.3), we can obtain $\mathbb{E}\left[P_{tot}^a(\theta_i^a)\right] \ge \mathbb{E}\left[P_{tot}^*(\theta_i^*)\right]$. This indicates that the average total power achieved by an arbitrary feasible solution of problem (15) is higher than that achieved by $\{N_S^*(\theta_i^*), P_T^*(\theta_i^*), i = 1,...,\ N_q\}$. This completes the proof. ■

APPENDIX C
PROOF OF THE OPTIMALITY OF PROBLEM (16) WHEN $E_C(\theta_i,\phi_i) = E_B(\theta_i)$

Proof: If problem (16) is feasible, the optimal solution exists, which is denoted as $\phi^* = \{N_S^*, P_T^*\}$ in this appendix for notational simplicity. To prove that the optimal solution is achieved when $E_C(\theta_i,\phi_i) = E_B(\theta_i)$, we only need to prove $E_C(\theta_i,\phi^*) = E_B(\theta_i)$. Because it is difficult to prove this directly, we assume $E_C(\theta_i,\phi^*) > E_B(\theta_i)$ and try to prove by contradiction.

As shown in Appendix A, both $f_p(x)$ and $g_p(x)$ are decreasing functions. Therefore, the effective capacity $E_C(\theta_i,\phi_i)$ is an increasing function of $P_T(\theta_i)$. As a result, there exists a transmit power $\tilde{P}_T$, $0 < \tilde{P}_T < P_T^*$ that satisfies

$$E_C\left(\theta_i,\tilde{\phi}_i\right) = N_S^* g_p\left\{\mathbb{E}_\xi\left[f_p\left(\frac{\tilde{P}_T}{N_S^*}\right)\right]\right\} = E_B(\theta_i),$$

where $\tilde{\phi}_i = \{N_S^*, \tilde{P}_T\}$. This suggests that $\{N_S^*, \tilde{P}_T\}$ is another feasible solution of problem (16), whose total power consumption is less than the total power consumption of the optimal solution $\{N_S^*, P_T^*\}$, which contradicts with the assumption that $\{N_S^*, P_T^*\}$ is the optimal resource allocation policy. ■

APPENDIX D
THE RELATION BETWEEN $(Q_{\max},\varepsilon_D)$ AND $(D_{\max},\varepsilon_D)$ WITH CONSTANT SERVICE RATE

If a source with random arrival rate $a(t)$ is served by a constant service rate $s(t) = s_0 = E_B(\bar{\theta})$, the queueing delay at time t will be $\frac{Q(t)}{E_B(\bar{\theta})}$. Then, to satisfy the delay bound $D_{\max}$, the queue length should not exceed $E_B(\bar{\theta})D_{\max}$, i.e., we should set $Q_{\max} = E_B(\bar{\theta})D_{\max}$. Moreover, for stationary and ergodic queueing systems, the delay violation probability can be expressed as $\Pr\{D_\infty > D_{\max}\} = \lim_{T\to\infty} \frac{\frac{1}{T}\int_0^T \mathbf{1}_{\{Q(t)>Q_{\max}\}}a(t)dt}{\frac{1}{T}\int_0^T a(t)dt} = \frac{\mathbb{E}\left[\mathbf{1}_{\{Q(t)>Q_{\max}\}}a(t)\right]}{\mathbb{E}[a(t)]}$, where $\mathbf{1}_{\{Q(t)>Q_{\max}\}}$ is the indicator function, which equals to "1" if $Q(t) > Q_{\max}$ and equals to "0" otherwise. For uncorrelated arrival process, $\mathbb{E}\{\{a(t)-\mathbb{E}[a(t)]\}\{a(t-\tau)-\mathbb{E}[a(t-\tau)]\}\} = 0$ for $\tau > 0$. Because $Q(t)$ only depends on the arrival process and service process in the interval $[0,t)$, we have $\mathbb{E}[a(t)Q(t)] = \mathbb{E}[a(t)]\mathbb{E}[Q(t)]$, and then $\Pr\{D_\infty > D_{\max}\} = \frac{\mathbb{E}\left[\mathbf{1}_{\{Q(t)>Q_{\max}\}}\right]\mathbb{E}[a(t)]}{\mathbb{E}[a(t)]} = \Pr\{Q_\infty > Q_{\max}\}$. This suggests that both $(Q_{\max},\varepsilon_D)$ and $(D_{\max},\varepsilon_D)$ can be guaranteed with constant service ability $s(t) = s_0 = E_B(\bar{\theta})$ by setting $Q_{\max} = E_B(\bar{\theta})D_{\max}$.

REFERENCES

[1] G. Wu, C. Yang, S. Li, and G. Li, "Recent advance in energy-efficient networks and its application in 5G systems," *IEEE Wireless Commun. Mag.*, vol. 22, no. 2, pp. 145 – 151, Apr. 2015.

[2] J. Wu, S. Zhou, and Z. Niu, "Traffic-aware base station sleeping control and power matching for energy-delay tradeoffs in green cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 8, pp. 4196–4209, 2013.

[3] J. Lee and N. Jindal, "Asymptotically optimal policies for hard-deadline scheduling over fading channels," *IEEE Trans. Inf. Theory*, vol. 59, no. 4, pp. 2482 – 2500, Apr. 2013.

[4] 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.

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

[6] 3GPP, *Further Advancements for E-UTRA Physical Layer Aspects*. TSG RAN TR 36.814 v9.0.0, Mar. 2010.

[7] 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.

[8] 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.

[9] Q. Du and X. Zhang, "Statistical QoS provisionings for wireless unicast/multicast of multi-layer video streams," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 3, pp. 420 – 433, Apr. 2010.

[10] J. Tang and X. Zhang, "Quality-of-service driven power and rate adaptation for multichannel communications over wireless links," *IEEE Trans. Wireless Commun.*, vol. 6, no. 12, pp. 4349–4360, Dec. 2007.

[11] L. Liu, P. Parag, J. Tang, W.-Y. Chen, and J.-F. Chamberland, "Resource allocation and quality of service evaluation for wireless communication systems using fluid models," *IEEE Trans. Inf. Theory*, vol. 53, no. 5, pp. 1767–1777, May 2007.

[12] 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.

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

[14] A. Helmy, L. Musavian, and T. Le-Ngoc, "Energy-efficient power adaptation over a frequency-selective fading channel with delay and power constraints," *IEEE Trans. Wireless Commun.*, vol. 12, no. 9, pp. 4529 – 4541, Sep. 2013.

[15] L. Musavian and T. Le-Ngoc, "Energy-efficient power allocation over Nakagami-m fading channels under delay-outage constraints," *IEEE Trans. Wireless Commun.*, vol. 13, no. 8, pp. 4081 – 4091, Aug. 2014.

[16] 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.

[17] 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*, Jan. 2012. [Online]. Available: https://www.ict-earth.eu/publications/deliverables/deliverables.html

[18] 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.

[19] *cdma2000 Evaluation Methodology: Revision A*. 3GPP2 C.R1002-A, 2009.

[20] V. G. Kulkarni, "Fluid models for single buffer systems," *Frontiers in Queueing: Models and Applications in Science and Engineering*, pp. 321–388, 1997. [Online]. Available: http://www.unc.edu/~vkulkarn/papers/fluid.pdf

[21] C.-S. Chang, "Stability, queue length, and delay of deterministic and stochastic queueing networks," *IEEE Trans. Autom. Control*, vol. 39, no. 5, pp. 913–931, May 1994.

[22] 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.

[23] C. Desset, *et al.*, "Flexible power modeling of LTE base stations," in *Proc. IEEE WCNC*, Apr. 2012.

[24] S. Boyd and L. Vandanberghe, *Convex optimization*. Cambridge Univ. Press, 2004.

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

[26] D. Liberzon, *Calculus of Variations and Optimal Control Theory*. Princeton University Press, 2012.

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

[28] M. A. Zafer and E. Modiano, "A calculus approach to minimum energy transmission policies with quality of service guarantees," in *Proc. IEEE INFOCOM*, Mar. 2005.

[29] ——, "Optimal rate control for delay-constrained data transmission over a wireless channel," *IEEE Trans. Inf. Theory*, vol. 54, no. 9, pp. 4020–4039, Sep. 2008.

[30] S. Tanwir and H. Perros, "A survey of VBR video traffic models," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 4, pp. 1778 – 1802, 2013.

[31] B. Maglaris, D. Anastassiou, P. Sen, G. Karlsson, and J. D. Robbins, "Performance models of statistical multiplexing in packet video communications," *IEEE Trans. Commun.*, vol. 36, no. 7, pp. 834–844, Jul. 1998.

[32] B. Soret, M. C. Aguayo-Torres, and J. T. Entrambasaguas, "Capacity with explicit delay guarantees for generic sources over correlated Rayleigh channel," *IEEE Trans. Wireless Commun.*, vol. 9, no. 6, pp. 1901–1911, Jun. 2010.

[33] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson, "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40 – 60, Jan. 2013.

[34] 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.

[35] N. Gunaseelan, L. Liu, J.-F. Chamberland, and G. H. Huff, "Performance analysis of wireless hybrid-arq systems with delay-sensitive traffic," *IEEE Trans. Commun.*, vol. 58, no. 4, pp. 1262–1272, Apr. 2010.