Energy Efficiency and Delay in Wireless Systems: Is Their Relation Always a Tradeoff?

Changyang She and Chenyang Yang

***Abstract*—It is well-known that the average transmit power can be traded off by average delay. This paper strives to study the relation between the maximal energy efficiency (EE) and the delay bound with a given violation probability for wireless systems serving randomly arrived traffic. We show that if the minimal average transmit and circuit powers consumed at a base station linearly increases with the required average service rate, i.e., the power-rate relation is linear, then a non-tradeoff region will appear in the EE-delay relation. By taking multi-input-multi-output system as an example, we show that the power-rate relation will be linear if the transmit power and bandwidth are jointly allocated and if the bandwidth constraint is inactive to support a required delay bound. The impacts of bandwidth constraint on the power-rate and EE-delay relations are then analyzed. To study fundamental EE-delay relation, a queue length dependent two-state policy is optimized. By further considering a compound Poisson arrival process in large number of transmit antennas asymptotics, we find the boundary of the tradeoff and non-tradeoff regions, and provide a lower bound of the Pareto optimal EE-delay relation in the tradeoff region, all with closed-form expressions. Our results show that the non-tradeoff region increases with the maximal bandwidth and the number of transmit antennas.**

***Index Terms*—EE-delay relation, power-rate relation, statistical QoS requirement**

I. Introduction

The fifth generation (5G) mobile networks are expected to support high throughputs for a wide variety of services with diverse quality-of-service (QoS) requirements, ranging from tactile internet with 1 ms latency to video streaming with much less stringent delay requirement [1, 2]. To support the ever-growing traffic demands with satisfactory user experience and to reduce the cost and global carbon dioxide emissions, energy efficiency (EE) has become one of the major design goals for 5G systems. To meet the possibly conflicting performance metrics, several fundamental tradeoffs need careful examination [3], among which the EE/power-delay tradeoff has drawn significant attention over the past decade. This is because delay is a representative QoS requirement that is more relevant to characterize user experience than a minimal data rate requirement [4]. Such a metric is especially important for delay sensitive traffic such as multimedia transmission [5, 6].

Manuscript received March 6, 2015; revised August 10, 2015, December 15, 2015 and May 26, 2016; accepted August 3, 2016. The associate editor coordinating the review of this paper and approving it for publication was M. C. Gursoy. The work is supported by China NSFC under Grant 61120106002 and 973 Program under Grant 2012CB316003.

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).

EE/power-delay tradeoff has long been believed as an inherent property for wireless communications bounded by Shannons channel capacity. Since revealed in the pioneering work in [7], it has been widely accepted that the relation between EE/power and delay is a tradeoff, no matter if the delay metric is the average delay, strict deadline or the delay bound with a small violation probability. Based on the fact that the power required to transmit a bit reliably is a strictly convex function of the transmission rate for a given fading channel state, the study in [7] shows that the average transmit power and the average queueing delay cannot be minimized at the same time, unless *both* the arrival rate *and* the channel are fixed for all time. The Pareto optimal power-delay tradeoff was characterized in [7] when the average delay approaches infinity and subsequently in [8] when the average delay approaches zero. When the delay performance is modelled as a strict deadline, the power-delay tradeoff was also observed and then exploited to save power by extending the delay deadline in [9, 10]. As pointed in [7], the average delay is relatively easy to analyze, but does not necessarily ensure the QoS required by delay sensitive applications. On the other hand, the strict delay deadline is too expensive to guarantee in wireless systems in terms of transmit power. The *statistical QoS requirement*, defined as a delay bound and a delay violation probability, is more relevant for wireless multimedia transmission [11] but is more difficult to analyze. In this context, the power-delay tradeoffs for several orthogonal-frequency-division-multiplexing (OFDM) systems were studied in [5], and the EE-delay tradeoffs were respectively observed for orthogonal frequency division multiple access (OFDMA) system in [12], narrow-band system in [13], and multi-input-multi-output (MIMO) systems in [14], all by simulations.

While these research efforts have provided important insights, only the transmit power was assumed to consume energy implicitly despite that a fixed circuit power has been taken into account in EE in several works, because only transmit power, modulation and coding schemes, and/or subcarrier allocation were adjusted to save power, while full bandwidth is used for transmission. In fact, this is the essential reason that leads to the EE/power-delay tradeoff. In prevalent practical systems, however, the energy consumed for running the circuits in a system is not negligible, hence other power-saving mechanisms become necessary [15] (and also viable in practice [16]), which not only reduce transmit power but

also reduce circuit power.[1] For example, we can further adapt the bandwidth or even turn a base station (BS) into idle mode to reduce the circuit power. While the importance of reducing circuit power has been widely recognized in green communication literature, the impact of a rate-dependent circuit power consumption model on the EE-delay relation is largely overlooked, which may change the characteristics of the EE/power-delay relation. Recently, it was observed in [6] that the average power is not a monotonically decreasing function of the average sojourn time when BS idling strategies are considered. Similar observation has been obtained in an earlier work [17] with transmission time as the delay metric.

A. Related Works

There have been a lot of works investigating transmit power or energy efficiency under delay constraints. In general, these works can be divided into two major categories in terms of goal and hence methodologies.

The goal of the first category of works is to reveal the fundamental relation between average transmit power and average delay requirement [7,8,18], which are from information-theoretic perspective. The fundamental studies aimed at finding the Pareto optimal power-delay tradeoff in [7, 8] imply that the power-delay tradeoff comes from the relationship between the transmit power and transmission/service rate, which is strictly convex according to the Shannon's capacity. An exceptional example in [18] demonstrates that if the average transmit power is a piecewise linear function of the required average rate, then the optimal power-delay tradeoff curve in [7] can be exceeded. To show how to achieve the Pareto optimal power-delay tradeoff, some transmission policies were provided in [7, 8, 18]. Since queueing delay is taken into account, all the policies in these works depend on both channel state information (CSI) and queue state information (QSI).

The goal of the second category of works is to design energy efficient policies with delay requirements, e.g., [5,6,9,10,12–14, 17, 19–25]. After optimizing a policy towards a specific objective function, the EE/power-delay relation that can be achieved by the policy is either analyzed theoretically or discussed via simulations, where the delay metric is hard deadline in [9, 10, 19–22], delay bound with a violation probability in [5,12–14], and real average delay achieved by the policies in [6,17,23–25]. Except in [6,17,25], all these works either do not consider or only consider a fixed rate-independent circuit power, and the power saving mechanism is power allocation that only reduces the transmit power. For all the works considering the statistical QoS requirement [5, 12–14], the optimized policies only adapt to CSI. This may be owing to the fact that effective capacity is powerful in designing the policies with statistical QoS provision, which can translate a cross-layer problem related to both QSI and CSI into a physical-layer issue only related to CSI. As indicated in [7], when *either* source *or* channel is stochastic, the resource allocation should depend on the QSI, otherwise the average power cannot be minimized for a target average delay. This implies that the EE/power-delay relations observed in [5,12–14] are not Pareto optimal.

B. Differences from Related Works and Contributions

In this paper, we strive to investigate fundamental EE-delay relation for wireless systems serving random sources with statistical QoS requirement and connect it with the power-rate relation. By " fundamental", we mean that our study is from information-theoretic perspective, along the same line with the first category of research in [7, 8, 18]. Nonetheless, our study differs from these works in the following two aspects. First, we consider statistical QoS requirement, hence the tool we used for analysis differs from theirs. Second, we reduce both transmit and circuit powers. Though this seems nothing but only more practical by accounting for the power consumed for operating systems, it is the rate-dependent circuit power consumption model that leads to the fundamental difference between EE-delay relation and power-delay relation.

To this end, we first derive the EE limit, which is the maximal EE approachable by a system when serving a traffic with any given average arrival rate and infinite delay bound, then find the optimal policy that is able to achieve the EE limit with finite delay bounds. This is in contrast to the second category of research that aims to find a policy to maximize or minimize a specific objective and then evaluates the EE-delay curve achieved by the policy. To maximize the EE of a system, we optimize the power-saving policy to reduce both the circuit power and the transmit power, which differs from the previous studies in [5,7–10,12–14,18–24]. To reflect the QoS requirement of delay sensitive traffic with various delay-bounded QoS provisioning, the delay metric is the delay bound with a small violation probability, which differs from the average delay bound in [7, 8, 18] and the *real delay* in serving a traffic in [6,17,23–25].[2]

The major contributions of this work are summarized as follows.

- We prove that the EE-delay relation can be divided into a tradeoff region and a non-tradeoff region according to the delay bound, which depends on the power-rate relation. We discover an important class of application scenarios referred to as the *linear case*, where the minimal average total power consumption linearly increases with the average service rate required to ensure a given QoS. In such a case, the EE-delay tradeoff vanishes even when

[1] It is worthy to note that the circuit power has been taken into account in [12,13], which is a constant independent of the transmission rate. However, only the mathematical property of the optimization problem is affected by the rate-independent circuit power, while the optimization variables (i.e., the transmit power and/or subcarrier allocation) cannot reduce the circuit power. Therefore, only transmit power is reduced, and the EE-delay relation is essentially the same as the power-delay relation.

[2] This delay metric is more appropriate for elastic traffic, whose delay *can be* traded off for energy reduction. The *delay bound* we considered is the maximal delay that a specific traffic can tolerate with a small violation probability in order not to compromise user experience. Such a statistical QoS provision is appropriate for delay sensitive traffic such as video conference, where a data packet becomes useless once its delay bound is violated, and the real delay to serve the traffic is not a concern.

both the arrival rate *and* the channel are random, which differs from the results in [7, 8].

- We prove that the EE limit can be achieved by a QSI based two-state policy when the required delay bound is finite and lies in the non-tradeoff region of the EE-delay relation, hence the obtained EE-delay relation in the non-tradeoff region is optimal in the sense of achieving the EE limit. This indicates that the EE of a system may not reduce when supporting delay sensitive services with respective to the delay tolerant services.
- To demonstrate whether and when the *linear case* exists in practice, we optimize a QSI dependent two-state policy by taking a downlink MIMO system as an example, where a BS serves multiple users with perfect CSI. We prove that if transmit power and used bandwidth are jointly optimized to maximize EE, the linear case will occur when the constraint on bandwidth is inactive to guarantee the delay bound required by a specific application. For a shorter delay bound, the required minimal bandwidth should be wider, which reflects an EE-delay-bandwidth tradeoff.
- By further taking the compound Poisson arrival process in large number of transmit antennas asymptotics as an example, we find the boundary between the tradeoff and non-tradeoff regions, and derive a lower bound of the Pareto optimal EE-delay relation in the tradeoff region, all in closed-form.

The rest of the paper is organized as follows. Section II describes queueing model and defines several notions. Section III proves that the EE-delay relation is determined by the power-rate relation. Section IV introduces the system and power consumption models for a MIMO system and formulates two problems to obtain the power-rate relation and the EE-delay relation, respectively. Section V and VI analyze the EE-delay relation and find the optimal two-state policies in single user and multi-user scenarios, respectively. Section VII provides simulation and numerical results to validate the analysis and to illustrate the EE-delay and power-rate relations. Section VIII concludes the paper.

II. Queueing Model and Definitions

A. Queueing Model and Statistical QoS Requirement

Consider a downlink multiuser system, where a BS serves K users with delay-sensitive services. The statistical QoS requirement of user k is defined as $(D_k^{\max}, \varepsilon_{D_k})$, where $D_k^{\max}$ is the delay bound and ε_{D_k} is the delay bound violation probability allowed by the service. In this paper, we consider the queuing delay and ignore the coding and transmission delay.

We consider a fluid queueing model [26, 27], which is valid when the time interval between arrived data packets and the interval of updating transmit policy (called the transmit time interval (TTI) of the system) are much shorter than the delay bound required by the traffic. At time t, the data of the kth user enters a first-in-first-out buffer of the BS at the arrival rate $a_k(t)$, and is transmitted to the user at the departure rate $b_k(t)$. We assume that the data for the K users wait in K queues, and denote the queue length of the data for the kth user as $Q_k(t)$. Then, the dynamics of the queue lengths can be expressed as [27],

$$\frac{dQ_k(t)}{dt} = a_k(t) - b_k(t) = \begin{cases} \max\{a_k(t) - s_k(t), 0\}, Q_k(t) = 0 \\ a_k(t) - s_k(t), \quad Q_k(t) > 0 \end{cases},$$

where $k = 1, ..., K$ and $s_k(t)$ is the transmission rate (also called service rate) of the kth user. Then, the throughput $b_k(t)$ is related with $a_k(t)$ and $s_k(t)$ as

$$b_k(t) = \begin{cases} \min\{a_k(t), s_k(t)\}, Q_k(t) = 0 \\ s_k(t), \quad Q_k(t) > 0 \end{cases}. \tag{1}$$

Under the assumption of infinite buffer size, $Q_k(t) = A_k(t) - B_k(t)$, where $A_k(t) \triangleq \int_0^t a_k(\tau)\, d\tau$ and $B_k(t) \triangleq \int_0^t b_k(\tau)\, d\tau$ are the amounts of data arrived and departed in the interval of $[0, t]$, respectively. Assume that the queues of the users are in steady state during the interval of $[0, t]$, and denote Q_k^{∞} as the steady state queue length of the kth user.

Effective bandwidth [28] and effective capacity [29] are powerful tools to design the systems with statistical QoS requirement. Denote $E_{B_k}(\theta_k)$ and $E_{C_k}(\theta_k)$ as the effective bandwidth of arrival process $\{a_k(t), t > 0\}$ and effective capacity of service process $\{s_k(t), t > 0\}$, where θ_k is the QoS exponent reflecting the performance in terms of queue length as [26]

$$\lim_{Q_k^{\max}\to\infty} -\frac{\ln \Pr(Q_k^{\infty} > Q_k^{\max})}{Q_k^{\max}} = \theta_k, \tag{2}$$

where $Q_k^{\max}$ is the maximal queue length. A large value of θ_k indicates a small value of $Q_k^{\max}$ with a given maximal queue length violation probability, and *vice versa*. For a given stationary arrival process with available effective bandwidth $E_{B_k}(\theta_k)$, the required QoS exponent θ_k^c can be obtained from $(D_k^{\max}, \varepsilon_{D_k})$ as follows [30, 31],

$$\Pr(D_k^{\infty} \geq D_k^{\max}) \approx \exp[-\theta_k E_{B_k}(\theta_k) D_k^{\max}] = \varepsilon_D, \tag{3}$$

where D_k^{∞} is the steady state delay of the kth user. To guarantee the QoS requirement characterized by $(\theta_1^c, \theta_2^c, ..., \theta_K^c)$, a transmit policy should satisfy [26]

$$E_{C_k}(\theta_k^c) \geq E_{B_k}(\theta_k^c), k = 1, 2, ..., K. \tag{4}$$

With given ε_{D_k}, when $D_k^{\max} \to \infty$ (i.e., $\theta_k^c \to 0$), (4) degenerates into

$$\mathbb{E}_h\{s_k(t)\} \geq \mathbb{E}\{a_k(t)\}, k = 1, 2, ..., K, \tag{5}$$

where $\mathbb{E}_x\{\cdot\}$ represents the expectation taken over x.

B. Two-state Transmit Policy

For randomly arrived process, the optimal transmit policy that minimizes the transmit power under delay constraint should depend on both QSI and CSI [7, 8]. To study the maximal EE achieved by a system with statistical QoS require-

ments, we introduce a QSI-based *two-state transmit policy*. When $Q_k^\infty > 0$, the policy allocates resources to the kth user depending on the user's own CSI, which is referred to as "ON" state of the user. When $Q_k^\infty = 0$, no resource is allocated to the kth user and hence $s_k(t) = 0$, which is referred to as "OFF" state of the user.

Remark 1. If the TTI is much shorter than the delay bound (i.e., the system can switch between "OFF" and "ON" states rapidly), then $(D_k^{\max}, \varepsilon_{D_k})$ can be guaranteed with a two-state policy satisfying $E_{C_k}(\theta_k^c) \geq E_{B_k}(\theta_k^c), Q_k^\infty > 0, k = 1, 2, ..., K$ [14].

C. EE-Delay Relation, EE Limit, and Power-Rate Relaion

The "bits per Joule" EE metric is defined as the ratio of average throughput[3] to average total power consumed at the BS [32]. With the two-state policy, the average total power consumption is $\mathbb{E}_{\boldsymbol{Q}^\infty,h}\{P_{tot}\}$, where $\boldsymbol{Q}^\infty = (Q_1^\infty, Q_2^\infty, ..., Q_K^\infty)$ is QSI, and h is CSI. When the queue is in steady state, $\mathbb{E}\{b_k(t)\} = \mathbb{E}\{a_k(t)\}$ [33, 34]. Then, for a system with any given traffic load characterized by $\sum_{k=1}^{K} \mathbb{E}\{a_k(t)\}$, the EE is

$$EE \triangleq \frac{\sum_{k=1}^{K} \mathbb{E}\{b_k(t)\}}{\mathbb{E}_{\boldsymbol{Q}^\infty,h}\{P_{tot}\}} = \frac{\sum_{k=1}^{K} \mathbb{E}\{a_k(t)\}}{\mathbb{E}_{\boldsymbol{Q}^\infty,h}\{P_{tot}\}}, \quad (6)$$

where P_{tot} is the total power consumption including transmit and circuit power consumptions depending on the specific system, which can be controlled by the two-state transmit policy.

It is worthy to note that such a definition of EE differs from the "delay sensitive EE" in [12–14], which is defined as $\sum_{k=1}^{K} E_{C_k}(\theta_k^c)/\mathbb{E}_h\{P_{tot}\}$. Maximizing the delay sensitive EE yields the policies only depending on CSI.

Definition 1: The *EE-delay relation* is denoted as $EE^{\max}(\theta_1^c, \theta_2^c, ..., \theta_K^c)$, which is defined as the maximal EE achieved by a system with any given traffic load satisfying the QoS requirements $(\theta_1^c, \theta_2^c, ..., \theta_K^c) \in \mathbb{R}_+^K$, where $\mathbb{R}_+^K$ is the positive real space of K-dimension.

Such an EE-delay relation is fundamental, which only depends on the system and the service. When only the transmit power is taken into account, the EE-delay relation degenerates to the power-delay relation, which will be the same as the power-delay relation in [7,8,18] if average delay requirement is considered. In [7,8,18], the relation is defined as the minimal average power required to ensure the average delay less than $\bar{D}$ (i.e., average delay bound).

Definition 2: The *EE limit* is defined as $EE^{\lim} \triangleq \lim_{\substack{\theta_k^c \to 0, \\ k=1,2,...,K}} EE^{\max}(\theta_1^c, \theta_2^c, ..., \theta_K^c)$. The corresponding average total power consumption is referred to as the *power limit*, which is denoted as $P_{tot}^{\lim} = \lim_{\substack{\theta_k^c \to 0, \\ k=1,2,...,K}} \mathbb{E}_{\boldsymbol{Q}^\infty,h}\{P_{tot}\}$.

The EE limit is an upper bound of $EE^{\max}(\theta_1^c, \theta_2^c, ..., \theta_K^c)$ for arbitrary $(\theta_1^c, \theta_2^c, ..., \theta_M^c) \in \mathbb{R}_+^K$, which is achievable when $D_k^{\max} \to \infty$, $k = 1, 2, ..., K$.

Definition 3: The *power-rate relation* is defined as $\mathbb{E}_h\{P_{tot}^{\min}\} - (\bar{s}_1, \bar{s}_2, ..., \bar{s}_K)$, where $\mathbb{E}_h\{P_{tot}^{\min}\}$ is the minimal average total power consumption required to support the average service rates $\mathbb{E}_h\{s_k(t)\} = \bar{s}_k$, $k = 1, 2, ..., K$.

The characteristic of the EE-delay relation is determined by the power-rate relation, as implied in [7, 8, 18]. This is because maximizing the EE is equivalent to minimizing the average total power consumption for any given traffic load $\sum_{k=1}^{K} \mathbb{E}\{a_k(t)\}$ according to the EE definition in (6). Besides, the required delay bound has a close relation with the required average service rate. Intuitively, ensuring a stringent delay bound needs a high service rate.

The power-rate relation depends on the power-saving mechanism and hence on whether circuit power can be reduced. In [7, 8, 18], the circuit power was not considered, thus the energy is saved by adjusting transmit power and transmission rate with adaptive modulation and coding. With the capacity-achieving coding, the power-rate relation, defined as the minimal *transmit power* required to support a service rate $s(t)$, is *strictly convex*, consequently the minimal average power is a strictly decreasing function of the average delay bound $\bar{D}$ [7, 8]. For modern systems where circuit power is not negligible but can be reduced by some new power-saving mechanisms, the power-rate relation and hence the resulting EE-delay relation behave quite differently. For example, for single antenna systems, we can adjust bandwidth together with transmit power to reduce the total power consumption. For multi-antenna systems, we can further adjust the number of antennas. This sharply differs from the analysis of only reducing transmit power, where all the available bandwidth and antennas should be used to save power. With more flexible power-saving mechanisms, it is possible that $\mathbb{E}_h\{P_{tot}^{\min}\}$ becomes a linear function of $(\bar{s}_1, \bar{s}_2, ..., \bar{s}_K)$ as illustrated later, which is more optimistic than the strict convex power-rate relation.

III. EE-DELAY RELATION AND POWER-RATE RELATION

According to the distinguished properties of the EE-delay relations, the power-rate relation can be divided into linear and nonlinear cases.

Linear case: This occurs for the scenario where the minimum average total power required to support an average service rate $\bar{s}_k$ can be expressed as $\mathbb{E}_h\{P_{tot}^{\min}\} = \sum_{k=1}^{K} c_k \bar{s}_k + c_0$, where $c_k, k = 1, 2, ..., K$ and c_0 are positive constants depending on specific system.

As implied from the discussion in [8], a necessary condition for a transmit policy to minimize transmit power is not to serve empty buffer. This suggests that in order for a two-state policy

[3]The throughput is the number of bits actually transmitted per second in a system for a given arrival process, rather than the maximal number of bits the system can transmit per second (i.e., capacity). Hence, the average throughput is $\sum_{k=1}^{K} \mathbb{E}\{b_k(t)\}$.

to achieve the EE-delay relation the following condition should be satisfied

$$\mathbb{E}_{\boldsymbol{Q}^{\infty},h}\{s_k(t)\} = \mathbb{E}\{a_k(t)\}. \tag{7}$$

From (5) and (7), the EE limit can be derived as (see Appendix A),

$$EE^{\lim} = \sum_{k=1}^{K} \mathbb{E}\{a_k(t)\} / (\sum_{k=1}^{K} c_k \mathbb{E}\{a_k(t)\} + c_0). \tag{8}$$

In the linear case, the following proposition indicates that for any given arrival process with average rate $\mathbb{E}\{a_k(t)\}$, the EE limit is achievable by a simple QSI-based policy even under finite delay requirement, and the EE-delay tradeoff vanishes (proved in Appendix B).

Proposition 1. In the linear case, the EE-delay relation does not depend on the delay bounds $D_k^{\max}, k = 1, ..., K$, and $EE^{\max}(\theta_1^c, \theta_2^c, ..., \theta_K^c) = EE^{\lim}$ can be achieved by a two-state policy.

The proposition suggests that: (i) if the power-rate relation is linear, then the EE-delay relation is not a tradeoff, and the maximal EE of the non-tradeoff region equals to the EE-limit, (ii) a two-state policy can achieve the EE limit.

Furthermore, the following corollary indicates that the EE-delay tradeoff may vanish as well when the average delay is considered as the delay metric (proved in Appendix C).

Corollary 1. In the linear case, the EE does not depend on the average delay bound $\bar{D}$.

Nonlinear case: For other scenarios where $\mathbb{E}_h\{P_{tot}^{\min}\}$ is a strictly convex function of $\bar{s}_k$, the tradeoff between average transmit power and average delay in single user systems has been studied in large-delay and small-delay regimes [7, 8], and the result in large-delay regime has been extended into multi-user systems [18]. In the context of statistical QoS requirement, we will find the maximal EE achieved by the two-state policy. Since a QSI-based policy with more than two states may achieve higher EE than the two-state policy, the EE achieved by the optimized two-state policy as a function of $D_k^{\max}$ is a lower bound of the Pareto optimal EE-delay tradeoff.

IV. MIMO: An Example System

To show when the linear and non-linear cases happen in real-world applications, we consider a downlink MIMO system as a concrete example in the subsequent sections. We employ frequency division multiple access to avoid multi-user interference, and maximum ratio transmission (MRT) for each user as precoding.

A. System Model

Consider a BS equipped with N_T antennas serves K single-antenna users under block fading channels. For notational simplicity, we consider flat fading channel. For frequency-selective channel, there is no fundamental difference on the EE-delay relation, as shown by simulation later. The spatial channel vector from the BS to the kth user during the lth channel fading block is $\mathbf{h}_{k,l} \in \mathbb{C}^{N_T \times 1}$, whose elements are assumed as independent and identically distributed (i.i.d.) Gaussian variables with zero mean and variance μ_k. Assume that the CSI is perfectly known at the BS. Then, the maximum achievable service/transmission rate of the kth user is

$$s_k(t) = W_k \log_2 \left(1 + \frac{\mu_k p_{k,l} g_{k,l}}{N_0 W_k}\right), (l-1)T_c < t \leq lT_c, \tag{9}$$

where W_k is the bandwidth allocated to the kth user, $p_{k,l}$ is the transmit power allocated to the kth user during the lth channel fading block, $g_{k,l} \triangleq \frac{1}{\mu_k}\mathbf{h}_{k,l}^H \mathbf{h}_{k,l}$ is the instantaneous channel power gain, T_c is the duration of each channel fading block, N_0 is single-sided spectrum density of white Gaussian noise, and $(\cdot)^H$ is complex conjugate transpose.

Denote $W^{\max}$ as the maximal available bandwidth of the system. Then, $\sum_{k=1}^{K} W_k \leq W^{\max}$. For i.i.d. block fading channel, the effective capacity can be expressed as follows [29],

$$E_{C_k}(\theta_k) = -\frac{1}{\theta_k^c T_c} \ln \mathbb{E}\{\exp\left[-\theta_k^c T_c s_k(t)\right]\}. \tag{10}$$

To save energy consumed for transmitting data and operating the system, the BS can adjust the service rate of each user by allocating transmit power and bandwidth.[4] Since the bandwidth affects the circuit power, from the in-depth analysis of the powers consumed for various modules including radio frequency transceivers and baseband processing in MIMO system [16], the total power consumption at the BS can be modeled as $P_{tot} = \frac{1}{\rho}\sum_{k=1}^{K} p_{k,l} + P_{CW}\sum_{k=1}^{K} W_k + P_0$, where P_{CW} is the circuit power consumed for baseband processing per unit bandwidth, P_0 is the rate-independent circuit power, and $\rho \in (0, 1]$ is the power amplifier efficiency. The values of P_{CW} and P_0 increase with the number of transmit antennas N_T [16]. Denote $\bar{P}_{T_k} = \mathbb{E}_h\{p_{k,l}\}$ as the average transmit power of the kth user. Then, the average total power consumption at the BS is

$$\mathbb{E}_h\{P_{tot}\} = \frac{1}{\rho}\sum_{k=1}^{K} \bar{P}_{T_k} + P_{CW}\sum_{k=1}^{K} W_k + P_0. \tag{11}$$

B. Problem Formulation

In order to study the *power-rate relation* and the *EE-delay relation* for the system, we formulate two problems.

According to Definition 3 and the power model in (11), the *power-rate relation* can be found from the following problem,

$$\min_{\substack{\bar{P}_{T_k}, W_k, \\ k=1,...,K}} \sum_{k=1}^{K} \bar{P}_{T_k} + \rho P_{CW}\sum_{k=1}^{K} W_k, \tag{12}$$

[4]The number of active antennas can also be controlled by switching off some unnecessary antennas to ensure QoS, which however leads to rather involved optimization. To simplify the analysis, we fix N_T in the following derivations. The impact of further jointly optimizing N_T on the EE-delay relation will be illustrated with numerical results later.

$$\text{s.t.} \quad \mathbb{E}_h\{s_k(t)\} = \bar{s}_k,\ k = 1, ..., K, \tag{12a}$$

$$\sum_{k=1}^{K} W_k \leq W^{\max}. \tag{12b}$$

To provide the EE-limit achieving policy in the linear case and find the lower bound of Pareto optimal *EE-delay tradeoff* in the nonlinear case, we need to find the EE-optimal two-state policy. According to (6), maximizing the EE is equivalent to minimizing the average total power consumption for any given arrival process, where $\sum_{k=1}^{K} \mathbb{E}\{a_k(t)\}$ is fixed. Therefore, we find the optimal two-state policy that minimizes the average total power under the QoS constraint.

Denote $\eta_k \triangleq \Pr(Q_k^{\infty} > 0)$, which is the non-empty probability of the buffer of the kth user. Then, the average total power consumed by the two-state policy can be expressed as,[5]

$$\mathbb{E}_{\boldsymbol{Q}^{\boldsymbol{\infty}},h}\{P_{tot}\} = \sum_{k=1}^{K} \eta_k \left(\frac{\bar{P}_{T_k}^{\text{on}}}{\rho} + P_{CW} W_k^{\text{on}} \right) + P_0, \tag{13}$$

where $\bar{P}_{T_k}^{\text{on}}$ and W_k^{on} are the average transmit power and the bandwidth allocated to the kth user when $Q_k^{\infty} > 0$ (i.e., in the "ON" state of the user), respectively.

Then, the EE-optimal two-state policy can be obtained from the following problem,

$$\min_{\substack{\bar{P}_{T_k}^{\text{on}}, W_k^{\text{on}}, \\ k=1,...,K}} \sum_{k=1}^{K} \eta_k \left(\frac{\bar{P}_{T_k}^{\text{on}}}{\rho} + P_{CW} W_k^{\text{on}} \right) + P_0, \tag{14}$$

$$\text{s.t.} \quad E_{C_k}(\theta_k^c) \geq E_{B_k}(\theta_k),\ \forall Q_k^{\infty} > 0, k = 1, ..., K, \tag{14a}$$

$$\sum_{k=1}^{K} W_k^{\text{on}} \leq W^{\max}. \tag{14b}$$

V. EE-delay Relationship in Single-user Scenario

In this section, a single user scenario is considered, where the index k in problems (12) and (14) can be omitted for notational simplicity. We first show that the minimal average total power consumption required to support an average service rate will linearly grow with the average service rate requirement if the average transmit power and bandwidth are jointly allocated and the constraint on $W^{\max}$ is inactive. Then, we provide the two-state policy to achieve the EE limit in the linear case. Next, the impacts of the bandwidth constraint are analyzed, and the boundary between tradeoff and non-tradeoff regions of the EE-delay relation are provided for a compound Poisson arrival process in large N_T asymptotics. Finally, the lower bound of the Pareto optimal EE-delay tradeoff is obtained in the nonlinear case.

[5]Analogous to [7,8,18], when we study the power-rate relation, the average total power consumption is $\mathbb{E}_h\{P_{tot}\}$, and when we find the two-state policy to achieve the EE-delay relation, the average total power consumption is $\mathbb{E}_{\boldsymbol{Q}^{\boldsymbol{\infty}},h}\{P_{tot}\}$.

A. EE-delay Relation in Linear Scenarios

We first find the relationship between $\mathbb{E}_h\{P_{tot}^{\min}\}$ and $\bar{s}$ for the application scenario where the constraints on $W^{\max}$ in (12b) and (14b) are not active. Then, we will discuss when the maximum bandwidth constraint is not active in Section V.B.

1) Power-Rate Relation: The service rate $s(t)$ in constraint (12a) relies on the instantaneous transmit power p_l allocated in the lth fading block as shown from (9), which depends on the instantaneous power allocation policy with given average transmit power $\bar{P}_T$ and bandwidth W, denoted as $f_p(\bar{P}_T, W, g_l)$. Then, problem (12) can be re-expressed as,

$$\min_{\bar{P}_T, W,} \bar{P}_T + \rho P_{CW} W, \tag{15}$$

$$\text{s.t.} \quad \mathbb{E}_h \left\{ W \log_2 \left[1 + \frac{\mu f_p(\bar{P}_T, W, g_l) g_l}{N_0 W} \right] \right\} = \bar{s}. \tag{15a}$$

To find the solution of problem (15), we first need to determine the optimal form of the function $f_p(\bar{P}_T, W, g_l)$ that minimizes the objective function in (15).

Finding the optimal form of a function belongs to the functional extreme problem, and the general method to solve such a problem is very difficult. In order to obtain a closed-form solution of $f_p(\bar{P}_T, W, g_l)$ for the succeeding optimization, in the sequel we employ an alternative way. Specifically, we first prove that the optimal instantaneous power allocation policy is water-filling, and then find the optimal solution of $\bar{P}_T, W$ with the optimized form of $f_p(\bar{P}_T, W, g_l)$.

As shown in [35], given the average transmit power $\bar{P}_T$ and bandwidth W, the instantaneous power allocation policy that can maximize the average service rate is a water-filling policy,

$$p_l = f_p^w\left(\bar{P}_T, W, g_l\right) \triangleq \begin{cases} \frac{N_0 W}{\mu}\left(\frac{1}{g^{\text{th}}} - \frac{1}{g_l}\right), g_l \geq g^{\text{th}}, \\ 0, \qquad\qquad g_l < g^{\text{th}}, \end{cases} \tag{16}$$

where the water level g^{th} can be obtained from

$$\int_{g^{\text{th}}}^{\infty} \frac{N_0}{\mu}\left(\frac{1}{g^{\text{th}}} - \frac{1}{g}\right) f_h(g) dg = \frac{\bar{P}_T}{W}, \tag{17}$$

and $f_h(g)$ is the distribution of channel gains, which follows the Wishart distribution [36] as,

$$f_h(g) = \frac{1}{(N_t - 1)!} g^{N_t - 1} e^{-g}, N_t > 1. \tag{18}$$

The following Proposition shows that (16) is the optimal instantaneous power allocation policy that minimizes the objective function in (15) (the proof is omitted since this proposition is a special case of Proposition 1 in [37]).

Proposition 2. Consider an arbitrary instantaneous power allocation policy $\tilde{f}_p(\bar{P}_T, W, g_l)$ differing from (16). The optimal solutions of problem (15) with policies $p_l = f_p^w\left(\bar{P}_T, W, g_l\right)$ and $p_l = \tilde{f}_p(\bar{P}_T, W, g_l)$ are denoted as $\{\bar{P}_T^w, W^w\}$ and $\{\tilde{\bar{P}}_T, \tilde{W}\}$, respectively. Then,

$$\bar{P}_T^w + \rho P_{CW} W^w \leq \tilde{\bar{P}}_T + \rho P_{CW} \tilde{W} \tag{19}$$

Note that problem (15) can be equivalently re-formulated as minimizing the ratio of the objective function in (15) to the average service rate under the constraint in (15a). Then, with the optimized form of $f_p(\bar{P}_T, W, g_l)$ in (16), problem (15) is equivalent to the following problem,

$$\max_{\bar{P}_T, W,} \frac{\frac{\bar{P}_T}{W} + \rho P_{CW}}{\mathbb{E}_h\left\{\log_2\left[1 + \frac{\mu f_p^w(\bar{P}_T, W, g_l) g_l}{N_0 W}\right]\right\}}, \quad \text{s.t.} \quad (15a). \tag{20}$$

Denote the average transmit power per unit bandwidth as $\bar{P}_{TW} \triangleq \frac{\bar{P}_T}{W}$.

From (16) and (17), we can see that the value of $\frac{f_p^w(\bar{P}_T, W, g_l)}{W}$ is determined by the value of $\bar{P}_{TW}$. Therefore, the objective function in (20) is a function of $\bar{P}_{TW}$ rather than a function of individual variables of $\bar{P}_T$ and W. Denote the average service rate per unit bandwidth as $\bar{R}_W(\bar{P}_{TW}) = \mathbb{E}_h\left\{\log_2\left[1 + \frac{\mu f_p^w(\bar{P}_T, W, g_l) g_l}{N_0 W}\right]\right\}$. Then, we can obtain the following proposition.

Proposition 3. $\bar{R}_W(\bar{P}_{TW})$ is strictly concave in $\bar{P}_{TW}$.

Proof: When $N_t = 1$, this proposition is the same as Proposition 3 in [37]. When $N_t > 1$, the distribution of instantaneous channel power gain in (18) is different from that when $N_t = 1$. It is not hard to see that $f_h(g)$ in (18) has no impact on the concavity of $\bar{R}_W(\bar{P}_{TW})$, thus $\bar{R}_W(\bar{P}_{TW})$ is strictly concave in $\bar{P}_{TW}$ for $N_t > 1$. This completes the proof. □

Since $\bar{R}_W(\bar{P}_{TW})$ is strictly concave in $\bar{P}_{TW}$, the objective function in (20) is strictly quasiconvex in $\bar{P}_{TW}$ [38]. Thus, the value of $\bar{P}_{TW}$ that maximizes the objective function is unique, which is denoted as $\bar{P}^*_{TW}$. According to the definition of $\bar{P}^*_{TW}$, the optimal solution of problem (20), $\bar{P}^*_T$ and W^*, should satisfy the following condition

$$\frac{\bar{P}^*_T}{W^*} = \bar{P}^*_{TW}. \tag{21}$$

Note that the value of $\bar{P}^*_{TW}$ is obtained without considering the average service rate requirement in (15a), hence it does not depend on $\bar{s}$. Further considering the service rate requirement in (15a), the optimal solution of problem (20) can be obtained as follows,

$$\bar{P}^*_T = \frac{\bar{P}^*_{TW}}{\bar{R}_W(\bar{P}^*_{TW})}\bar{s} \quad \text{and} \quad W^* = \frac{1}{\bar{R}_W(\bar{P}^*_{TW})}\bar{s}. \tag{22}$$

Substituting (22) into (11), we can derive that

$$\mathbb{E}_h\{P^{\min}_{tot}\} = \left[\frac{\bar{P}^*_{TW}}{\rho\bar{R}_W(\bar{P}^*_{TW})} + \frac{P_{CW}}{\bar{R}_W(\bar{P}^*_{TW})}\right]\bar{s} + P_0 = c_1\bar{s} + c_0, \tag{23}$$

where $c_1 = \frac{\bar{P}^*_{TW}}{\rho\bar{R}_W(\bar{P}^*_{TW})} + \frac{P_{CW}}{\bar{R}_W(\bar{P}^*_{TW})}$ and $c_0 = P_0$. It is shown that the minimal average transmit power linearly increases with the average service rate.

Remark 2. One may suppose that such a linear relation is an artifact of the considered power consumption model, where the total power is a linear function of W. However, this is not true. In fact, the linear relation is attributed to the flexible power-saving mechanisms in reducing both transmit and circuit powers, i.e., the joint transmit power and bandwidth allocation here. Later, we will use numerical result to show that the power-rate relation is still linear when N_T is jointly optimized with $\bar{P}_T$ and W, where the total power consumption is a non-linear function of N_T.

2) EE-Delay Relation: In the linear case, according to Proposition 1, the EE-delay relation $EE^{\max}(\theta^c)$ equals to the EE limit $EE^{\lim}$ for all values of θ^c, i.e., there is no EE-delay tradeoff.

When $\theta^c \to 0$, from (8) and with the constants c_1 and c_0 defined in (23), the EE limit and the corresponding power limit of the MIMO system can be respectively expressed as

$$EE^{\lim} = \mathbb{E}\{a(t)\}/\mathbb{E}_h\{P^{\lim}_{tot}\} = \mathbb{E}\{a(t)\}/(c_1\mathbb{E}\{a(t)\} + c_0), \tag{24}$$

$$\mathbb{E}_h\{P^{\lim}_{tot}\} = \left[\frac{\bar{P}^*_{TW}}{\rho\bar{R}_W(\bar{P}^*_{TW})} + \frac{P_{CW}}{\bar{R}_W(\bar{P}^*_{TW})}\right]\mathbb{E}\{a(t)\} + P_0. \tag{25}$$

3) EE-Limit Achieving Policy: In what follows, we find the two-state policy to achieve the EE limit from a degenerated version of problem (14).

With a two-state policy, which is a special QSI-dependent policy, the necessary condition in (7) reduces to

$$\begin{aligned}\mathbb{E}\{a(t)\} &= \mathbb{E}_{Q^\infty,h}\{s(t)\} = \Pr(Q^\infty > 0)\mathbb{E}_h\left\{s(t)\,|Q^\infty > 0\right\} \\ &= \eta\mathbb{E}_h\left\{s(t)\,|Q^\infty > 0\right\},\end{aligned} \tag{26}$$

from which we obtain the non-empty probability of the buffer as $\eta = \Pr(Q^\infty > 0) = \frac{\mathbb{E}\{a(t)\}}{\mathbb{E}_h\{s(t)|Q^\infty>0\}}$. Then, the average total power consumption of the two-state policy can be expressed as

$$\mathbb{E}_{Q^\infty,h}\{P_{tot}\} = \eta\bar{P}^{\text{on}}_{tot} + (1-\eta)\bar{P}^{\text{off}}_{tot}, \tag{27}$$

where the average is taken over both queue length and instantaneous channel power gain, and $\bar{P}^{\text{on}}_{tot}$ and $\bar{P}^{\text{off}}_{tot}$ are the average total power consumptions in "ON" and "OFF" states, respectively.

Further considering (11), the average total power consumption can be obtained as,

$$\mathbb{E}_{Q^\infty,h}\{P_{tot}\} = \frac{\mathbb{E}\{a(t)\}}{\mathbb{E}_h\{s(t)|Q^\infty > 0\}}\left(\frac{\bar{P}^{\text{on}}_T}{\rho} + P_{CW}W^{\text{on}}\right) + P_0. \tag{28}$$

The optimal two-state policy that minimizes $\mathbb{E}_{Q^\infty,h}\{P_{tot}\}$ under the statistical QoS constraint reflected by θ^c can be found from the degenerated version of problem (14) without the bandwidth constraint, which is

$$\min_{\bar{P}^{\text{on}}_T, W^{\text{on}}} \frac{\frac{\bar{P}^{\text{on}}_T}{\rho} + P_{CW}W^{\text{on}}}{\mathbb{E}_h\{s(t)|Q^\infty > 0\}}, \tag{29}$$

$$\text{s.t.} \quad E_C(\theta^c) \geq E_B(\theta^c), \quad \forall Q^\infty > 0. \tag{29a}$$

Again, the effective capacity $E_C(\theta^c)$ for $Q^\infty > 0$ depends on the instantaneous power allocation policy with given average transmit power $\bar{P}_T^{\rm on}$ and bandwidth $W^{\rm on}$. This suggests that the optimal two-state policy also includes an instantaneous power allocation implicitly except the average transmit power and bandwidth allocation.

The objective function in (29) has the same form as the objective function in (20). Thus, if the constraint in (29a) is not considered, the minimum of (29) can be achieved by the instantaneous power policy in (16) that minimizes the objective function in (20), and the average transmit power and bandwidth in the "ON" state similar to that in (22),

$$\begin{aligned}\bar{P}_T^{\rm on^*} &= \frac{\bar{P}_{TW}^*}{\bar{R}_W(\bar{P}_{TW}^*)}\mathbb{E}_h\{s(t)|Q^\infty>0\}\\ W^{\rm on^*} &= \frac{1}{\bar{R}_W(\bar{P}_{TW}^*)}\mathbb{E}_h\{s(t)|Q^\infty>0\}.\end{aligned} \tag{30}$$

Substituting (30) into (28), we can derive that $\mathbb{E}_{Q^\infty,h}\{P_{tot}\}$ is the same as the power limit in (25). This indicates that the power limit can be achieved by using $\bar{P}_T^{\rm on^*}$ and $W^{\rm on^*}$.

Now we find the two-state policy that can achieve the power limit under the statistical QoS requirement in (29a). To this end, we need to derive the effective capacity. With the water-filling policy in the "ON" state, the effective capacity can be expressed as follows,

$$\begin{aligned}&E_C\left(\theta^c\right)\\ &= -\frac{\ln\int_0^\infty\left[1+\frac{\mu f_p^w(\bar{P}_{TW}^* W^{\rm on^*},W^{\rm on^*},g)g}{N_0 W^{\rm on^*}}\right]^{-\beta} f_h\left(g\right)dg}{\theta_k^c T_c},\end{aligned} \tag{31}$$

where $\beta = \frac{\theta_k^c T_c W^{\rm on^*}}{\ln 2}$. It is not hard to see (31) increases with $W^{\rm on^*}$. Substituting (31) into $E_C\left(\theta^c\right) = E_B\left(\theta^c\right)$, the minimal bandwidth required to ensure the statistical QoS requirement, $\min(W^{\rm on^*})$, can be obtained numerically.

Remark 3. From $E_C\left(\theta^c\right) = E_B\left(\theta^c\right)$, we can show that $\min(W^{\rm on^*})$ increases with θ^c. In other words, in order to achieve the EE limit with a shorter delay bound $D_k^{\max}$, the required minimal bandwidth in the "ON" state increases. This essentially reflects an EE-delay-bandwidth tradeoff.

From $\min(W^{\rm on^*})$, we can obtain the minimal average transmit power required to achieve the power limit under the statistical QoS requirement, which is $\min(\bar{P}_T^{\rm on^*}) = \min(W^{\rm on^*})\bar{P}_{TW}^*$. Then, the instantaneous power allocation to achieve the power limit under the QoS requirement is,

$$p_l^* = \begin{cases}\frac{N_0\min(W^{\rm on^*})}{\mu}\left(\frac{1}{g^{\rm th}}-\frac{1}{g_l}\right) \triangleq p_l^{\rm on^*}, g_l \geq g^{\rm th}, Q^\infty>0,\\ 0, \qquad \text{otherwise},\end{cases} \tag{32}$$

which depends on both QSI and CSI, and $p_l^{\rm on^*}$ is in the form of water-filling, where $g^{\rm th}$ can be obtained from $\int_{g^{\rm th}}^\infty \frac{N_0}{\mu}\left(\frac{1}{g^{\rm th}}-\frac{1}{g}\right)f_h(g)dg = \bar{P}_{TW}^*$.

Remark 4. $p_l^{\rm on^*}$ in (32) is different from the optimal power allocation in [14, 39] that maximizes effective capacity with given $\mathbb{E}_h\{P_T\}$ and W (and hence can maximize the ratio of effective capacity to $\mathbb{E}_h\{P_{tot}\}$). When using the power allocation in [39], the average transmit power in the "ON" state of the queue is less than $p_l^{\rm on^*}$ under the same constraint $E_C(\theta^c) \geq E_B(\theta^c)$. However, minimizing the average transmit power in the "ON" state (i.e., $\mathbb{E}_h\{P_T\}$) is not equivalent to minimizing $\mathbb{E}_{Q^\infty,h}\{P_{tot}\}$ (with expression in (28)). By using the optimal two-state policy (i.e., $\min(W^{\rm on^*}), \min(\bar{P}_T^{\rm on^*}), p_l^*$), $\mathbb{E}_{Q^\infty,h}\{P_{tot}\}$ can be minimized. Given $\bar{P}_T^{\rm on}$ and $W^{\rm on}$, the water-filling power allocation in the "ON" state $p_l^{\rm on^*}$ can maximize $\mathbb{E}_h\{s(t)|Q^\infty>0\}$, and hence can minimize $\Pr(Q^\infty > 0) = \frac{\mathbb{E}\{a(t)\}}{\mathbb{E}_h\{s(t)|Q^\infty>0\}}$ with given $\mathbb{E}\{a(t)\}$.

B. Boundary Between the Tradeoff and Non-tradeoff Regions

If $\min(W^{\rm on^*})$ exceeds the maximal bandwidth $W^{\max}$, then the feasible solution of problem (29) cannot achieve the EE limit. This indicates that the required delay bound lies in the non-tradeoff region of the EE-delay relation when $\min(W^{\rm on^*}) \leq W^{\max}$. To obtain the closed-form expression of the boundary between the tradeoff and non-tradeoff regions, we consider the large N_T asymptotics in the rest of this subsection.

When $N_T \to \infty$, the service rate in (9) can be re-expressed as follows [40],

$$s(t) = W\log_2\left(1+\frac{\mu N_T \bar{P}_T}{N_0 W}\right) \triangleq C, \tag{33}$$

which can be achieved by the MRT when $N_T \gg K$ [40]. Due to channel hardening, the transmit power $p_l = \bar{P}_T$, and the service rate is constant when the transmit power and bandwidth are given. Then, $E_C(\theta) = \lim_{t\to\infty} -\frac{1}{\theta^c t}\ln\mathbb{E}\left\{\exp\left[-\theta^c t C\right]\right\} = C$. Therefore, the constraint of problem (29) degenerates into

$$C^{\rm on} \geq E_B(\theta), \tag{34}$$

where $C^{\rm on} \triangleq \mathbb{E}_h\{s(t)|Q^\infty>0\} = W^{\rm on}\log_2\left(1+\frac{\mu N_T \bar{P}_T^{\rm on}}{N_0 W^{\rm on}}\right)$.

After substituting $\mathbb{E}_h\{s(t)|Q^\infty>0\} = C^{\rm on}$, the power limit achieving average transmit power and bandwidth allocation policy in (30) can be simplified as $\bar{P}_T^{\rm on^*} = \bar{P}_{TW}^* C^{\rm on}/\bar{R}_W(\bar{P}_{TW}^*)$ and $W^{\rm on^*} = C^{\rm on}/\bar{R}_W(\bar{P}_{TW}^*)$. To ensure (34), the minimal bandwidth can be expressed as

$$\min(W^{\rm on^*}) = E_B(\theta)/\bar{R}_W(\bar{P}_{TW}^*). \tag{35}$$

To achieve the EE limit, $\min(W^{\rm on^*}) \leq W^{\max}$ should be satisfied. Thus, if the effective bandwidth of a certain arrival process satisfies $E_B(\theta)/\bar{R}_W(\bar{P}_{TW}^*) = \min(W^{\rm on^*}) \leq W^{\max}$, then the EE limit can be achieved and the EE-delay tradeoff will vanish.

To help understand the impacting factors on the boundary, in the sequel we provide a closed-form expression of the boundary by taking compound Poisson arrival process as an example. The inter-arrival time interval between packets and the packet size for the compound Poisson arrival process are exponential distributed with parameters λ^a and λ^u, respectively. The effective bandwidth of this arrival process can be

expressed as follows [41],

$$E_B(\theta^c) = \frac{\lambda^a}{\lambda^u - \theta^c}, \theta^c < \lambda^u. \tag{36}$$

At the boundary of the non-tradeoff region, the QoS exponent and the related delay bound requirement are denoted as $\theta^{\rm th}$ and $D_{\rm th}^{\max}$. Then, $\theta^{\rm th}$ can be obtained from $E_B(\theta)/\bar{R}_W^*(\bar{P}_{TW}^*) = W^{\max}$. From (3), the delay bound can be obtained as $D_{\rm th}^{\max} = \frac{\ln(1/\varepsilon_D)}{\theta^{\rm th} E_B(\theta^{\rm th})}$. For compound Poisson process, we have

$$D_{\rm th}^{\max} = \frac{(\lambda^u - \theta^{\rm th})\ln(1/\varepsilon_D)}{\lambda^a \theta^{\rm th}},$$
$$\theta^{\rm th} = \lambda^u - \frac{\lambda^a}{W^{\max}\log_2\left(1 + \frac{\mu N_T}{N_0}\bar{P}_{TW}^*\right)}. \tag{37}$$

As shown in (37), $D_{\rm th}^{\max}$ decreases with $\theta^{\rm th}$, and $\theta^{\rm th}$ increases with $W^{\max}$ and N_T. Therefore, by increasing $W^{\max}$ or N_T, the non-tradeoff region of the EE-delay relation can be widened. Again, this comes from the *EE-delay-bandwidth tradeoff* (as mentioned in Remark 3), or an *EE-delay-antenna tradeoff*. It suggests that the EE of a system will not reduce when supporting delay sensitive services with respective to the delay tolerant service if the bandwidth and/or the number of antennas can be increased.

C. EE-delay Relation in the Strictly Convex Case

If the delay bound required by a service is stringent such that $\min(W^{\rm on^*}) \geq W^{\max}$, then in order to guarantee $E_C(\theta^c) \geq E_B(\theta^c)$ the maximal bandwidth constraints in (14b) will be active, i.e., $W^{\rm on^*} = W^{\max}$. With given bandwidth, the effective capacity increases with average transmit power, and hence the minimal average transmit power $\min(\bar{P}_T^{\rm on^*})$ can be obtained from $E_C(\theta^c) = E_B(\theta^c)$ numerically.

To obtain closed-form results, we consider the large N_T asymptotics again in this subsection. When $D^{\max} < D_{\rm th}^{\max}$, $W^{\rm on^*} = W^{\max}$. As shown in what follows, $C^{\rm on}$ lies in the strictly convex region of the power-rate relation.

1) Power-rate Relation: Denote the service rate when $W = W^{\max}$ and $\bar{P}_T = \bar{P}_{TW}^* W^{\max}$ as $C_{\rm th} \triangleq W^{\max}\log_2\left(1 + \frac{\mu N_T}{N_0}\bar{P}_{TW}^*\right)$. To support higher service rate, the system needs to increase transmit power. From the maximum achievable rate in (33), we can obtain the minimal transmit power to support a service rate C that is higher than $C_{\rm th}$ as $\bar{P}_T^* = \frac{N_0 W^{\max}}{\mu N_T}\left(2^{C/W^{\max}} - 1\right)$. Substituting $\bar{P}_T^*$ and $W^* = W^{\max}$ into (11), the power-rate relation can be obtained as

$$P_{tot}^{\min}(C) = \frac{N_0 W^{\max}}{\rho \mu N_T}\left(2^{C/W^{\max}} - 1\right) + P_{CW} W^{\max} + P_0, \tag{38}$$

which is strictly convex in C. With the *strictly convex* power-rate relation, the corresponding maximal EE increases with the required delay bound, i.e., the EE can be traded off by delay.

2) Lower Bound of the Pareto Optimal EE-delay Tradeoff: In the following, we find the maximal EE achieved by a two-state policy, which can serve as a lower bound of the EE-delay relation. We first find the optimal two-state policy from problem (14) by setting $K = 1$. To gain useful insight, we again take the compound Poisson arrival process as an example.

To obtain the two-state policy, we need the following proposition (proved in Appendix D).

Proposition 4. Given the value of $W^{\rm on}$ (or $\bar{P}_T^{\rm on}$), the average total power consumption in (28) first decreases and then increases with $\bar{P}_T^{\rm on}$ (or $W^{\rm on}$), and achieves a unique minimal value at $\bar{P}_T^{\rm th}$ (or $W^{\rm th}$). Moreover, $\bar{P}_T^{\rm th} = \bar{P}_{TW}^* W^{\rm on}$ (or $W^{\rm th} = \bar{P}_T^{\rm on}/\bar{P}_{TW}^*$).

From Proposition 4, we can obtain the following corollary (proved in Appendix E).

Corollary 2. If $D^{\max} < D_{\rm th}^{\max}$, the optimal bandwidth in "ON" state will be $W^{\rm on^*} = W^{\max}$.

Since $W^{\rm on^*} = W^{\max}$, we only need to solve $\bar{P}_T^{\rm on^*}$. As shown in Proposition 4, the average total power consumption increases with $\bar{P}_T^{\rm on}$ when $\bar{P}_T^{\rm on} > \bar{P}_{TW}^* W^{\max}$. Besides, $C^{\rm on}$ also increases with $\bar{P}_T^{\rm on}$. For compound Poisson process, to minimize the average power consumption and satisfy the constraint $C^{\rm on} \geq \frac{\lambda^a}{\lambda^u - \theta^c}$, $\bar{P}_T^{\rm on^*}$ can be obtained from $C^{\rm on} = \frac{\lambda^a}{\lambda^u - \theta^c}$ as

$$\bar{P}_T^{\rm on^*} = \frac{N_0 W^{\max}}{\mu N_T}\left[2^{\frac{\lambda^a}{(\lambda^u - \theta^c)W^{\max}}} - 1\right]. \tag{39}$$

Then, the optimal two-state policy employs the transmit power in (39) and full bandwidth $W^{\max}$ in "ON" state. Substituting (36) into (3), the required QoS exponent can be obtained as $\theta^c = \frac{\lambda^u \ln(1/\varepsilon_D)}{\lambda^a D^{\max} + \ln(1/\varepsilon_D)}$. Upon substituting θ^c into (39), and substituting (39) and $W^{\rm on^*} = W^{\max}$ into (28), the average power consumption of the two-state policy can be obtained as follows,

$$\mathbb{E}_{Q^\infty}\{P_{tot}^*\} = \frac{\frac{W^{\max}N_0}{\rho\mu N_T}\left\{e^{\frac{\lambda^a D^{\max} + \ln(1/\varepsilon_D)}{W^{\max}\lambda^u D^{\max}}\ln 2} - 1\right\} + P_{CW}W^{\max}}{1 + \ln(1/\varepsilon_D)/(\lambda^a D^{\max})} + P_0, \tag{40}$$

where $\mathbb{E}\{a(t)\} = \frac{\lambda^a}{\lambda^u}$ and $\mathbb{E}_h\{s(t)|Q^\infty > 0\} = \frac{\lambda^a}{\lambda^u - \theta^c}$ are applied. Considering the EE definition in (6), the lower bound of the EE-delay relation can be obtained as $EE^{LB} = \frac{\lambda^a}{\lambda^u \mathbb{E}_{Q^\infty}\{P_{tot}^*\}}$.

VI. EE-Delay Relation in Multi-User Scenario

As in the single user scenario, we first consider the case without the bandwidth constraint and provide the EE limit that is an upper bound of the EE-delay relation $EE^{\max}(\theta_1^c, \theta_2^c, ..., \theta_K^c)$. Then, we consider the case with the bandwidth constraint and study the maximal EE achieved by a two-state policy, which is a lower bound of $EE^{\max}(\theta_1^c, \theta_2^c, ..., \theta_K^c)$. Finally, the boundary between EE-delay tradeoff and non-tradeoff regions is briefly discussed.

A. EE Limit

When the required delay bounds are large such that the constraint on $W^{\max}$ is not active, the transmit policy for each user does not affect the policies for others. Then, the multi-user system can be decomposed into K independent single user systems, and the power limit can be directly extended from (25) as, $\mathbb{E}_h\{P_{tot}^{\lim}\} = \sum_{k=1}^{K}\left[\frac{\bar{P}^*_{TW_k}}{\rho\bar{R}_{W_k}(\bar{P}^*_{TW_k})}+\frac{P_{CW}}{\bar{R}_{W_k}(\bar{P}^*_{TW_k})}\right]\mathbb{E}\{a_k(t)\}+P_0$, where $\bar{P}^*_{TW_k}$ and $\bar{R}_{W_k}(\bar{P}^*_{TW_k})$ are the optimal average transmit power and the average service rate per unit bandwidth of the kth user, respectively.

It is not hard to show that the power-rate relation is linear. According to Proposition 1, the EE-delay relation always equals to the EE limit, which is $EE^{\lim} = \sum_{i=1}^{M}\mathbb{E}\{a_k(t)\}/\mathbb{E}_h\{P_{tot}^{\lim}\}$.

B. A Two-state Policy and the Boundary of Non-tradeoff Region

In what follows, we first provide a two-state policy with the bandwidth constraint, from which the lower bound of the Pareto optimal EE-delay tradeoff can be obtained. Since the transmit policy of each user has two QSI states, the policy of the system with K users has 2^K states. To derive the average power consumption, we need to obtain the power consumed in each of the 2^K states and the probability that the system stays in each state. As a result, it is rather involved to find the optimal two-state policy from problem (14) when K is large. To tackle this difficulty, we find a policy by exploiting the structure of EE limit-achieving policy.

As shown in (30), the EE limit can be achieved with $\bar{P}_T^{\text{on}^*} = \bar{P}^*_{TW}W^{\text{on}^*}$. Recalling from Proposition 4 that the average power consumption increases with $\left|\bar{P}_T^{\text{on}} - \bar{P}^*_{TW}W^{\text{on}}\right|$ in large N_T asymptotic. Intuitively, the average total power consumption in multi-user case will increase with $\sum_{k=1}^{K}\left(\bar{P}_{T_k}^{\text{on}} - \bar{P}^*_{TW_k}W_k^{\text{on}}\right)^2$.[6] Hence, we formulate the optimization problem to find the two-state policy as follows,

$$\min_{\substack{\bar{P}_{T_k}^{\text{on}},W_k^{\text{on}}\\ k=1,2,\ldots,K}} \sum_{k=1}^{K}\left(\bar{P}_{T_k}^{\text{on}} - \bar{P}^*_{TW_k}W_k^{\text{on}}\right)^2 \quad (41)$$
$$\text{s.t.} \quad (14a) \text{ and } (14b).$$

The objective function in problem (41) is convex. From the discussion in [12], it is not hard to know that $E_{C_k}(\theta_k^c)$ is jointly concave in average transmit power and bandwidth, and hence (14a) is convex. Furthermore, the constraint in (14b) are linear. Therefore, the problem is convex, whose optimal solution can be solved with standard tools such as the interior-point method [38]. Due to the same reason as in the single user case, the optimal two-state policy also implicitly includes an instantaneous power allocation as in (32) except the average transmit power and bandwidth allocation.

When the constraint in (14b) is inactive, the problem can be decoupled into several single-user problems, and the minimum of the objective function in (41) is zero, which can be achieved with $\bar{P}_{T_k}^{\text{on}^*} = \bar{P}^*_{TW_k}W_k^{\text{on}^*}, k = 1, 2, ..., K$. This indicates that the resulting two-state policy can achieve $EE^{\lim}$.

When the constraint in (14b) is active, the obtained two-state policy does not minimize the average power consumption, and hence the EE achieved by the policy is a lower bound of the EE-delay relation.

For a set of requirements $(\theta_1^c, \theta_2^c, ..., \theta_K^c)$, if the solution of problem (41) satisfies $\sum_{k=1}^{K}W_k^{\text{on}^*} < W^{\max}$, the bandwidth constraint will not be active, and then $EE^{\lim}$ can be achieved by the two-state policy. This set of requirements are in the *non-tradeoff region*.

For a set of requirements $(\theta_1^c, \theta_2^c, ..., \theta_K^c)$, if the solution of problem (41) is obtained when the constraint is active, i.e., $\sum_{k=1}^{K}W_k^{\text{on}^*} = W^{\max}$, then the objective function in (41) is positive. Such a set of requirements are in the *tradeoff region*.

VII. Simulation and Numerical Results

In this section, we first validate that the statistical QoS requirement can be satisfied with the two-state policy in practical systems by comparing simulation and numerical results. Then, we demonstrate the power-rate relation and the EE-delay curve, and show the EE-delay curve when the number of transmit antennas are jointly adjusted with the transmit power and bandwidth via numerical results.

We take the MIMO-OFDMA system as an example to illustrate the EE-delay relation, where the subcarrier separation is $W_S = 15$ kHz. We consider K symmetric users served by one BS, all have the same QoS requirement and the same distance from the BS, since there is no fundamental difference for more general multi-user scenarios. Frequency-selective channel is considered in the simulation with i.i.d. channel gains on all the subcarriers. Denote the number of subcarriers allocated to the kth user as N_{S_k}. Since a symmetric scenario is considered, the K-user problem can be decomposed into K single-user problems with constraint on the number of subcarriers $N_{S_k} \le W^{\max}/(W_S K)$. The power-rate relation is obtained by solving problem (12), where $W_k = W_S N_{S_k}$. The two-state policy and the EE achieved by it are obtained by solving problem (14), where $W_k^{\text{on}^*} = W_S N_{S_k}^{\text{on}^*}$. The methods to solve problem (12) and problem (14) over frequency-selective channel are similar to that over flat fading channel. The only difference lies in the expression of $s_k(t)$, which is $s_k(t) = \sum_{j=1}^{N_{S_k}} W_S \log_2\left(1+\frac{\mu_k p_{k,l,j} g_{k,l,j}}{N_0 W_S}\right), (l-1)T_c < t \le lT_c$. The packet arrival process of each user is a compound Poisson random process with average packet arrival rate λ^a and average packet size $1/\lambda^u$. The circuit power model and parameters in [16] are used, where the circuit power per unit bandwidth

[6] $\left|\bar{P}_{T_k}^{\text{on}} - \bar{P}^*_{TW_k}W_k^{\text{on}}\right|$ is not differentiable. Therefore, we employ $\sum_{k=1}^{K}\left(\bar{P}_{T_k}^{\text{on}} - \bar{P}^*_{TW_k}W_k^{\text{on}}\right)^2$ as the objective function for mathematical tractability, which will not change the optimal solution.

P_{CW} and the fixed circuit power consumption P_0 increase with the number of transmit antennas N_T. The parameters used in the following simulation and numerical results are listed in Table I. Unless otherwise specified, the above setup will be used in the sequel.

TABLE I
PARAMETERS [11, 16]

Maximal available bandwidth $W^{\max}$	20 MHz
Efficiency of power amplifier ρ	38 %
Circuit power per unit bandwidth P_{CW}	$(72N_T + N_T^2)$ mW/MHz
Rate-independent circuit power P_0	$2N_T + 1$ W
Distance between users to BS d	200 m
Path loss model (dB)	$35.3 + 37.6 \log_{10} d$
Power spectral density of noise N_0	-174 dBm/Hz

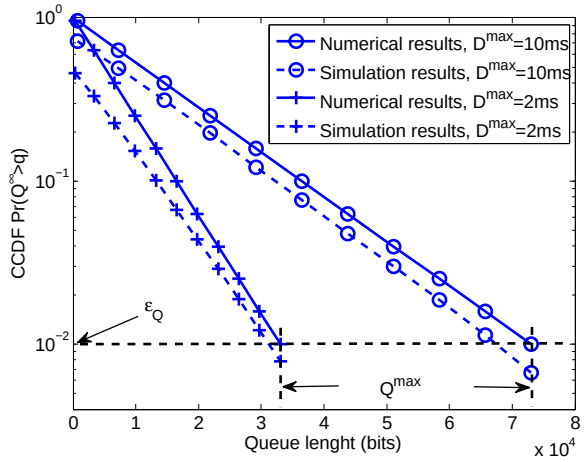


Fig. 1. Complementary CDF of the queue length, $K = 10$, $N_T = 4$, $\lambda^a = 1000$ packets/s, $1/\lambda^u = 5$ kbits, and TTI =1 ms.

Figure 1 shows the complementary *cumulative distributed function* (CDF) of Q^{∞}. As discussed in [42, 43], the QoS requirement $(D^{\max}, \varepsilon_D)$ is equivalent to $(Q^{\max}, \varepsilon_Q)$, where $Q^{\max} = E_B(\theta^c)D^{\max}$ and $\varepsilon_Q = \varepsilon_D$. The value of θ^c can be obtained from $\exp\left[-\theta^c E B(\theta^c) D^{\max}\right] = \varepsilon_D$ [31]. The simulation results are obtained by computing the queue length in every TTI (which is 1 ms [11]) during the simulation time of 100 s, where the compound Poisson arrival process is served by the two-state policy. The numerical results are obtained from $\Pr\left(Q^{\infty} > q\right) \leq \exp\left(-\theta^c q\right)$ [31], where $\eta \leq 1$ is applied. It is shown that the numerical results are upper bounds of simulation results. This indicates that the statistical QoS requirement can be guaranteed by using the two-state policy, which validates Remark 1. Besides, the results show that when $D^{\max} = 2$ ms, the maximum queue length, $Q^{\max} = E_B(\theta^c)D^{\max}$, is large enough such that the exponential decay rate of $\Pr\left(Q^{\infty} > q\right)$ is close to θ^c for the compound Poisson process.

Figure 2 shows the power-rate relation and the EE-delay curves achieved by the two-state policy. For comparison, the EE-delay curves achieved by two existing QSI-independent policies are also provided, which consider statistical QoS requirement. The first one is the policy in [5], where the total transmit power of the K users is minimized under constraints $E_{C_k}(\theta_k^c) \geq E_{B_k}(\theta_k^c)$ and $N_{S_k} \leq W^{\max}/(W_S K)$ (with legend "power minimizing policy"). The second is the policy in [12], where the ratio of $\sum_{k=1}^{K} E_{C_k}(\theta_k^c)$ to the total power consumption is maximized under the same constraints as in [5] (with legend "delay-sensitive EE policy"). Since the results in [13, 14] are similar to that in [12], we do not compare with [13, 14].

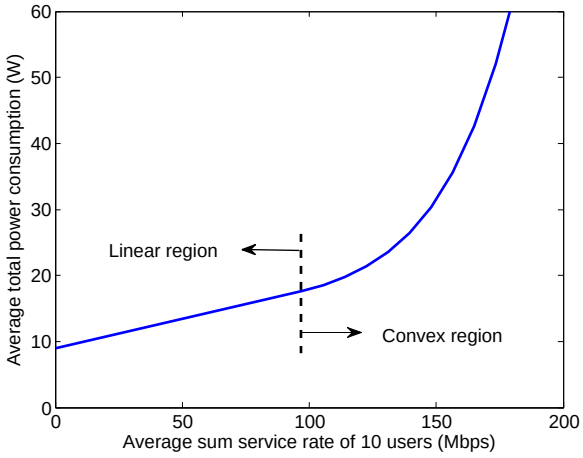
The power-rate relation in Fig 2(a) includes two segments, a linear region and a convex region. In the linear region, the optimal policy to support $\mathbb{E}\{s(t)\} = \bar{s}$ satisfies $\bar{P}_{T_k}^* = \bar{P}_{TW}^* W_S N_{S_k}^* < \bar{P}_{TW}^* W^{\max}/K$, and the bandwidth constraint is inactive. In the convex region, $\bar{P}_{T_k}^* > \bar{P}_{TW}^* W^{\max}/K$ and $W_S N_{S_k}^* = W^{\max}/K$. In the boundary of linear and convex regions, $\bar{P}_{T_k}^* = \bar{P}_{TW}^* W^{\max}/K$ and $W_S N_{S_k}^* = W^{\max}/K$.

Each EE-delay curve achieved by the two-state policy in Fig. 2(b) includes a tradeoff region and a non-tradeoff region. The vertical dash line is the boundary of these two regions. In the non-tradeoff region, $P_{T_k}^{\text{on}^*} = P_{TW}^* W_S N_{S_k}^{\text{on}^*}$, corresponding to the linear region in Fig. 2(a). In the tradeoff region, $P_{T_k}^{\text{on}^*} > P_{TW}^* W^{\max}/K$ and $W_S N_{S_k}^{\text{on}^*} = W^{\max}/K$, corresponding to the convex region in Fig. 2(a). The non-tradeoff region exists when the delay bound is large, within which supporting more stringent delay bound $D_k^{\max}$ does not reduce the EE, in contrast to the traditional belief. In the tradeoff region, small additional $D_k^{\max}$ leads to substantial EE increase, which is consistent with the results in priori studies [5, 7–10, 18].

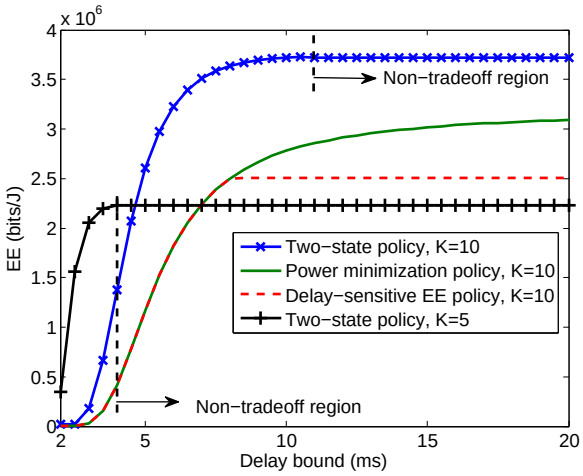
The EE achieved by the "power minimizing policy" always increases with the delay bound, which however is much lower than the EE achieved by the two-state policy. This is because the policy in [5] is QSI-independent (such that transmit power cannot be minimized) and only the transmission power and rate are optimized (such that the circuit power cannot be reduced). The EE achieved by the "power minimizing policy" is higher than the "delay-sensitive EE policy" when the required $D_k^{\max}$ is large, because the latter provides higher service rate than required by the QoS (i.e., $E_{C_k}(\theta_k) > E_{B_k}(\theta_k)$) that leads to a waste of energy. The EE achieved by the "delay-sensitive EE policy" also grows with the delay bound, though not obvious in the figure.

From Fig. 2(b), we can also observe the impact of the number of users. With more users, non-tradeoff region shrinks because the bandwidth for each user decreases.

Figure 3 illustrates what happens when the number of transmit antennas are jointly adjusted with the transmit power and bandwidth. To obtain the EE-delay curve, we find the two-state policy with different values of N_T, and then selecting N_T that maximizes the EE. The achieved EE decreases with N_T when the delay bound is large and increases with N_T when the delay bound is short. We can see that when N_T is jointly allocated with $P_{T_k}^{\text{on}}$ and $N_{S_k}^{\text{on}}$, the EE-delay curve also includes tradeoff and non-tradeoff regions, despite that the circuit power consumption is not a linear function of N_T as



(a) Power-rate relation, $K = 10$.



(b) EE-delay curves achieved by different policies for compound Poisson arrival process, $\varepsilon_D = 0.01$.

Fig. 2. Power-rate relation and EE-delay curves, $N_T = 4$, $\lambda^a = 500$ packets/s, $1/\lambda^u = 10$ kbits.

shown in Table I. Numerical results show that the power-rate relation is linear when the resource constraints are inactive, which is not provided for conciseness.

VIII. Conclusion

In this paper, we studied fundamental EE-delay relation for wireless systems serving randomly arrived data with statistical QoS requirements. Different from the widely accepted wisdom that there is an inherent tradeoff between the EE and delay, we showed that the EE-delay relation includes tradeoff region and non-tradeoff region. We illustrated that for a system with adjustable circuit power consumption, the required minimal average total power consumption may linearly increase with the average service rate. We proved that in such a *linear case*, the EE-delay tradeoff vanishes, and the EE-limit can be achieved by a two-state policy. By taking MIMO system as an example, we demonstrated when the linear case will occur. Specifically, we proved that the power-rate relation will be linear when the transmit power and bandwidth are jointly optimized in single-user scenario if the bandwidth constraint

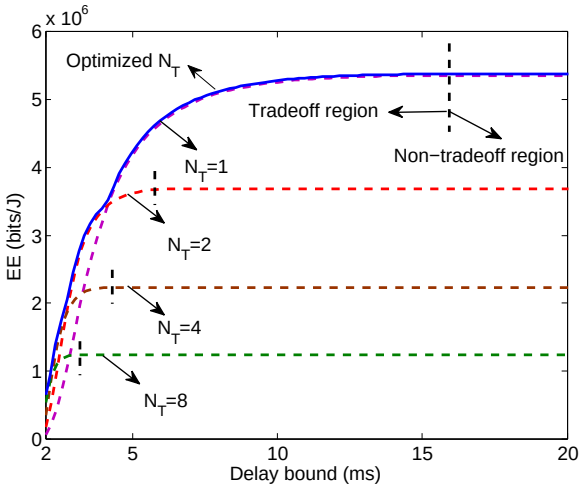


Fig. 3. EE-delay curves achieved by the two-state policy with optimized N_T (the solid line) and given values of N_T (the dash lines), $K = 5$, $\lambda^a = 500$ packets/s, $1/\lambda^u = 10$ kbits, and $\varepsilon_D = 0.01$.

is not active. We then obtained the EE-delay relation and the EE-limit achieving policy in the linear case. Furthering considering the compound Poisson arrival process in large number of transmit antennas asymptotics, the closed-form boundary between the tradeoff and non-tradeoff regions was derived, and a lower bound of the Pareto optimal EE-delay tradeoff in the nonlinear case was obtained. Finally, the extension to the multi-user scenario was briefly addressed. Analytical and numerical results showed that increasing the maximal bandwidth or the number of transmit antennas can broaden the non-tradeoff region, within which supporting more stringent delay bound does not reduce the EE. This comes from the EE-delay-bandwidth tradeoff or EE-delay-antenna tradeoff. Besides, the EE-delay curve achieved by the two-state policy in the tradeoff region, i.e., the lower bound, is much higher than those achieved by existing policies that also consider statistical QoS requirement.

Appendix A
Proof of EE limit (8)

Proof: For any QSI based transmit policy that satisfies the necessary condition in (7), the average total power consumption in linear case can be expressed as $P_{tot}^{LB} = \mathbb{E}_{\boldsymbol{Q}\infty}\{\mathbb{E}_h\{P_{tot}^{\min}\}\} = \mathbb{E}_{\boldsymbol{Q}\infty}\{\sum_{k=1}^{K} c_k \mathbb{E}_h[s_k(t)] + c_0\} = \sum_{k=1}^{K} c_k \mathbb{E}_{\boldsymbol{Q}\infty,h}\{s_k(t)\} + c_0 = \sum_{k=1}^{K} c_k \mathbb{E}\{a_k(t)\} + c_0$. Because a power limit achieving policy needs to further satisfy (5), $P_{tot}^{\lim}$ is no less than P_{tot}^{LB}. When $D_k^{\max} \to \infty$, $k = 1, 2, ..., K$, both the QoS constraint (5) and the necessary condition in (7) can be satisfied with a transmit policy that satisfies $\mathbb{E}_h\{s_k(t)\} = \mathbb{E}\{a_k(t)\}$, i.e., P_{tot}^{LB} is achievable. Therefore, $P_{tot}^{\lim} = P_{tot}^{LB}$, and $EE^{\lim} = \sum_{k=1}^{K} \mathbb{E}\{a_k(t)\} / (\sum_{k=1}^{K} c_k \mathbb{E}\{a_k(t)\} + c_0)$. □

APPENDIX B
PROOF OF PROPOSITION 1

Proof: For a two-state policy, if $Q_k(t)=0$, then $s_k(t)=0$. From (1), we have $b_k(t)=\min\{s_k(t),a_k(t)\}=s_k(t)$. If $Q_k(t)>0$, from (1), we also have $b_k(t)=s_k(t)$. Hence, with the two-state policy, $b_k(t)=s_k(t),\forall Q_k(t)$. In steady state, $\mathbb{E}\{b_k(t)\}=\mathbb{E}_{Q_k^\infty,h}\{s_k(t)\}$.

To guarantee the QoS requirement, $E_{C_k}(\theta_k)\geq E_{B_k}(\theta_k)$, $Q_k^\infty>0$ should be satisfied. The expectations of $s_k(t)$ under conditions $Q_k^\infty>0$ and $Q_k^\infty=0$ are denoted as $\mathbb{E}_h\{s_k(t)|Q_k^\infty>0\}$ and $\mathbb{E}_h\{s_k(t)|Q_k^\infty=0\}$, respectively.

Further considering that $\mathbb{E}\{a_k(t)\}=\mathbb{E}\{b_k(t)\}$ in steady state, we have

$$\mathbb{E}\{a_k(t)\}=\mathbb{E}\{b_k(t)\}=\mathbb{E}_{Q_k^\infty,h}\{s_k(t)\},k=1,2,...,K. \tag{B.1}$$

Denote $EE^{ts}(\theta_1^c,\theta_2^c,...,\theta_K^c)=\sum_{k=1}^{K}\mathbb{E}\{a_k(t)\}/\mathbb{E}_{\boldsymbol{Q}^\infty}\{\mathbb{E}_h(P_{tot}^{\min})\}$ as the EE achieved by the two-state policy. In the linear case, i.e., $\mathbb{E}_h\{P_{tot}^{\min}\}=\sum_{k=1}^{K}c_k\mathbb{E}_h\left\{s_k\left(t\right)\right\}+c_0$, the minimal average power consumption achieved by the two-state policy can be obtained from the following expression,

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{Q}^\infty}\{\mathbb{E}_h\{P_{tot}^{\min}\}\}\\
=&\Pr\left(Q_k^\infty>0\right)\mathbb{E}_{\mathbf{Q}^\infty}\{\mathbb{E}_h\{P_{tot}^{\min}\}|Q_k^\infty>0\}\\
&+\Pr\left(Q_k^\infty=0\right)\mathbb{E}_{\mathbf{Q}^\infty}\{\mathbb{E}_h\{P_{tot}^{\min}\}|Q_k^\infty=0\}\\
=&\sum_{k=1}^{K}c_k\Pr\left(Q_k^\infty>0\right)\mathbb{E}_h\left\{s_k\left(t\right)|Q_k^\infty>0\right\}\\
&+\sum_{k=1}^{K}c_k\Pr\left(Q_k^\infty=0\right)\mathbb{E}_h\left\{s_k\left(t\right)|Q_k^\infty=0\right\}+c_0\\
=&\sum_{k=1}^{K}c_k\mathbb{E}_{Q_k^\infty,h}\{s_k(t)\}+c_0=\sum_{k=1}^{K}c_k\mathbb{E}\{a_k(t)\}+c_0.
\end{aligned} \tag{B.2}$$

From (B.2), we can obtain that

$$EE^{ts}(\theta_1^c,\theta_2^c,...,\theta_K^c)=\sum_{k=1}^{K}\mathbb{E}\{a_k(t)\}/(\sum_{k=1}^{K}c_k\mathbb{E}\{a_k(t)\}+c_0), \tag{B.3}$$

which is exactly the same as $EE^{\lim}$ in (8). This indicates that a two-state policy can achieve $EE^{\lim}$ for arbitrary set of delay requirements $(\theta_1^c,\theta_2^c,...,\theta_K^c)\in\mathbb{R}_+^K$. Further considering that $EE^{ts}(\theta_1^c,\theta_2^c,...,\theta_K^c)\leq EE^{\max}(\theta_1^c,\theta_2^c,...,\theta_K^c)\leq EE^{\lim}$, the proposition is proved. □

APPENDIX C
PROOF OF COROLLARY 1

Proof: For a finite QoS exponent θ_k^c, the tail probability of the steady state queue length of the kth user Q_k^∞ can be approximated as $\Pr(Q_k^\infty>Q_{th})\approx e^{-\theta_k^cQ_{th}}$ [26]. Then, the average of Q_k^∞ is $\mathbb{E}\{Q_k^\infty\}=\int_0^\infty q d\Pr\left(Q_k^\infty\leq q\right)\approx\frac{1}{\theta_k^c}$. According to the Little's Law, the average delay of the kth user can be approximated as $\frac{1}{\theta_k^c\mathbb{E}\{a_k(t)\}}$. Therefore, to ensure the average delay $\bar{D}$, a transmit policy should satisfy $\theta_k^c=\frac{1}{\mathbb{E}\{a_k(t)\}\bar{D}}$. As shown in Proposition 1, the EE achieved by the two-state policy is independent of θ_k^c, and hence is independent of $\bar{D}$. □

APPENDIX D
PROOF OF PROPOSITION 4

Proof: Since $\mathbb{E}\{a(t)\}$ and P_0 are fixed, minimizing the average power consumption (28) is equivalent to minimizing the objective function in (29). In the sequel, we analyze *the reciprocal of the objective function* in (29), which is

$$\varphi\left(W^{\text{on}},\bar{P}_T^{\text{on}}\right)=\frac{W^{\text{on}}\log_2\left(1+\frac{\mu N_T\bar{P}_T^{\text{on}}}{N_0W^{\text{on}}}\right)}{\frac{\bar{P}_T^{\text{on}}}{\rho}+P_{CW}W^{\text{on}}}, \tag{E.1}$$

where (33) is used. In this appendix, we only prove the case when W^{on} is given. For the other case, i.e., $\bar{P}_T^{\text{on}}$ is given, the proof is similar and hence is omitted for conciseness.

To find the optimal $\bar{P}_T^{\text{on}}$ with given W^{on}, we take the first order derivative of (E.1) as

$$\frac{\partial\varphi\left(W^{\text{on}},\bar{P}_T^{\text{on}}\right)}{\partial\bar{P}_T^{\text{on}}}=\frac{\frac{W^{\text{on}}}{\rho\ln 2}\varphi_1\left(W^{\text{on}},\bar{P}_T^{\text{on}}\right)}{\left(\frac{\bar{P}_T^{\text{on}}}{\rho}+P_{CW}W^{\text{on}}\right)^2},$$

where $\varphi_1\left(W^{\text{on}},\bar{P}_T^{\text{on}}\right)=\frac{\left(\frac{\bar{P}_T^{\text{on}}}{W^{\text{on}}}+\rho P_{CW}\right)\frac{\mu N_T}{N_0}}{1+\frac{\mu N_T\bar{P}_T^{\text{on}}}{N_0W^{\text{on}}}}-\ln\left(1+\frac{\mu N_T\bar{P}_T^{\text{on}}}{N_0W^{\text{on}}}\right)$.

It is not hard to show that $\varphi_1\left(W^{\text{on}},0\right)>0$, $\varphi_1\left(W^{\text{on}},\infty\right)<0$, and

$$\frac{\partial\varphi_1\left(W^{\text{on}},\bar{P}_T^{\text{on}}\right)}{\partial\bar{P}_T^{\text{on}}}<\frac{\frac{\mu N_T}{N_0W^{\text{on}}}}{\left(1+\frac{\mu N_T\bar{P}_T^{\text{on}}}{N_0W^{\text{on}}}\right)^2}-\frac{\frac{\mu N_T}{N_0W^{\text{on}}}}{1+\frac{\mu N_T\bar{P}_T^{\text{on}}}{N_0W^{\text{on}}}}<0,$$

which means that $\varphi_1\left(W^{\text{on}},\bar{P}_T^{\text{on}}\right)$ strictly decreases with $\bar{P}_T^{\text{on}}$. Thus, the equation $\varphi_1\left(W^{\text{on}},\bar{P}_T^{\text{on}}\right)=0$ has a unique solution, denoted as $\bar{P}_T^{\text{th}}$.

If $P_T^{\text{on}}<P_T^{\text{th}}$, then $\varphi_1\left(W^{\text{on}},\bar{P}_T^{\text{on}}\right)>0$, and hence $\varphi\left(W^{\text{on}},\bar{P}_T^{\text{on}}\right)$ increases with $\bar{P}_T^{\text{on}}$. Since $\varphi\left(W^{\text{on}},\bar{P}_T^{\text{on}}\right)$ is the reciprocal of the objective function in (29), the average power consumption decreases with $\bar{P}_T^{\text{on}}$. Similarly, if $\bar{P}_T^{\text{on}}>\bar{P}_T^{\text{th}}$, the average power consumption increases with $\bar{P}_T^{\text{on}}$. Therefore, $\bar{P}_T^{\text{th}}$ is the global optimal value. Moreover, the optimal average transmit power per unite bandwidth is $\bar{P}_{TW}^*$. Therefore, we have $\bar{P}_T^{\text{th}}=\bar{P}_{TW}^*W^{\text{on}}$. □

APPENDIX E
PROOF OF COROLLARY 2

Proof: The delay requirement $D^{\max}<D_{\text{th}}^{\max}$ can not be guaranteed with any policy that satisfies $W^{\text{on}}\leq W^{\max}$ and $\bar{P}_T^{\text{on}}\leq\bar{P}_{TW}^*W^{\max}$ (otherwise, $D^{\max}$ lies in the non-tradeoff region, i.e., $D^{\max}\geq D_{\text{th}}^{\max}$). Then, to guarantee the delay requirement, $\bar{P}_T^{\text{on}}>\bar{P}_{TW}^*W^{\max}$, and hence $W^{\max}<P_T^{\text{on}}/P_TW^*$. According to Proposition 4, when $W^{\text{on}}\leq W^{\max}<\bar{P}_T^{\text{on}}/\bar{P}_{TW}^*$, the average total power consumption

decreases with $W^{\rm on}$. Hence for any given $P_T^{\rm on} > \bar{P}_{TW}^* W^{\max}$, the average total power consumption is minimized with $W^{\rm on} = W^{\max}$. Therefore, $W^{\rm on^*} = W^{\max}$. □

REFERENCES

[1] R. Q. Hu and Y. Qian, "An energy efficient and spectrum efficient wireless heterogeneous network framework for 5G systems," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 94–101, May 2014.

[2] G. Fettweis and S. Alamouti, "5G: Personal mobile internet beyond what cellular did to telephony," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 140–145, Feb. 2014.

[3] Y. Chen, S. Zhang, S.-G. Xu, and G. Y. Li, "Fundamental trade-offs on green wireless networks," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 30 – 37, Jun. 2011.

[4] Q. Wu, W. Chen, M. Tao, J. Li, H. Tang, and J. Wu, "Resource allocation for joint transmitter and receiver energy efficiency maximization in downlink OFDMA systems," *IEEE Trans. Commun.*, vol. 63, no. 2, pp. 416 – 430, Feb. 2015.

[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] Z. Niu, X. Guo, S. Zhou, and P. R. Kumar, "Characterizing energy-delay tradeoff in hyper-cellular networks with base station sleeping control," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 4, pp. 641–650, Apr. 2015.

[7] 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.

[8] R. A. Berry, "Optimal power-delay tradeoffs in fading channels—small-delay asymptotics," *IEEE Trans. Inf. Theory*, vol. 59, no. 6, pp. 3939–3952, Jun. 2013.

[9] E. Uysal-Biyikoglu, B. Prabhakar, and A. E. Gamal, "Energy-efficient packet transmission over a wireless link," *IEEE/ACM Trans. Networking*, vol. 10, no. 4, pp. 487 – 499, Aug. 2002.

[10] D. Rajan, A. Sabharwal, and B. Aazhang, "Delay-bounded packet scheduling of bursty traffic over wireless channels," *IEEE Trans. Inform. Theory*, vol. 50, no. 1, pp. 125–144, Jan. 2004.

[11] 3GPP, *Further Advancements for E-UTRA Physical Layer Aspects*. TSG RAN TR 36.814 v9.0.0, Mar. 2010.

[12] 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.

[13] 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.

[14] W. Cheng, X. Zhang, and H. Zhang, "Joint spectrum and power efficiencies optimization for statistical QoS provisionings over SISO/MIMO wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 5, pp. 903–915, May 2013.

[15] Z. Xu, C. Yang, G. Y. Li, S. Zhang, Y. Chen, and S. Xu, "Energy-efficient configuration of spatial and frequency resources in MIMO-OFDMA systems," *IEEE Trans. Commun.*, vol. 61, no. 2, pp. 564 – 575, Feb. 2013.

[16] B. Debaillie, C. Desset, and F. Louagie, "A flexible and future-proof power model for cellular base stations," in *Proc. IEEE VTC Spring*, 2015.

[17] S. Cui, A. J. Goldsmith, and A. Bahai, "Energy-constrained modulation optimization," *IEEE Trans. Wireless Commun.*, vol. 4, no. 5, pp. 2149–1621, Sep. 2005.

[18] M. J. Neely, "Optimal energy and delay tradeoffs for multi-user wireless downlinks," *IEEE Trans. Inf. Theory*, vol. 53, no. 9, pp. 3095–3113, Sep. 2007.

[19] M. A. Zafer and E. Modiano, "Minimum energy transmission over a wireless channel with deadline and power constraints," *IEEE Trans. Automatic Control*, vol. 54, no. 12, pp. 2841–2852, Dec. 2009.

[20] ——, "A calculus approach to energy-efficient data transmission with quality-of-service constraints," *IEEE/ACM Trans. Networking*, vol. 17, no. 3, pp. 898–911, Jun. 2009.

[21] Z. Nan, X. Wang, and W. Ni, "Energy-efficient transmission of delay-limited bursty data packets under non-ideal circuit power consumption," in *Proc. IEEE ICC*, 2014.

[22] X. Wang and Z. Li, "Energy-efficient transmissions of bursty data packets with strict deadlines over time-varying wireless channels," *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, pp. 2533–2543, May 2013.

[23] Y. Li, M. Sheng, Y. Shi, X. Ma, and W. Jiao, "Energy efficiency and delay tradeoff for time-varying and interference-free wireless networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 11, pp. 5921–5931, Nov. 2014.

[24] J. Choi, "Energy-delay tradeoff comparison of transmission schemes with limited CSI feedback," *IEEE Trans. Wireless Commun.*, vol. 12, no. 4, pp. 1762 – 1773, Apr. 2013.

[25] 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.

[26] 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.

[27] 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

[28] C. 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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

[33] M. C. Gursoy, D. Qiao, and S. Velipasalar, "Analysis of energy efficiency in fading channels under QoS constraints," *IEEE Trans. Wireless Commun.*, vol. 8, no. 8, pp. 4252–4263, Aug. 2009.

[34] C. S. Chang and T. Zajic, "Effective bandwidths of departure processes from queues with time varying capacities," in *Proc. IEEE INFOCOM*, Apr. 1995.

[35] A. Goldsmith, *Wireless Communications*. Cambridge University Press, 2005.

[36] I. E. Telatar, *Capacity of multi-antenna Gaussian channels*, 1995.

[37] C. She and C. Yang, "Context aware energy efficient optimization for video on-demand service over wireless networks," in *Proc. IEEE ICCC*, 2015.

[38] S. Boyd and L. Vandanberghe, *Convex Optimization*. Cambridge Univ. Press, 2004.

[39] J. Tang and X. Zhang, "Quality-of-service driven power and rate adaptation over wireless links," *IEEE Trans. Wireless Commun.*, vol. 6, no. 8, pp. 3058–3068, Aug. 2007.

[40] 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.

[41] F. Kelly, "Notes on effective bandwidths," *Stochastic networks: theory and applications*, 1996.

[42] C. She, C. Yang, and L. Liu, "Energy-efficient resource allocation for MIMO-OFDM systems serving random sources with statistical QoS requirement," *IEEE Trans. Commun.*, vol. 63, no. 11, pp. 4125–4141, Nov. 2015.

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