Optimal EE-Delay Relation in Wireless Systems

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

***Abstract*—It is widely accepted that a tradeoff exists between transmit power and average delay. In this paper, we consider wireless systems transmitting randomly arrived traffic over fading channels with statistical quality-of-service requirement, characterized by a delay bound and a delay bound violation probability. We study the relation between the maximal energy efficiency (EE) and the delay bound with given delay violation probability. We prove that the EE-delay tradeoff vanishes if the average total power consumption, including transmit and circuit powers of the base station, linearly increases with the average service/transmission rate. By taking massive multi-input-multi-output (MIMO) system as an example, we show that if the required total power consumption is a linear function of the service rate, the maximal EE is independent of the delay bound. If the required total power is strictly convex in the service rate, then the EE can be improved by extending the delay.**

I. Introduction

The fifth generation (5G) mobile networks are expected to support a variety of services with diverse quality-of-service (QoS) requirements [1]. To provide high throughput with satisfactory user experience meanwhile reduce the cost and global carbon dioxide emissions, energy efficiency (EE) has become one of the major design goals. To optimize towards several possibly conflicting performance metrics, the fundamental relations between these metrics need careful examination [2,3], among which the EE-delay relation has drawn significant attention, which is especially important for multimedia traffic with various delay-bounded QoS provisioning.

It is widely accepted that a tradeoff exists between transmit power and average delay (and therefore between EE and delay) [2,4–7]. This suggests that EE can be increased by extending the delay. According to Shannon's capacity, the transmit power is a strictly convex function of the transmission rate (also called service rate in the sequel) for a given channel state. Based on this fact, the pioneering study in [4] shows that the average transmit power and the average delay cannot be minimized at the same time, unless *both* source *and* channel are *not random*. When the QoS requirement is modeled by a hard deadline, the power-delay tradeoff also exists when transmit power is strictly convex in service rate [5]. An exceptional example in [6] demonstrates that if the required average transmit power is a piecewise linear function of the average rate, then the optimal power-delay tradeoff in [4] can be exceeded. These studies indicate that the power-delay tradeoff depends on the relation between the power and rate.

Compared with average delay and hard deadline, *statistical QoS requirement*, defined as a delay bound and a delay violation probability, is more relevant for wireless multimedia transmission [8]. In the context of statistical QoS requirement, the power-delay tradeoffs of several systems were studied in [7], where only transmit power was considered. However, the policies in [7] are independent of queue state information (QSI), and thus the achieved power-delay tradeoffs are not optimal for randomly arrived data, as implied in [4].

Furthermore, adjusting transmit power is the only power-saving mechanism considered in the prior studies. Noting that the power consumed for running the circuits is not negligible in prevalent systems, other power saving mechanisms become necessary [9]. For example, we can further adjust the bandwidth to reduce the circuit power after ensuring the QoS. As a consequence, the relation between average total power consumption and average service rate (i.e., power-rate relation) and the resulting EE-delay relation may change.

In this paper, we investigate the optimal EE-delay relation for wireless systems serving the traffic with random arrivals and statistical QoS requirements, where both transmit and circuit powers are taken into account. We show that for any system if the required average total power consumption linearly increases with the average service rate, the maximal EE is independent of the delay bound and can be achieved by a simple two-state policy. To demonstrate whether such a system exists in practice, we take massive multi-input-multi-output (MIMO) system as a concrete example. Our analyses show that the maximal EE cannot be traded off by the delay bound when the required service rate lies in the linear region of the power-rate relation. When the required service rate lies in the strictly convex region of the power-rate relation, which happens when the delay bound is stringent, there exists a tradeoff between the maximal EE and the delay bound.

II. Statistical QoS Requirement and Definitions

A. Queueing Model and Statistical QoS Requirement

Consider a downlink multiuser system, where a base station (BS) serves K users with delay-sensitive services. The statistical QoS requirement of the kth user is defined as $(D_k^{\max}, \varepsilon_k)$, where $D_k^{\max}$ is the delay bound and ε_k is the delay bound violation probability allowed by the service, $k = 1, ..., K$. In this paper, we consider the queuing delay and ignore the coding and transmission delay as in [2,4–7].

As in [10], we consider a fluid queueing model. Such a mode is accurate when the inter-arrival time between packets of the data source and the time interval for a system to update transmit policy is much shorter than $D_k^{\max}$. At time t, the data of the kth user enters a first-in-first-out buffer of the BS at arrival rate $a_k(t)$, and is sent to the user at departure rate $b_k(t)$, which are related as

$$b_k\left(t\right) = \left\{ \begin{array}{ll} \min\{a_k\left(t\right), s_k(t)\}, & Q_k\left(t\right) = 0 \\ s_k\left(t\right), & Q_k\left(t\right) > 0 \end{array} \right., \quad (1)$$

where $s_k(t)$ is the service/transmission rate and $Q_k(t)$ is the queue length of the kth user. When the buffer size is infinite, $Q_k(t) = \int_0^t a_k\left(\tau\right) - b_k\left(\tau\right) d\tau$. We assume that the queues are in steady state when $t > 0$, and denote Q_k^∞ as the steady state queue length of the kth user.

Effective bandwidth [11] and effective capacity [12] are useful tools to study statistical QoS requirement. The effective bandwidth of arrival process $\{a_k(t), t > 0\}$ and the effective capacity of service process $\{s_k(t), t > 0\}$ are respectively defined as [11, 12],

$$E_{B_k}\left(\theta_k\right) = \lim_{t\to\infty} \frac{1}{\theta_k t} \ln \mathbb{E}\left[e^{\theta_k \int_0^t a_k(\tau) d\tau}\right], \quad (2)$$

$$E_{C_k}\left(\theta_k\right) = -\lim_{t\to\infty} \frac{1}{\theta_k t} \ln \mathbb{E}\left[e^{-\theta_k \int_0^t s_k(\tau) d\tau}\right], \quad (3)$$

where θ_k is the QoS exponent. To guarantee statistical QoS requirement $(D_k^{\max}, \varepsilon_k)$, θ_k should satisfy [3]

$$\Pr\left(D_k^\infty > D_k^{\max}\right) \approx \exp\left[-\theta_k E_{B_k}(\theta_k) D_k^{\max}\right] = \varepsilon_k, \quad (4)$$

where D_k^∞ is the queueing delay when the queue stays in steady state. Since $E_{B_k}(\theta_k)$ increases with θ_k [11], given the delay violation probability ε_k, a large value of θ_k indicates a small $D_k^{\max}$, and *vice versa*. To ensure the statistical QoS requirement, a resource allocation policy should satisfy [10]

$$E_{C_k}(\theta_k) \geq E_{B_k}(\theta_k). \quad (5)$$

B. Several Definitions

For easy exposition, we first define several notions to be used throughout the paper.

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

To support the average service rates $\bar{s}_1, ..., \bar{s}_K$, the policy that minimizes average total power is independent of QSI [4]. Thus, the expectation is only taken over channel gain.

EE can be defined as the ratio of average throughput $\sum\limits_{k=1}^{K} \mathbb{E}[b_k(t)]$ to the average total power [13]. For ergodic arrival process and service process, when (5) is satisfied, $\mathbb{E}[a_k(t)] = \mathbb{E}[b_k(t)]$ (the proof is omitted due to the lack of space). Then, the EE can be expressed as

$$EE \triangleq \frac{\sum\limits_{k=1}^{K} \mathbb{E}\left[b_k\left(t\right)\right]}{\mathbb{E}_{\boldsymbol{Q}^\infty,h}\left(P_{tot}\right)} = \frac{\sum\limits_{k=1}^{K} \mathbb{E}\left[a_k\left(t\right)\right]}{\mathbb{E}_{\boldsymbol{Q}^\infty,h}\left(P_{tot}\right)} \quad \text{(bits/J)}. \quad (6)$$

To maximize the EE of a system with random arrivals, the policy should adapt to both the QSI, $\boldsymbol{Q}^\infty = (Q_1^\infty, ..., Q_K^\infty)$, and the channel state information (CSI) [4, 14]. Hence, the expectation is taken over both queue length and channel gain.

Definition 2: *Optimal EE-delay relation* is defined as $EE^{\max}(D_1^{\max}, ..., D_K^{\max})$, which denotes the maximal EE under delay requirement $(D_1^{\max}, ..., D_K^{\max})$ with given $\mathbb{E}[a_k(t)]$ and ε_k, $k = 1, ..., K$.

The optimal EE-delay relation is fundamental for a system, which is not an EE-delay curve achieved by a given policy. According to (6), with given traffic load $\sum_{k=1}^{K} \mathbb{E}[a_k(t)]$, to maximize EE, we can equivalently minimize the total transmit and circuit power consumption.

Definition 3: *Power limit* is defined as the minimal average total power consumption that is achieved for infinite delay bound, i.e., $P_{tot}^{\lim} = \lim\limits_{D_k^{\max}\to\infty, k=1,...,K} \mathbb{E}_{\boldsymbol{Q}^\infty,h}(P_{tot})$.

Such a limit is a lower bound of the average total power consumption for a system with arbitrary delay bounds.

To prove that the optimal EE-delay tradeoff vanishes in some scenarios, we will show that $P_{tot}^{\lim}$ can be achieved with finite $D_k^{\max}$. To this end, we introduce a QSI-dependent policy.

Two-state policy: When $Q_k^\infty = 0$, to avoid serving empty buffer, no resource is allocated to the kth user and hence $s_k(t) = 0$, which is referred to as "OFF" state. When $Q_k^\infty > 0$, to guarantee $(D_k^{\max}, \varepsilon_k)$, the resource is allocated such that (5) is satisfied, which is referred to as "ON" state.

III. Optimal EE-Delay Relation

According to the distinguished EE-delay relations, we study the *power-rate relation* in linear and strictly convex scenarios. In the linear scenario, the EE-delay tradeoff vanishes, since the *power limit* can be achieved with arbitrary delay requirement, as indicated in the sequel.

Linear scenario: This occurs for the scenario when $\mathbb{E}_h(P_{tot}^{\min}) = \sum\limits_{k=1}^{K} c_k \bar{s}_k + c_0$, where $c_k, k = 0, 1, ..., K$ are positive constants.

In this scenario, it is not hard to show that the *power limit* of a system is $P_{tot}^{\lim} = \sum\limits_{k=1}^{K} c_k \mathbb{E}[a_k(t)] + c_0$.

Proposition 1. *In linear scenario, $P_{tot}^{\lim}$ of a system can be achieved with arbitrary delay requirement $(D_k^{\max}, \varepsilon_k)$ by a two-state policy.*

Proof: For $K = 1$, the proof can be found in Appendix A. For $K > 1$, the proof is omitted due to the lack of space. ■

Strictly Convex Scenario: This occurs for the scenario when $\mathbb{E}_h(P_{tot}^{\min})$ is strictly convex in average service rate. In this scenario, the EE-delay relation is very hard to obtain. We will show that the strictly convex *power-rate relation* leads to a tradeoff between the maximal EE and the delay bound.

IV. Optimal EE-Delay Relation of Massive MIMO

To show when the different scenarios happen in real-world applications, we consider a single cell downlink massive MIMO system in this section.

A. System and Power Consumption Models

Consider a BS equipped with N_T antennas serves K single-antenna users. The spatial channel vector from the BS to the kth user is $\mathbf{h}_k \in \mathbb{C}^{N_T \times 1}$, whose elements are assumed as independent and identically distributed (i.i.d.) Gaussian variables with zero mean and variance μ_k. Then, the receive signal of the kth user can be expressed as

$$y_k = \mathbf{h}_k^H \left(\sum_{i=1}^{K} \mathbf{x}_i \right) + n_k, k = 1, ..., K, \tag{7}$$

where $\mathbf{x}_i \in \mathbb{C}^{N_T \times 1}$ are the signal vector transmitted to the ith user, and $n_k \in \mathbb{C}^{1 \times 1}$ is white Gaussian noise.

Assume that CSI is perfectly known at the BS. When N_T is large enough, the channel vectors of multiple users are asymptotically orthogonal, and the maximum achievable service rates of the users can be expressed as [15]

$$C_k = W_k \log_2 \left(1 + \frac{\mu_k N_T P_{T_k}}{N_0 W_k} \right), k = 1, ..., K, \tag{8}$$

where N_0 is the single-sided noise spectral density, P_{T_k} and W_k are the transmit power and bandwidth allocated to the kth user, respectively.

Since the service rate of each user does not depend on the instantaneous channel, $E_{C_k}(\theta_k) = \mathbb{E}_h[s_k(t)] = C_k$. Then, constraint (5) degenerates into a constraint on service rate as

$$C_k \geq E_{B_k}(\theta_k). \tag{9}$$

and the *power-rate relation*, $\mathbb{E}_h(P_{tot}^{\min}) - (\bar{s}_1, ..., \bar{s}_K)$, can be simplified into $P_{tot}^{\min}(\bar{s}_1, ..., \bar{s}_K)$. If the elements of $\mathbf{h}_k$ are not i.i.d. and intercell interference is considered, the expression of $E_{C_k}(\theta_k)$ will be very hard to obtain. Nonetheless, the final conclusion will not change.

To save energy with ensured QoS, the BS can adjust the service rate of each user by allocating transmit power and bandwidth (e.g., by allocating the number of active subcarriers in OFDM systems [9]). Denote $P_T^{\max}$ and $W^{\max}$ as the maximal transmit power of the BS and the maximal available bandwidth of the system. Considering that the users are spatially well-separated, they can be served simultaneously in the same frequency. Then, $\sum_{k=1}^{K} P_{T_k} \leq P_T^{\max}$ and $W_k \leq W^{\max}$.

As an illustration, we only consider the impact of bandwidth on circuit power. From the in-depth analysis of the powers consumed for various modules in massive MIMO system [9], the total power consumption at the BS can be modeled as

$$P_{tot} = \frac{1}{\rho} \sum_{k=1}^{K} P_{T_k} + P_{cw} \sum_{k=1}^{K} W_k + P_0, \tag{10}$$

where P_{cw} is the circuit power consumed for baseband processing per unit bandwidth, P_0 is the fixed circuit power independent of bandwidth, and $\rho \in (0, 1]$ is the power amplifier efficiency. The values of P_{cw} and P_0 depend on N_T.[1]

B. Problem Formulation

To find the *power-rate relation* and the *optimal EE-delay relation* for the system, we formulate two problems.

According to the power model in (10) and Definition 1, the *power-rate relation* can be found from the following problem,

$$\min_{\substack{P_{T_k}, W_k, \\ k=1,...,K}} \sum_{k=1}^{K} P_{T_k} + \rho P_{cw} \sum_{k=1}^{K} W_k, \tag{11}$$

$$\text{s.t.} \quad C_k = \bar{s}_k, \ k = 1, ..., K, \tag{11a}$$

$$\sum_{k=1}^{K} P_{T_k} \leq P_T^{\max} \text{ and } W_k \leq W^{\max}. \tag{11b}$$

To obtain the *optimal EE-delay relation*, we need to find the optimal two-state policy that minimizes the average total power consumption under the QoS constraint. Then from the maximal EE achieved by the optimal policy, we can find the optimal EE-delay relation.

Denote $\eta_k = \Pr(Q_k^\infty > 0)$, 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,

$$\mathbb{E}_{\boldsymbol{Q}^\infty}(P_{tot}) = \sum_{k=1}^{K} \eta_k \left(\frac{P_{T_k}^{\text{on}}}{\rho} + P_{cw} W_k^{\text{on}} \right) + P_0, \tag{12}$$

where $P_{T_k}^{\text{on}}$ and W_k^{on} is the transmit power and the bandwidth allocated to the kth user when $Q_k^\infty > 0$ (i.e., the "ON" state of the user). Then, the optimal two-state policy can be obtained from the following problem,

$$\min_{\substack{P_{T_k}^{\text{on}}, W_k^{\text{on}}, \\ k=1,...,K}} \sum_{k=1}^{K} \eta_k \left(\frac{P_{T_k}^{\text{on}}}{\rho} + P_{cw} W_k^{\text{on}} \right) + P_0, \tag{13}$$

$$\text{s.t.} \quad C_k^{\text{on}} \geq E_{B_k}(\theta_k), \ \forall Q^\infty > 0, k = 1, ..., K, \tag{13a}$$

$$\sum_{k=1}^{K} P_{T_k}^{\text{on}} \leq P_T^{\max} \text{ and } W_k^{\text{on}} \leq W^{\max}, \tag{13b}$$

where (13a) are the service rate constraints from (9) for the "ON" state. It is not hard to prove that constraints in (9) and (13a) are equivalent in the sense that they can guarantee the same statistical QoS requirements. Due to the lack of space, the proof is omitted.

In the following two sections, we only consider single user case for easy exposition and notational simplification. Then, the index k in problems (11) and (13) can be omitted. For multi-user case, the conclusions are similar, as will be illustrated by numerical results.

C. Optimal EE-Delay Relation in Linear Scenario

In this subsection, we solve problems (11) and (13) for the services whose required delay bounds are large (i.e., the required service rates are low) such that the constraints on transmit power and bandwidth are inactive. We can show that the required minimal total power $P_{tot}^{\min}(\bar{s})$ linearly grows with $\bar{s}$, i.e., this is a *linear scenario*. Then, we discuss when the

[1] When N_T is jointly allocated with P_{T_k} and W_k, the conclusion of this paper will not change, although P_{tot} is a non-linear function of N_T [9].

resource constraints are inactive, from which we can find the boundary of the EE-delay non-tradeoff region.

1) Power-Rate Relation: By solving problem (11) without resource constraints (11b), we can obtain the following proposition (proved in Appendix B).

Proposition 2. The optimal solution of problem (11), P_T^* and W^*, satisfies $\frac{P_T^*}{W^*} = P_{tw}^*$, where P_{tw}^* is the optimal transmit power per unit bandwidth, which is independent of $\bar{s}$.

When transmitting with P_{tw}^*, from (8) the optimal service rate per unit bandwidth can be expressed as

$$r_w^* = \log_2\left(1 + \frac{\mu N_T}{N_0} P_{tw}^*\right). \tag{14}$$

Then, the optimal bandwidth and transmit power that minimize the required total power to support $\bar{s}$ can be expressed as

$$W^* = \frac{\bar{s}}{r_w^*}, \text{ and } P_T^* = P_{tw}^* \frac{\bar{s}}{r_w^*}. \tag{15}$$

Upon substituting into (10), the power-rate relation can be obtained as,

$$P_{tot}^{\min}(\bar{s}) = \left(\frac{P_{tw}^*}{\rho r_w^*} + \frac{P_{cw}}{r_w^*}\right)\bar{s} + P_0, \tag{16}$$

which is a linear function of $\bar{s}$ since P_{tw}^* and r_w^* are independent of $\bar{s}$. In other words, this is a *linear scenario*.

2) Optimal EE-Delay Relation: To show that the EE-delay tradeoff vanishes, we first find the power limit. Then, we show that the minimal average total power achieved by the optimal two-state policy for any delay bound equals to the power limit.

In massive MIMO systems, when $\theta \to 0$ (i.e., $D^{\max} \to \infty$), $E_B(\theta) = \mathbb{E}[a(t)]$ [11]. Therefore, constraint (13a) can be re-expressed as $C \geq \mathbb{E}[a(t)]$. Then, $P_{tot}^{\lim}$ can be achieved by a policy that minimizes the total power consumption under this constraint. It is easy to see that the power-minimizing policy should adjust the resource such that $C = \bar{s} = \mathbb{E}[a(t)]$. From the minimal total power consumption required to support $\bar{s}$ in (16), the power limit can be expressed as,

$$P_{tot}^{\lim} = \left(\frac{P_{tw}^*}{\rho r_w^*} + \frac{P_{cw}}{r_w^*}\right)\mathbb{E}[a(t)] + P_0. \tag{17}$$

In what follows, we find the optimal policy from problem (13) and derive the average total power achieved by the policy. By rewriting (A.1) as $\mathbb{E}[a(t)] = \mathbb{E}_{Q^\infty,h}[s(t)] = \eta\mathbb{E}_h[s(t)|Q^\infty > 0]$, we have $\eta = \frac{\mathbb{E}[a(t)]}{\mathbb{E}_h[s(t)|Q^\infty > 0]}$. In massive MIMO systems, $\mathbb{E}_h[s(t)|Q^\infty > 0] = C^{\text{on}}$. Then, the average total power consumption in (13) can be expressed as

$$\frac{\mathbb{E}[a(t)]}{C^{\text{on}}}\left(\frac{P_T^{\text{on}}}{\rho} + P_{cw}W^{\text{on}}\right) + P_0. \tag{18}$$

Because $\mathbb{E}[a(t)]$ is fixed for any given traffic, minimizing (18) is equivalent to minimizing

$$\frac{P_{tot}^{\text{on}} - P_0}{C^{\text{on}}}, \tag{19}$$

where $P_{tot}^{\text{on}} \triangleq P_T^{\text{on}}/\rho + P_{cw}W^{\text{on}} + P_0$ is the total power consumption in "ON" state.

According to the analysis in section IV.C.1, to support any given $C^{\text{on}} = \bar{s}$, the minimal total power consumption in "ON" state can be expressed as $P_{tot}^{\text{on}^*} = P_{tot}^{\min}(\bar{s})$, where $P_{tot}^{\min}(\bar{s})$ is expressed in (16). Moreover, the power-minimization policy should satisfy the following condition

$$P_T^{\text{on}^*} = P_{tw}^* W^{\text{on}^*}, \tag{20}$$

under which $C^{\text{on}^*} = W^{\text{on}^*} r_w^*$, where r_w^* is expressed in (14). Substituting C^{on^*} into the QoS constraint (13a), we obtain

$$W^{\text{on}^*} \geq E_B(\theta)/r_w^*. \tag{21}$$

It follows that any policy that satisfies (20) and (21) can minimize the total power in "ON"' state meanwhile ensure the QoS. Next we show that the average total power consumption achieved by such policies is equal to the power limit, which is the lower bound of the minimal average total power. In this way, we know that these policies are the optimal solution of problem (13).

Substituting $P_{tot}^{\text{on}^*}$ and $C^{\text{on}} = \bar{s}$ into (19), the minimum value of (19) can be expressed as $\frac{P_{tw}^*}{\rho r_w^*} + \frac{P_{cw}}{r_w^*}$, and then the minimum value of (18) is

$$\mathbb{E}_{Q^\infty}(P_{tot}) = \left(\frac{P_{tw}^*}{\rho r_w^*} + \frac{P_{cw}}{r_w^*}\right)\mathbb{E}[a(t)] + P_0, \tag{22}$$

which is exactly the same as $P_{tot}^{\lim}$ in (17).

This indicates the average total power achieved by the optimal two-state policy for the delay bounds equals to the power limit. In other words, there is no EE-delay tradeoff.

3) Boundary of EE-delay Non-tradeoff Region: From the maximum resource constraints, we can find the boundary of non-tradeoff region. Substituting (15) into (11b), we can obtain

$$\bar{s} \leq \min\left(P_T^{\max} r_w^*/P_{tw}^*, W^{\max} r_w^*\right) \triangleq \bar{s}_{\text{th}}, \tag{23}$$

where $\bar{s}_{\text{th}}$ is the boundary of the linear region of power-rate relation. If the first term in $\min(\cdot,\cdot)$ is larger, i.e., $P_T^{\max} > P_{tw}^* W^{\max}$, the boundary is

$$\bar{s}_{\text{th}} = W^{\max}\log_2\left(1 + \frac{\mu N_T}{N_0}P_{tw}^*\right), \tag{24}$$

where $P_{tw}^* = P_T^*/W^{\max}$. If the second term is larger, the result is similar and hence is not shown.

Considering the maximum resource constraints in (13b), the policy satisfying (20) and (21) exists when $E_B(\theta)/r_w^* \leq W^{\text{on}^*} \leq W^{\max}$ and $P_{tw}^* E_B(\theta)/r_w^* \leq P_T^{\text{on}^*} \leq P_T^{\max}$, i.e.,

$$E_B(\theta) \leq \min\left(P_T^{\max} r_w^*/P_{tw}^*, W^{\max} r_w^*\right). \tag{25}$$

If the required delay bound of a service satisfies (25), the EE-delay tradeoff vanishes. From (4), the related delay bound can be obtained from $D_{\text{th}}^{\max} = \frac{-\ln\varepsilon}{\theta^{\text{th}} E_B(\theta^{\text{th}})}$, where $E_B(\theta^{\text{th}}) = \bar{s}_{\text{th}}$. When $D^{\max} \geq D_{\text{th}}^{\max}$, the maximum EE is independent of the delay bound.

Note that the right hand sides of (23) and (25) are identical. This means that if the required service rate for a traffic lies in the linear region of the power-rate relation, then the EE-delay tradeoff vanishes.

D. EE-Delay Tradeoff in Strictly Convex Scenario

In this subsection, we consider the scenario when $D^{\max} \leq D_{\text{th}}^{\max}$, i.e., $\bar{s} > \bar{s}_{\text{th}}$. For such kind of traffic, the power-rate relation is strictly convex as shown in the sequel.

1) Power-Rate Relation: We study the case where the first term in $\min(\cdot,\cdot)$ in (23) is larger (i.e., $P_T^{\max} > P_{tw}^{*} W^{\max}$). For the other case where $P_T^{\max} \leq P_{tw}^{*} W^{\max}$, the results are similar and hence are omitted.

When $\bar{s} \geq \bar{s}_{\text{th}}$, the constraint on $W^{\max}$ is active. To support higher server rate $\bar{s}$, the system needs to increase transmit power, and thus $P_T > P_{tw}^{*} W^{\max}$. From the maximum achievable rate in (8), we can obtain the minimal transmit power to support service rate $\bar{s}$ as $P_T^{*} = \frac{N_0 W^{\max}}{\mu N_T}\left(2^{\bar{s}/W^{\max}} - 1\right)$. Substituting P_T^{*} and $W^{*} = W^{\max}$ into (10), the power-rate relation can be expressed as

$$P_{tot}^{\min}\left(\bar{s}\right) = \frac{N_0 W^{\max}}{\rho\mu N_T}\left(2^{\bar{s}/W^{\max}} - 1\right) + P_{cw} W^{\max} + P_0, \tag{26}$$

which is strictly convex in $\bar{s}$. In other words, this is a *strictly convex scenario*.

2) EE-Delay Tradeoff: In the sequel, we show that if the required service rate lies in the strictly convex region of the power-rate relation, then there will be a tradeoff between the maximal EE and the delay bound.

Similar to the linear scenario, the *power limit* can be obtained from the related power-rate relation in (26) with $C = \bar{s} = \mathbb{E}[a(t)]$, i.e., $P_{tot}^{\text{lim}} = P_{tot}^{\min}\left\{\mathbb{E}[a(t)]\right\}$.

Denote $C(Q^{\infty})$ as the service rate of a QSI dependent policy (e.g., the two-state policy). When the power-rate relation is strictly convex, the average total power consumption of this policy satisfies

$$\mathbb{E}_{Q^{\infty}}\left\{P_{tot}^{\min}\left[C(Q^{\infty})\right]\right\} \geq P_{tot}^{\min}\left\{\mathbb{E}_{Q^{\infty}}\left[C(Q^{\infty})\right]\right\} \tag{27}$$

$$\geq P_{tot}^{\min}\left\{\mathbb{E}[a(t)]\right\}, \tag{28}$$

where (27) comes from the Jensen's inequality, and (28) comes from $\mathbb{E}_{Q^{\infty}}\left[C(Q^{\infty})\right] \geq \mathbb{E}[a(t)]$, which is the necessary condition for a policy to guarantee finite delay [6].

The equality in (27) will hold if and only if the service rate is independent of QSI according to the property of the Jensen's inequality. Denote $C(Q^{\infty}) = C_0$, $\forall Q^{\infty} \geq 0$ as the service rate of a QSI independent policy, with which the statistical QoS requirement can be satisfied if $C_0 \geq E_B(\theta)$ [7]. When the arrival rate is random, with positive θ (i.e., finite $D^{\max}$), $E_B(\theta) > \mathbb{E}[a(t)]$ [3] and hence $C_0 > \mathbb{E}[a(t)]$, i.e., the equality in (28) does not hold.

It follows that the power limit can not be achieved, no matter with QSI dependent or independent policy. This implies that the EE-delay relation is a tradeoff.

The analysis indicates that under the statistical QoS requirement, the tradeoff between maximal EE and delay bound stems from the strictly convex power-rate relation. This is consistent with prior results [4,5], where average delay and strict deadline requirements were considered.

V. Numerical Results

In this section, we illustrate the power-rate relation and the EE-delay curve achieved by the optimal two-state policy via numerical results.

For comparison, the EE-delay curve achieved by a QSI independent policy in [7] is provided, where the transmit power is minimized (with legend "Existing policy").

We consider $K = 10$ users, served by a BS with $N_T = 100$ antennas over bandwidth $W^{\max} = 20$ MHz. With 10 users, the EE-delay relation will be a 11-dimensional curve. To capture the essence of the problem but without loss of generality, all users are set with identical delay bound and identical distance to the BS of $d = 200$ m. The conclusion will not change in more practical scenarios. Then, the maximum transmit power for each user is $P_T^{\max}/K$, and the multi-user policy can be decomposed into K single user policies. $P_T^{\max}$ is set as 41 dBm. The path loss model is $35.3 + 37.6\log_{10} d$ dB, and the noise power spectral density is $N_0 = -174$ dBm/Hz. The arrival process of each user is a compound Poisson process with average packets arrival rate 2000 packets/s and average packet size 50 kbits. For other kinds of sources, the results will not change. The parameters in (10) are as follows, $\rho = 50$ %, $P_{cw} = 0.075$ W/MHz, and $P_0 = 13.6$ W, which are predicted by GreenTouch for the year of 2020 [9].

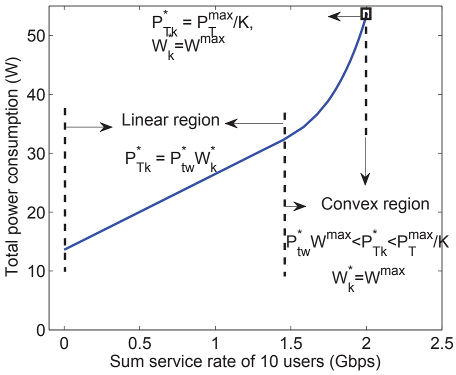

Fig. 1. Power-rate relation.

Figure 1 shows the power-rate relation. In the linear region, $P_{T_k} < P_T^{\max}/K$ and $W_k < W^{\max}$, and $P_{T_k} = P_{tw}^{*} W_k$ is satisfied. In the convex region, the maximal bandwidth is used. To increase the service rate of each user, the system can only increase the transmit power.

Figure 2 shows the EE-delay curve achieved by the optimal two-state policy. In the non-tradeoff region, $P_{T_k}^{\text{on}^*} = P_{tw}^{*} W_k^{\text{on}}$, corresponding to the linear region of Fig. 1. In the tradeoff region, $W_k^{\text{on}^*} = W^{\max}$ and $P_{T_k}^{\text{on}^*} > P_T^{\text{th}}$, corresponding to the convex region of Fig. 1. This confirms that the EE-delay relation is determined by the power-rate relation. It is shown that the maximal EE achieved by the optimal two-state policy is much higher than that of existing policy in the two regions of the delay bounds.

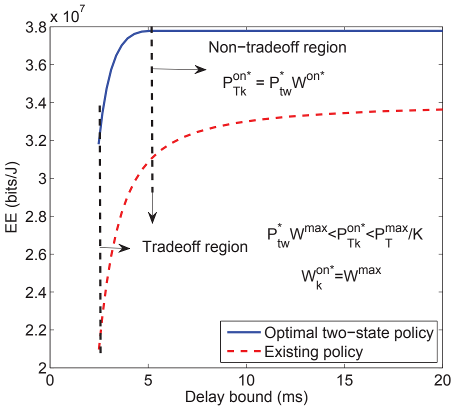

Fig. 2. EE-delay relation with 10 compound Poisson sources, where $\varepsilon = 0.01$ and $\mathbb{E}[a_k(t)] = 0.1$ Gbps.

VI. Conclusion

In this paper, we studied the optimal EE-delay relation for wireless systems serving the traffic with random arrivals and statistical QoS requirements. We proved that when the required minimal average total power consumption of a system linearly increases with the average service rate, the EE-delay tradeoff vanishes. By taking massive MIMO system as an example, we showed that the optimal EE-delay relation includes a non-tradeoff region when the desired delay bound is large and a tradeoff region when the delay bound is small. The existence of such an EE-delay non-tradeoff region implies that we cannot improve the maximal EE by extending the delay bound when the delay bound lies in the non-tradeoff region. Moreover, under the statistical QoS requirement, the EE-delay tradeoff stems from the strictly convex relation between average total power and average service rate.

Appendix A
Proof of Proposition 1

Proof: When the two-state policy is applied, from (1) we know that $s_1(t) = b_1(t)$. Moreover, it is not hard to show that $\mathbb{E}[a_1(t)] = \mathbb{E}[b_1(t)]$ when (5) is satisfied, i.e., $(D_1^{\max}, \varepsilon_1)$ is ensured. Therefore, we have

$$\mathbb{E}[a_1(t)] = \mathbb{E}[b_1(t)] = \mathbb{E}_{Q_1^\infty,h}[s_1(t)]. \tag{A.1}$$

In the linear scenario, the average power consumption of the two-state policy can be expressed as

$$\begin{aligned}\mathbb{E}_{Q_1^\infty}[\mathbb{E}_h(P_{tot}^{\min})] =& \Pr(Q_1^\infty > 0)\{c_1\mathbb{E}_h[s_1(t)|Q_1^\infty > 0] + c_0\} \\ &+ \Pr(Q_1^\infty = 0)\{c_1\mathbb{E}_h[s_1(t)|Q_1^\infty = 0] + c_0\} \\ =& c_1\mathbb{E}_{Q^\infty,h}[s_1(t)] + c_0 \\ =& c_1\mathbb{E}[a_1(t)] + c_0,\end{aligned}$$

which equals to the power limit. ■

Appendix B
Proof of Proposition 2

Proof: To prove Proposition 2, we analyze the Karush-Kuhn-Tucker (KKT) conditions, which are the necessary conditions that the optimal solution of problem (11) should satisfy. From the KKT conditions of this problem, we can derive that

$$\frac{\frac{\mu N_T}{N_0}\left(\rho P_{cw} + \frac{P_T}{W}\right)}{1 + \frac{\mu N_T P_T}{N_0 W}} - \ln\left(1 + \frac{\mu N_T P_T}{N_0 W}\right) = 0, \tag{B.1}$$

from which we can obtain the solution of the transmit power per unit bandwidth, $P_{tw} \triangleq \frac{P_T}{W}$. To find this solution of P_{tw}, we define $g\left(P_{tw}\right) \triangleq \frac{\frac{\mu N_T}{N_0}(\rho P_{cw}+P_{tw})}{1+\frac{\mu N_T}{N_0}P_{tw}} - \ln\left(1 + \frac{\mu N_T}{N_0}P_{tw}\right)$. It is easy to show that $g(0) > 0$, $g(\infty) < 0$, and $g'(P_{tw}) < 0, \forall P_{tw} \in [0, \infty)$. Therefore, $g(0) > 0$, $g(\infty) < 0$, and $g(P_{tw})$ strictly decreases with P_{tw}. It follows that the equation in (B.1) has a unique solution, P_{tw}^*, which is independent of $\bar{s}$. Therefore, the optimal solution of problem (11) should satisfy the following condition: $\frac{P_T^*}{W^*} = P_{tw}^*$. ■

References

[1] X. Zhang, W. Cheng, and H. Zhang, "Heterogeneous statistical QoS provisioning over 5G mobile wireless networks," *IEEE Network*, vol. 28, no. 6, pp. 46–53, Nov. 2014.

[2] 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.

[3] B. Soret, P. Mogensen, K. I. Pedersen, and M. C. Aguayo-Torres, "Fundamental tradeoffs among reliability, latency and throughput in cellular networks," in *IEEE Globecom Workshop*, Dec. 2014.

[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] 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.

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

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

[8] 3GPP, *Further Advancements for E-UTRA Physical Layer Aspects*. TSG RAN TR 36.814 v9.0.0, Mar. 2010.

[9] C. Desset, B. Debaillie, and F. Louagie, "Modeling the hardware power consumption of large scale antenna systems," in *Proc. IEEE OnlineGreenComm*, Nov. 2014.

[10] 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.

[11] 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.

[12] 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.

[13] 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

[14] 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.

[15] 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.