

On the Queuing Model of the Energy-Delay Tradeoff in Wireless Links With Power Control and Link Adaptation

Ege Orkun Gamgam, Caglar Tunc, *Student Member, IEEE*, and Nail Akar^{ID}, *Member, IEEE*

Abstract—A transmission profile refers to a transmission power and a modulation and coding scheme to be used for packet transmissions over a wireless link. The goal of this paper is to develop transmission profile selection policies so as to minimize the average power consumption on a wireless link while satisfying a certain delay constraint given in terms of a delay violation probability. Toward the assessment of profile selection policies, a multi-regime Markov fluid queue model is proposed to obtain the average power consumption and the queue waiting time distribution which allows one to analyze the energy-delay tradeoff in queuing systems for which the packet transmission duration is allowed to depend on the delay experienced by the packet until the beginning of service. Numerical examples are presented with transmission profiles obtained from realistic LTE simulations. Several transmission profile selection policies are proposed and subsequently compared using the analytical model.

Index Terms—Wireless networks, link adaptation, power control, energy efficiency, queuing analysis, Markov fluid queues.

I. INTRODUCTION

MODELING power consumption in cellular networks and techniques for reducing it have been important research topics in recent years [1], [2]. It has been shown in [3] that the largest share of overall power consumption stems from the Base Station (BS) in cellular networks. The power consumption model in [1] reveals that the total power consumption P_{in} of the BS can approximately be written as the sum of two terms for each transmit/receive antenna: a fixed term corresponding to the power consumption in various units including the baseband unit, the RF unit, active cooling, losses incurred by the DC-DC power supply and mains supply, etc., and an additional load-dependent power amplifier term that

linearly depends on the BS load P_{out}/P_{max} , where P_{out} and P_{max} represent the actual RF output power radiated at the antenna element (called the transmission power throughout the paper) and maximum output power, respectively. Therefore, a power reduction ΔP_{out} at the output of the antenna element leads to an overall reduction $\Delta_P \Delta P_{out}$ in the overall power consumption where Δ_P is termed as the power gradient [2]. It was shown in [1, Table 2] that in an LTE macro BS with 10 MHz bandwidth, 3 sectors, 2x2 MIMO configuration and $P_{max} = 20$ Watts, the power gradient Δ_P equals around 4.7 and the latter load-dependent term contributes above 40 percent of the total power whereas the fixed term is more dominant in low power BSs such as pico and femto cells. Therefore, it is crucial to develop techniques to reduce P_{out} in relatively high power BSs by appropriate energy-efficient transmission techniques. A subset of the proposed energy-efficient techniques give rise to increased packet delays and the resulting energy-delay trade-off has been studied extensively in the literature. A key mechanism to play the energy-delay trade-off is transmission profile selection with a profile comprising the following two attributes (i) the transmission power, and (ii) the modulation and coding scheme (the latter also known as link adaptation), when a packet gets to be transmitted. Typically, higher service rate profiles reduce queuing delays but they lead to increased energy consumption per packet. On the other hand, lower service rate profiles reduce the per-packet energy consumption at the expense of increased queuing delays. The goal of this work is to choose appropriate transmission profiles so as to minimize the average power consumption while meeting queuing delay constraints.

A. System Setup

In this paper, we consider the following setup. The system model consists of a wireless transmitter with packets having statistical delay constraints given in terms of delay violation probabilities. The packet arrival process to the transmitter is Poisson and packets join a FIFO buffer before being transmitted to a single receiver. Given the channel conditions, a finite set of K transmission profiles, denoted by $\mathcal{K} = \{1, 2, \dots, K\}$ is assumed to be available. Each transmission profile $k \in \mathcal{K}$ is characterized with the pair (P_k, μ_k) where P_k and μ_k denote the transmission power (Watts) and service rate (packets/sec), respectively, of profile k . When the Head of Line (HoL) packet just gets to be transmitted, a decision is to be made on which of the K transmission profiles is to be used for transmitting

Manuscript received April 13, 2018; revised August 21, 2018 and November 28, 2018; accepted January 26, 2019. Date of publication February 5, 2019; date of current version May 15, 2019. This research was supported in part by the Scientific and Technological Research Council of Turkey (Tübitak) grant no: EEEAG-115E360. The associate editor coordinating the review of this paper and approving it for publication was N. Pappas. (*Corresponding author: Nail Akar.*)

E. O. Gamgam is with the Electrical and Electronics Engineering Department, Bilkent University, 06800 Ankara, Turkey, and also with Aselsan, 06370 Ankara, Turkey (e-mail: gamgam@ee.bilkent.edu.tr).

C. Tunc is with the Department of Electrical and Computer Engineering, New York University Tandon School of Engineering, New York, NY 11201 USA (e-mail: ct1909@nyu.edu).

N. Akar is with the Electrical and Electronics Engineering Department, Bilkent University, 06800 Ankara, Turkey (e-mail: akar@ee.bilkent.edu.tr).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TCOMM.2019.2897700

this packet. If the transmission profile selection is to be made on the basis of the number of packets in the queue, then conventional Markov Chain (MC) techniques could be used to study the resulting queuing system; see for example [4]. However, there has been a recent trend of using the actual queuing delay information in active queue management algorithms as opposed to using the number of packets; see for example the CoDel active queue management algorithm in [5], [6] by means of which tuning of the algorithm parameters is made independent from link speeds, types, etc. Moreover, the constraint that needs to be met in the current work is in terms of delays and not in terms of packets. Motivated by these two observations, we consider transmission profile selection policies that are allowed to depend on the queuing delay rather than the number of packets and consequently, conventional MC techniques cannot be used. For exponentially distributed packet lengths, this setting gives rise to an M/M/1 queuing model with delay-dependent service times and the approach we take is to reduce its solution to that of an appropriate Multi-Regime Markov Fluid Queue (MRMFQ) whose numerically efficient and stable solutions are available in [7]. Expressions are then given for the average consumed power and the delay distribution of packets for the delay-dependent M/M/1 queuing model. Thanks to the matrix-analytic nature of the MRMFQ model, extension of the results to more general service times with phase type (PH-type) distributions is also presented; see [8] for PH-type distributions and their properties.

B. Our Contributions

The main contributions of our work are stated as follows:

- Obtaining the MRMFQ model for the delay-dependent queuing system of interest is one of our main contributions. The MRMFQ model for a similar queuing system with two service rates has recently been proposed in [9] in the context of an overlay cognitive radio system. In the current study, the MRMFQ model is extended to the case of multiple service rates and also PH-distributed service times along with average power consumption expressions. Moreover, numerical solutions are sought in contrast with the closed-form expressions in [9] that are only valid for the case of two service rates.
- We study transmission profile selection policies given in terms of just one or a few parameters making it possible to repeatedly solve the MRMFQ throughout the low-dimensional parameter space and subsequently obtain the optimum parameter setting for the proposed policies.
- In the numerical examples, the transmission profiles are obtained from physical layer LTE simulations. Therefore, the findings of the current work are expected to have real world applications in LTE networks or in similar settings.

C. Organization

The paper is organized as follows. Related work is given in Section II. MRMFQs are briefly described in Section III. The M/M/1 queue with delay-dependent service times and its MRMFQ solution are presented in Section IV along with

extensions to PH-type service times. The proposed transmission profile selection policies are described in Section V. Numerical results in the context of wireless transmission of packets using LTE transmission profiles are presented in Section VI. Finally, we conclude.

II. RELATED WORK

Energy-efficient transmission techniques to reduce the transmission power have been studied in [10]–[12] and the references therein. The off-line energy-efficient wireless transmission problem under deadline constraints has been studied in [13] where the packet arrival epochs are assumed to be known in advance. An optimal lazy scheduling scheme is obtained in this work for the case of infinitely many transmission profiles with a fixed power-rate relationship. Scheduling schemes have also been proposed in [13] for the on-line case without elaborating on their optimality. Similarly, the work of [14] considers the optimal transmission of a number of packets within their deadlines in a time-varying channel using continuous-time stochastic control. Reference [15] studies the minimization of the delay bound violation probability subject to constraints on average power, arrival rate, and delay bound. In [16], Chen *et al.* study energy efficient transmission techniques with individual packet delay constraints.

An important application of service rate adjustment is the so-called speed scaling which adapts the speed of a computer system to trade off energy and performance [17]. In static speed scaling, a single speed is employed unless the system is idle and is put into a sleep mode when idle [17]. In dynamic speed scaling, the speed is adapted continuously based on the instantaneous state, e.g., number of packets in the system. Modern processors and computer systems allow dynamic speed scaling which leads the way to investigate its impact on various performance measures; see [18]–[22] and the references therein.

There have been quite a few studies on queues with state-dependent service rates. One of the earlier works is [23] in which the service rate is made a function of the instantaneous queue occupancy. The resulting system with continuous service rate adjustment is a birth-and-death process and is quite straightforward to solve. The problem addressed in [24] is the selection of optimal service rates for a single server queue with state-dependent Poisson arrivals and continuous service rate adjustment. The case of the service time being dependent on the number of customers at the epoch of service start has been studied in [25]. This system is not a birth-and-death process and can be analyzed by embedded Markov chain techniques [25]. The service rate is adjusted at service start epochs (as opposed to continuous adjustment) and therefore the server profile is kept intact for the entire service duration of the packet in these systems. An M/G/1 queue with adaptable service speed based on the amount of work right after customer arrivals is studied in [26]. Reference [27] studies the case of service speed adaptations taking place only at the arrival instants of an external Poisson observer. The work [28] studies a queuing system where the arrival rate and/or speed of the server continuously depend on the amount of the present workload. Reference [29] describes a

workload-dependent M/G/1 system with a two-stage service policy. The work in [30] studies a bi-level hysteretic control of an M/M/1-type system in which there is a change of service rate when the queue length exceeds a given threshold and then this service rate remains in effect until the queue length is reduced back to another lower threshold. A similar hysteretic queuing system with more general Lévy inputs is studied in [31].

With the deployment of 5G and applications that have stringent delay requirements, such as augmented and virtual reality, the trade-off between energy and resource efficiency and QoS constraints is becoming a more prominent issue which has been widely studied in the literature [32]–[40]. In [32], optimal power control and rate adaptation is derived that maximizes the effective capacity of the channel under power and statistical QoS constraints. Reference [33] considers frequency-selective channels in an OFDM system in which channel states and power consumed in the transmitter circuitry affect the optimal power allocation across the channels and the modulation. Reference [34] proposes scheduling strategies that minimize packet retransmissions while satisfying a deadline constraint for a given queue size for various limited CSI (Channel State Information) feedback models. In [35], Sinaie *et al.* optimize an energy efficiency metric with delay threshold and maximum feasible power as the constraints which are defined based on a queuing model of the wireless link. Centralized and decentralized power control algorithms for 5G wireless communication systems are developed in [36] to optimize energy efficiency under rate constraints. In [37], a power control approach is proposed to jointly optimize energy and delay constraints in wireless networks using game theory. Reference [38] proposes an energy efficient cross-layer design for transmitting bursty traffic over Nakagami-m fading channels with delay demands. In [39], Chen *et al.* investigate a cognitive shared access network with energy harvesting-based opportunistic secondary nodes with the aim of maximizing the secondary throughput with primary delay constraints. Reference [40] investigates a queuing model to assess the performance of a BS fully powered by renewable energy sources.

Fluid queue models have been used in the context of energy efficient communication in wireless channels in several studies including [41]–[44]. In [41], effects of different hybrid automatic repeat request (HARQ) schemes are investigated under outage, deadline, and queuing constraints for different arrival processes, including an on-off source modeled as a Markov fluid process. Reference [42] characterizes the maximum average arrival rate under queuing constraints for a similar Markov fluid source. References [43] and [44] define optimal power control policies for fading channels with Markovian sources, including the Markov fluid source, and queuing constraints.

III. MULTI-REGIME MARKOV FLUID QUEUES

In fluid queue models, a fluid acts as the input to and output of a buffer. In particular, Markov Fluid Queues (MFQ) are described by a joint Markovian process $(X(t), Z(t))$, $t \geq 0$ where $X(t)$ represents the fluid level (or buffer content) or the modulated process [45]. On the other hand,

$Z(t)$ is an underlying finite state-space continuous-time Markov chain that determines the drift, i.e., the rate at which the buffer content $X(t)$ changes. The process $Z(t)$ is called the modulating process of the MFQ. MRMFQs are generalizations of single-regime MFQs in the sense that the buffer space in MRMFQs is partitioned into a finite number of non-overlapping intervals which are called the regimes of the MRMFQ [7], [46]. In MRMFQs, the infinitesimal generator of the background CTMC as well as the drift into the buffer depend on the regime at which the buffer level resides. The material below for the brief description of MRMFQs and their notation is based on [7]. In an infinite-buffer MRMFQ, the buffer is partitioned into $K > 1$ regimes with the boundaries $0 = T^{(0)} < T^{(1)} < \dots < T^{(K-1)} < T^{(K)} = \infty$. The case of $T^{(K)} < \infty$ is referred to as a finite-buffer MRMFQ but is outside the scope of this paper. If $T^{(k-1)} < X(t) < T^{(k)}$, the system is said to be in regime k at time t . Let $X(t) \in [0, \infty)$ and $Z(t) \in \{0, 1, \dots, N-1\}$ denote the buffer content and the background process, respectively, at time t , as in usual MFQs. We denote the infinitesimal generator and drift matrices associated with regime k by $Q^{(k)}$ and $R^{(k)}$, respectively, for $1 \leq k \leq K$. The regime- k drift matrix $R^{(k)}$ is the diagonal matrix

$$R^{(k)} = \text{diag}(r_0^{(k)}, r_1^{(k)}, \dots, r_{N-1}^{(k)}),$$

where $r_i^{(k)}$ is the net drift of the buffer at state i and regime k . Note that $Q^{(k)}$ and $R^{(k)}$ are fixed within a given regime. Similar to $Q^{(k)}$ and $R^{(k)}$, we define $\tilde{Q}^{(k)}$ and $\tilde{R}^{(k)}$ as the infinitesimal generator and drift matrices associated with the boundary $T^{(k)}$ (or simply boundary- k) for $0 \leq k \leq K-1$, where the drift of state i at boundary- k is denoted by $\tilde{r}_i^{(k)}$. We define the joint probability density function (pdf) vector $f^{(k)}(x)$ for regime- k when $T^{(k-1)} < x < T^{(k)}$ as follows:

$$f_i^{(k)}(x) = \lim_{t \rightarrow \infty} \frac{d}{dx} \Pr\{X(t) \leq x, Z(t) = i\}, \quad (1)$$

$$f^{(k)}(x) = [f_0^{(k)}(x) \ f_1^{(k)}(x) \ \dots \ f_{N-1}^{(k)}(x)]. \quad (2)$$

Similarly, the steady-state probability mass accumulation (pma) vector $c^{(k)}$ is defined for each boundary- k for $0 \leq k \leq K-1$ as follows:

$$c_i^{(k)} = \lim_{t \rightarrow \infty} \Pr\{X(t) = T^{(k)}, Z(t) = i\}, \quad (3)$$

$$c^{(k)} = [c_0^{(k)} \ c_1^{(k)} \ \dots \ c_{N-1}^{(k)}]. \quad (4)$$

A matrix-analytic algorithm has been proposed in [7] to obtain the joint pdf vector in (2) for each regime- k and the joint pma vector in (4) for each boundary- k . This numerical algorithm requires the solution of a linear matrix equation of at most size $N(2K+1)$ for an MRMFQ with N states and K regimes. The computational complexity of the proposed algorithm can be reduced to $\mathcal{O}(N^3 K)$ on the basis of the observation that the linear matrix equation is in block tridiagonal form [47].

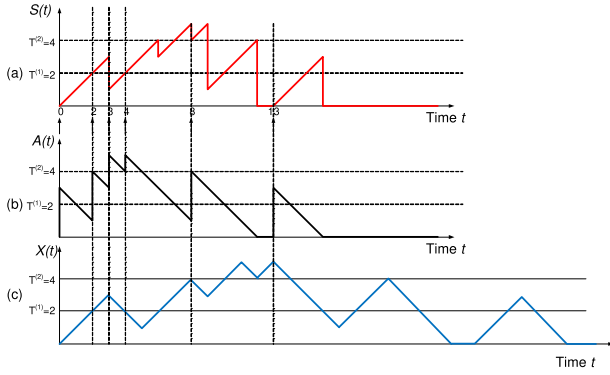


Fig. 1. Sample paths of the following processes: (a) $S(t)$, (b) $A(t)$, and (c) $X(t)$.

IV. THE QUEUING MODEL FOR THE WIRELESS LINK WITH DELAY-DEPENDENT SERVICE TIMES

We first describe the system of interest. Then, we provide the MRMFQ model. Subsequently, we provide expressions on how to obtain the related performance measures of interest.

A. System Description

We first assume a single server FIFO queue with packets exponentially distributed in length arriving at the wireless link according to a Poisson process with rate λ packets/sec. A finite number of transmission profiles indexed by $k = 1, 2, \dots, K$ are assumed to be available to serve the given packet. The server profile $k, 1 \leq k \leq K$, is characterized with service rate μ_k and power P_k with $\mu_i < \mu_j$ when $i < j$ without loss of generality. Let $D(t)$ denote the delay already experienced by the HoL packet at service start time t and a transmission profile selection is to be made for the HoL packet at the FIFO queue at time t . K thresholds are defined satisfying $0 = T^{(0)} < T^{(1)} < \dots < T^{(K-1)} < T^{(K)} = \infty$ to describe the operation of the transmission profile selection policy. Particularly, when $T^{(k-1)} \leq D(t) < T^{(k)}$, then the packet is to be served with server profile k . This particular choice stems from the fact that larger service rates should be used with increased delays towards the satisfaction of statistical delay constraints. We call this system as the delay-dependent M/M/1 queue. All the profile selection policies to be proposed in the next section can well be studied within this general system framework.

In order to obtain the distribution of $D(t)$, we need to define the sojourn time $S(t)$ that is the overall time spent in the system including service for the packet being served by the server. If there are no packets being served at time t , then $S(t) = 0$. Moreover, let the virtual waiting time $A(t)$ denote the amount of time to drain all waiting packets (also including service) in the system at time t . It is clear that a packet arriving at the system at time t with $T^{(k-1)} \leq A(t) < T^{(k)}$ is to be eventually served at rate μ_k . The sample paths for the two processes $S(t)$ and $A(t)$ are given in figures 1a and 1b, respectively, for an example scenario with two thresholds $T^{(1)} = 2$ and $T^{(2)} = 4$ and for the case of packet arrivals occurring at $t = 0, 2, 3, 4, 8, 13$. For the sake of convenience, the service times in regimes 1, 2, and 3, are deterministically set to 3, 2, and 1, respectively, for this example.

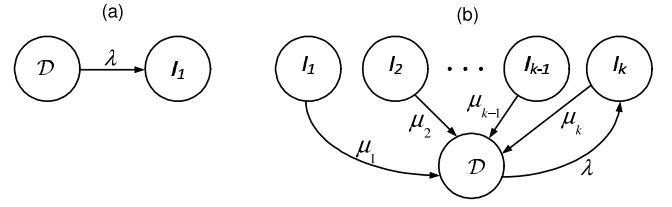


Fig. 2. State transitions (a) for $X(t) = 0$ and (b) for regime $k, k = 1, \dots, K$.

B. MRMFQ Model

The abrupt jumps in the sample paths of $S(t)$ and $A(t)$ correspond to drifts of $-\infty$ and $+\infty$, respectively. Therefore, these processes cannot be represented directly by a Markov fluid queue since the drifts need to be finite in this framework. Fig. 1c depicts an auxiliary process $X(t)$ which is obtained by replacing the abrupt downward jumps in $S(t)$ by linear decrements corresponding to a drift of minus 1. The sample path followed by $X(t)$ can indeed be modeled as the modulated process of an MRMFQ with K regimes and $K + 1$ states. Moreover, it is clear from sample path arguments that the steady-state distribution of the process $S(t)$ ($A(t)$) can be derived from that of $(X(t), Z(t))$ by censoring out the states corresponding to negative (positive) drifts. Therefore, we will first focus on the MRMFQ model for $X(t)$. For this purpose, we define the service state I_k for regime k for $k = 1, \dots, K$ during which the packet is being served by profile k with rate μ_k and $X(t)$ is increased with a unit drift. When the service of the current packet completes in state I_k , the system transits into a state denoted by \mathcal{D} during which $X(t)$ is decreased with a unit drift for an exponentially distributed amount of time with mean $1/\lambda$ so that the delay of the new HoL packet is reduced by an amount corresponding to its inter-arrival time. If $T^{(k-1)} \leq X(t) < T^{(k)}$ for some $k \leq K$, the system transits into state I_k and so on. Moreover, $X(t)$ may hit zero in state \mathcal{D} meaning that there are no packets waiting in the queue. When $X(t) = 0$, once a packet arrives at the system, the server selects profile 1 with a service rate of μ_1 for this new packet. Hence, the only transition at the boundary $X(t) = 0$ occurs out of state \mathcal{D} into state I_1 with rate λ . With states \mathcal{D} and I_i for $i = 1, \dots, K$, the background process, denoted by $Z(t)$, has $K + 1$ states in total. State transitions for the possible cases are illustrated in Fig. 2. Moreover, with the states ordered as $I_K, I_{K-1}, \dots, I_1, \mathcal{D}$, the infinitesimal generator matrix of regime- j , denoted by $Q^{(j)}$, for $j = 1, \dots, K$, is written as follows:

$$\begin{matrix}
 & I_K & \cdots & I_{j+1} & I_j & I_{j-1} & \cdots & I_1 & \mathcal{D} \\
 I_K & \left[\begin{array}{cccccccc}
 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\
 \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
 I_{j+1} & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\
 I_j & 0 & \cdots & 0 & -\mu_j & 0 & \cdots & 0 & \mu_j \\
 I_{j-1} & 0 & \cdots & 0 & 0 & -\mu_{j-1} & \cdots & 0 & \mu_{j-1} \\
 \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
 I_1 & 0 & \cdots & 0 & 0 & 0 & \cdots & -\mu_1 & \mu_1 \\
 \mathcal{D} & 0 & \cdots & 0 & \lambda & 0 & \cdots & 0 & -\lambda
 \end{array} \right. & .
 \end{matrix} \quad (5)$$

Note that since $X(t)$ increases in the service state I_k for $k = 1, \dots, j-1$, there may be transitions from state I_k to state \mathcal{D} in regime- j for $k \leq j$. We set $\tilde{Q}^{(j)} = Q^{(j+1)}$ as the generator at boundary- j for $j = 1, \dots, K$. $\tilde{Q}^{(0)}$ is similar to $Q^{(1)}$ except that there is no transition from state I_1 to state \mathcal{D} at boundary 0 and the only transition is from state \mathcal{D} to state I_1 . Moreover, the drift matrices at regime- k and boundary- k , denoted by $R^{(k)}$ and $\tilde{R}^{(k)}$, respectively, are written as follows:

$$\begin{aligned} R^{(k)} &= \mathbf{diag}(\mathbf{I}, -1), \quad 1 \leq k \leq K, \\ \tilde{R}^{(k)} &= \begin{cases} R^{(k+1)}, & 1 \leq k < K, \\ \mathbf{max}(0, R^{(1)}), & k = 0, \end{cases} \end{aligned} \quad (6)$$

where \mathbf{max} is the element-wise operator and \mathbf{I} denotes an identity matrix of appropriate size. This concludes the construction of the MRMFQ model.

C. Performance Metrics

Since the virtual waiting time $A(t)$ dictates the amount of delay that a virtual packet arrival will experience, the steady-state probability distribution of state \mathcal{D} is to be used to obtain the performance metrics of interest including the average power consumption, and the queuing delay distribution, as a direct consequence of the PASTA (Poisson Arrivals See Time Averages) property. For the purpose of obtaining the steady-state distribution of $A(t)$ from that of the fluid process $(X(t), Z(t))$, we censor out all the states $I_k, k = 1, \dots, K$. Mathematically, we have the following identity:

$$\lim_{t \rightarrow \infty} \Pr\{A(t) \leq x\} = \lim_{t \rightarrow \infty} \frac{\Pr\{Z(t) = \mathcal{D}, X(t) \leq x\}}{\Pr\{Z(t) = \mathcal{D}\}}. \quad (7)$$

We denote the probability that a packet is served with rate μ_k by q_k for $k = 1, \dots, K$:

$$q_k = \lim_{t \rightarrow \infty} \Pr\{T^{(k-1)} \leq A(t) < T^{(k)}\}, \quad 1 \leq k \leq K. \quad (8)$$

Moreover, we denote the probability that a newly arriving packet finds the queue empty by p_0 , i.e., $p_0 = \lim_{t \rightarrow \infty} \Pr\{A(t) = 0\}$. With these definitions, the average power consumption P can be written as:

$$P = p_0 P_I + \frac{(1-p_0)}{\sum_{k=1}^K \frac{q_k}{\mu_k}} \sum_{k=1}^K \frac{q_k P_k}{\mu_k}, \quad (9)$$

where P_I is the power consumed when the wireless link is idle, i.e., there is no transmission. The cumulative distribution function of the steady-state queuing delay $D(t)$, denoted by $F_D(\cdot)$, is also equal to that of $A(t)$ from the PASTA property [48]:

$$F_D(x) = \lim_{t \rightarrow \infty} \Pr\{D(t) \leq x\} = \lim_{t \rightarrow \infty} \Pr\{A(t) \leq x\}. \quad (10)$$

Since the solution to the M/M/1-type queue with K profiles has been reduced to the steady-state solution of an MRMFQ with $N = K + 1$ states and K regimes, the computational complexity of the overall algorithm to find the performance metrics of interest is $\mathcal{O}(N^3 K)$ or $\mathcal{O}(K^4)$ (see [47]) making it possible to rapidly evaluate a profile selection policy with tens of profiles.

D. Extension to PH-Type Service Times

In this subsection, we present the extensions required to handle the more general PH-type service time distribution scenario. To describe a PH-type distribution, a continuous-time MC is defined on the state space $\{1, \dots, l, l+1\}$ with state $l+1$ being absorbing, other states being transient, initial probability vector $(v, 0)$ and an infinitesimal generator of the form

$$\begin{bmatrix} S & h \\ 0 & 0 \end{bmatrix},$$

where v is a row vector of size l , the sub-generator S is $l \times l$, e denotes a column vector of ones of appropriate size and $h = -Se$ is a column vector of size l [8]. The time till absorption into the absorbing state $l+1$ denoted by Y is said to be of PH-type or order l , i.e., $Y \sim PH(v, S)$. The pdf of Y denoted by $f_Y(y)$ and $E[Y]$ are then given as:

$$f_Y(y) = -ve^{Sy}Se, \quad y \geq 0, \quad E[Y] = -vS^{-1}e. \quad (11)$$

Let Y denote the PH-distributed packet transmission time and let $Y \sim PH(v, S)$ of order l normalized so that $E[Y] = 1$ sec and $h = -Se$. Then the packet transmission time when using profile k will be denoted by $Y_k \sim (v, \mu_k S)$ with $E[Y_k] = 1/\mu_k$. For this delay-dependent M/PH/1 model, the MRMFQ constructed for exponential service times requires a slight modification. For this purpose, we define the set of states $\mathbf{I}_k = \{I_{k,1}, I_{k,2}, \dots, I_{k,l}\}$ where the individual state $I_{k,i}$ refers to when the packet is being served at rate μ_k while the service state being i . With this description, the infinitesimal generator matrix for regime- j , denoted by $Q^{(j)}$ of the modified MRMFQ, for $j = 1, \dots, K$, is then written as follows:

$$\begin{array}{cccccccc} & \mathbf{I}_K & \cdots & \mathbf{I}_{j+1} & \mathbf{I}_j & \mathbf{I}_{j-1} & \cdots & \mathbf{I}_1 & \mathcal{D} \\ \mathbf{I}_K & \left[\begin{array}{cccccccc} 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots \\ \mathbf{I}_{j+1} & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ \mathbf{I}_j & 0 & \cdots & 0 & \mu_j S & 0 & \cdots & 0 & \mu_j h \\ \mathbf{I}_{j-1} & 0 & \cdots & 0 & 0 & \mu_{j-1} S & \cdots & 0 & \mu_{j-1} h \\ \vdots & \vdots & & \vdots & \vdots & & \ddots & \vdots & \vdots \\ \mathbf{I}_1 & 0 & \cdots & 0 & 0 & 0 & \cdots & \mu_1 S & \mu_1 h \\ \mathcal{D} & 0 & \cdots & 0 & \lambda v & 0 & \cdots & 0 & -\lambda \end{array} \right] \\ & & & & & & & & \end{array} \quad (12)$$

Similar to the M/M/1 case, one can set $\tilde{Q}^{(j)} = Q^{(j+1)}$ as the generator at boundary- j for $j = 1, \dots, K$ and $\tilde{Q}^{(0)}$ is an all-zeros generator except for the last row which is the same as that of $Q^{(1)}$ given in (12). The drifts in all the I states and the \mathcal{D} state are set to $+1$ and -1 , respectively, as before. The way to compute all the related performance metrics is again via conditioning on the \mathcal{D} state as in (7) and using the same expressions (8), (9), and (10) for finding the quantities q_k , P , and $F_D(x)$, respectively, which completes the extension to PH-type service times. Since the modified MRMFQ has $N = Kl + 1$ states, the additional computational complexity introduced for the delay-dependent M/PH/1 queue will be $\mathcal{O}(l^3)$ when compared to the M/M/1 case. In the numerical examples, the focus will however be on exponential service times only.

V. TRANSMISSION PROFILE SELECTION POLICIES

Consider a wireless link with a given transmission profile set $\mathcal{N} = \{U_1, U_2, \dots, U_N\}$. Each profile U_i in this set is represented by the pair $(P^{(i)}, \mu^{(i)})$ where $P^{(i)}$ is the transmit power in Watts and $\mu^{(i)}$ is the service rate in packets/sec with $\mu^{(i)} > \mu^{(j)}$ for $i > j$. It may not be desirable to use all of the profiles in this set since some of the profiles may not be as energy efficient as others, or the policy needs to be constructed only with a few profiles for reduction of implementation complexity. A transmission profile selection policy is therefore governed by the choice of a particular subset $\mathcal{K} \subset \mathcal{N}$ to be used for power control and link adaptation along with the delay thresholds.

The delay violation probability is defined as $p_v = \lim_{t \rightarrow \infty} \Pr\{D(t) > D_0\}$, for a given delay bound D_0 . A transmission profile selection policy needs to be tuned to satisfy the statistical delay constraint which is given as $p_v < \varepsilon$ for a tolerance parameter ε . Next, we introduce various profile selection policies each of which involves the choice of the subset $\mathcal{K} = \{1, 2, \dots, K\} \subseteq \mathcal{N}$ and the thresholds $T^{(1)}, \dots, T^{(K-1)}$ such that profile $k \in \mathcal{K}$ with power P_k and rate μ_k is selected for the transmission for the HoL packet at the service start epoch if the queuing delay $D(t)$ experienced by the packet turns out to reside in the interval $[T^{(k-1)}, T^{(k)})$ where $T^{(0)} = 0$ and $T^{(K)} = \infty$.

A. Shortest Delay Policy (SDP)

The shortest delay policy aims at minimizing the expected delay by resorting to the single profile in \mathcal{N} with the largest service rate. Therefore, in *SDP*, $K = 1$ with $(P_1, \mu_1) = (P^{(N)}, \mu^{(N)})$. The performance metrics of *SDP* can be obtained with the conventional M/M/1 queuing model due to the use of a single regime. The average power consumption of *SDP*, denoted by P_{SDP} , is then written as

$$P_{SDP} = (1 - \rho)P_I + \rho P^{(N)} \quad (13)$$

where $\rho = \lambda/\mu^{(N)} < 1$. In the numerical examples, we only focus on the values of the arrival rate $\lambda < \lambda_M$ such that *SDP* satisfies the delay constraint, i.e., $p_v = \rho e^{-(\mu^{(N)} - \lambda_M)D_0}$ in the M/M/1 queue for the particular value $\lambda = \lambda_M$ of the arrival rate, equals ε [48].

B. Single Threshold Policy (STP)

STP is a binary rate adjustment policy in which the service rate is set to the maximum (minimum) possible rate when $D(t)$ is above (below) a single threshold value denoted by T_{STP} . Therefore, there are $K = 2$ regimes in *STP* which can be expressed as follows:

$$(P_k, \mu_k) = \begin{cases} (P^{(1)}, \mu^{(1)}), & k = 1, \\ (P^{(N)}, \mu^{(N)}), & k = 2, \end{cases} \quad (14)$$

and $T^{(1)} = T_{STP}$. The performance metrics of *STP* can be obtained with the MRMFQ model with three states and two regimes. The value of T_{STP} for which the average power consumption is minimized while satisfying the delay constraint is denoted by T_{STP^*} which is obtained by exhaustive search.

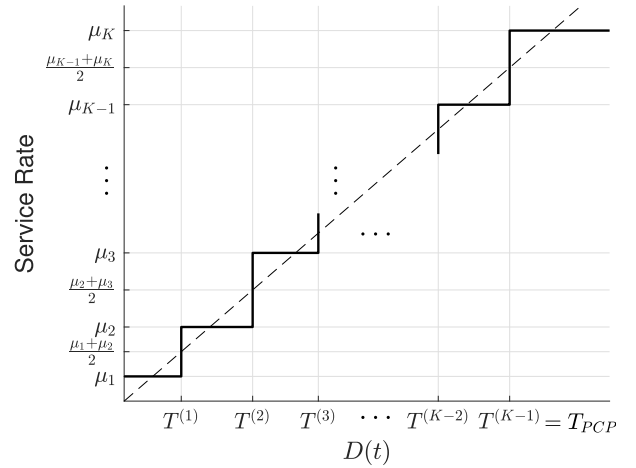


Fig. 3. Service rate selection for the HoL packet in *PCP*.

We denote the *STP* employing the particular threshold value T_{STP^*} by *STP**.

C. Proportional Control Policy (PCP)

In *PCP*, all of the available profiles from the set \mathcal{N} are used for rate adjustment, i.e., $K = N$, and the following identity holds:

$$(P_k, \mu_k) = (P^{(k)}, \mu^{(k)}), \quad k = 1, 2, \dots, N. \quad (15)$$

In *PCP*, a threshold value T_{PCP} is defined for the queuing delay $D(t)$ above which the service rate μ_K is to be selected. Moreover, similar to *STP*, when $D(t) = 0$, the service rate is set to μ_1 . When $0 < D(t) < T_{PCP}$, the service rate is selected from the set $\{\mu_1, \mu_2, \dots, \mu_K\}$ in a way that the service rate is linearly proportional with $D(t)$ as shown in Fig. 3. Mathematically, when $0 < D(t) < T_{PCP}$, proportional control is applied as follows:

$$T^{(k)} = \begin{cases} \frac{T_{PCP}(\mu_k + \mu_{k+1})}{\mu_{K-1} + \mu_K}, & 0 < k < K - 1, \\ T_{PCP}, & k = K - 1. \end{cases} \quad (16)$$

The performance metrics of *PCP* are obtained with the MRMFQ model with $N + 1$ states and N regimes. Similar to *STP*, the value of T_{PCP} for which the average power consumption is minimized while satisfying the delay constraint is denoted by T_{PCP^*} giving rise to *PCP** representing the particular *PCP* that employs the threshold value T_{PCP^*} .

D. Energy-Efficient PCP (EPCP)

Some of the profiles in \mathcal{N} may not be as effective as others in terms of energy efficiency. Therefore, we propose an enhanced policy, namely *EPCP*, in which a subset of \mathcal{N} is first selected that contains relatively energy-efficient profiles. Then, *PCP* is applied on this subset with the threshold value T_{EPCP} , rather than the entire set \mathcal{N} as is the case for the ordinary *PCP*.

Consider the profiles U_f , U_h and U_j in a transmission profile set \mathcal{N} given that $f < h < j$. We define the profile

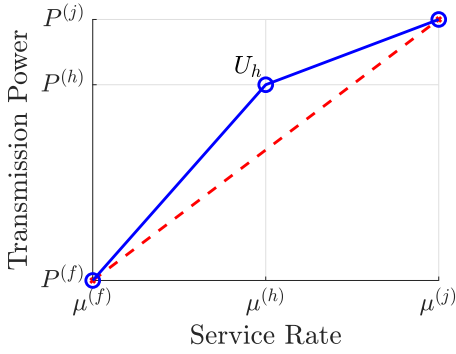


Fig. 4. An example of a relatively energy-inefficient profile.

U_h as a relatively energy-inefficient profile if there exists any pair (f, j) such that the following inequality holds:

$$\frac{(P^{(j)} + P^{(h)})(\mu^{(j)} - \mu^{(h)}) + (P^{(h)} + P^{(f)})(\mu^{(h)} - \mu^{(f)})}{(P^{(j)} + P^{(f)})(\mu^{(j)} - \mu^{(f)})} > 1. \quad (17)$$

An example of a relatively energy-inefficient profile U_h is shown in Fig. 4. Suppose for a given time, the profile U_f is being used. The strategy of switching to either U_h or U_j results in an increase in throughput. However, the throughput increase per Watt is less for the strategy of switching to U_h from U_f than it is for switching to U_j . Therefore, a relatively energy-inefficient profile can be identified if there exists another profile that gives more throughput at the expense of lesser power requirement per bit than that particular profile, i.e., U_h in Fig. 4.

EPCP aims at constructing the subset \mathcal{V} such that all relatively energy-inefficient profiles in the set \mathcal{N} are excluded. For this purpose, we propose an energy inefficiency index for a given profile subset $\mathcal{L} \subseteq \mathcal{N}$, denoted by $\Gamma(\mathcal{L})$, which is given by:

$$\Gamma(\mathcal{L}) = \sum_{l=2}^L (P^{(l)} + P^{(l-1)})(\mu^{(l)} - \mu^{(l-1)}). \quad (18)$$

We obtain the profile subset \mathcal{V} of size V , which minimizes the energy inefficiency metric over all possible subsets containing the two profiles U_1 and U_N :

$$\mathcal{V} = \arg \min_{\{U_1, U_N\} \subseteq \mathcal{L} \subseteq \mathcal{N}} \Gamma(\mathcal{L}) \quad (19)$$

The value of the threshold T_{EPCP} for which the average power consumption is minimized while satisfying the delay constraint is denoted by T_{EPCP^*} and *EPCP** denotes the corresponding *EPCP*.

For each of the three policies $p \in \{STP, PCP, EPCP\}$ (called basic policies hereafter), the profile to be used in the first regime is always fixed to the minimum service rate profile U_1 . Relaxing this fixed choice allows one to obtain an extended policy for each of the three basic policies. For this purpose, we use all the subsets $\{U_m, U_{m+1}, \dots, U_N\} \subseteq \mathcal{N}, m \in \{1, 2, \dots, N-1\}$ indexed by m as the starting profile set and for each such subset, we apply the methodology described above. Using a two-dimensional exhaustive

search, we propose to use the particular subset m_{p^*} and the corresponding threshold parameter T_{p^*} which minimizes the average power consumption while satisfying the delay constraint. The resulting policies are named as *STP**, *PCP**, and *EPCP**, respectively.

For each of the six proposed policies (three basic and three extended) policies, denoted by policy p , we define a percentage energy gain relative to *SDP* as follows:

$$G_p = 100 \frac{(P_{SDP} - P_p)}{P_{SDP}}. \quad (20)$$

VI. NUMERICAL RESULTS

In this section, we will first outline the simulation results from the LTE physical layer performance study for the Physical Downlink Shared Channel (PDSCH) detailed in [49]. Then, the construction process of the universal profile set \mathcal{N} from physical layer simulations is described. Subsequently, the transmission profile set $\mathcal{K} \subseteq \mathcal{N}$ is obtained for all the six proposed policies. Finally, the analytical model is used to compare the energy gains of the six proposed policies with respect to the baseline policy *SDP* for a wide range of system parameters.

A. System Setup for the Numerical Examples

We assume that the packet arrival process is Poisson with rate λ and packet sizes are exponentially distributed with mean $\beta = 500$ Bytes. We assume no power consumption in the transmitter when idle, i.e., $P_I = 0$. For the multi-path fading model, we consider the Extended Pedestrian A model with Doppler frequency of 5 Hz (EPA5), MIMO configuration is assumed to be 2×2 spatial multiplexing, and perfect channel estimator is assumed as in [49]. LTE-TDD frame structure is assumed as in [49] where each Physical Resource Block (PRB) consists of 12 sub-carriers with 15 kHz carrier spacing. We fix the number of PRBs to $N_B = 50$ that is allocated to the wireless link of interest within a given sub-frame with a duration 1 ms. For other parameters of the physical layer simulation setup, we refer to the study [49]. For a given average Signal-to-Noise Ratio (SNR) at the receiver, denoted by α (in dB), we denote the throughput by $r(\alpha, I_M)$ in bits/PRB and the Block Error Rate (BLER) by $e(\alpha, I_M)$, where I_M denotes the Modulation and Coding Scheme (MCS) index. The optimal I_M value (denoted by I_M^*) is selected in such a way that it will maximize the throughput while meeting a target Block Error Rate (BLER) denoted by e_b :

$$\tau(\alpha, I_M) = \begin{cases} r(\alpha, I_M), & e(\alpha, I_M) \leq e_b \\ 0, & \text{otherwise} \end{cases} \quad (21)$$

$$I_M^* = \arg \max_{I_M} \tau(\alpha, I_M) \quad (22)$$

The optimal throughput $r^*(\alpha)$ (bits/PRB) for a particular SNR value α is obtained by using the particular MCS index I_M^* :

$$r^*(\alpha) = r(\alpha, I_M^*) \quad (23)$$

Physical layer simulations have been conducted in [49] from which the performance metrics $e(\alpha, I_M)$ and $r(\alpha, I_M)$ are

TABLE I
BLER $e(\alpha, I_M)$ AS A FUNCTION OF THE SNR α AND MCS INDEX I_M

α/I_M	0	1	2	3	4	5	6	7	8	9	10	11
1	0.043	0.098	0.154	0.281	0.409	0.536	0.662	0.772	0.881	0.94	1	1
2	0.014	0.056	0.098	0.199	0.3	0.426	0.553	0.678	0.803	0.901	1	1
3	0.005	0.032	0.06	0.133	0.206	0.325	0.444	0.575	0.707	0.831	0.956	0.999
4	0	0.015	0.03	0.081	0.131	0.237	0.343	0.474	0.604	0.735	0.866	0.936
5	0	0.006	0.012	0.044	0.076	0.166	0.256	0.381	0.507	0.635	0.762	0.85
6	0	0	0	0.02	0.04	0.112	0.185	0.299	0.414	0.539	0.663	0.764
7	0	0	0	0.006	0.012	0.07	0.128	0.228	0.328	0.446	0.564	0.671
8	0	0	0	0.001	0.002	0.044	0.086	0.169	0.253	0.363	0.473	0.585
9	0	0	0	0	0.001	0.028	0.055	0.123	0.19	0.289	0.389	0.503
10	0	0	0	0	0	0.014	0.029	0.083	0.136	0.225	0.313	0.428
11	0	0	0	0	0	0.007	0.014	0.054	0.095	0.171	0.247	0.364
12	0	0	0	0	0	0	0	0.033	0.066	0.127	0.188	0.305
13	0	0	0	0	0	0	0	0.023	0.047	0.093	0.14	0.247
14	0	0	0	0	0	0	0	0.017	0.034	0.067	0.1	0.194
15	0	0	0	0	0	0	0	0.012	0.025	0.048	0.071	0.15
16	0	0	0	0	0	0	0	0.009	0.018	0.034	0.051	0.111
17	0	0	0	0	0	0	0	0.006	0.013	0.024	0.036	0.079
18	0	0	0	0	0	0	0	0.004	0.008	0.017	0.025	0.056
19	0	0	0	0	0	0	0	0.002	0.004	0.011	0.018	0.042
20	0	0	0	0	0	0	0	0	0.001	0.007	0.013	0.032

TABLE II
THE THROUGHPUT $r(\alpha, I_M)$ IN BITS/PRB AS A FUNCTION OF THE SNR α AND MCS INDEX I_M

α/I_M	0	1	2	3	4	5	6	7	8	9	10	11
1	52.98	64.88	74.92	82.03	85.64	81.5	69.58	56.52	33.1	18.98	0	0
2	54.55	67.94	79.94	91.49	101.45	100.69	92.15	79.69	54.69	31.36	0	0
3	55.07	69.62	83.25	98.95	114.98	118.46	114.64	105.19	81.64	53.79	13.94	7.81
4	55.31	70.88	85.96	104.98	125.84	133.89	135.44	130.37	110.15	84.56	42.77	34.54
5	55.34	71.54	87.55	109.16	133.86	146.41	153.42	153.28	137.36	116.66	75.77	67.76
6	55.36	71.99	88.62	111.93	139.12	155.89	168.21	173.63	163.11	147.28	107.48	100.23
7	55.36	72	88.63	113.55	143.22	163.3	179.8	191.31	187.22	177.03	139.33	133.96
8	55.36	72	88.64	114.13	144.67	167.91	188.57	205.86	208.08	203.51	168.37	164.83
9	55.36	72	88.64	114.18	144.81	170.68	194.85	217.43	225.61	227.03	195.31	194.04
10	55.36	72	88.64	114.24	144.96	173.05	200.22	227.33	240.6	247.7	219.44	220.35
11	55.36	72	88.64	114.24	144.96	174.43	203.46	234.4	252.11	264.93	240.69	243.24
12	55.36	72	88.64	114.24	144.96	175.63	206.29	239.68	260.18	278.85	259.28	263.77
13	55.36	72	88.64	114.24	144.96	175.66	206.34	242.11	265.56	289.76	274.92	282.44
14	55.36	72	88.64	114.24	144.96	175.68	206.4	243.77	269.22	298.17	287.56	298.64
15	55.36	72	88.64	114.24	144.96	175.68	206.4	244.86	271.67	304.17	296.75	311.52
16	55.36	72	88.64	114.24	144.96	175.68	206.4	245.7	273.55	308.54	303.33	321.96
17	55.36	72	88.64	114.24	144.96	175.68	206.4	246.37	275.05	311.77	308.07	330.17
18	55.36	72	88.64	114.24	144.96	175.68	206.4	246.92	276.29	314.16	311.42	335.95
19	55.36	72	88.64	114.24	144.96	175.68	206.4	247.43	277.43	315.9	313.61	339.63
20	55.36	72	88.64	114.24	144.96	175.68	206.4	247.83	278.35	317.36	315.47	342.43

tabulated as a function of the SNR parameter α (in 1 dB granularity) and the MCS index I_M in tables I and II, respectively. In our system model, we consider a relatively low value for $e_b = 0.02$ in order to reduce the effect of HARQ retransmissions on the service time distribution of packets. For this BLER constraint, $r^*(\alpha)$ values obtained using (23) are marked in bold text in Table II.

For a given channel condition and receiver sensitivity, α is a function of transmit power, which makes it possible to adjust the parameter α by varying the transmit power, which cannot exceed the maximum limit P_{max} . We define the maximum attainable average SNR, denoted by α_m , as the value of α obtained when the transmit power is P_{max} . It is possible to further reduce the transmit power (thus reduce the SNR) as long as the target BLER is satisfied. Also note that when $\alpha = 1$ dB, there does not exist any MCS which satisfies the target BLER. In line with the simulation results of the study [49] which are presented in 1 dB granularity, we reduce the transmit power by 1 dBm at each step to construct the transmission profiles. In particular, for each value

of the receiver SNR $\alpha \in \{2, 3, \dots, \alpha_m\}$, we obtain a different transmission profile. For a profile corresponding to a particular value α , power (in dBm) and service rate attributes of the constructed profile can be written as:

$$P(\alpha) = P_{max} - (\alpha_m - \alpha), \quad \mu(\alpha) = \frac{1000r^*(\alpha)N_B}{8\beta}. \quad (24)$$

In our study, we assume $P_{max} = 46$ dBm which is typical for PDSCH. The power and service rate attributes of the constructed profiles in \mathcal{N} are provided in Table III for the case $\alpha_m = 12$ dB and $e_b = 0.02$, and of those used for *STP*, *PCP*, and *EPCP* are illustrated in Fig. 5. Note that *PCP* uses all the profiles in \mathcal{N} and *EPCP* eliminates all the relatively energy-inefficient profiles from the set \mathcal{N} by means of minimizing the area under the curve in Fig. 5.

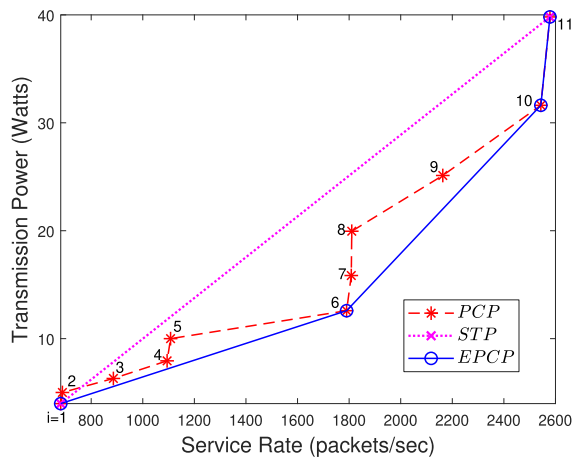
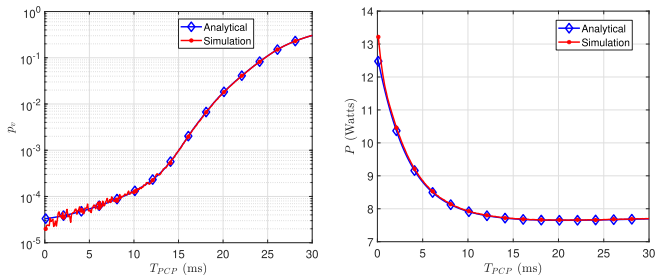
B. Model Validation

In this section, frame level simulations are performed to validate the proposed analytical model. In our simulations,

TABLE III

PROFILES IN THE SET \mathcal{N} FOR THE CASE $\alpha_m = 12$ dB AND $e_b = 0.02$.

i	$P^{(i)}$	$\mu^{(i)}$	i	$P^{(i)}$	$\mu^{(i)}$
1	3.98	681.9	7	15.84	1808.4
2	5.01	688.4	8	19.95	1810.2
3	6.30	886	9	25.11	2163.1
4	7.94	1094.3	10	31.62	2543.3
5	10	1107.8	11	39.81	2578.6
6	12.58	1790.2			

Fig. 5. Transmission power and service rate attributes of the profiles used for STP , PCP , and $EPCP$, when $\alpha_m = 12$ dB and $e_b = 0.02$.(a) Delay violation probability p_v (b) Average power consumption P Fig. 6. Performance metrics of the particular policy PCP as a function of the parameter T_{PCP} .

sub-frame packet structure for LTE is considered such that the required byte paddings are performed to align the size of payload to the nearest Transport Block size [50]. The total number of packet arrivals is set to 10^7 for each simulation. For a specific example scenario, we consider $D_0 = 15$ ms, $\alpha_m = 12$ dB, and $\lambda = 1000$. The delay violation probability and the average power consumption of the particular policy PCP with respect to its threshold parameter T_{PCP} are depicted in Fig. 6 for both the analytical model and simulations. It can be concluded that the analytical results are in line with the simulation results. When the performance parameter T_{PCP} is close to zero, the average power consumption appears to be slightly higher than for the analytical model. The reason is that the energy spent for the padded bytes increases when

the packets are transmitted with profiles using higher I_m values more frequently for relatively small threshold parameter T_{PCP} . However, the effect of padding on power savings can be considered negligible to none depending on system parameters. Therefore, for the rest of the paper, we will use only the proposed analytical model for evaluating the proposed profile selection policies.

C. Performance Evaluation of the Proposed Policies

In this section, the energy gain performance of the proposed profile selection policies with respect to the three system parameters λ , α_m , and D_0 are evaluated.

In the first two examples, we study a specific scenario when $D_0 = 15$ ms, $\alpha_m = 12$ dB, and the tolerance parameter ε is set to 0.001. For the purpose of laying out the methodology, we first fix $\lambda = 1000$ in which case $P_{SDP} = 15.44$ Watts using (13). The delay violation probability and the average power consumption of the three proposed policies with respect to their threshold parameter T_p , $p \in \{STP, PCP, EPCP\}$ are depicted in Fig. 7(a) and Fig. 7(b), respectively. The delay violation probability of PCP appears to be lower than that of $EPCP$ (and also STP) as seen in Fig. 7a(a) which indicates that the exclusion of some of the intermediate profiles for energy efficiency purposes slightly reduces the delay performance. On the other hand, the average power consumption of $EPCP$ is much lower than of PCP (and also STP). For a given policy $p \in \{STP^*, PCP^*, EPCP^*\}$, the optimum threshold value T_p is marked on Fig. 7(a) and the corresponding power consumption figures are illustrated in Fig. 7(b).

In the second numerical example, the energy gains of the six proposed policies are obtained with respect to the packet arrival rate $\lambda < \lambda_M = 2130.766$ is presented in Fig. 8. For relatively lower values of the arrival rate, all the six proposed policies appear to be serving at the profile with the lowest possible service rate for majority of the time and their power consumption figures are similar with an around 62% gain over SDP . This is because when the system load is relatively low, STP_e^* , PCP_e^* and $EPCP_e^*$ select the parameter $m_p^* = 1$ which makes their energy gains same as their basic versions STP^* , PCP^* and $EPCP^*$, respectively. For relatively higher arrival rates in the vicinity of λ_M , the extended versions provide higher energy gains. We note that the particular policy $EPCP_e^*$ consistently outperforms all other policies for all values of the arrival rate whereas a simple-to-implement policy STP_e^* using only two profiles is able to provide acceptable energy gains with slightly degraded performance when compared with $EPCP_e^*$.

In the third numerical example, we fix $\lambda = 1000$, $\alpha_m = 12$ dB, and study the effect of the choice of the delay bound D_0 on the energy gain performance. For this particular scenario, SDP does not satisfy the delay constraint for $D_0 < 3.776$ ms. In Fig. 9, percentage energy gains of the proposed policies are depicted with respect to the delay bound $D_0 > 3.776$ ms. We observe that for a wide range of relatively low D_0 values, the basic versions of the proposed policies provide no energy gain at all while their extended versions still provide

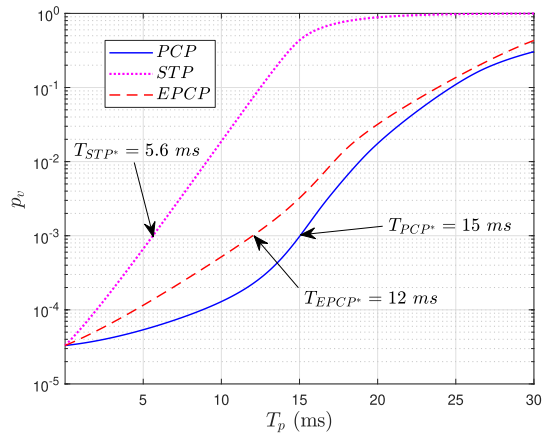
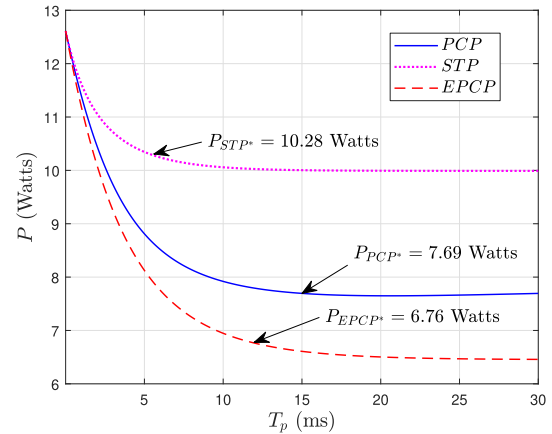
(a) Delay violation probability p_v (b) Average power consumption P

Fig. 7. Performance metrics of the three proposed policies STP , PCP , and $EPCP$ as a function of the parameter T_p when $\lambda = 1000$ packets/sec, the delay bound $D_0 = 15$ ms, and the maximum attainable SNR $\alpha_m = 12$ dB.

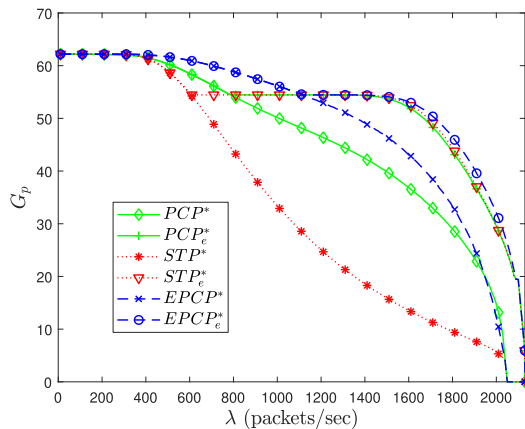


Fig. 8. The percentage energy gain G_p as a function of the arrival rate λ for the six proposed policies.

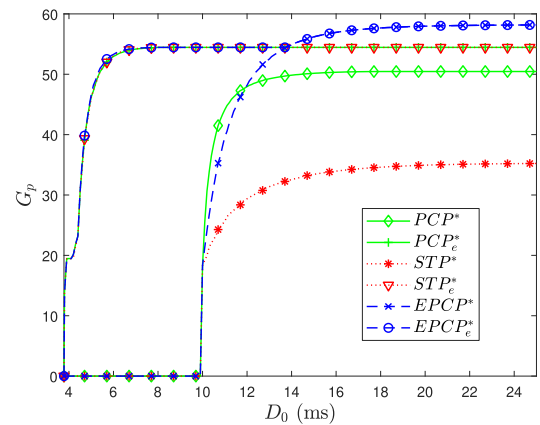


Fig. 9. The percentage energy gain G_p as a function of the delay bound D_0 for the six proposed policies.

substantial gains. This shows that the flexibility of adjusting the minimum service rate allows one to obtain energy gains even for lower values of the delay bound D_0 . Also note that $EPCP_e^*$ provides the maximum energy gain for all values of the delay bound parameter when compared to other policies whereas it outperforms STP_e^* again only slightly for this example.

In the next example, we set $\lambda = 1000$, $D_0 = 15$ ms, and we obtain G_p for the six policies with respect to varying maximum attainable SNR α_m which is depicted in Fig. 10. It can be concluded that energy gain of all the policies increases as α_m increases but $EPCP_e^*$ consistently outperforming all others again with STP_e^* lagging slightly behind. Also note that $EPCP^*$, $EPCP_e^*$, STP^* and STP_e^* turn out to use the same profile set \mathcal{K} for $\alpha_m = 7$ dB giving rise to the same gain for this particular value of α_m .

Finally, we plot the absolute saved power $P_p^{saved} = P_{SDP} - P_p$ at the output of the antenna element for any of the six proposed policies indexed by p in Fig. 11 as a function of the arrival rate λ . We observe that for all the applied policies, the absolute saved power vanishes for very low arrival rates

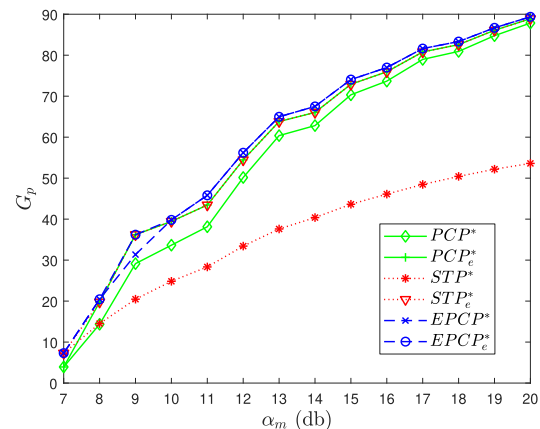


Fig. 10. The impact of the maximum attainable average SNR α_m on the percentage energy gain G_p for the six proposed policies.

but it peaks for medium to higher arrival rates. For arrival rates close to λ_{max} , this saving is brought down to zero as expected since only the highest service rate profile would be used in

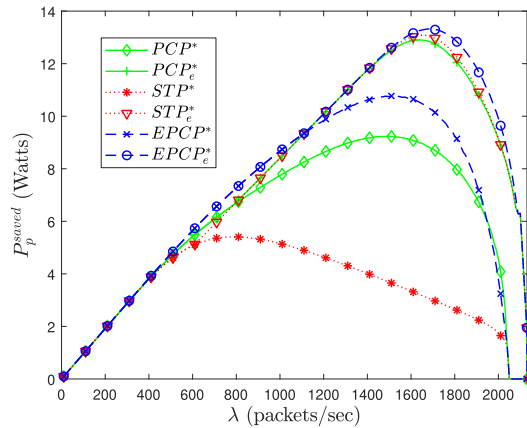


Fig. 11. The absolute saved power P_p^{saved} as a function of the arrival rate λ for the six proposed policies.

this situation. We also observe that the particular choice of $p = EPCP_e^*$ offers the maximum saving for all the arrival rates. Also note that the total power saving at the BS would be proportional with P_p^{saved} with the proportionality constant governed by the number of transmit/receive antennas per site and the power gradient term ΔP .

VII. CONCLUSIONS

In this paper, we study delay-dependent transmission profile selection policies for a wireless link in order to minimize the average transmission power consumption while keeping the delay violation probability below a certain tolerance value. Under the assumption of Poisson packet arrivals and exponentially distributed service times, the wireless link is first modeled as a single server M/M/1 queue with delay-dependent service times. A multi-regime Markov fluid queue model has been introduced to numerically solve the resulting queuing system so as to study the energy-delay trade-off for six transmission profile selection policies proposed in this paper. Moreover, we also provide the MRMFQ model required to generalize the methodology to more general PH-type service times as well. Out of all the proposed policies studied in this paper, the policy $EPCP_e^*$ is shown to consistently outperform all the other studied policies through a wide range of system parameters in terms of average power consumption while satisfying delay constraints. On the other hand, the policy STP_e^* slightly lags $EPCP_e^*$ in energy performance but the use of binary control with only two profiles in STP_e^* is an apparent advantage in terms of reduced implementation complexity.

REFERENCES

- [1] G. Auer *et al.*, "How much energy is needed to run a wireless network?" *IEEE Trans. Wireless Commun.*, vol. 18, no. 5, pp. 40–49, Oct. 2011.
- [2] H. Holtkamp, G. Auer, V. Giannini, and H. Haas, "A parameterized base station power model," *IEEE Commun. Lett.*, vol. 17, no. 11, pp. 2033–2035, Nov. 2013.
- [3] A. Fehske, G. Fettweis, J. Malmodin, and G. Biczok, "The global footprint of mobile communications: The ecological and economic perspective," *IEEE Commun. Mag.*, vol. 49, no. 8, pp. 55–62, Aug. 2011.
- [4] C. Gunaratne, K. Christensen, B. Nordman, and S. Suen, "Reducing the energy consumption of Ethernet with adaptive link rate (ALR)," *IEEE Trans. Comput.*, vol. 57, no. 4, pp. 448–461, Apr. 2008.

- [5] K. Nichols and V. Jacobson, "Controlling queue delay," *Queue*, vol. 10, no. 5, pp. 42–50, May 2012.
- [6] K. Nichols, V. Jacobson, A. McGregor, and J. Iyengar, "Controlled Delay Active Queue Management," document RFC 8289, Internet Requests for Comments, Jan. 2018. [Online]. Available: <https://tools.ietf.org/rfc/rfc8289.txt>
- [7] H. E. Kankaya and N. Akar, "Solving multi-regime feedback fluid queues," *Stochastic Models*, vol. 24, pp. 425–450, Aug. 2008.
- [8] G. Latouche and V. Ramaswami, *Introduction to Matrix Analytic Methods in Stochastic Modeling*, 1st ed. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 1999.
- [9] K. A. Mehr, J. M. Niya, and N. Akar, "Queue management for two-user cognitive radio with delay-constrained primary user," *Comput. Netw.*, vol. 142, pp. 1–12, Sep. 2018.
- [10] G. Li, S. Jin, F. Zheng, X. Gao, and X. Wang, "Energy efficient link adaptation for downlink transmission of LTE/LTE-A systems," in *Proc. IEEE 78th Veh. Technol. Conf. (VTC Fall)*, Sep. 2013, pp. 1–5.
- [11] Y. Zhang, H.-M. Wang, T.-X. Zheng, and Q. Yang, "Energy-efficient transmission design in non-orthogonal multiple access," *IEEE Trans. Veh. Technol.*, vol. 66, no. 3, pp. 2852–2857, Mar. 2017.
- [12] Y. Niu *et al.*, "Energy-efficient scheduling for mmWave backhauling of small cells in heterogeneous cellular networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 3, pp. 2674–2687, Mar. 2017.
- [13] E. Uysal-Biyikoglu, B. Prabhakar, and A. El Gamal, "Energy-efficient packet transmission over a wireless link," *IEEE/ACM Trans. Netw.*, vol. 10, no. 4, pp. 487–499, Aug. 2002.
- [14] M. Zafer and E. Modiano, "Optimal rate control for delay-constrained data transmission over a wireless channel," *IEEE Trans. Inf. Theory*, vol. 54, no. 9, pp. 4020–4039, Sep. 2008.
- [15] X. Li, X. Dong, and D. Wu, "On optimal power control for delay-constrained communication over fading channels," *IEEE Trans. Inf. Theory*, vol. 57, no. 6, pp. 3371–3389, Jun. 2011.
- [16] W. Chen, M. J. Neely, and U. Mitra, "Energy-efficient transmissions with individual packet delay constraints," *IEEE Trans. Inf. Theory*, vol. 54, no. 5, pp. 2090–2109, May 2008.
- [17] A. Wierman, L. L. H. Andrew, and M. Lin, "Speed scaling: An algorithmic perspective," in *Handbook of Energy-Aware and Green Computing*. Boca Raton, FL, USA: CRC Press, Jan. 2012, pp. 385–406.
- [18] N. Bansal, T. Kimbrel, and K. Pruhs, "Speed scaling to manage energy and temperature," *J. ACM*, vol. 54, no. 1, pp. 3:1–3:39, Mar. 2007.
- [19] J. M. George and J. M. Harrison, "Dynamic control of a queue with adjustable service rate," *Oper. Res.*, vol. 49, no. 5, pp. 720–731, Sep./Oct. 2001.
- [20] T. V. Dinh, L. L. H. Andrew, and Y. Nazarathy, "Architecture and robustness tradeoffs in speed-scaled queues with application to energy management," *Int. J. Syst. Sci.*, vol. 45, no. 8, pp. 1728–1739, Aug. 2014.
- [21] M. Elahi, C. Williamson, and P. Woelfel, "Decoupled speed scaling: Analysis and evaluation," *Perform. Eval.*, vol. 73, pp. 3–17, Mar. 2014.
- [22] C. Tunc and N. Akar, "Performance modeling of delay-based dynamic speed scaling," in *Proc. 9th Int. Conf. Matrix-Analytic Methods Stochastic Models (MAM9)*, Jun. 2016, pp. 7–13.
- [23] R. W. Conway and W. L. Maxwell, "A queuing model with state dependent service rates," *J. Ind. Eng.*, vol. 12, no. 2, pp. 132–136, 1962.
- [24] W. K. Grassmann, X. Chen, and B. R. Kashyap, "Optimal service rates for the state-dependent M/G/1 queues in steady state," *Oper. Res. Lett.*, vol. 29, no. 2, pp. 57–63, 2001.
- [25] C. M. Harris, "Queues with state-dependent stochastic service rates," *Oper. Res.*, vol. 15, no. 1, pp. 117–130, 1967.
- [26] R. Bekker and O. J. Boxma, "An M/G/1 queue with adaptable service speed," *Stochastic Models*, vol. 23, no. 3, pp. 373–396, 2007.
- [27] R. Bekker, O. J. Boxma, and J. A. C. Resing, "Queues with service speed adaptations," *Statist. Neerlandica*, vol. 62, no. 4, pp. 441–457, 2008.
- [28] R. Bekker, S. C. Borst, O. J. Boxma, and O. Kella, "Queues with workload-dependent arrival and service rates," *Queueing Syst.*, vol. 46, no. 3, pp. 537–556, 2004.
- [29] J. Lee and J. Kim, "A workload-dependent M/G/1 queue under a two-stage service policy," *Oper. Res. Lett.*, vol. 34, no. 5, pp. 531–538, 2006.
- [30] R. F. Gebhard, "A queuing process with bilevel hysteretic service-rate control," *Naval Res. Logistics Quart.*, vol. 14, no. 1, pp. 55–67, 1967.
- [31] R. Bekker, "Queues with Lévy input and hysteretic control," *Queueing Syst.*, vol. 63, pp. 281–299, Dec. 2009.
- [32] 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.

- [33] G. Miao, N. Himayat, and G. Y. Li, "Energy-efficient link adaptation in frequency-selective channels," *IEEE Trans. Commun.*, vol. 58, no. 2, pp. 545–554, Feb. 2010.
- [34] B. Tomasi and J. C. Preisig, "Energy-efficient transmission strategies for delay constrained traffic with limited feedback," *IEEE Trans. Wireless Commun.*, vol. 14, no. 3, pp. 1369–1379, Mar. 2015.
- [35] M. Sinaie, A. Zappone, E. A. Jorswieck, and P. Azmi, "A novel power consumption model for effective energy efficiency in wireless networks," *IEEE Wireless Commun. Lett.*, vol. 5, no. 2, pp. 152–155, Apr. 2016.
- [36] A. Zappone, L. Sanguinetti, G. Bacci, E. Jorswieck, and M. Debbah, "Energy-efficient power control: A look at 5G wireless technologies," *IEEE Trans. Signal Process.*, vol. 64, no. 7, pp. 1668–1683, Apr. 2016.
- [37] A. Zappone, L. Sanguinetti, and M. Debbah, "Energy-delay efficient power control in wireless networks," *IEEE Trans. Commun.*, vol. 66, no. 1, pp. 418–431, Jan. 2018.
- [38] K. Wang and W. Chen, "Delay-aware energy-efficient communications over Nakagami-m fading channel with MMPP traffic," in *Proc. IEEE Int. Conf. Commun. Workshop (ICCW)*, Jun. 2015, pp. 2750–2755.
- [39] Z. Chen, N. Pappas, and M. Kountouris, "Energy harvesting in delay-aware cognitive shared access networks," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, May 2017, pp. 168–173.
- [40] I. Dimitriou, S. Alouf, and A. Jean-Marie, "A Markovian queueing system for modeling a smart green base station," in *Proc. Eur. Workshop Perform. Eng. (Lecture Notes in Computer Science)*, vol. 9272. Madrid, Spain: Springer, 2015, pp. 3–18.
- [41] Y. Li, G. Ozcan, M. C. Gursoy, and S. Velipasalar, "Energy efficiency of hybrid-ARQ under statistical queuing constraints," *IEEE Trans. Commun.*, vol. 64, no. 10, pp. 4253–4267, Oct. 2016.
- [42] M. Ozmen and M. C. Gursoy, "Wireless throughput and energy efficiency with random arrivals and statistical queuing constraints," *IEEE Trans. Inf. Theory*, vol. 62, no. 3, pp. 1375–1395, Mar. 2016.
- [43] M. Ozmen and M. C. Gursoy, "Energy-efficient power control in fading channels with Markovian sources and QoS constraints," *IEEE Trans. Commun.*, vol. 64, no. 12, pp. 5349–5364, Dec. 2016.
- [44] G. Ozcan, M. Ozmen, and M. C. Gursoy, "QoS-driven energy-efficient power control with random arrivals and arbitrary input distributions," *IEEE Trans. Wireless Commun.*, vol. 16, no. 1, pp. 376–388, Jan. 2017.
- [45] D. Anick, D. Mitra, and M. M. Sondhi, "Stochastic theory of a data-handling system with multiple sources," *Bell Syst. Tech. J.*, vol. 61, pp. 1871–1894, Oct. 1982.
- [46] M. Mandjes, D. Mitra, and W. Scheinhardt, "Models of network access using feedback fluid queues," *Queueing Syst.*, vol. 44, no. 4, pp. 365–398, 2003.
- [47] M. A. Yazici and N. Akar, "The workload-dependent MAP/PH/1 queue with infinite/finite workload capacity," *Perform. Eval.*, vol. 70, no. 12, pp. 1047–1058, Dec. 2013.
- [48] D. Gross and C. M. Harris, *Fundamentals of Queueing Theory*, 2nd ed. New York, NY, USA: Wiley, 1985.
- [49] W.-B. Yang and M. Souryal, "LTE physical layer performance analysis," NIST, Gaithersburg, MD, USA, Tech. Rep. 7986, May 2014.
- [50] *Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Layer Procedures*, document TS36.213, version 10.5.0., 3rd Generation Partnership Project (3GPP), Mar. 2012,



Ege Orkun Gavgam received the B.S. and M.S. degrees in electrical and electronics engineering from Bilkent University, Ankara, Turkey, in 2015 and 2018, respectively, where he is currently pursuing the Ph.D. degree. He is also a Design Engineer at ASELSAN, Ankara, a military communications company. His current research interests include the design and analysis of wireless communication systems and the performance modeling of wireless networks.



Caglar Tunc received the B.S. and M.S. degrees in electrical and electronics engineering from Bilkent University, Ankara, Turkey, in 2013 and 2016, respectively. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, New York University Tandon School of Engineering. His current research interests are on the broad area of wireless communications and stochastic modeling of networked systems.



Nail Akar received the B.S. degree from Middle East Technical University, Turkey, in 1987, and the M.S. and Ph.D. degrees from Bilkent University, Ankara, Turkey, in 1989 and 1994, respectively, all in electrical and electronics engineering. From 1994 to 1996, he was a Visiting Scholar and a Visiting Assistant Professor with the Computer Science Telecommunications Program, University of Missouri–Kansas City, USA. He joined the Long Distance Division, Technology Planning and Integration Group, Sprint, Overland Park, KS, USA, in 1996, where he held a senior member of technical staff position from 1999 to 2000. Since 2000, he has been with Bilkent University, where he is currently a Professor with the Electrical and Electronics Engineering Department. He visited the School of Computing, University of Missouri–Kansas City, as a Fulbright Scholar, in 2010, for a period of six months. His research interests include the performance analysis of computer and communication systems and networks, queuing models and tools, wireless networks, Internet of Things.