Modeling and Optimizing the LTE Discontinuous Reception Mechanism Under Self-Similar Traffic

Ke Wang, Xi Li, Hong Ji, *Senior Member, IEEE*, and Xiaojiang Du, *Senior Member, IEEE*

***Abstract*—The discontinuous reception (DRX) mechanism is adopted in Long-Term Evolution (LTE) systems as a core technology to prolong the battery lifetime of user equipment (UE). With the development of mobile Internet, there is an increasingly urgent need to optimize the DRX performance to accommodate the emerging applications. In this paper, to describe the self-similarity exhibited by the applications, a truncated-Pareto-distributed arrival traffic model is introduced into the LTE-DRX modeling framework. With this premise, a DRX analytical model based on a discrete-time semi-Markov process (DTSMP) is established. Using the proposed model, the performance of DRX operations under certain configurations can be evaluated precisely. To deploy it in practical use, we have designed an online power-saving strategy (OPSS) to improve the energy efficiency of the UE. The OPSS is conducted in two phases: estimation and optimization. In the first phase, several derived statistical estimators are deployed to capture the fluctuations of the traffic conditions and DRX operations. It is proved that these estimators could unbiasedly estimate the statistics within just 1 s. In the second phase, the DRX configuration is optimized by considering the trade off between the packet delay and the power-saving performance. Solid simulations are conducted to verify the accuracy of the DRX analytical model and to evaluate the efficiency of the OPSS. The well-matched results demonstrate that the analytical model is correctly derived. Moreover, we have proved that the proposed OPSS could outperform the conventional LTE DRX mechanism in terms of both energy conservation and packet delay.**

***Index Terms*—Discontinuous reception (DRX), discrete-time semi-Markov process (DTSMP), Long-Term Evolution (LTE), multimedia, power saving, self-similar traffic.**

NOMENCLATURE

t_I	Threshold of the Inactive timer.
t_{SD}	Length of short DRX cycle.
t_{LD}	Length of long DRX cycle.
t_{ON}	Threshold of On-duration timer.
N_{SD}	Maximum number of short DRX cycle.
α	Distribution shape parameter of truncated Pareto distribution.
x_m	Location parameter of truncated Pareto distribution.
m	Truncated value of truncated Pareto distribution.
h_i	Sojourn time distribution of state i.
U_i	Mean sojourn time in state i.
t_{busy}	Length of packets' continuous arriving during *active period*.
f_{ij}	Conditional distribution of the sojourn time in state j, given the previous state transition is from state i.
q_{ij}	Discrete-time semi-Markov kernel of state j, given the previous state transition is from state i.
PS	Power-saving factor.
D_S	Expected wake-up delay in *light sleep period*.
D_L	Expected wake-up delay in *deep sleep period*.
$\mathbb{E}(D)$	Expected wake-up delay.
$\hat{X}_m$	Unbiased estimator of x_m.
α^*	Unbiased estimator of α.
$\hat{U}_i(M)$	Estimator of U_i.
$\hat{\nu}(i)$	Estimator of ν_i.

I. INTRODUCTION

NOWADAYS, along with the rapid development of smart mobile devices, wireless communication technologies and various mobile Internet applications, with multimedia as one of the representatives, have emerged and gained huge popularity. This calls for larger system capacity, higher transmission rate, and better user experience, to drive the evolution of new technologies such as high-order modulation, beamforming, channel equalization, and multiple-input–multiple-output antennas. These physical-layer techniques not only could significantly improve spectrum efficiency but could increase the complexity of receivers' computational circuitry as well, thus draining more the battery power of user equipment (UE). Substantial improvements of power-saving operation mechanisms are necessary for UE design [1]. For this purpose, the Third-Generation Partnership Project (3GPP) LTE standard has specified new discontinuous reception (DRX) based on the Universal Mobile Telecommunications System (UMTS) DRX mechanism, namely LTE-DRX. It is worth noting that another important standard, i.e., Worldwide Interoperability for Microwave Access (WiMAX), has also introduced a novel sleep/idle transmission mechanism to prolong the UE's battery lifetime.

The basic idea of DRX is to allow UE to turn off its wireless transceiver to save energy from unnecessary consumption.

Manuscript received September 17, 2013; revised January 20, 2014 and March 3, 2014; accepted April 29, 2014. Date of publication June 3, 2014; date of current version July 14, 2016. This work was supported in part by the National Natural Science Foundation of China under Grant 61271182 and Grant 61302080 and in part by the U.S. National Science Foundation under Grant CNS-1065444. The review of this paper was coordinated by Dr. N.-D. Dao.

K. Wang, X. Li, and H. Ji are with the Key Laboratory of Universal Wireless Communications of the Ministry of Education and School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: wangke@bupt.edu.cn; lixi@bupt.edu.cn; jihong@bupt.edu.cn).

X. Du is with the Department of Computer and Information Sciences, Temple University, Philadelphia, PA 19122 USA (e-mail: dxj@ieee.org).

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/TVT.2014.2328232

In LTE, the evolved NodeB (eNB) controls the UE's DRX operation by a radio resource control (RRC) entity via the physical downlink control channel (PDCCH). With DRX, the UE is allowed to discontinuously monitor the PDCCH on which the downlink transmission grants are assigned. LTE defines two modes of DRX, depending on whether the UE has an active session, rather than defining power-saving classes corresponding to different kinds of traffic in WiMAX [2]. When there are no active sessions, LTE DRX works in `RRC_Idle` mode, and the UE performs similarly as other DRX mechanisms [3]. During the active sessions, the `RRC_Connected`-mode LTE DRX is activated. Different from other DRXs, LTE DRX in `RRC_Connected` mode allows the UE to enter sleep mode while it is registered with eNB. LTE DRX is one of the most effective power-saving mechanism and is already supported by the latest LTE phones (e.g. iPhone 5 and HTC One).

However, the present LTE DRX still has more room for improvement. Existing LTE DRX mechanism specified in the 3GPP Release 8 [4] is based on static sleep mode, which leads to inevitable performance degradation. To provide a *universal and easy-to-implement* method, it abandons the mechanism that is based on traffic types (used by WiMAX [2]) and adopts a static sleep mode that is expected to meet most of the traffic QoS requirements. Nevertheless, with the development of mobile Internet, a *universal* method that fits all traffic does not exist. It is necessary to consider different applications in LTE DRX. Recently, in 3GPP Release 11, work item "RAN Enhancements of Diverse Data Applications" [5] has been conducted to study the battery drain due to emerging mobile Internet applications, which also confirms the given conclusion. Therefore, the optimization of LTE DRX based on emerging mobile Internet applications could significantly influence the energy efficiency of LTE systems.

The main approach of existing works on optimizing LTE DRX is to adjust the *DRX parameters*. The 3GPP LTE standard has specified several DRX parameters to precisely control the UE's behavior, such as when to sleep, when to monitor the PDCCH, when to activate, and how long each state should take. In fact, the early discussion on the LTE DRX parameters described in 3GPP technical reports [6], [7] has already shown that 1) flexible configurable DRX parameters could obtain better performance than fixed parameters and 2) that different traffic conditions could influence the LTE DRX performance. Therefore, the DRX parameters should be adaptively optimized with changing traffic conditions.

Over the past few years, this area of research has attracted much interest. The related research can be divided into two directions: optimization based on concrete traffic types [6], [8], [9] and optimization based on general traffic analytical models [3], [10], [11]. The studies in the first direction are always with the aid of extensive simulations and realistic measurements. They focus on the traffic with specific operating patterns, such as hypertext markup language [6], voice over IP [8], [9], [12], and Hypertext Transfer Protocol [13]. In these works, the trade-off between power saving and delay is achieved by adjusting the DRX parameters. Meanwhile, the best DRX parameter configurations for different arrival rates are also given. They have the advantage of precisely fitting the specific traffic and are very practical to deploy. The drawback is also obvious: The scope of their solutions is narrow, and they cannot deal with various mobile Internet applications running simultaneously on the UE. Hence, the second direction is attracting a lot of attention these days.

Generally speaking, the efforts in the second direction focus on DRX analytical modeling and analytical-model-based optimization algorithms. The DRX analytical model is the foundation of optimization. All of the existing LTE DRX models [10], [11], [14] follow the method and assumptions in [3], in which an analytical model of UMTS DRX is provided for the first time. All these models assume that the packet arrival intervals and transmission times follow exponential and general distributions, respectively. Furthermore, they all adopt the semi-Markov chain as the main mathematical tool, and most of the statistics are derived by utilizing the memoryless property of exponential distribution. These models can predict power-saving performance and expected delay of LTE DRX under non-real-time traffic. Most existing analytical studies for optimizing traffic-based power-saving mechanisms are conducted in the WiMAX system. In [15], the sleep cycle optimization method of WiMAX in [16] is applied to LTE, and the LTE DRX model proposed in [10] is adopted. There are two common drawbacks of the existing works in the second direction. First, as shown in [17] and [18], the Poisson traffic model cannot accurately reflect the multimedia/mixed traffic in mobile Internet. Second, the traffic estimation algorithm is required to observe and record traffic conditions for a long time and tends to consuming many more eNB's resources.

The complexity of multimedia traffic is a natural consequence of integrating a diverse range of traffic resources such as video, voice, and data, which significantly differ in their traffic patterns and requirements [18]. In [19] and [20], it is demonstrated that broadband multimedia services and Internet traffic exhibit properties of self-similarity and long-range dependence (LRD). In general, self-similar traffic shows identical statistical characteristics over a wide range of time scales, which may have a significant impact on network performance. To the best of our knowledge, there is no existing work that has considered the self-similar traffic model in LTE DRX modeling. However, with the emergence of mobile Internet, we believe that considering the self-similar traffic in DRX studies is necessary.

In this paper, we first introduce and set up a self-similar traffic model, and then, we establish a new LTE DRX analytical model by using the discrete-time semi-Markov process (DTSMP). Considering practical deployment, we propose an online power-saving strategy (OPSS), by which the traffic condition and the statistics of DRX operations can be unbiasedly estimated and the optimal DRX parameter configuration can be eventually obtained.

Our first contribution is that this is the first work to introduce a self-similar traffic model into DRX modeling and to construct a complete analytical model. Since the interarrival times follow the truncated Pareto distribution, which do not have the memoryless property of the exponential distribution, the DRX modeling is quite different from previous works. We find that it is difficult to analyze the LTE DRX under self-similar traffic by using the methods in [10], [11], and [14].

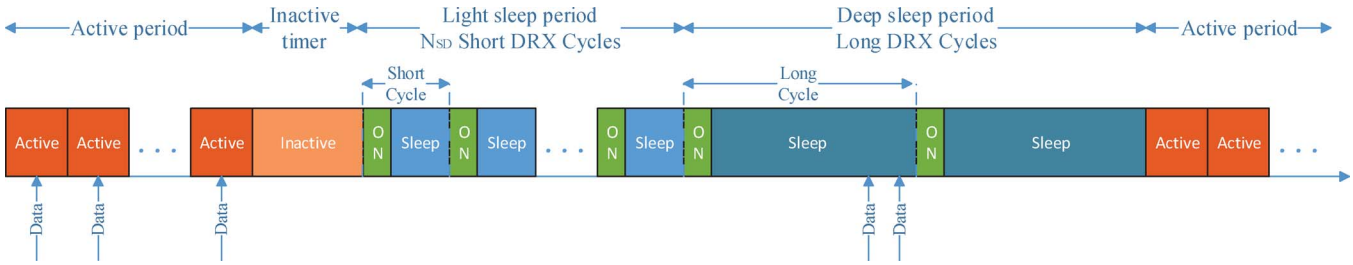


Fig. 1. 3GPP LTE DRX mechanism.

Furthermore, the continuous semi-Markov process, which is widely used in previous works, is not suitable for describing the discrete-time property of LTE DRX. Hence, we adopt the DTSMP as the main mathematical tool, and based on this the statistical parameters such as sojourn time in each state and transition probability are derived with the probability density function (pdf) of the truncated Pareto distribution. Moreover, following the suggestion in [21], we remove the self-transitions in the embedded Markov chain (EMC) and consider the state's self-transition as *sojourn in that state*. Simulation results show that, by this adjustment, our deduction meets the definition of the semi-Markov process better and is more suitable for the LTE DRX modeling, as compared with existing works.

Our second contribution is to propose a novel practical OPSS based on the proposed analytical model, which is different from estimating only the packet arrival rate in previous Poisson-traffic-based DRX studies [22]. In this paper, we design online estimators for both traffic conditions and LTE DRX operations. Specifically, by using the maximum likelihood (ML) method and the properties of gamma distribution, we derive an unbiased estimator for traffic conditions and prove its effectiveness. Based on the recent progress in estimating the DTSMP parameters [23], we propose an unbiased estimator for DRX operations. Extensive simulations and analysis show that the estimators can make unbiased estimations based on the random samples collected in just 1 s. Compared with other estimation algorithms, our methods can greatly reduce the eNB's resource consumption. Furthermore, we divide the estimation work between eNB and UE, where eNB estimates the traffic, and the UE estimates the statistics of the DRX operations. It is worth nothing that the idea of a UE-assisted power-saving mechanism has already been adopted in the 3GPP release 11 [5].

Extensive simulations has been conducted to verify the LTE-DRX analytical model and the proposed OPSS. The simulation results show a good match between the proposed model and the LTE DRX mechanism, and demonstrate that the power-saving efficiency of the DRX operation can be enhanced by our OPSS, while the delay requirements of UE is still preserved.

The remainder of this paper is organized as follows. In Section II, we introduce the background technology used in LTE DRX modeling. In Section III, we present our analytical model of LTE DRX. In Section IV, we present the online statistical estimators used in our optimization strategy. In Section V, the design of the OPSS is given in detail. In Section VI, we evaluate the analytical models and the online strategy with simulations. Section VII concludes this paper.

II. Background Technology

Here, we discuss two important fundamental conceptions: LTE `RRC_Connected`-mode DRX and self-similar traffic, which are the bases of the analytical model in the following.

A. *RRC_Connected-Mode DRX*

As mentioned in Section I, the difference between LTE DRX and the previous DRX is mainly concentrated on `RRC_Connected` mode. By this mode, the DRX parameters are allowed to be modified during the data transmission. Therefore, the `RRC_Connected` DRX is able to provide quick adaptation to traffic status changes and to obtain higher energy efficiency [10]. Therefore, in this paper, we only focus on the `RRC_Connected`-mode DRX. For simplicity, the "DRX" mentioned in the remainder of this paper stands for the LTE DRX.

In LTE, any data transmission requires that the UE is in "high power" RRC-connected state. With all data applications, there are often short moments when no data are sent or received, and during those moments, `RRC_Connected`-mode DRX can save energy. `RRC_Connected`-mode DRX cyclically wakes up and shuts down the receiver circuits to save energy. As shown in Fig. 1, the whole data transmission process can be divided into three states, namely *active period*, *light sleep period*, and *deep sleep period*. In *active period*, the packet arrival intervals are very short; therefore, the UE should stay awake to monitor the PDCCH and receive data as a non-DRX system. The other two states constitute the `RRC_Connected`-mode DRX, and the 3GPP LTE standard [4] specifies the following parameters to control the UE's behavior in these states.

- *Inactive timer* t_I denotes the maximum duration after the last transmission at which UE shall remain turning on the receiver to monitor the PDCCH. This timer is reset to zero and enabled immediately after successful reception of the PDCCH (resource grant or allocation).
- *Short DRX cycle* (t_{SD}) specifies the number of consecutive subframes that the UE shall follow in one DRX cycle during the *light sleep period*.
- *DRX short cycle timer* (N_{SD}) is expressed in integer numbers of short DRX cycles. N_{SD} indicates the maximum number of short DRX cycles in *light sleep period* before transitioning to the *deep sleep period*.
- *Long DRX cycle* (t_{LD}) specifies the number of consecutive subframes that the UE shall follow in one DRX cycle during the *deep sleep period*. A typical long DRX cycle is at least ten times longer than a short DRX cycle.

- *DRX Long cycle timer* (N_{LD}) is expressed in integer numbers of long DRX cycles. N_{LD} indicates the maximum number of long DRX cycles in *deep sleep period* before the eNB indicates the release of UE's RRC connection or, in other words, initiation of the `RRC_Idle` mode DRX.
- *On-duration timer* (t_{ON}) denotes the time period for which the UE shall stay awake in each DRX cycle. The ONduration is part of the DRX cycle and has the same length for both short and long DRX cycles by default. To facilitate the later discussion on t_{ON}, we define two new parameters η and η' to represent the proportion of t_{ON} in t_{SD} and t_{LD}, respectively. Hence, we can get $t_{\mathrm{ON}} = \eta t_{\mathrm{SD}} = \eta' t_{\mathrm{LD}}$ and $\eta' = \eta t_{\mathrm{SD}}/t_{\mathrm{LD}}$.

All these adjustable parameters are communicated by means of higher layer (RRC) signaling to each UE during the bearer setup and when parameter updates are needed. With the help of these parameters, we can describe the `RRC_Connected`-mode DRX precisely. Inactive timer equals to the time interval between the reception of packets during the *active period*. At the moment when this value is larger than t_I, the UE switches to *light sleep period*, during which only the short DRX cycle can be enabled. In each short DRX cycle, the UE first opens its receiver to monitor the PDCCH for t_{ON} subframes; then, regardless of the packet arrival is detected or not, the UE falls asleep for the remaining $t_{\mathrm{SD}} - t_{\mathrm{ON}}$ subframes. If the PDCCH indicates a downlink transmission, the UE transits from the *light sleep period* into the *active period* and cancels its DRX short cycle timer. Otherwise, the UE goes into the succeeding DRX short cycle and increases its DRX short cycle timer by 1. At the moment when the DRX short cycle timer is equal to $t_{\mathrm{SD}} + 1$, the UE switches to *deep sleep period*, and the long DRX cycle is activated. Except for the time duration after t_{ON} taking $t_{\mathrm{LD}} - t_{\mathrm{ON}}$ subframes and the maximum value of DRX long cycle timer being N_{LD}, the UE's behavior during the long DRX cycle is exactly the same as that during the short DRX cycle.

Since t_{SD} is smaller than t_{LD}, the short DRX cycle can be employed in higher frequency, which makes it fit for bursty traffic. However, when there exists services with aperiodic silent period or occasional packet arrivals, the long DRX cycle is more suitable [14]. In fact, the two-level DRX conception is a compromise with system complexity because more flexibility could be gained to fit different traffic patterns with more DRX cycle levels. To be practical, the DRX mechanism is always employed in a semi-static manner, which keeps UE's DRX parameters static for long duration and when updating is needed. Tuning the timer length to reach a suitable and functional length is a tough task. In this paper, we utilize the nature of semi-static by optimizing the DRX parameters based on the observation of the UE's traffic condition and DRX performance periodically.

B. Self-Similar Traffic

The traffic condition, particularly the interarrival time distribution, could significantly impact the performance of the DRX mechanism. As mentioned in Section I, multimedia traffic explicitly exhibits *self-similarity* properties. Thus, we need to figure out what property does self-similar traffic have and how could we use them to model the interarrival time distribution.

The self-similar traffic, in general, implies LRD [18], [19]. Mathematically, LRD is defined as the autocorrelations of sequence decaying hyperbolically with increasing lag. That is, a covariance stationary process $X = (X_t : t = 0, 1, 2, \ldots)$, with mean μ and variance σ^2, is LRD if it has an autocorrelation function $r(k),\ k \geq 0$, of the following form:

$$r(k) \sim k^{-\beta} L(t) \quad \text{as} \quad k \to \infty \tag{1}$$

where $0 < \beta < 1$, and $L(t)$ is slowly varying at infinity, i.e., $\lim_{t\to\infty} L(tx)/L(t) = 1$ for all $x > 0$. This definition implies that, for such a sequence, the autocorrelations have no finite sum, i.e., $\sum_{k=0}^{\infty} r(k) = \infty$. The parameter β of the autocorrelation function is related to the Hurst parameter by $H = 1 - \beta/2$. The Hurst parameter H is taken to be a measure of the self-similarity of the series. For self-similar traffic, the Hurst parameter should be greater than 1/2; typically $H \approx 0.7$, or 0.8. Conventional traffic models converge to white noise as the time scale increases by only a couple of orders of magnitude, whereas self-similar traffic retains similar characteristics over a wide range of time scales, hence the term "self-similar." Self-similar sequences are sometimes referred to as *fractal-like*. Fractals are an example of *deterministic* self-similarity in that they repeat exactly when scaled. However, our concern here is *stochastic* self-similarity, i.e., random sequences that remain statistically the same when scaled. The pioneering work on self-similarity and its application to communications was by Mandelbrot [20]; he argued that an interarrival time distribution of the form $P(x) = \Pr(interval \geq x) = Cx^{-k}$ would produce a self-similar arrival process. Here, $0 < k < 1$ is a fixed parameter of the distribution, and x is defined between limits that approach 0 and ∞. The distribution is normalized by scaling constant C. This distribution has a *heavy tail* so that arbitrarily long interarrival times may occur with finite probability. It also has an asymptote at zero, allowing (nearly) zero interarrival times to occur, providing bursts of closely spaced arrivals. This means that the arrival count process is invariant when scaled up from even the smallest initial time intervals. This combination of bursts and spaces is a characteristic of self-similar series.

The best known heavy-tailed distribution is the Pareto distribution, which also fits well the packet interval time distribution of self-similar traffic [17]. The pdf of type I Pareto distribution takes the following form:

$$f(x) = \frac{\alpha x_m^{\alpha}}{x^{\alpha+1}} \quad \text{for} \quad x \geq x_m \tag{2}$$

where α is the distribution shape parameter, and x_m is the location parameter, with both being positive numbers. Specifically, if $\alpha \leq 2$, then the distribution has infinite variance; if $\alpha \leq 1$, then the distribution has infinite mean. For self-similar traffic, α should be between 1 and 2. Then, the distribution has slowly decaying variances, and the correlated Hurst parameter is $H = (3 - \alpha)/2$. This is an intuitively pleasing result as when $\alpha = 2$, $H = 1/2$, which indicates that the process is no longer self-similar. Moreover, when $\alpha = 1$, $H = 1$, which is the largest possible value in the self-similar range of H.

In real-world scenarios, the packet interarrival time cannot be arbitrarily long; there is always an upper bound of the probability tail. Thus, the truncated Pareto distribution, which has been adopted in the 3GPP traffic model [24], is used in this paper. Compared with type I Pareto distribution, the pdf of the truncated Pareto distribution is much more complicated and, of course, brings about redundantly mathematical derivation. Therefore, we intuitively seek a method to degrade the pdf of truncated Pareto distribution to a simple form. The 3GPP traffic model provides a good way, as given in the following. The pdf of the truncated Pareto distribution is in the following form:

$$f(x) = \begin{cases} \frac{\alpha x_m^\alpha}{x^{\alpha+1}}, & x_m \le x < m \\ \left(\frac{x_m}{m}\right)^\alpha, & x = m \end{cases} \tag{3}$$

where m is the value of the truncated point that is equal to the maximum value of packet interarrival time; the other parameters are the same with (2). The cumulative distribution function is then defined as

$$F_X(x) = \begin{cases} \left(\frac{x_m}{m}\right)^\alpha, & x = m \\ 1 - \left(\frac{x_m}{x}\right)^\alpha, & x_m \le x < m \\ 0, & x < x_m. \end{cases} \tag{4}$$

It can be seen that the probability of m is a constant, and we could treat other points in the field of definition by following the type I Pareto distribution. It is worth noting that this degradation brings slight inaccuracy at the probability of m because its value is brutally equal to the summation of the probability of values greater than or equal to m. Fortunately, the packet interarrival time in real traffic has a large variation range; therefore, we can assume that the value of m is much larger than the location parameter x_m, which will be shown in subsequent analysis where the assumption makes the inaccuracy at m to insignificantly affect the analytical results.

III. Analytical Model For the RRC_Connected-Mode Discontinous Reception Mechanism

A. Self-Similar Traffic Model

The homogeneous Poisson process is most frequently used to illustrate the packet-interarrival-time property in former research on DRX modeling (i.e., in [10], [11], and [14]). To better fit the actual multimedia traffic, we adopt the truncated Pareto distribution to illustrate the packet interarrival times. This choice has two main advantages in general. First, as discussed in Section II, it could well fit the self-similar property of multimedia traffic. Second, it unifies different types of packet interarrival times defined by the Poisson process framework; therefore, the modeling complexity is reduced. Specifically, such distribution could generate bursts of closely spaced arrivals, which is corresponding to the conception of "interpacket call idle time" defined in [10] and [14], and it also has an unignorable probability to generate sparse spaced arrivals, which is corresponding to the conception of "intersession idle time" defined in [10] and [14].

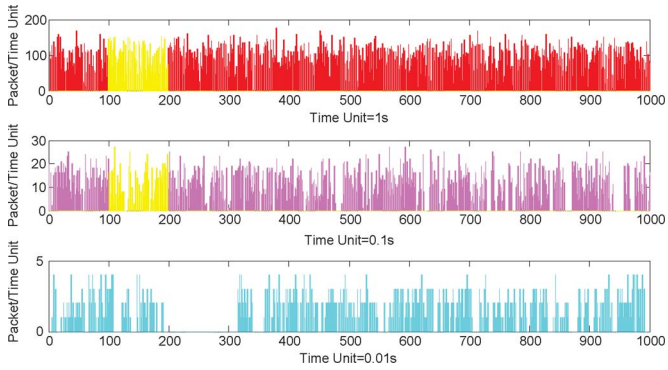


Fig. 2. Self-similar traffic trace ($H = 0.8$) generated by (5) on three different time scales. Different gray levels indicate the same segments of traffic on the different time scales.

Random samples that has the pdf of (3) can be generated by using *inverse transform sampling* [25]. Given a random variate U drawn from the uniform distribution on the interval $((x_m/m)^\alpha, 1]$, the formula to generate the truncated Pareto distribution is

$$x = \frac{x_m}{U^{1/\alpha}}. \tag{5}$$

Fig. 2 shows the different scaling behaviors of self-similar traffic traces generated by (5). These traces show the property of LRD, which exhibits fluctuations over a wide range of time scales, as discussed earlier.

In this paper, to study the effect of different DRX parameter configurations on the performance directly, we simplify the traffic model under the following assumptions.

- *Exhaustive service*. We assume that the buffer is emptied as soon as the UE wakes up. This assumption is feasible in practical LTE systems since the peak data rate could achieve more than 120 Mb/s for specific UE. Therefore, we could ignore the transmission time and focus on the delay that is only caused by the DRX mechanism. This leads to the fact that the packet length is *meaningless* in this paper. Accordingly, the *packet interarrival time* is defined as the time between the instants of the *entire packet* entering the eNB's buffer.
- *Neglecting impact of the scheduling and resource-allocation algorithms*. As we know, the scheduling and resource-allocation algorithms could significantly affect the transmission time and the waiting time as well. To provide an insight on the effect of different DRX parameter configurations on the DRX performance directly, we neglect the impact of scheduling and resource-allocation algorithms. We will take them into account in future work.
- *Single eNB-UE model*. The cell load and the interference among UE will affect the data transmission; therefore, we only study on the single eNB-UE model in this paper for the same reasons stated previously.
- *Downlink-only traffic*. Since, in this paper, we concentrate on the DRX mechanism, only the downlink traffic has been considered. However, as mentioned in [13], the modeling procedures for the DTX is similar with DRX modeling. Therefore, the analytical model proposed in this paper

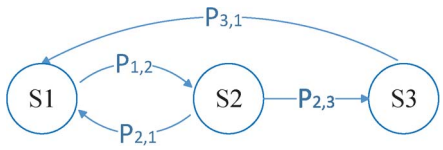


Fig. 3. DTSMP for the RRC_Connected -mode DRX analysis.

is also effective in DTX modeling, which considers the uplink traffic only.

B. Semi-Markov-Process-Based DRX Analytic Model

Semi-Markov processes are generalizations of Markov processes in which the time intervals between transitions have an arbitrary distribution rather than an exponential distribution. It has been proved to be a powerful mathematical tool for precisely representing the DRX mechanism in UMTS [26] and LTE [10], [11] with Poisson traffic. In this paper, we also adopt a semi-Markov process as the main method but in a different way. First, we introduce the concept of *sojourn time distribution* to calculate the mean sojourn time in each state, rather than utilizing the memoryless property of the exponential as in [3] and [11]. Second, since the smallest time unit in LTE is the transmission time interval (TTI) and the DRX mechanism is working on cycles (integer multiple of TTIs), we adopt the DTSMP [27] instead of the continuous-time process used in previous studies.

As shown in Fig. 3, the DTSMP consists of three states, which are relevant to the three periods showed in Fig. 1. $S1$, $S2$, and, $S3$ correspond to *active period, light sleep period, and deep sleep period*, respectively. The transitions among states strictly obeys the mechanism described in Section II.

By means of the DTSMP, we are going to deduce the *power-saving factor* and the *expectation of wake-up delay*, which have been considered the two most important performance output measures of DRX [3]. Before that, we should calculate the following necessary statistical parameters: the transition probability and the stationary distribution of *EMC*,[1] which also mean the sojourn time of each state and the limit distribution of the DTSMP.

Let UE's DRX state space $E = \{1, 2, 3\}$, whose evolution in time is governed by a stochastic process $Z = (Z_n; n \in \mathbb{N})$. Then, let $J = (J_n; n \in \mathbb{N})$ be a stochastic process with state space $E = \{1, 2, 3\}$, where J_n is the system state at the nth jump time $\mathbb{N} = \{0, 1, \ldots\}$. The relation between process Z and process J of the successively visited states is given by $Z_k = J_{N(k)}$, or, equivalently, $J_n = Z_{S_n}, n, k \in \mathbb{N}$, where $N(k) = \max\{n \in \mathbb{N} | S_n \leq k\}$ is the discrete-time counting process of the number of jumps in $[1, k] \subset \mathbb{N}$. The stochastic process $S = \{S_n, n \in \mathbb{N}\}$ with state space $\mathbb{N}$, where S_n is the nth jump time. We suppose $S_0 = 0$ and $0 < S_1 < S_2 < \ldots < S_n < S_{n+1} < \ldots$. The stochastic process $D = (D_n; n \in \mathbb{N}^*)$ with state space $\mathbb{N}^* = \mathbb{N} - \{0\}$, where D_n is the sojourn time in state J_{n-1} before the nth jump. Thus, we have, for all $n \in \mathbb{N}^*$, $D_n = S_n - S_{n-1}$. The stochastic process $(J, S) = ((J_n, S_n); n \in \mathbb{N})$ is said to be a discrete-time Markov renewal process (DTMRP). We assume that (J, S) is homogeneous in time. The important notion associated to the DTMRP is the *discrete-time semi-Markov kernel* $q = \{q_{ij}(k), k \in \mathbb{N}\} \in \mathcal{M}_E$, where $\mathcal{M}_E$ is the set of nonnegative matrices on $E \times E$, which is defined as

$$q_{ij}(k) = \mathbb{P}(J_{n+1} = j, D_{n+1} = k | J_n = i) \tag{6}$$

where $i, j \in E$, and $q_{ij}(0) = 0$. It satisfies the relation

$$q_{ij}(k) = p_{ij} f_{ij}(k),\ i,\ j \in E,\ k \in \mathbb{N}. \tag{7}$$

$p_{ij} = \mathbb{P}(J_{n+1} = j | J_n = i)$ is the transition probability of EMC, and $f_{ij}(k)$ is the conditional distribution of the sojourn time, which are defined as

$$f_{ij}(k) = \mathbb{P}(D_{n+1} = k | J_n = i, J_{n+1} = j), i, j \in E, k, n \in \mathbb{N}. \tag{8}$$

It is shown in Fig. 3 that our DTSMP models the self-transition $p_{ii} = 0$, $\forall i \in E$. One reason for this is that such transitions are invisible in Z. Another is that, with this assumption, the sample function of Z and the joint sample functions of J and D uniquely specify each other. According to (6), we can get $q_{ii}(k) = 0$, $\forall i \in E$, $\forall k \geq 0$.

Moreover, for all $i \in E$, we denote $h_i(\cdot)$ as the sojourn time distribution in state i, which is defined as

$$h_i(k) = \mathbb{P}(D_{n+1} = k | J_n = i) = \sum_{j \in E} q_{ij}(k), \quad k \in \mathbb{N}. \tag{9}$$

Finally, we denote U_i as the mean sojourn time in a state $i \in E$, which is defined as

$$U_i = \mathbb{E}(S_1 | J_0 = i) = \sum_{k \geq 1} k h_i(k). \tag{10}$$

It is obvious that, to obtain the statistical parameter of the DTSMP, the EMC $(J_n; n \in \mathbb{N})$ should be first analyzed. Denote random variable x as the packet interval time that follows the truncated Pareto distribution, and the state transition probabilities of EMC can be derived as follows.

- State 1 to State 2: Since there is only one transition out from state "S1" to "S2," we have $p_{12} = 1$.
- State 2 to State 1: As shown in Fig. 4, if the packet arrival happens between t_1 and t_2, the UE can get the resource allocation grants by decoding the PDCCH. Then, it will wake up at t_3 and enter "S1"; thus, the sojourn time is t_{SD}. Else, if the packet arrival happens between t_2 and t_4, UE will wake up after another sleep period and then enter "S1"; therefore, the sojourn time is $2t_{\mathrm{SD}}$.

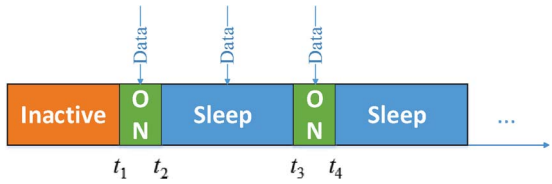


Fig. 4. Termination of the DRX cycles.

[1]The EMC is a process in which we only focus on the time of state transitioning, and we ignore the holding time in each state.

According to the balance equation of every state, this transition happens with the following probability[2]:

$$p_{21} = 1 - p_{23}. \tag{11}$$

- State 2 to State 3: If the UE has not detected any packet arrival before the last ON duration in "S2", the transition from "S2" to "S3" takes place. Hence, from (3), we have

$$\begin{aligned} p_{23} &= \int_{t_I}^{t_I+(N_{\mathrm{SD}}-1+\eta)t_{\mathrm{SD}}} f(x)dx \\ &= \left[\frac{t_I}{t_I+(N_{\mathrm{SD}}-1+\eta)t_{\mathrm{SD}}}\right]^{\alpha}. \end{aligned} \tag{12}$$

- State 3 to State 1: We could obtain $p_{31} = 1$ for the same reason as p_{12}.

The transition probability matrix $\mathcal{P}$ of the EMC then can be given as

$$\mathcal{P} = \begin{bmatrix} 0 & 1 & 0 \\ p_{21} & 0 & p_{23} \\ 1 & 0 & 0 \end{bmatrix}. \tag{13}$$

Let the row vector $\nu = (\nu_n; n \in \{1,2,3\})$ be the stationary distribution of the EMC. By applying $\sum_{n=1}^{3} \nu_n = 1$ and the balance equations $\nu_i = \sum_{j=1}^{3} \nu_j p_{ji}$, we can solve the stationary distribution and obtain

$$\begin{cases} \nu_1 = \frac{1}{2+p_{23}} \\ \nu_2 = \frac{1}{2+p_{23}} \\ \nu_3 = \frac{p_{23}}{2+p_{23}}. \end{cases} \tag{14}$$

Now, all the interested statistical parameters of the EMC have been obtained. It is easy to find that the EMC is irreducible and positive recurrent. According to [22, Lemma 6.8.1], the expected holding interval at each state exists with probability of 1. Then, the mean sojourn time U_i at each state can be derived as follows.

U_1*:* We assume in state "S1" that the UE experiences n new packet arrivals (denoted as t_{busy}) and then encounters an *inactive timer* expiration. Thus, we have

$$U_1 = \mathbb{E}(t_{\mathrm{busy}}). \tag{15}$$

The service time of packets in the *active period* is a constant with a value of one TTI, which is shorter than the packet interarrival time. Hence, t_{busy} can be calculated by the sum of $n-1$ packet interarrival times. Since the interarrival time between packets are i.i.d., the increment for t_{busy} follows geometric distribution. Now, we present the corresponding increment probability ψ_c as

$$\psi_c = (1 - p_{>t_I})^c \, p_{>t_I} \tag{16}$$

where $c \geq 0$. i indicates how many times the inactive timer restart until its expiration. $p_{>t_I}$ denotes the probability of packet interarrival time that is longer than t_I. $p_{>t_I}$ can be obtained as follows:

$$\begin{aligned} p_{>t_I} &= \int_{t_I}^{m} \frac{\alpha x_m^{\alpha}}{x^{\alpha+1}} dx + \left(\frac{x_m}{m}\right)^{\alpha} \\ &= \left(\frac{x_m}{t_I}\right)^{\alpha}. \end{aligned} \tag{17}$$

The expected packet interarrival time during t_{busy}, which is denoted as $\hat{x}$, is given by

$$\begin{aligned} \mathbb{E}(\hat{x}) &= \left\lceil \int_{x_m}^{t_I} \frac{\alpha x_m^{\alpha}}{x^{\alpha}} dx \right\rceil \\ &= \left\lceil \frac{\alpha x_m^{\alpha} t_I^{1-\alpha} - \alpha x_m}{1-\alpha} \right\rceil \end{aligned} \tag{18}$$

where $\lceil \cdot \rceil$ stands for round-up function.[3] Thereupon, we have the mean sojourn time in state "S1" by

$$\begin{aligned} U_1 &= \sum_{c=1}^{\infty} c\psi_c \mathbb{E}(\hat{x}) \\ &= \left\lceil \frac{\alpha x_m^{\alpha} t_I^{1-\alpha} - \alpha x_m}{1-\alpha} \right\rceil \left[\left(\frac{t_I}{x_m}\right)^{\alpha} - 1 \right]. \end{aligned} \tag{19}$$

U_2*:* Since deriving U_2 in a similar way with U_1 is exhausting and seems to be intriguing ambiguity, we introduce the DTSMP to solve this problem. According to the DTSMP [27], the answer could be gained after four steps. First, derive the conditional distribution of sojourn time $f_{21}(k)$ and $f_{23}(k)$ by (8). Second, calculate the corresponding discrete-time semi-Markov kernels $q_{21}(k)$ and $q_{23}(k)$ by (7). Third, substitute the obtained kernels into (9) to get the sojourn time distribution $h_2(k)$. Finally, by (10), the mean sojourn time U_2 could be gained. One important property that we will use repeatedly later is that, in state "S2", the sojourn time is the integer times of t_{SD}, and the maximum value is $N_{\mathrm{SD}}t_{\mathrm{SD}}$ (as shown in Fig. 4). A detailed description of each step is explained as follows. By (8), we have

$$\begin{aligned} f_{21}(lt_{\mathrm{SD}}) &= \mathbb{P}(D_{n+1} = lt_{\mathrm{SD}} | J_n = 2, J_{n+1} = 1) \\ &\qquad l \in \{1, 2, \ldots N_{\mathrm{SD}}\} \\ &= \begin{cases} \frac{\mathbb{P}[t_I \leq x \leq \eta t_{\mathrm{SD}} + t_I]}{\mathbb{P}[x \geq t_I]}, & l = 1 \\ \frac{\mathbb{P}[t_I \leq x \leq (l-1)t_{\mathrm{SD}} + \eta t_{\mathrm{SD}}]}{\mathbb{P}[x \geq t_I]}, & 1 < l \leq N_{\mathrm{SD}} \end{cases} \\ &= \begin{cases} 1 - (\eta t_{\mathrm{SD}} + t_I)^{-\alpha} t_I^{\alpha}, & l = 1 \\ \left\{[(l-2+\eta)t_{\mathrm{SD}} + t_I]^{-\alpha} \right. \\ \quad \left. - [(l-1+\eta)t_{\mathrm{SD}} + t_I]^{-\alpha}\right\} t_I^{\alpha}, & 1 < \leq N_{\mathrm{SD}} \end{cases} \end{aligned} \tag{20}$$

and

$$\begin{aligned} f_{23}(N_{\mathrm{SD}}t_{\mathrm{SD}}) &= \mathbb{P}(D_{n+1} = N_{\mathrm{SD}}t_{\mathrm{SD}} | J_n = 2, J_{n+1} = 3) \\ &= \left[\frac{x_m}{(N_{\mathrm{SD}} - 1 + \eta)t_{\mathrm{SD}}}\right]^{\alpha}. \end{aligned} \tag{21}$$

[2]In this paper, we assume the truncated value m of truncated-Pareto satisfies $N_{\mathrm{SD}}t_{\mathrm{SD}} < m - t_I$. If the converse is true, the state "S2" could never transit to "S3" in *Situation 1*, which implies that the two-level DRX mechanism works ineffectively under such bursty traffic. This fact has been observed in [28], which studies on the DRX mechanism performance under video streaming.

[3]As stated previously, the smallest time unit in DRX is one TTI, which implies that, even if the packet enters the buffer in noninteger times of TTI, eNB will transmit the packet to UE in the next integer times of TTI.

Notice that the truncated Pareto distribution is not memoryless; therefore, we need to transform the conditional probability function into division forms by the *Bayes theorem*, as shown in (20). The corresponding discrete-time semi-Markov kernels can be now obtained by $q_{21} = p_{21}f_{21}$ and $q_{23} = p_{23}f_{23}$.

Finally, by (10), the mean sojourn time in "S2" could be given as

$$U_2 = \sum_{l=1}^{N_{\rm SD}} l t_{\rm SD} q_{21}(l t_{\rm SD}) + N_{\rm SD} t_{\rm SD} q_{23}. \tag{22}$$

U_3: Since the deviation for U_3 is almost the same as U_2, we directly give the results to make it brief.

The conditional distribution of the sojourn time is given by (23), shown at the bottom of the page, where $\lfloor\cdot\rfloor$ is the round-down function, and $l \in \{1, 2, \ldots n_m = \lfloor (m - t_I - N_{\rm SD}t_{\rm SD})/t_{\rm LD})\rfloor\}$. Then, we get the semi-Markov kernel as $q_{31} = p_{31}f_{31} = f_{31}$. The mean sojourn time in state "S3" can be thus given as follows:

$$U_3 = \sum_{l=1}^{n_m} l t_{\rm LD} q_{31}(l t_{\rm LD}). \tag{24}$$

Through this analysis, we can come to a conclusion that the DTSMP is irreducible and the mean sojourn time is finite for any state $j \in E$. According to [28, Proposition 8], the limit distribution of the DTSMP could be obtained as follows:

$$\begin{cases} \pi_1 = \frac{\nu(1)U_1}{\sum_{i\in E}\nu(i)U_i} \\ \pi_2 = \frac{\nu(2)U_2}{\sum_{i\in E}\nu(i)U_i} \\ \pi_3 = \frac{\nu(3)U_3}{\sum_{i\in E}\nu(i)U_i}. \end{cases} \tag{25}$$

C. Evaluation Metrics

Power-Saving Factor: The power-saving factor is introduced to indicate the sleeping time ratio, compared with the overall operation time [3]. It can be given by

$$\text{PS} = \frac{(1-\eta)U_2\nu(2) + (1-\eta')U_3\nu(3)}{\sum_i U_i\nu(i)}, \quad i \in E. \tag{26}$$

Substituting (14), (19), (22), and (24) into (26), we derive the closed-form equation for the power-saving factor PS.

Expectation of Wake-Up Delay: As aforementioned, the packet that arrives in DRX cycles should be waiting in the buffer until the UE transits from "S2"(or "S3") to "S1." In the following, we focus on obtaining the expectation of this waiting time, namely *wake-up delay*. It is not difficult to understand that the probability of the packet arrival could be detected in the lth ON duration that is equal to the already obtained $f_{21}(l t_{\rm SD})$ and $f_{31}(l t_{\rm LD})$. Then, we need to derive the expected delay of whether the packet arrival is detected in the lth ON duration. We denote $D_S(l)$ and $D_L(l)$ as the delays that happen in *light sleep period* and *deep sleep period*, respectively. $D_S(l)$ could be calculated as

$$\begin{aligned} D_S(l) &= \begin{cases} t_{\rm SD} - \int_{t_I}^{t_I+\eta t_{\rm SD}} \frac{\alpha x_m^\alpha}{x^\alpha} dx, & l = 1 \\ 2t_{\rm SD} - \int_{t_I+(l-1)\eta t_{\rm SD}}^{t_I+l\eta t_{\rm SD}} \frac{\alpha x_m^\alpha}{x^\alpha} dx, & 1 < l < N_{\rm SD} \end{cases} \\ &= \begin{cases} t_{\rm SD} - \frac{\alpha x_m^\alpha}{1-\alpha}\left[(t_I+\eta t_{\rm SD})^{(1-\alpha)} - t_I^{(1-\alpha)}\right], & l = 1 \\ 2t_{\rm SD} - \frac{\alpha x_m^\alpha}{1-\alpha}\Big\{(t_I + l\eta t_{\rm SD})^{(1-\alpha)} \\ \qquad\qquad - [t_I + (l-1)\eta t_{\rm SD}]^{(1-\alpha)}\Big\}, & 1 < l < N_{\rm SD}. \end{cases} \end{aligned} \tag{27}$$

Similarly, we could get (28), shown at the bottom of the page, where n_m is defined the same as in (23). Therefore, we can obtain the expected wake-up delay in the form of

$$\begin{aligned} \mathbb{E}(D) &= \mathbb{E}\left[D_S(l)\right] + \mathbb{E}\left[D_L(l)\right] \\ &= \pi_2 \sum_{l=1}^{N_{\rm SD}} f_{21}(l t_{\rm SD}) D_S(l) + \pi_3 \sum_{l=1}^{n_m} f_{31}(l t_{\rm LD}) D_L(l). \end{aligned} \tag{29}$$

Substituting (20), (23), (25), (27), and (28) into (29), we derive the closed-form equation for the expectation of wake-up delay.

$$\begin{aligned} f_{31}(l t_{\rm LD}) &= \mathbb{P}(D_{n+1} = l t_{\rm LD} | J_n = 3, J_{n+1} = 1) = \begin{cases} \frac{\mathbb{P}[t_I + N_{\rm SD}t_{\rm SD} \le x \le t_I + N_{\rm SD}t_{\rm SD} + \eta' t_{\rm LD}]}{\mathbb{P}(x > t_I + N_{\rm SD}t_{\rm SD})}, & l = 1 \\ \frac{\mathbb{P}[t_I + N_{\rm SD}t_{\rm SD} \le x \le t_I + N_{\rm SD}t_{\rm SD} + (\eta' + l - 1)t_{\rm LD}]}{\mathbb{P}(x > t_I + N_{\rm SD}t_{\rm SD})}, & l > 1 \end{cases} \\ &= \begin{cases} 1 - (t_I + N_{\rm SD}t_{\rm SD})^\alpha \cdot (t_I + N_{\rm SD}t_{\rm SD} + \eta' t_{ld})^{-\alpha}, & l = 1 \\ \left\{[t_I + N_{\rm SD}t_{\rm SD} + (\eta' + l - 2)t_{\rm LD}]^{-\alpha} - [t_I + N_{\rm SD}t_{\rm SD} + (\eta' + l - 1)t_{\rm LD}]^{-\alpha}\right\} \cdot (t_I + N_{\rm SD}t_{\rm SD})^\alpha, & l > 1 \end{cases} \end{aligned} \tag{23}$$

$$\begin{aligned} D_L(l) &= \begin{cases} t_{\rm LD} - \int_{t_I+N_{\rm SD}t_{\rm SD}}^{t_I+N_{\rm SD}t_{\rm SD}+\eta' t_{\rm LD}} \frac{\alpha x_m^\alpha}{x^\alpha} dx, & l = 1 \\ 2t_{\rm LD} - \int_{t_I+N_{\rm SD}t_{\rm SD}+(l-1)\eta' t_{\rm SD}}^{t_I+N_{\rm SD}t_{\rm SD}+l\eta' t_{\rm SD}} \frac{\alpha x_m^\alpha}{x^\alpha} dx, & 1 < l < n_m \end{cases} \\ &= \begin{cases} t_{\rm LD} - \frac{\alpha x_m^\alpha}{1-\alpha}\left[(t_I + N_{\rm SD}t_{\rm SD} + \eta' t_{\rm LD})^{(1-\alpha)} - (t_I + N_{\rm SD}t_{\rm SD})^{(1-\alpha)}\right], & l = 1 \\ 2t_{\rm LD} - \frac{\alpha x_m^\alpha}{1-\alpha}\left\{(t_I + N_{\rm SD}t_{\rm SD} + l\eta' t_{\rm SD})^{(1-\alpha)} - [t_I + N_{\rm SD}t_{\rm SD} + (l-1)\eta' t_{\rm SD}]^{(1-\alpha)}\right\}, & 1 < l < n_m \end{cases} \end{aligned} \tag{28}$$

IV. Online Estimation of Statistical Parameters

We present the LTE-DRX analytical model not only to evaluate the power-saving operation performances but to provide a practical way to optimize the DRX parameters for the on-going system as well. As stated previously, in practical systems, the dynamic DRX mechanism is deployed in a semi-static manner, which gives the eNB a chance to estimate the traffic conditions and power-saving performance of a specific user. Therefore, here, we focus on the following practical issues, including 1) which statistical parameter needs to be estimated, 2) how do we estimate them from the random samples, and 3) how long should the parameters to be estimated unbiasedly be?

A. Estimation of Traffic Conditions

Apparently, our traffic model is based on the truncated Pareto distribution. Thus, the traffic condition estimation (TCE) problem could be transformed into a truncated-Pareto-distribution parameter estimation problem. In [29], the main methods of estimating the parameters of Pareto distribution from which a random sample comes were compared. The compared methods include *method of moments*, *ML method*, and *quantile regression method*. The numerical results show that the *method of moments* behaves horribly in estimating the shape parameter α, but it is able to estimate location parameter x_m unbiasedly. On the other hand, the *quantile regression method* has shown to be the least biased estimator for estimation α, but it generally had greater than 20% standard error than the *ML method*. Meanwhile, the *ML method* yields smaller MSE than the *quantile regression method* for estimating α. We seek to find an online TCE strategy that could unbiasedly estimate the parameters within numerable random samples. Hence, based on the work in [29], we adopt the *method of moments* to estimate x_m and utilize the *ML method* to estimate α. To get unbiased estimation of α, we go a step further. By means of transforming Pareto distribution into an exponential distribution, the properties of gamma distribution moments could be utilized to further improve the ML estimator's (MLE) accuracy. It is particularly necessary to point out that, since the random samples could rarely achieve the truncated value,[4] it is impossible to estimate the truncated value during an observing window. Therefore, we assume the truncated point value m is predefined as a large value. Then, we begin by constructing the MLE for the location parameter.

Suppose we have a random sample of observed packet interarrival times $\{x_1, x_2, \ldots, x_n\}$, which has been obtained by truncated Pareto distribution. We denote the corresponding joint random variables as $\{\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_n\}$. Then, the likelihood function L for the Pareto distribution has the following form:

$$L(x_m, \alpha|x) = \prod_{i=1}^{n} \frac{\alpha x_m^{\alpha}}{x_i^{\alpha+1}}, \quad 0 < x_m \leq \min\{x_i\}, \alpha > 0. \tag{30}$$

[4]The tradeoff between the observing window length and the ML's computing complexity should be considered here. In the following, we always tune the observing window length smaller than the truncated value; thus, it is impossible to observe a whole truncated valued packet interarrival time during an observing window.

By analyzing the likelihood function of a series of i.i.d Pareto random variables, the MLE of the scale parameter can be shown to be

$$\hat{x}_m = x_{(1)} = \min\{x_1, x_2, \ldots, x_n\}. \tag{31}$$

This result is convincing since it is plausible that the finite endpoint of the distribution support should be estimated from the smallest observation in a random sample from the distribution.

It can also be demonstrated that the conditional distribution of the random sample $\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_n$, which is conditioned on the event that the random variable $\tilde{X}_{(1)} = x$, does not depend on $\hat{x}_m$, implying that the minimum $\tilde{X}_{(1)}$ is a sufficient statistic for $\hat{x}_m$. Furthermore, the Pareto distribution belongs to the exponential family of distributions, which is complete (will be proved later). This implies that we can construct the minimum-variance unbiased estimator. By observing that $\mathbb{E}(X_{(1)}) = n\alpha x_m/(n\alpha - 1)$, we can then choose the random variable:

$$\hat{X}_m = \frac{(n\alpha - 1)X_{(1)}}{n\alpha}. \tag{32}$$

Since this is a function of a complete sufficient statistic, it is a minimum-variance unbiased estimator for its mean, which is clearly x_m by construction. For large sample sizes, the statistic $\hat{X}_m$ is asymptotically the same as the minimum $X_{(1)}$; therefore, this can be used to remove dependence on α in the estimator (32).

Next, we examine the MLE for the shape parameter. To find the ML estimate of α, calculus is appropriate. Since L is nonnegative, we can take its logarithm. We do this because it is easier to differentiate $\log L$ than L itself. Logarithms are bijective functions; therefore, the value of α that maximizes L also maximizes $\log L$. The brief process is as follows:

$$\begin{aligned} &\log L(x_m, \alpha|x) \\ &\quad = \sum_{i=1}^{n} \log\left(\frac{\alpha x_m^{\alpha}}{x_i^{\alpha+1}}\right) \\ &\quad = n\log(\alpha) + \alpha n \log(x_m) - (\alpha+1)\sum_{i=1}^{n}\log(x_i). \end{aligned} \tag{33}$$

Therefore

$$\frac{d\log L(\cdot)}{d\alpha} = n/\alpha + n\log(x_m) - \sum_{i=1}^{n}\log(x_i). \tag{34}$$

Setting the derivative equal to zero and using a little algebra and an omitted second derivative check confirm that we maximize L rather than minimize L, yielding

$$\hat{\alpha} = \frac{n}{\sum_{i=1}^{n}\left[\log(x_i) - \log(\hat{X}_m)\right]}. \tag{35}$$

Then, we prove that the Pareto distribution belongs to the exponential family as follows. If X is Pareto distributed with parameters x_m and α, then we define a random variable $Y = \log(X/x_m)$. It is not difficult to show that the constructed distribution function of Y is an exponential distribution with parameter α. This means that a Pareto $P(\alpha, x_m)$ random variable is statistically equivalent to $x_m e^Y$, where $Y \overset{\circ}{=} \exp(\alpha)$.

Due to the fact that $\log(X_i/\hat{X}_m)$ is exponential distributed, it can be shown that $T = \sum_{i=1}^{n}[\log(X_i) - \log(\hat{X}_m)]$ has a gamma distribution with density function

$$f_T(t) = \frac{\alpha^n}{(n-1)!} t^{n-1} e^{-\alpha t}. \tag{36}$$

As $\hat{\alpha} = n/T$, we obtain the following:

$$\mathbb{E}(\hat{\alpha}) = \frac{n\alpha}{(n-1)} \int_0^\infty \frac{\alpha^{n-1}}{(n-2)!} t^{n-2} e^{-\alpha t} dt = \frac{n}{n-1}\alpha \tag{37}$$

$$\mathbb{E}(\hat{\alpha}^2) = \frac{n^2\alpha^2}{(n-1)(n-2)} \int_0^\infty \frac{\alpha^{n-2}}{(n-3)!} t^{n-3} e^{-\alpha t} dt$$

$$= \frac{n^2}{(n-1)(n-2)}\alpha^2. \tag{38}$$

Hence, the variance of $\hat{\alpha}$ is

$$\mathrm{Var}(\hat{\alpha}) = \frac{n^2\alpha^2}{n-1}\left(\frac{1}{n-2} - \frac{1}{n-1}\right) = \frac{n^2}{(n-1)^2(n-2)}\alpha^2. \tag{39}$$

The MLE $\hat{\alpha}$ of α is not unbiased; therefore, the estimator is

$$\alpha^* = \frac{n-1}{T} = \frac{n-1}{\sum_{i=1}^{n}\left[\log(x_i) - \log(\hat{X}_m)\right]} \tag{40}$$

with T as defined earlier. Furthermore

$$\mathrm{Var}(\alpha^*) = \frac{1}{n-2}\alpha^2 < \mathrm{Var}(\hat{\alpha}). \tag{41}$$

Thus, α^* is a better estimator of α than $\hat{\alpha}$ and is unbiased for α. By constructing joint conditional distributions, it can be demonstrated that α^* is a complete sufficient statistic for α, and since it is unbiased, it is also the minimum-variance unbiased estimator for α (also due to the completeness, as shown earlier). Hence, (32) can be used to estimate the position parameter (using a large sample approximation), whereas (40) is used with this to estimate the shape parameter.

Next, we study on "how long the estimation should take to be sufficient" by numerical methods. First, we generate 1.5×10^5 packet interarrival times by (5). The statistical parameter of the truncated Pareto distribution is $\alpha = 1.1$, $x_m = 2$ ms, and $m = 2$ s, respectively. Then, we generate $(1.5 \times 10^5 + 1)$ packet arrival times, where the first packet arrives at "instant 0," and each packet interarrival time is corresponding to one of the obtained samples. Thus, the nth $(n > 1)$ packet arrival time is equal to the summation of the earlier $n-1$ packet interarrival times. Finally, based on the packets arriving during different sizes of the observing window T_{ob}, i.e., 50 ms, 100 ms, 1 s, and 10 s, the estimation of x_m and α is carried out, respectively.

Each subgraph in Figs. 5 and 6 demonstrates the estimation of α and x_m obtained in each T_{ob}. The total duration of each measurement is 100 s. When there are no packets arriving during T_{ob}, the corresponding estimation value is set to zero. Since the biggest packet interarrival time is equal to the truncated value $m = 2$ s, there should be zero point as long as T_{ob} is less than 2 s. This case is shown both in Figs. 5 and 6.

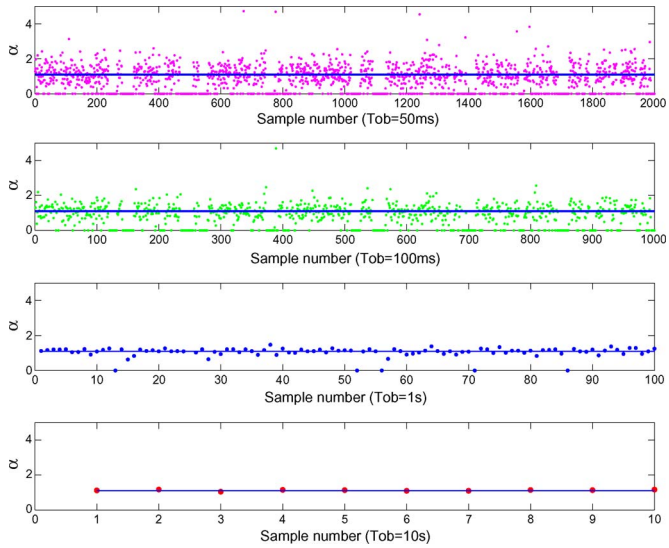


Fig. 5. Estimation of α with different sizes of T_{ob}. The line is an indicator of the true value.

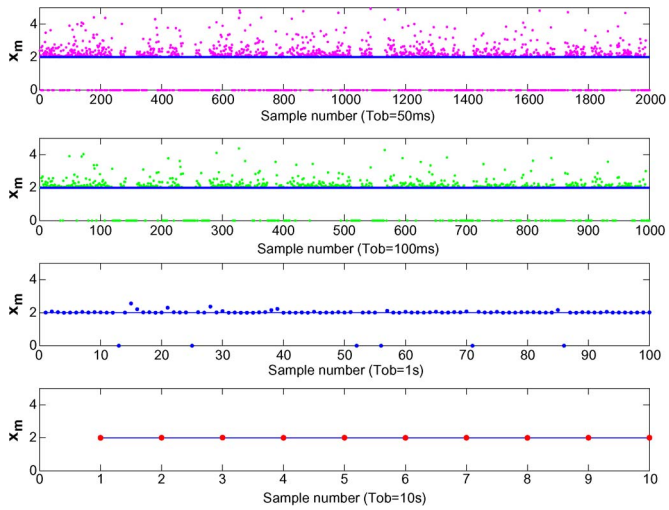


Fig. 6. Estimation of x_m with different sizes of T_{ob}. The line is an indicator of the true value.

These figures also reveals that the estimated quantities converge toward the true values for increasing T_{ob}.

The statistical characters of estimation results are collected in Table I. The convergence of estimated quantities can also be clearly observed. After comparing the statistical characters, we recommend 1 s as one good choice of T_{ob} for two reasons. First, the bias is acceptable (less than 2%), and the estimation is stable (MSE $< 3\%$). Second, by simulation, we find that the average number of packet arrivals is 87.7067 and 877.0667 for 1 s and 10 s, respectively. We should notice that the location parameter α here is 1.1. It will be shown later that the number of packet arrivals increases with the growth of α. It is impractical for the eNB to record thousands of packet arrival times for one specific UE, and the ML for such a good many samples will take quite a bit of computing and storage resources. Therefore, $T_{\mathrm{ob}} = 1$ s will bring the eNB sufficient samples, which are neither too little nor too much.

TABLE I
STATISTICAL CHARACTERS OF ESTIMATION RESULTS FOR PARETO DISTRIBUTION ($x_m = 2, \alpha = 1.1$)

Param	T_{ob}	Mean	Bias	MSE
α	50ms	1.1935	0.0935	0.2890
	100ms	1.1223	0.0223	0.1620
	1000ms	1.1188	0.0188	0.0208
	10000ms	1.0989	0.0011	0.0016
x_m	50ms	2.3696	0.3696	0.3676
	100ms	2.2385	0.2385	0.1854
	1000ms	2.0369	0.0369	0.0087
	10000ms	2.0027	0.0027	1.0912e-5

B. Estimation of DRX Operations

The DRX performance are evaluated by PS and $\mathbb{E}(D)$, but real-time computation of (26) and (29) is still quite difficult. The motivation of answering the question of "how can we estimate the statistical parameters in the expressions of PS and $\mathbb{E}(D)$ directly" is obviously strong as the answer will help us to find a easier way to obtained current UE's DRX performance. In (26) and (29), we can see that, in addition to α, the other statistical parameters are all associated with the DTSMP. Thus, the essential problem is transformed into a DTSMP parameter estimation problem, which is also an important issue in reliable systems and has been studied in [23]. Through [23], some useful estimators for DTSMP parameters are provided. In the following, we first introduce the estimators into our problem and then demonstrate their performance by numerical evaluations. Note that all the discussions here on the DTSMP still follow the same definition in Section III.

Let us assume now that we have an observation of the DTSMP, censored at fixed arbitrary time $M \in \mathbb{N}^*$, an observation of the associated Markov renewal chain $(J_n, S_n)_{n\in\mathbb{N}}$, $(J_0, D_1, \ldots, J_{N(M)-1}, D_{N(M)}, J_{N(M)}, u_M)$, where $u_M = M - S_{N(M)}$ is the censored sojourn time in the last visited state $J_{N(M)}$.

For all states $i, j \in E$, let us introduce

$$N_i(M) = \sum_{n=0}^{M} \mathbb{N}_{\{J_n=i, S_{n+1}\le M\}} \tag{42}$$

where $N_i(M)$ stands for the number of visits to state i of the EMC $(J_n)_{n\in\mathbb{N}}$ up to time M, and

$$N_{ij}(M) = \sum_{n=1}^{M} \mathbb{N}_{\{J_{n-1}=i, J_n=j, S_{n+1}\le M\}} \tag{43}$$

where $N_{ij}(M)$ stands for the number of transitions of the EMC $(J_n)_{n\in\mathbb{N}}$ from i to j up to time M. Define D_{ik} as the sojourn time of the kth visiting to state i. Then, we could define the empirical estimator of the stationary distribution of the EMC $(J_n)_{n\in\mathbb{N}}$ by

$$\hat{\nu}(i, M) = \frac{N_i(M)}{N(M)}, \quad i \in E. \tag{44}$$

The estimator for U_i is given by

$$\hat{U}_i(M) = \frac{1}{N_i(M)} \sum_{k=1}^{N_i(M)} D_{ik}. \tag{45}$$

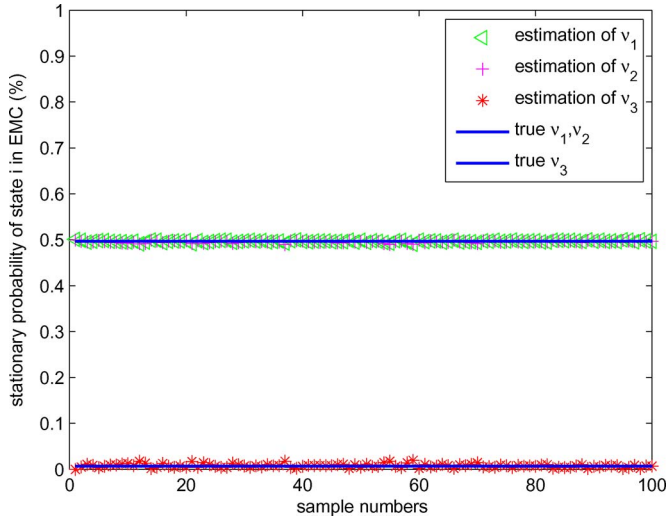


Fig. 7. Estimation of stationary distribution of EMC for $T_{\rm ob} = 10$ s.

Consequently, an estimator of the mean sojourn time of the DTSMP $\bar{U}$ is

$$\hat{\bar{U}}(M) = \frac{1}{N(M)} \sum_{k=1}^{N(M)} D_k \tag{46}$$

and we get the following estimator of the stationary distribution of the DTSMP:

$$\hat{\pi}_i(M) = \frac{1}{\hat{\bar{U}}(M) N(M)} \sum_{k=1}^{N_i(M)} D_{ik}, \quad i \in E. \tag{47}$$

According to [24, Lemma 1] the following holds: For any state $i \in E$ of the DTSMP, the estimators $\hat{\nu}(i, M)$, $\hat{U}_i(M)$, $\hat{\bar{U}}(M)$, and $\hat{\pi}_i(M)$ proposed in (44)–(47) are strongly consistent, as M tends to infinity. Next, we will check their performance and also answer the question of "how long the estimation should take to be sufficient" by numerical methods.

First, we generate a traffic trace that lasts for 1000 s, as we described earlier. Second, we input the traffic trace into an LTE-DRX entity, which is written via a MATLAB event-driven simulator (which will be further discussed in Section VI). Third, we record the $N_i(M)$, $N_{i,j}(M)$, D_{ik}, and D_k, and obtain the true value of $\nu(i), U_i, \bar{U}, \pi_i, (i, j \in E)$, respectively. Finally, we verify the estimator's performance at different time scales. Due to space constraints, we cannot demonstrate all the estimator's performance but can only show the estimation traces of $\nu(i)$ and U_i in the figures. Since if $\hat{\nu}(i, M)$, $\hat{U}_i(M)$ are given, the value of $\hat{\bar{U}}(M)$ and $\hat{\pi}_i(M)$ could be obtained by using a little algebraic computation. In addition to $T_{\rm ob} = 10$ s, we also directly give the statistical characters of the estimation results while the $T_{\rm ob}$ is equal to other values.

Figs. 7 and 8 have proved that the estimators can effectively estimate most of DTSMP's statistical parameters. However, in Fig. 8, it is obvious that such estimator cannot estimate mean sojourn time in state "S3" correctly. The reason is that the Pareto distribution is a kind of power-law probability distribution, which means that there is little chance to see an extremely big packet interarrival time during the $T_{\rm ob}$. By multiplying the

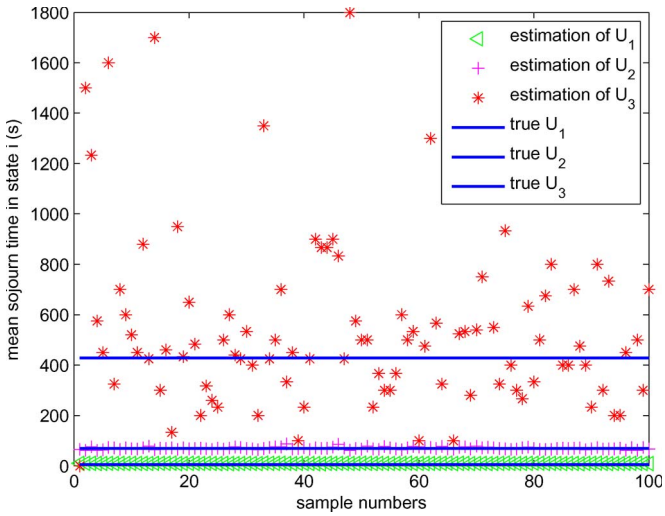


Fig. 8. Estimation of mean sojourn time for $T_{\rm ob} = 10$ s.

TABLE II
STATISTICAL CHARACTERS OF ESTIMATION RESULTS FOR STATIONARY DISTRIBUTION OF EMC ($\nu_1 = \nu_2 = 0.4968, \nu_3 = 0.0065$)

Param	T_{ob}	Mean	Bias	MSE
ν_1	500ms	0.5202	0.0238	0.0031
	1000ms	0.5099	0.0131	0.0016
	2000ms	0.5199	0.0151	0.0015
	10000ms	0.4970	2.0e-4	4.3105e-6
ν_2	500ms	0.4809	0.0159	0.0013
	1000ms	0.4878	0.0090	3.5580e-4
	2000ms	0.5199	0.0151	8.3681e-4
	10000ms	0.4955	0.0013	7.2626e-5
ν_3	500ms	0.1126	0.1061	0.0232
	1000ms	0.0405	0.0340	0.0035
	2000ms	0.0249	0.0184	0.0016
	10000ms	0.0077	0.0012	1.8743e-5

TABLE III
STATISTICAL CHARACTERS OF ESTIMATION RESULTS FOR MEAN SOJOURN TIME ($U_1 = 9.7722, U_2 = 69.6624$)

Parameter	T_{ob}	Mean	Bias	MAD
U_1	500ms	9.8570	0.0848	1.9991
	1000ms	9.8887	0.1165	1.5187
	2000ms	9.8328	0.0606	1.5868
	10000ms	9.8667	0.0945	0.4444
U_2	500ms	85.8975	16.2351	26.3857
	1000ms	78.3620	8.6996	16.7216
	2000ms	80.5486	10.8862	19.4432
	10000ms	69.6624	4.9096e-5	3.2887

large interarrival time and its probability as (45) does, the large fluctuation shown in Fig. 8 is inevitable. For this reason, we commend to estimate the U_3 by our analytical model, which is well coinciding with the observation. (The proof is given in Section VI.)

"MAD" in Table III's heading stands for *median absolute deviation*, which is a robust measure of the variability of a univariate sample of quantitative data. In Tables II and III, we can conclude that $T_{\rm ob} = 1$ s is still a good choice. Therefore, the two parts of the statistical parameter estimation algorithm can be deployed simultaneously. Furthermore, to reduce the eNB's burden, we recommend UE to estimate the DRX performance itself and to inform the eNB at the first ON duration after the "estimation phase" via the physical uplink control channel. Then, eNB should enter the "optimization phase" and optimize specific UE's DRX parameters based on the strategy discussed in the following.

V. ONLINE POWER-SAVING STRATEGY

Here, we develop an OPSS. In general, this strategy is divided into "estimation phase" and "optimization phase." The "estimation phase" has been introduced in detail earlier. Thus, here, we will first discuss the algorithm adopted in the "optimization phase" and then give out all of the procedures of the OPSS.

The optimization algorithm should consider the maximum delay $t_{\max}$ constraint of the application and the minimum power saving $\mathrm{PS}_{\rm th}$ imposed by the UE, to decide a balance tradeoff between power saving and delay. In practice, there are two common scenarios while deploying the DRX optimization algorithm. Accordingly, we introduce two parameters γ_d and $\gamma_{\rm ps}$ to quantify their difference as in [9].

- Scenario 1: $\gamma_d > \gamma_{\rm ps}$, which indicates that the latency performance is more important than power saving. In this case, we first should find the feasible ranges for values of DRX parameters that can guarantee the specified delay constraint $t_{\max}$. Then, the parameters that could maximize the power saving will be selected. This scenario is defined for the real-time applications such as video streaming.
- Scenario 2: $\gamma_{\rm ps} > \gamma_d$, which indicates that the power-saving performance is more important than latency. In this case, feasible ranges of values for DRX parameters satisfying specified UE's power-saving constraint $\mathrm{PS}_{\rm th}$ are identified. The parameters should be tuned to minimize the delay within the calculated feasible ranges. This scenario is defined for the UE in the "power saving mode," which is activated when the UE has a stringent battery power constraint or for non-real-time applications such as social networking applications.

Calculating the PS and $\mathbb{E}(D)$ for all the DRX parameter combinations is time-consuming; this also makes it impossible to compute them in real time. Therefore, we utilize the tradeoff between computation and storage, which means offline computing all DRX performance (e.g., PS and $\mathbb{E}(D)$) based on every possible pair of traffic conditions and DRX parameter configurations and then storing them into a table called Ω. Every entry of Ω should include the values of $\{t_I, t_{\rm SD}, N_{\rm SD}, t_{\rm LD}, t_{\rm ON}, \alpha, x_m, PS, \mathbb{E}(D)\}$. The entry is denoted as κ in this paper. Note that, here, we include all the configurable parameters in the entries, which may be much fewer in practical systems [7], [9], [11]. Finally, we summarize the main procedures of the OPSS in *Algorithm 1* (as shown in the Appendix).

It is worth noting that, in our optimization algorithm, we compare the numerical calculating optimum performance with the current DRX performance. The reason is that, while the traffic is stable, the DRX parameter configurations that have been optimized in the last cycle may still be the optimum configuration at present. Therefore, if there is no need to change the DRX parameters, the eNB will not sent the DRX

configuration updating information to UE to increase spectrum efficiency. The method of quickly finding out the $\kappa_{\max}$ from Ω is beyond the scope of this paper and will be demonstrated in our future work. The searching method adopted in this paper is based on the brute-force research algorithm.

VI. Performance Evaluation

Simulations are conducted to evaluate the proposed analytical model and the OPSS. All the simulations are implemented via a MATLAB event-driven simulator, and a single eNB/UE pair with DL self-similar traffic is considered in the simulation environment. We implement the DRX mechanism following the 3GPP release 10 specification [4], which includes all the required procedures and functions of both the eNB and UE. For ease of description, we name the entire implemented DRX mechanism as "the DRX entity" in the following. Furthermore, our proposed OPSS is also implemented in the test bed. Specifically, we implement the traffic condition estimation and online DRX parameter configuration optimization in eNB, and the DRX operation estimation mechanism in UE, respectively. The required parameter negotiation and update functions are established based on the discrete-event simulation method. Each simulation run is executed with a traffic trace generated by the method described in Section IV-A. Moreover, each presented simulation result is in the average of 100 simulation runs.

A. Evaluation of the Analytical Models

First, we verify the accuracy of our proposed analytical models. For the results here, the parameters used in the simulations are $t_I = 5$ ms, $t_{\rm SD} = 20$ ms, $t_{\rm LD} = 100$ ms, $t_{\rm ON} = 2$ ms, $N_{\rm SD} = 10$, $x_m = 2$ ms, and $m = 2000$ ms. We generate the traffic trace by gradually increasing the α with the step of 0.1. Then, the generated traffic traces are enetered into the DRX entity one by one, and the simulations are deployed for each trace. After each simulation, the statistics of DRX operations are collected. Based on these statistics, we could calculate 1) stationary distribution of EMC, 2) mean sojourn time in each state, 3) limit distribution of DTSMP, 4) PS, and 5) $\mathbb{E}(D)$ by the definitions described in Section III. We also calculate the listed numerical characteristics by using the proposed theoretical models. The comparison between the simulation results and analytical results are shown in Figs. 9–13. The well-matched results demonstrating the numerical equations are correctly derived.

With the help of these insight numerical characteristics, we can find out what happens when the traffic condition is changing, and meanwhile, the UE deploys a fixed DRX configuration. Fig. 9 shows the accuracy of the analytical model of the stationary distribution at each state in EMC. It also shows that our adjustment of previous work (*wiping off* the state self-transition in EMC; see [10] and [11]) is correct. It is not surprising to observe the increase in the steady-state probability of "S1" and "S2", and the decrease in the stationary probability of "S3." With the increase in α, the packet interarrival times are gathered in the small range near zero by (5), which implies that there are much more packets arrivals before the t_I and $N_{\rm SD}$

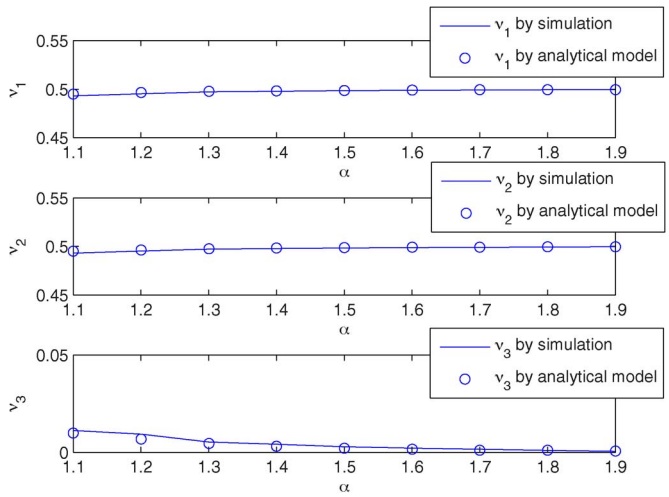


Fig. 9. Validation of the analytical model of $\nu_i, (i \in E)$.

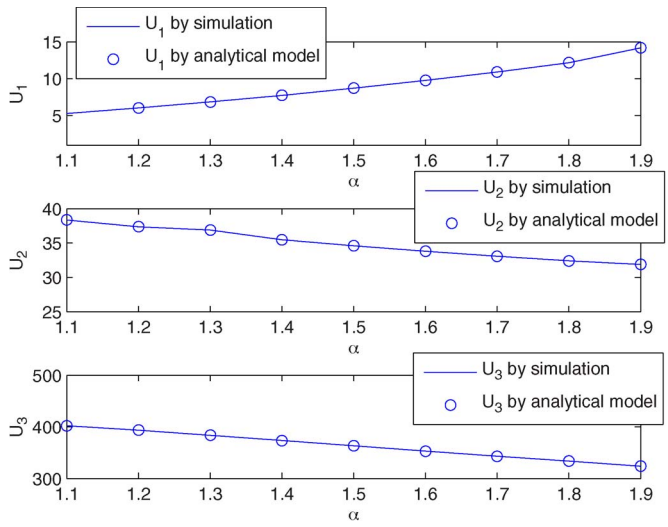


Fig. 10. Validation of the analytical model of $U_i, (i \in E)$.

expired. This leads to the results that UE is always staying in the *active period* and *light sleep period*. Therefore, the probability of the packet interarrival time larger than the $N_{\rm SD}t_{\rm SD} + t_I$ is highly reduced, which leads to the decrease in the stationary probability of *deep sleep period*. In Fig. 10, we can clearly observe that, with the increase in α, the mean sojourn time is increased by 53% for *active period* and decreased by 21% and 25% for *light sleep period* and *deep sleep period*, respectively. According to the definition of limit distribution of the DTSMP (25), the variation trend of limited distribution is decided by the trend of EMC's stationary probability and mean sojourn time in each state. Thus, Fig. 11 reveals the comprehensive situation of the DRX operations with the increasing traffic. The limit steady-state probability of *light sleep period* and *deep sleep period* reduce to 10% and 96%, whereas α increases from 1.1 to 1.9. It suggests that the fixed DRX parameter configurations cannot obtain good power-saving performance for varying traffic conditions, which can be directly seen by the degradation of PS in Fig. 12. Since we do not consider the buffer and service delay in this paper, the $\mathbb{E}(D)$ only reflects the delay caused by the DRX mechanism. As expected, while deploying

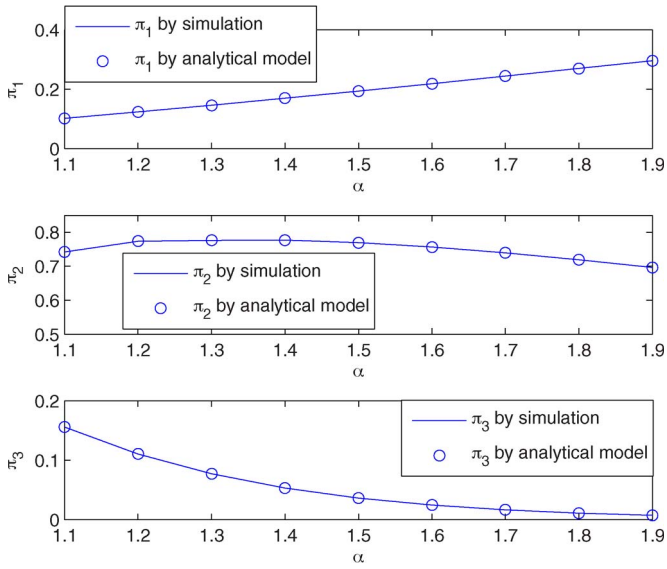


Fig. 11. Validation of the analytical model of $\pi_i, (i \in E)$.

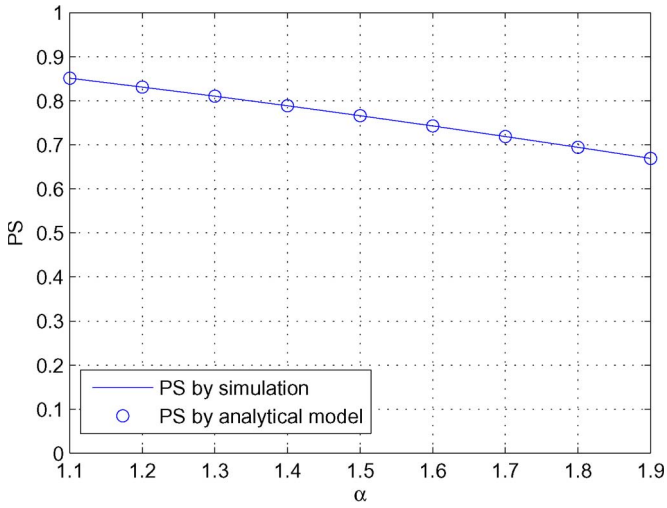


Fig. 12. Validation of the analytical model of PS.

the fixed DRX parameter configuration, the DRX mechanism is increasingly ineffective with the increase in traffic; therefore, the decrease in $\mathbb{E}(D)$ takes place, as shown in the Fig. 13. We can conclude that, to obtain a good performance for varying traffic, the tradeoff between PS and $\mathbb{E}(D)$ must be made.

B. Evaluation of the Online Power-Saving Strategy

To evaluate the performance of our strategy, we compare the OPSS with three fixed DRX cases and measure PS and $\mathbb{E}(D)$ for various α. The results are reported in Fig. 14. Specifically, the fixed DRX parameter sets for cases 1–3 are summarized in Table IV. The selectable parameters are defined as follows: $t_I \in \{1, 2, \ldots, 20\}$, $t_{\mathrm{SD}} \in \{10, 15, \ldots 50\}$, $t_{\mathrm{LD}} \in \{100, 150, \ldots 1000\}$, and $N_{\mathrm{SD}} \in \{5, 10, 15, 20\}$. In addition, the η for all cases equals to 0.1. The traffic trace is generated following the parameters of $x_m = 2$ and $m = 2000$. We assume that the $\gamma_d \leq \gamma_{\mathrm{ps}}$, whereas the $\alpha \in [1.1, 1.5]$, which indicates the OPSS should pick out the DRX parameters that satisfies

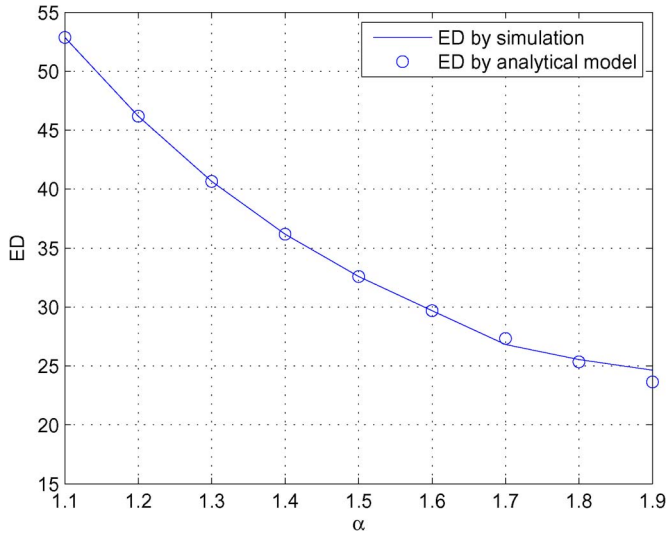


Fig. 13. Validation of the analytical model of $\mathbb{E}(D)$.

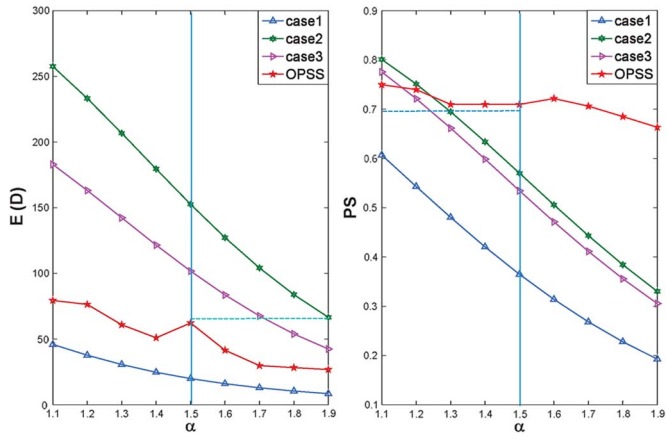


Fig. 14. Evaluation of the OPSS.

TABLE IV
FIXED DRX PARAMETER SETS FOR CASES 1, 2, AND 3

case	t_I	t_{SD}	t_{LD}	N_{SD}
1	15	20	100	15
2	10	10	200	5
3	10	10	150	5

$\mathrm{PS_{th}}$ and meanwhile minimize the delay. We also assume $\gamma_{\mathrm{ps}} < \gamma_d$ while $\alpha \in (1.5, 1.9]$, which indicates that the OPSS should pick out the DRX parameters that satisfy $t_{\max}$ and meanwhile maximize the PS. In particular, the $\mathrm{PS_{th}}$ is set as 0.7, and $t_{\max}$ is set as 75 ms in the simulation.

First, we discuss the situation of $\alpha \in [1.1, 1.5]$ (the left half of each subgraph). In cases 1 and 2, although PS is higher than $\mathrm{PS_{th}}$ while $\alpha < 1.3$, the mean packet delay is very high. On the other hand, case 3 can obtain good delay performance, but PS is much lower than $\mathrm{PS_{th}}$. It can be seen that the performance of the proposed OPSS can satisfy the $\mathrm{PS_{th}}$ with feasible DRX parameters. Meanwhile, the OPSS can provide a dynamic adjustment of DRX parameters under different traffic loads, which could maintain the mean packet delay at a low level.

Next, we discuss the situation of $\alpha \in [1.6, 1, 9]$ (the right half of each subgraph). It is impressive that, the OPSS not only could satisfy the delay constraint but also obtain good power-saving performance, which is much better than the fixed parameter strategies. It is worth noting that the advantage of the OPSS comes from the potentials of a large number of DRX parameter combinations. It also shows that, by taking all the DRX parameters into consideration, the tradeoff between PS and mean packet delay is not as simple, as shown in [11] and [14]. There can be an optimum DRX parameter configuration under specific traffic conditions, which could achieve best power-saving (or delay) performance, and have good delay (or power-saving) performance at the same time. It is left as the future work as how to efficiently find the optimum DRX parameter configuration from the feasible parameter set.

VII. Conclusion

In this paper, we have studied modeling and optimizing LTE DRX operations under self-similar traffic. We have presented an analytical model for evaluating the DRX operation by using the DTSMP. Based on the analytical model, we obtained accurate power-saving factors and wake-up delays without using sophisticated mathematical techniques. Furthermore, based on the analytical model, we also designed an OPSS. The OPSS could unbiasedly estimate the traffic condition and DRX operations within 1 s, and optimize the DRX parameter configurations. Therefore, the OPSS could enhance the DRX operations with great energy saving while still meet the packet delay constraint. Extensive MATLAB simulations showed the correctness of our analytical model and the good performance of the OPSS.

Appendix

Algorithm 1 On-line Power Saving Strategy

Estimation phase:

During the observing window $T_{\rm ob}$:

- eNB records the samples of packet interarrival times $\{x_1, x_2, \ldots, x_n\}$ for the specific UE.
- UE records N_i and N_{ij} $(i, j \in E)$.

After $T_{\rm ob}$:

- UE estimates $\hat{\nu}_i, (i \in E)$, $\hat{U}_i, (i \in E - \{3\})$ by (44) and (45), respectively, and then sends them to eNB in the first ON duration.
- Based on the updates from UE, the eNB estimates the U_3, $\bar{U}$, π_i by (24), (46) and (47), respectively. Meanwhile, eNB also estimates $\hat{X}_m$ and α^* by (32) and (40), respectively. Then, eNB obtains $\mathrm{PS}_{\rm now}$ and $\mathbb{E}(D)_{\rm now}$ by (26) and (29), respectively, at the end of this phase. We denote the current DRX parameters set as $\kappa_{\rm now}$

Optimization phase:

Update γ_d, $\gamma_{\rm ps}$

if $\gamma_d > \gamma_{\rm ps}$ **then**

- Update $t_{\max}$;
- Look up the table Ω, and find out the feasible DRX configuration set ω, in which all entries κ could satisfies: $\mathbb{E}(D) < t_{\max}$.
- Find out the largest PS in the feasible set, denoted as $\mathrm{PS}_{\max}$, and denote the corresponding DRX parameters set as $\kappa_{\max}$. Denote the optimized DRX parameters set as $\kappa_{\rm opt}$.

if $\mathrm{PS}_{\rm now} \geq \mathrm{PS}_{\max}$ **then**
  if $\mathbb{E}(D) \leq t_{\max}$ **then**
    $\kappa_{\rm opt} \leftarrow \kappa_{\rm now}$
  else
    $\kappa_{\rm opt} \leftarrow \kappa_{\max}$
  end if
else
  $\kappa_{\rm opt} \leftarrow \kappa_{\max}$
end if
else

- Update $\mathrm{PS}_{\rm th}$
- Look up the table Ω, and find out the feasible set ω that satisfies: $\mathrm{PS} > \mathrm{PS}_{\rm th}$
- Find out the smallest $\mathbb{E}(D)$ in the feasible set, denote as $\mathbb{E}(D)_{\min}$, and denote the corresponding DRX parameters set as $\kappa_{\max}$. Moreover, denote the optimized DRX parameters set as $\kappa_{\rm opt}$.

if $\mathrm{PS}_{\rm now} \geq \mathrm{PS}_{\rm th}$ **then**
  if $\mathbb{E}(D) \leq \mathbb{E}(D)_{\min}$ **then**
    $\kappa_{\rm opt} \leftarrow \kappa_{\rm now}$
  else
    $\kappa_{\rm opt} \leftarrow \kappa_{\max}$
  end if
else
  $\kappa_{\rm opt} \leftarrow \kappa_{\max}$
end if
  end if

References

[1] S. Chen and J. Zhao, "The requirements, challenges, and technologies for 5g of terrestrial mobile telecommunication," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 36–43, May 2014.

[2] *IEEE Standard for Local and Metropolitan Area Networks Part 16: Air Interference for Fixed Broadband Wireless Access Systems*, IEEE Std. 802.16-2004, 2004.

[3] S.-R. Yang, S.-Y. Yan, and H.-N. Hung, "Modeling UMTS power saving with bursty packet data traffic," *IEEE Trans. Mobile Comput.*, vol. 6, no. 12, pp. 1398–1409, Dec. 2007.

[4] *Medium Access Control (MAC) Protocol Specification*, 3GPP TS 36.321 Std. 10.2.0 (Rel. 10), 2011.

[5] "LTE RAN enhancements for diverse data application," Sophia-Antipolis, France, 3GPP TR 36.822 Tech. Rep. 11.0.0 (Rel. 11), 2012.

[6] "DRX parameters in LTE," Espoo, Finland, Tech. Rep. R2-071285, 2007.

[7] "On the need for flexible DRX," Espoo, Finland, Tech. Rep. R2-071286, 2007.

[8] M. Polignano, D. Vinella, D. Laselva, J. Wigard, and T. B. Sorensens, "Power savings and QoS impact for VoIP application with DRX/DTX feature in LTE," in *Proc. 73rd IEEE VTC Spring*, 2011, pp. 1–5.

[9] S. C. Jha, A. T. Koç, and R. Vannithamby, "Optimization of Discontinuous Reception (DRX) for mobile Internet applications over LTE," in *Proc. IEEE VTC Fall*, 2012, pp. 1–5.

[10] L. Zhou *et al.*, "Performance analysis of power saving mechanism with adjustable DRX cycles in 3GPP LTE," in *Proc. 68th VTC-Fall*, 2008, pp. 1–5.

[11] S. Fowler, R. S. Bhamber, and A. Mellouk, "Analysis of adjustable and fixed DRX mechanism for power saving in LTE/LTE-advanced," in *Proc. IEEE ICC*, 2012, pp. 1964–1969.

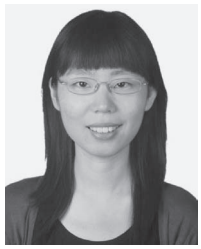
[12] J. Wigard, T. Kolding, L. Dalsgaard, and C. Coletti, "On the user performance of lte ue power savings schemes with discontinuous reception in LTE," in *Proc. IEEE ICC Workshops*, 2009, pp. 1–5.

[13] S. Alouf, V. Mancuso, and N. C. Fofack, "Analysis of power saving and its impact on web traffic in cellular networks with continuous connectivity," *Pervasive Mobile Comput.*, vol. 8, no. 5, pp. 646–661, Oct. 2012.
[14] S. Jin and D. Qiao, "Numerical analysis of the power saving in 3GPP LTE advanced wireless networks," *IEEE Trans. Veh. Technol.*, vol. 61, no. 4, pp. 1779–1785, May 2012.
[15] Y.-P. Yu and K.-T. Feng, "Traffic-based DRX cycles adjustment scheme for 3GPP LTE systems," in *Proc. 75th IEEE VTC Spring*, 2012, pp. 1–5.
[16] C.-H. Hsu, K.-T. Feng, and C.-J. Chang, "Statistical control approach for sleep-mode operations in IEEE 802.16 m systems," *IEEE Trans. Veh. Technol.*, vol. 59, no. 9, pp. 4453–4466, Nov. 2010.
[17] V. Paxson and S. Floyd, "Wide area traffic: The failure of Poisson modeling," *IEEE/ACM Trans. Netw.*, vol. 3, no. 3, pp. 226–244, Jun. 1995.
[18] Z. Sahinoglu and S. Tekinay, "On multimedia networks: Self-similar traffic and network performance," *IEEE Commun. Mag.*, vol. 37, no. 1, pp. 48–52, Jan. 1999.
[19] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the self-similar nature of ethernet traffic (extended version)," *IEEE/ACM Trans. Netw.*, vol. 2, no. 1, pp. 1–15, Feb. 1994.
[20] B. Mandelbrot, "Self-similar error clusters in communication systems and the concept of conditional stationarity," *IEEE Trans. Commun. Technol.*, vol. COM-13, no. 1, pp. 71–90, Mar. 1965.
[21] R. G. Gallager, *Discrete Stochastic Processes*, vol. 101. Boston, MA, USA: Kluwer, 1996.
[22] S. Jin, X. Chen, D. Qiao, and S. Choi, "Adaptive sleep mode management in IEEE 802.16 m wireless metropolitan area networks," *Comput. Netw.*, vol. 55, no. 16, pp. 3774–3783, Nov. 2011.
[23] V. Barbu, J. Bulla, and A. Maruotti, "Estimation of the stationary distribution of a semi-Markov chain," *J. Reliab. Stat. Stud.*, vol. 5, pp. 15–26, 2012.
[24] "Feasibility study for OFDM for UTRAN enhancement," Sophia-Antipolis, USA, 3GPP TR 25.892 Tech. Rep. 2.0.0 (Rel. 6), 2004.
[25] H. Tanizaki, *Computational Methods in Statistics and Econometrics*, vol. 172. Boca Raton, FL, USA: CRC, 2004.
[26] *Universal Mobile Telecommunications System (UMTS): Selection procedures for the choice of radio transmission technologies of the UMTS*, ETSI Std. 30.03 TR 3.2.0, 1998.
[27] V. Barbu, M. Boussemart, and N. Limnios, "Discrete-time semi-Markov model for reliability and survival analysis," *Commun. Stat., Theory Methods*, vol. 33, no. 11, pp. 2833–2868, 2004.
[28] C. Bontu and E. Illidge, "DRX mechanism for power saving in LTE," *IEEE Commun. Mag.*, vol. 47, no. 6, pp. 48–55, Jun. 2009.
[29] J. L. Petersen, "Estimating the parameters of a Pareto distribution," Univ. Montana, Missoula, MT, USA, Tech. Rep., 2000.

Ke Wang is currently working toward the Ph.D. degree with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications (BUPT), Beijing, China.

He is also currently with the Key Laboratory of Universal Wireless Communications, Ministry of Education, BUPT. His research interests include scheduling algorithms and energy-efficient communications.

Mr. Wang has served as a Reviewer for several journals and conference proceedings, including IEEE COMMUNICATION LETTERS, the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, *Elsevier Computer Communications*, and the 2013 IEEE Global Communications Conference. He received a postgraduate innovation fund from Rohde & Schwarz–Beijing University of Posts and Telecommunications in 2013.



Xi Li received the B.E. and Ph.D. degrees in communication and information systems from Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2005 and 2010, respectively.

She is currently an Associate Professor with the School of Information and Communication Engineering, BUPT, where is also with the Key Laboratory of Universal Wireless Communications of the Ministry of Education. Her current research interests include broadband wireless communications and heterogeneous networks.

Dr. Li served as the Special Track Chair on cognitive test beds at the 2011 International Conference on Communications and Networking in China and as a member of the Technical Program Committees of the Medium Access Control (MAC) Layer Track at the 2012 and 2013 IEEE Wireless Communications and Networking Conference, as well as the MAC and Cross-Layer Design Track at the 2012 IEEE International Symposium on Personal, Indoor, and Mobile Radio Communications.

Hong Ji (SM'09) received the B.S. degree in communication engineering and the M.S. and Ph.D. degrees in information and communication engineering from Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 1989, 1992, and 2002, respectively.

From June to December 2006, she was a Visiting Scholar with the University of British Columbia, Vancouver, BC, Canada. She is currently a Professor with the Key Laboratory of Universal Wireless Communication, Ministry of Education, BUPT. She is also engaged with national science research projects, including the Hi-Tech Research and Development Program of China (863 program) and the National Natural Science Foundation of China. Her research interests include heterogeneous networks, peer-to-peer protocols, cognitive radio networks, relay networks, Long-Term Evolution/fifth generation, and cooperative communications.

Dr. Ji serves on the editorial boards of several journals, including the *Wiley International Journal of Wireless Communications and Mobile Computing* and the *Wiley International Journal of Communication Systems*.

Xiaojiang Du (SM'03) received the B.S. and M.S. degrees from Tsinghua University, Beijing, China, in 1996 and 1998, respectively, and the M.S. and Ph.D. degrees from the University of Maryland, College Park, MD, USA, in 2002 and 2003, respectively, all in electrical engineering.

From August 2004 to July 2009, he was an Assistant Professor with the Department of Computer Science, North Dakota State University, Fargo, ND, USA. He is currently an Associate Professor with the Department of Computer and Information Sciences, Temple University, Philadelphia, PA, USA. He is the author of over 120 journal and conference papers. His research interests are security, cloud computing, wireless networks, and computer networks and systems.

Dr. Du is a Life Member of the Association for Computing Machinery. He is serving or has served as a Technical Program Committee (TPC) member for several premier Association for Computing Machinery/IEEE conferences, such as the IEEE International Conference on Computer Communications (2007–2015), the IFIP/IEEE International Symposium on Integrated Network Management, IEEE International Conference on Communications, IEEE Global Communications Conference, IEEE Wireless Communications and Networking Conference, and BroadNet. He serves on the editorial boards of four international journals. He received the Excellence in Research Award from North Dakota State University in May 2009.