Modeling UMTS Power Saving with Bursty Packet Data Traffic

Shun-Ren Yang*, Sheng-Ying Yan, and Hui-Nien Hung

Abstract

The universal mobile telecommunications system (UMTS) utilizes the discontinuous reception (DRX) mechanism to reduce the power consumption of mobile stations (MSs). DRX permits an idle MS to power off the radio receiver for a predefined sleep period, and then wake up to receive the next paging message. The sleep/wake-up scheduling of each MS is determined by two DRX parameters: the inactivity timer threshold and the DRX cycle. In the literature, analytic and simulation models have been developed to study the DRX performance mainly for Poisson traffic. In this paper, we propose a novel semi-Markov process to model the UMTS DRX with bursty packet data traffic. The analytic results are validated against simulation experiments. We investigate the effects of the two DRX parameters on output measures including the power saving factor and the mean packet waiting time. Our study provides inactivity timer and DRX cycle value selection guidelines for various packet traffic patterns.

Keywords: bursty packet data traffic, discontinuous reception, power saving, universal mobile telecommunications system (UMTS)

---

*Corresponding Author: Shun-Ren Yang, Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan, R.O.C.; Email: sryang@cs.nthu.edu.tw

1 Introduction

The third generation mobile cellular system *universal mobile telecommunications system* (UMTS) offers high data transmission rates to support a variety of mobile applications including voice, data and multimedia. In order to fulfill the high-bandwidth requirement of these different services, the *mobile station* (MS) power saving is a crucial issue for the UMTS network operation. Since the data bandwidth is significantly restricted by the battery capacity, most existing wireless mobile networks (including UMTS) employ *discontinuous reception* (DRX) to conserve the power of MSs. DRX allows an idle MS to power off the radio receiver for a predefined period (called the *DRX cycle*) instead of continuously listening to the radio channel. Some typical DRX mechanisms are briefly described as follows.

- In MOBITEX [15], all sleeping MSs are required to synchronize with a specific $\langle$SVP6$\rangle$ frame and wake up immediately before the $\langle$SVP6$\rangle$ transmission starts. When some MSs experience high traffic loads, the network may decide to shorten the $\langle$SVP6$\rangle$ announcement interval to reduce the frame delay. As a result, the low traffic MSs will consume extra unnecessary power budget.

- In CDPD [5, 12] and IEEE 802.11 [8], a sleeping MS is not forced to wake up at every announcement instant. Instead, the MS may choose to skip some announcements for further reducing its power consumption. A wake-up MS has to send a receiver ready (RR) frame to notify the network of its capability to receive the pending frames. However, such RR transmissions may collide with each other if the MSs tend to wake up at the same time. Thus, RR retransmissions may occur and extra power is unnecessarily consumed.

- UMTS DRX [2, 4] improves the aforementioned mechanisms by allowing an MS to negotiate its DRX cycle length with the network. Therefore, the network is aware of the sleep/wake-up scheduling of each MS, and only delivers the paging message when the MS wakes up.

In the literature, DRX mechanisms have been studied. Lin *et al.* [12] proposed simulation models to investigate the CDPD DRX mechanism. In [10], an analytic model was developed to investigate the CDPD DRX mechanism. This model does not provide close-form solution. Furthermore, the model is not validated against simulation experiments. In our previous work [21], we proposed a variant of the $M/G/1$ vacation model to explore the performance of the UMTS DRX. We derived the close-form equations for the output measures based on the Poisson assumption. However, the Poisson distribution has been proven to be impractical when modeling bursty packet data traffic [20]. In [11, 23], simulation and analysis were utilized to examine the UMTS DRX mechanism. The authors studied the impact of inactivity timer on energy consumption for both real-time and non-real-time traffic. However, they do not consider the mean packet waiting time under DRX. This paper proposes a novel semi-Markov process to model the UMTS DRX for bursty packet data applications. The analytic results are validated against simulation experiments. Based on the proposed analytic and simulation models, the DRX performance is investigated by numerical examples. Specifically, we consider the following two performance measures:

- power saving factor: the probability that the MS receiver is turned off when exercising the UMTS DRX mechanism; this factor indicates the percentage of power saving in the DRX (compared with the case where DRX is not exercised);
- mean packet waiting time: the expected waiting time of a packet in the UMTS network buffer before it is transmitted to the MS.

2 UMTS DRX Mechanism

As illustrated in Figure 1, a simplified UMTS architecture consists of the *core network* and the *UMTS terrestrial radio access network* (UTRAN). The core network is responsible for switching/routing calls and data connections to the external networks, while the UTRAN handles all radio-related functionalities. The UTRAN consists of *radio network controllers* (RNCs) and *Node Bs* (i.e., base stations) that are connected by an *asynchronous transfer mode* (ATM) network.

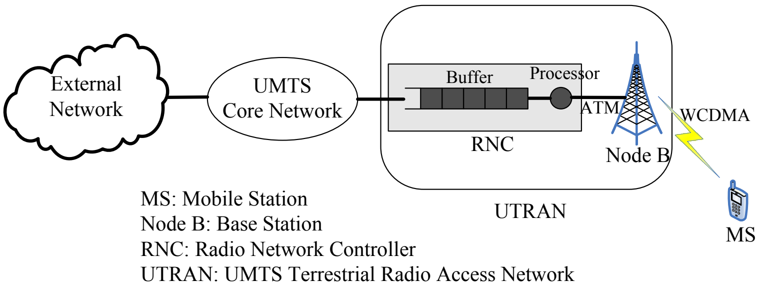

Figure 1: A simplified UMTS network architecture

An MS communicates with Node Bs through the radio interface based on the wideband CDMA (WCDMA) technology [7].

The UMTS DRX mechanism is realized through the *radio resource control* (RRC) finite state machine exercised between the RNC and the MS [1]. There are two modes in this finite state machine (see Figure 2). In the **RRC Idle** mode, the MS is tracked by the core network without the help of the UTRAN. When an RRC connection is established between the MS and its serving RNC, the MS enters the **RRC Connected** mode. In this mode, the MS could stay in one of the following four states:

- In the **Cell_DCH** state, the MS occupies a dedicated traffic channel;
- In the **Cell_FACH** state, the MS is allocated a common or shared traffic channel;
- In the **Cell_PCH** state, no uplink access is possible, and the MS monitors paging messages from the RNC;
- In the **URA_PCH** state, the MS eliminates the location registration overhead by performing URA updates instead of cell updates.

In the **Cell_DCH** and **Cell_FACH** states, the MS receiver is always turned on to receive packets. These states correspond to the *power active mode*. In the **RRC Idle** mode, **Cell_PCH** and **URA_PCH** states, the DRX is exercised to conserve the MS power budget. These states/mode correspond to the *power saving mode*. Under DRX, the MS receiver activities could be described in terms of three periods (see Figure 3):

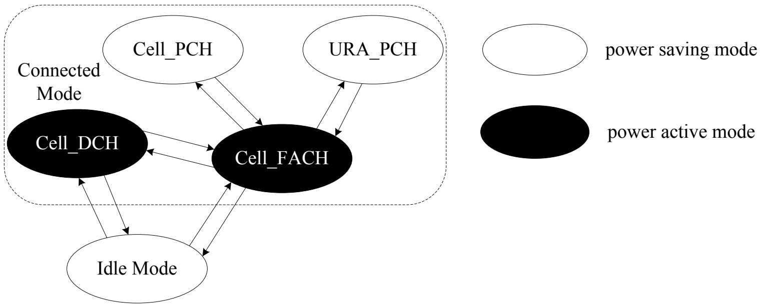

Figure 2: The RRC state diagram

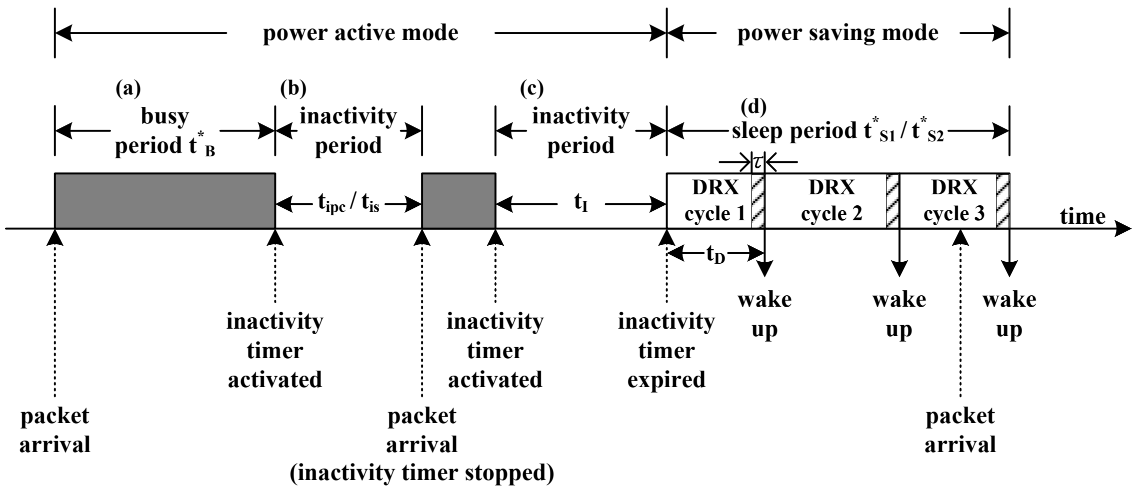

Figure 3: The timing diagram for UMTS MS receiver activities

- In the **busy period** (see Figure 3(a)), the MS is in the power active mode, and the UMTS core network delivers packets to the MS through the RNC and Node B in the *first in-first out* (FIFO) order. Compared with WCDMA radio transmission, ATM is much faster and more reliable. Therefore the ATM transmission delay is ignored in this paper, and the RNC and the Node B are regarded as a FIFO server. Furthermore, due to high error-rate and low bit-rate nature of radio transmission, the *Stop-And-Wait Hybrid Automatic Repeat re-Quest* (SAW-Hybrid ARQ) flow control algorithm [3] is executed to guarantee successful radio packet delivery: when the Node B sends a packet to the MS, it waits for a positive acknowledgment (ack) from the MS before it can transmit the next packet. Hybrid ARQ was originally proposed for the *High Speed Downlink Packet Access* (HSDPA) system, and has also been adopted by next-generation wireless networks including IEEE 802.16 WiMAX system. SAW-Hybrid ARQ is one of the simplest forms of ARQ requiring very little overhead. Hybrid ARQ using this stop-and-wait mechanism offers significant improvements by reducing the overall bandwidth demanded for signaling and the MS memory. Due to it's simplicity, SAW-Hybrid ARQ could also be implemented in the earlier UMTS releases without HSDPA support.

- In the **inactivity period** (see Figure 3(b)), the RNC buffer is empty, and the RNC inactivity timer is activated. If any packet arrives at the RNC before the RNC inactivity timer expires, the timer is stopped. The RNC processor starts another busy period to transmit packets. In the inactivity period, the MS receiver is turned on, and the MS is still in the power active mode.

- If no packet arrives within the threshold t_I of the RNC inactivity timer (see Figure 3(c)), the MS turns off its radio receiver and enters the **sleep period** to save power (see Figure 3(d)). The MS sleep period contains at least one DRX cycles t_D. At the end of a DRX cycle, the MS wakes up to listen to the paging channel. If the paging message indicates that some packets have arrived at the RNC during the last DRX cycle, the MS starts to receive packets and the sleep period terminates. Otherwise, the MS returns to sleep until the end of the next

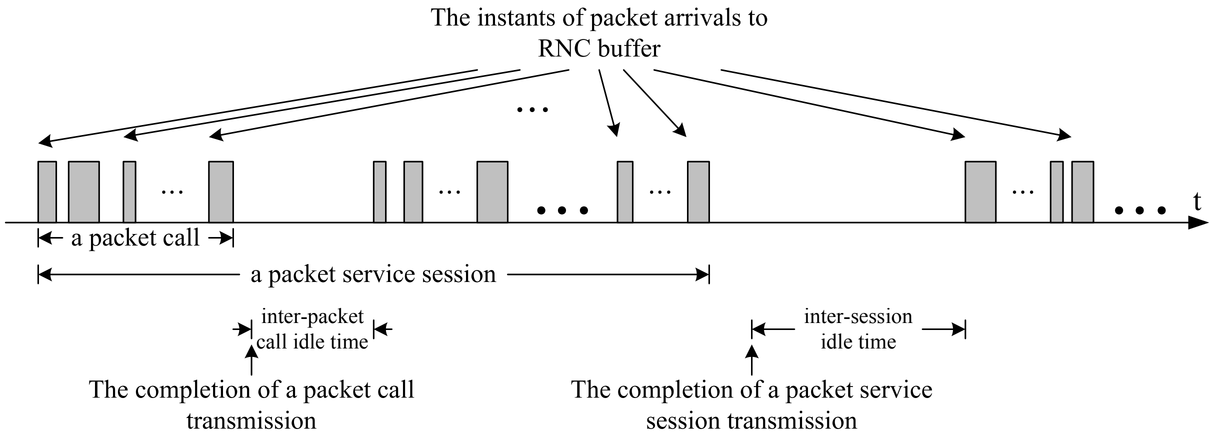

Figure 4: ETSI packet traffic model

DRX cycle. In the power saving mode, the RNC processor will not transmit any packets to the MS.

3 ETSI Packet Traffic Model

The validity of traditional queueing analyses depends on the Poisson nature of the data traffic. However, in many real-world cases, it has been found that the predicted results from these queueing analyses differ substantially from the actual observed performance. In recent years, a number of studies have demonstrated that for some environments, the data traffic pattern is self-similar [20] rather than Poisson. Compared with traditional Poisson traffic models which typically focus on a very limited range of time scales and are thus short-range dependent in nature, self-similar traffic exhibits burstiness and correlations across an extremely wide range of time scales (i.e., possesses long-range dependence). It has also been shown that heavy-tailed distributions such as Pareto and Weibull distributions are more appropriate when modeling data network traffic [14]. In this paper, we adopt the ETSI packet traffic model [6], where the packet size and the packet transmission time are assumed to follow the truncated Pareto distribution.

As shown in Figure 4, we assume that the packet data traffic consists of packet service sessions. Each packet service session contains one or more packet calls depending on the applications. For example, the streaming video may comprise one single packet call for a packet session, while a web surfing packet session includes a sequence of packet calls. An MS/mobile user initiates a packet call when requesting an information element (e.g., the downloading of a WWW page). If the request is permitted, a burst of packets (e.g., as a whole constituting a video clip in the WWW page) will be transmitted to the MS through the RNC and Node B. When the RNC receives the positive acknowledgment for the last packet from the MS, the current packet call transmission has completed. The time interval between the end of this packet call transmission and the beginning of the next packet call transmission is referred to as the *inter-packet call idle time* t_{ipc}. Having received all packets of the ongoing packet service session, the MS will then experience an even longer *inter-session idle time* t_{is}. The t_{is} period represents the time interval between the end of the packet session and the beginning of the next packet session.

The statistical distributions of the parameters in our traffic model follow the recommendation in [6] and are summarized as follows. Note that, since we consider continuous time scale in this paper, the exponential distribution is used to replace the geometric distribution for continuous random variables.

- The **inter-session idle time** t_{is} is modeled as an exponentially distributed random variable with mean $1/\lambda_{is}$.
- The **number of packet calls** N_{pc} within a packet service session is assumed to be a geometrically distributed random variable with mean μ_{pc}.
- The **inter-packet call idle time** t_{ipc} is an exponential random variable with mean $1/\lambda_{ipc}$.
- The **number of packets** N_p within a packet call follows a geometric distribution with mean μ_p.
- The **inter-packet arrival time** t_{ip} within a packet call is drawn from an exponential distribution with mean $1/\lambda_{ip}$.
- The truncated (or cut-off) Pareto distribution is used to model the **packet size**. Pareto distribution [9] has been found to match very well with the actual data traffic measurements [20].

A Pareto distribution has two parameters: the shape parameter β and the scale parameter l, where β describes the "heaviness" of the tail of the distribution. The probability density function is

$$f_x(x) = \left(\frac{\beta}{l}\right)\left(\frac{l}{x}\right)^{\beta+1} \quad \text{and the expected value is} \quad E[x] = \left(\frac{\beta}{\beta-1}\right) l.$$

If β is between 1 and 2, the variance for the distribution becomes infinity. We follow the suggestion in [6] and define the packet size S_d with the following formula:

$$\text{Packet Size } S_d = \min(P, m),$$

where P is a normal Pareto distributed random variable with $\beta = 1.1$ and $l = 81.5$ bytes, and $m = 66666$ bytes is the maximum allowed packet size. According to the above parameter values, the average packet size is calculated as 480 bytes. The above β, l, and m parameter settings for the packet size distribution have been validated by the ETSI technical bodies [6]. Many telecommunications vendors and operators adopted these settings to conduct the UMTS field trials. These configurations were also followed by a number of analytic and simulation studies in the literature [11, 18] to investigate the performance of UMTS networks.

- Let the **packet service time** t_x denote the time interval between when the packet is transmitted by the RNC processor and when the corresponding positive ack is received by the RNC processor. The t_x distribution has mean value $1/\lambda_x$. In our model, we suppose that t_x is proportional to the packet size S_d and is defined as

  $$t_x = \frac{\text{Packet Size } S_d}{\text{Transmission Bit Rate}}.$$

  Six types of transmission bit rates are proposed in [6] for the WWW surfing service: 8 kbit/s, 32 kbit/s, 64 kbit/s, 144 kbit/s, 384 kbit/s and 2048 kbit/s.

The ETSI model for bursty packet data traffic with long-range dependence can be justified by [14]. Appendix D of [14] has shown that the $M/G/\infty$ model with infinite-variance Pareto distribution

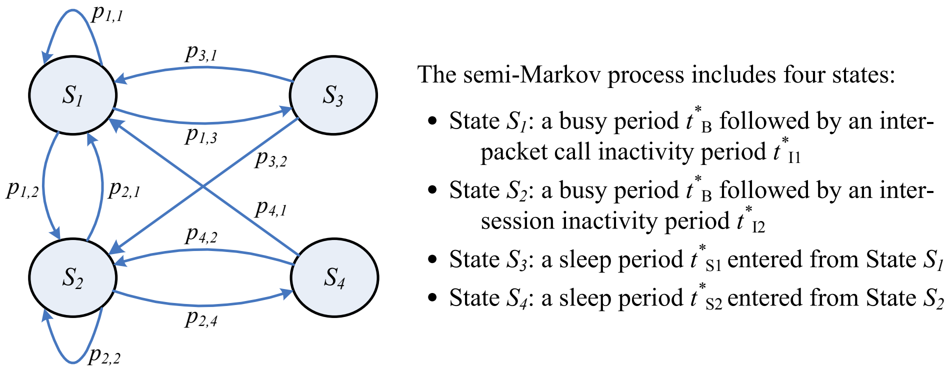

Figure 5: A semi-Markov process for UMTS power saving analysis

can be used to generate self-similar traffic. In the ETSI packet traffic model, the packet arrival process is governed by the exponential and geometric distributions with memoryless property, while the packet service time t_x is a truncated Pareto random variable.

4 An Analytic Model for UMTS Power Saving

Based on the ETSI packet traffic model defined in the last section, this section proposes an analytic model to study the UMTS power saving mechanism. The notation used in the analytic model is listed in Appendix A. Let the two UMTS DRX parameters t_I and t_D be of fixed values $1/\lambda_I$ and $1/\lambda_D$, respectively. We first describe a *semi-Markov process* [13, 17]. We then show how to use this semi-Markov process to investigate the performance of the UMTS power saving mechanism (including the power saving factor and the mean packet waiting time). As illustrated in Figure 5, this semi-Markov process consists of four states:

- State S_1 includes a busy period t^*_B and then an inter-packet call inactivity period t^*_{I1}.
- State S_2 includes a busy period t^*_B and then an inter-session inactivity period t^*_{I2}.
- State S_3 includes a sleep period t^*_{S1} which is entered from state S_1.
- State S_4 includes a sleep period t^*_{S2} which is entered from state S_2.

If we view this semi-Markov process only at the times of state transitions, then we could obtain an embedded Markov chain with state transition probabilities $p_{i,j}$ (where $i, j \in \{1, 2, 3, 4\}$). These state transition probabilities are derived as follows.

- $p_{1,1}$ and $p_{1,2}$: In state S_1, the RNC inactivity timer is activated at the end of the busy period t_B^*, and then the MS enters the inter-packet call inactivity period t_{I1}^*. Note that, based on the burstiness nature, our analytic model assumes that a busy period corresponds to the transmission duration of a packet call. Namely, the inter-packet arrival time t_{ip} within a packet call is significantly shorter than the packet service time t_x, and a busy period will not terminate until the end of the corresponding packet call delivery. When the first packet of the next packet call arrives at the RNC before the inactivity timer expires (with probability $q_1 = \Pr[t_{ipc} < t_I] = 1 - e^{-\lambda_{ipc}/\lambda_I}$), the timer is stopped and another busy period begins. In this case, if the new arriving packet call is the last one of the ongoing session (with probability $q_2 = 1/\mu_{pc}$; the memoryless property of geometric distributions), then the MS enters state S_2. Otherwise (with probability $1 - q_2$), the ongoing session continues and the MS enters state S_1 again. From the above discussion, we have

$$p_{1,1} = q_1(1 - q_2) = (1 - e^{-\frac{\lambda_{ipc}}{\lambda_I}})(1 - \frac{1}{\mu_{pc}})$$

  and

$$p_{1,2} = q_1 q_2 = (1 - e^{-\frac{\lambda_{ipc}}{\lambda_I}})\frac{1}{\mu_{pc}}.$$

- $p_{2,1}$ and $p_{2,2}$: The derivations of $p_{2,1}$ and $p_{2,2}$ are exactly the same as that of $p_{1,1}$ and $p_{1,2}$ except that the inter-packet call idle period t_{ipc} is replaced by the inter-session idle period t_{is} and q_1 is replaced by $q_3 = \Pr[t_{is} < t_I] = 1 - e^{-\lambda_{is}/\lambda_I}$. Therefore, we have

$$p_{2,1} = q_3(1 - q_2) = (1 - e^{-\frac{\lambda_{is}}{\lambda_I}})(1 - \frac{1}{\mu_{pc}})$$

  and

$$p_{2,2} = q_3 q_2 = (1 - e^{-\frac{\lambda_{is}}{\lambda_I}})\frac{1}{\mu_{pc}}.$$

- $p_{1,3}$ and $p_{2,4}$: In state S_1, if no packet arrives before the inactivity timer expires (with probability $1-q_1$), then the MS enters the sleep period t^*_{S1} (state S_3; i.e., the power saving mode). Therefore,

$$p_{1,3} = 1 - q_1 = e^{-\frac{\lambda_{ipc}}{\lambda_I}}. \tag{1}$$

Similarly, $p_{2,4}$ can be derived by substituting q_3 for q_1 in (1), and we have

$$p_{2,4} = 1 - q_3 = e^{-\frac{\lambda_{is}}{\lambda_I}}.$$

- $p_{3,1}$ and $p_{3,2}$: In state S_3, if the next packet call terminates the ongoing session (with probability q_2), then the MS will move to S_2 at the next state transition. Otherwise (with probability $1-q_2$), the MS will switch to state S_1. Thus,

$$p_{3,1} = 1 - q_2 = 1 - \frac{1}{\mu_{pc}} \quad \text{and} \quad p_{3,2} = q_2 = \frac{1}{\mu_{pc}}.$$

- $p_{4,1}$ and $p_{4,2}$: Similar to $p_{3,1}$ and $p_{3,2}$, the transition from state S_4 to state S_1 or S_2 also depends only on whether the coming packet call is the end of the packet session. Therefore, we conclude that

$$p_{4,1} = p_{3,1} = 1 - q_2 = 1 - \frac{1}{\mu_{pc}} \quad \text{and} \quad p_{4,2} = p_{3,2} = q_2 = \frac{1}{\mu_{pc}}.$$

The transition probability matrix $P = (p_{i,j})$ of the embedded Markov chain can thus be given as

$$P = \begin{pmatrix} q_1(1-q_2) & q_1q_2 & 1-q_1 & 0 \\ q_3(1-q_2) & q_3q_2 & 0 & 1-q_3 \\ 1-q_2 & q_2 & 0 & 0 \\ 1-q_2 & q_2 & 0 & 0 \end{pmatrix}.$$

Let π_i^e ($i \in \{1,2,3,4\}$) denote the probability that the embedded Markov chain will stay at S_i in the steady state. By using $\sum_{i=1}^4 \pi_i^e = 1$ and the balance equations $\pi_i^e = \sum_{j=1}^4 \pi_j^e p_{j,i}$, we can solve the stationary distribution $\Pi^e = (\pi_i^e)$ and obtain

$$\Pi^e = \begin{cases} \pi_1^e = \dfrac{1-q_2}{1+(1-q_1)(1-q_2)+q_2(1-q_3)} \\ \pi_2^e = \dfrac{q_2}{1+(1-q_1)(1-q_2)+q_2(1-q_3)} \\ \pi_3^e = \dfrac{(1-q_1)(1-q_2)}{1+(1-q_1)(1-q_2)+q_2(1-q_3)} \\ \pi_4^e = \dfrac{q_2(1-q_3)}{1+(1-q_1)(1-q_2)+q_2(1-q_3)} \end{cases}. \tag{2}$$

Let H_i ($i \in \{1,2,3,4\}$) be the holding time of the semi-Markov process at state S_i. We proceed to derive $E[H_i]$.

- $E[H_1]$: In state S_1, the MS experiences a busy period t_B^* and then an inter-packet call inactivity period t_{I1}^*. Hence,

$$E[H_1] = E[t_B^*] + E[t_{I1}^*]. \tag{3}$$

Since a busy period is identical to the duration of a packet call delivery, a t_B^* consists of N_p packet service times t_x. From Wald's theorem [13, Th. 5.18], we have

$$E[t_B^*] = E[N_p]E[t_x] = \frac{\mu_p}{\lambda_x}. \tag{4}$$

As shown in Figure 3, $t_{I1}^* = \min(t_{ipc}, t_I)$. If the next packet arrives before the inactivity timer expires (i.e., $t_{ipc} < t_I$), then $t_{I1}^* = t_{ipc}$, and the next busy period follows (see Figure 3(b)). Otherwise (the next packet arrives after the inactivity timer has expired; i.e., $t_{ipc} \geq t_I$), $t_{I1}^* = t_I$, and the next sleep period follows (see Figure 3(c)). Therefore,

$$\begin{aligned}
E[t_{I1}^*] &= E[\min(t_{ipc}, t_I)] \\
&= \int_{x=0}^{\infty} \Pr[\min(t_{ipc}, t_I) > x]dx \\
&= \int_{x=0}^{1/\lambda_I} \Pr[t_{ipc} > x]dx \\
&= \int_{x=0}^{1/\lambda_I} e^{-\lambda_{ipc}x}dx \\
&= \left(\frac{1}{\lambda_{ipc}}\right)\left[1 - e^{-\frac{\lambda_{ipc}}{\lambda_I}}\right].
\end{aligned} \tag{5}$$

Substitute (4) and (5) into (3) to yield

$$E[H_1] = \frac{\mu_p}{\lambda_x} + \left(\frac{1}{\lambda_{ipc}}\right)\left[1 - e^{-\frac{\lambda_{ipc}}{\lambda_I}}\right]. \tag{6}$$

- $E[H_2]$: State S_2 contains a busy period t_B^* and an inter-session inactivity period t_{I2}^*. Therefore,

$$E[H_2] = E[t_B^*] + E[t_{I2}^*]. \tag{7}$$

Similar to the derivation of $E[t^*_{I1}]$, $E[t^*_{I2}]$ is

$$E[t^*_{I2}] = E[\min(t_{is}, t_I)] = \left(\frac{1}{\lambda_{is}}\right)\left[1 - e^{-\frac{\lambda_{is}}{\lambda_I}}\right]. \tag{8}$$

Substituting (4) and (8) into (7), $E[H_2]$ is expressed as

$$E[H_2] = \frac{\mu_p}{\lambda_x} + \left(\frac{1}{\lambda_{is}}\right)\left[1 - e^{-\frac{\lambda_{is}}{\lambda_I}}\right]. \tag{9}$$

- $E[H_3]$: State S_3 contains a sleep period t^*_{S1} from state S_1. Suppose that there are N_{d1} DRX cycles in a t^*_{S1} period. Due to the memoryless property of the exponential t_{ipc} distribution, N_{d1} has geometric distribution with mean $1/p_{d1}$. p_{d1} is the probability that packets arrive during a DRX cycle, and is derived as follows

$$p_{d1} = \Pr[t_{ipc} \le t_D] = 1 - e^{-\frac{\lambda_{ipc}}{\lambda_D}}. \tag{10}$$

Since N_{d1} is a stopping time, from (10) and Wald's theorem, we have

$$E[H_3] = E\left[\sum_{i=1}^{N_{d1}} t_D\right] = E[N_{d1}]t_D = \left(\frac{1}{1 - e^{-\frac{\lambda_{ipc}}{\lambda_D}}}\right)\left(\frac{1}{\lambda_D}\right). \tag{11}$$

- $E[H_4]$: State S_4 comprises a sleep period t^*_{S2} from state S_2. Assume that t^*_{S2} consists of N_{d2} DRX cycles. Likewise, N_{d2} is a geometric random variable with mean $1/p_{d2}$, where

$$p_{d2} = \Pr[t_{is} \le t_D] = 1 - e^{-\frac{\lambda_{is}}{\lambda_D}}.$$

Thus, we obtain

$$E[H_4] = E\left[\sum_{i=1}^{N_{d2}} t_D\right] = E[N_{d2}]t_D = \left(\frac{1}{1 - e^{-\frac{\lambda_{is}}{\lambda_D}}}\right)\left(\frac{1}{\lambda_D}\right). \tag{12}$$

Based on the semi-Markov process, we derive the power saving factor P_s and the mean packet waiting time $E[t_w]$ in the following two subsections.

4.1 Power Saving Factor P_s

The power saving factor P_s is equal to the probability that the semi-Markov process is at S_3 or S_4 (i.e., the sleep period or power saving mode) in the steady state. We note that at the end of every

DRX cycle, the MS must wake up for a short period τ so that it can listen to the paging information from the network. Therefore, the "power saving" period in a DRX cycle is $t_D - \tau$. Let $E[H'_3]$ and $E[H'_4]$ be the mean "effective" sleep periods in states S_3 and S_4, respectively. Then, $E[H'_3]$ and $E[H'_4]$ can be obtained by replacing the t_D in (11) and (12) with $t_D - \tau$, and we have

$$E[H'_3] = \left(\frac{1}{1 - e^{-\frac{\lambda_{ipc}}{\lambda_D}}}\right)\left(\frac{1}{\lambda_D} - \tau\right) \tag{13}$$

and

$$E[H'_4] = \left(\frac{1}{1 - e^{-\frac{\lambda_{is}}{\lambda_D}}}\right)\left(\frac{1}{\lambda_D} - \tau\right). \tag{14}$$

From [17, Th. 4.8.3],

$$\begin{aligned} P_s &= \lim_{t\to\infty} \Pr[\text{the MS receiver is turned off at time } t] \\ &= \frac{\pi_3^e E[H'_3] + \pi_4^e E[H'_4]}{\sum_{i=1}^{4} \pi_i^e E[H_i]}. \end{aligned} \tag{15}$$

Substituting (2), (6), (9) and (11)-(14) into (15), we derive the close-form equation for the power saving factor P_s.

4.2 Mean Packet Waiting Time $E[t_w]$

In order to derive the mean packet waiting time $E[t_w]$, we first need to compute the expected total number of packets $E[N_t]$ that are processed in states S_1 and S_2, and the expected total waiting time $E[W_t]$ of all these packet arrivals. (Note that no packets are processed in states S_3 and S_4.) Then, $E[t_w]$ can be expressed as

$$E[t_w] = \frac{E[W_t]}{E[N_t]}. \tag{16}$$

Let $E[N_{i,j}]$ ($i \in \{1,2\}$ and $j \in \{1,2,3,4\}$) be the mean number of packets delivered to the MS in state S_i, given that the previous state transition is from state S_j. Denote $E[W_{i,j}]$ as the mean total waiting time of the associated $N_{i,j}$ packet arrivals. Using expectation by conditioning technique [16], we have that

$$E[N_t] = \sum_{i=1}^{2}\sum_{j=1}^{4} \pi_j^e p_{j,i} E[N_{i,j}] \text{ and } E[W_t] = \sum_{i=1}^{2}\sum_{j=1}^{4} \pi_j^e p_{j,i} E[W_{i,j}]. \tag{17}$$

We proceed to derive $E[N_{i,j}]$ and $E[W_{i,j}]$.

- $E[N_{1,1}]$ and $E[W_{1,1}]$: Given the previous state S_1, the transmitted $N_{1,1}$ packets in state S_1 correspond to a packet call and have geometric N_p distribution with mean μ_p and variance $\mu_p(\mu_p - 1)$. Therefore,

$$E[N_{1,1}] = \mu_p. \tag{18}$$

Since these $N_{1,1}$ packet transmissions constitute the busy period t_B^* in state S_1, we have the mean total waiting time

$$\begin{aligned} E[W_{1,1}] &= E\left[\sum_{i=1}^{N_p-1}(it_x - it_{ip})\right] \\ &= E\left[\frac{N_p(N_p-1)}{2}\right]E[t_x - t_{ip}] \\ &= \left[\frac{\mu_p^2 + \mu_p(\mu_p - 1) - \mu_p}{2}\right]E[t_x - t_{ip}] \\ &= \mu_p(\mu_p - 1)\left(\frac{1}{\lambda_x} - \frac{1}{\lambda_{ip}}\right). \end{aligned} \tag{19}$$

- $E[N_{1,2}]$ and $E[W_{1,2}]$: If the previous state is S_2, the number of transmitted packets $N_{1,2}$ in state S_1 also has geometric N_p distribution. Thus,

$$E[N_{1,2}] = E[N_{1,1}] = \mu_p \tag{20}$$

and

$$E[W_{1,2}] = E[W_{1,1}] = \mu_p(\mu_p - 1)\left(\frac{1}{\lambda_x} - \frac{1}{\lambda_{ip}}\right). \tag{21}$$

- $E[N_{1,3}]$ and $E[W_{1,3}]$: To derive $E[N_{1,3}]$ and $E[W_{1,3}]$, besides those $N''_{1,3}$ packets that arrive during state S_1, we also need to consider the $N'_{1,3}$ packets that are accumulated during the sleep period of the previous state S_3. Suppose that the first of the $N'_{1,3}$ packets arrives at time t_Λ within the DRX cycle t_D. Due to the memoryless property, t_Λ has the truncated exponential t_{ipc} distribution with the following density function

$$f_\Lambda(t) = \left(\frac{1}{1 - e^{-\frac{\lambda_{ipc}}{\lambda_D}}}\right)\lambda_{ipc}e^{-\lambda_{ipc}t}, \tag{22}$$

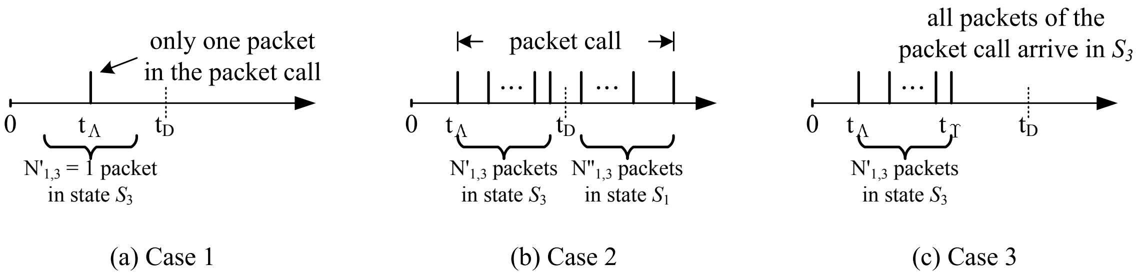

Figure 6: Three cases for the $N_{1,3}$ packet arrivals

where $0 \leq t \leq \frac{1}{\lambda_D}$. Since the inter-packet call idle time t_{ipc} (several hundred seconds; see the suggested value in [6]) is significantly longer than the suitable t_D periods (will be elaborated on in Numerical Examples section), we also assume that at most one packet call could appear in a t_D period. Under these assumptions, three cases for the $N_{1,3}$ packet arrivals are possible (see Figure 6): in case 1, the packet call contains only one packet; in case 2, there are $N'_{1,3}$ and $N''_{1,3}$ packets in state S_3 and state S_1, respectively; in case 3, all packets of the packet call arrive in state S_3. For given t_Λ, let θ_i ($1 \leq i \leq 3$) be the probability that case i occurs. It is clear that

$$\theta_1 = \frac{1}{\mu_p}. \tag{23}$$

Denote t_H as the interval between the first packet arrival and the last packet arrival of the packet call. Then t_H has the Erlang-N_p distribution with rate λ_{ip}, and θ_2 can be derived as follows

$$\begin{aligned}
\theta_2 &= \left(1 - \frac{1}{\mu_p}\right) \Pr[t_H > t_D - t_\Lambda] \\
&= \left(1 - \frac{1}{\mu_p}\right) \sum_{n=1}^{\infty} \Pr[N_p = n] \Pr[t_{H|N_p=n} > t_D - t_\Lambda] \\
&= \left(1 - \frac{1}{\mu_p}\right) \sum_{n=1}^{\infty} \frac{1}{\mu_p} \left(1 - \frac{1}{\mu_p}\right)^{n-1} \left\{ e^{-\lambda_{ip}(\frac{1}{\lambda_D} - t_\Lambda)} \sum_{k=0}^{n-1} \frac{[\lambda_{ip}(\frac{1}{\lambda_D} - t_\Lambda)]^k}{k!} \right\} \\
&= \left(1 - \frac{1}{\mu_p}\right) \frac{1}{\mu_p} e^{-\lambda_{ip}(\frac{1}{\lambda_D} - t_\Lambda)} \sum_{k=0}^{\infty} \frac{[\lambda_{ip}(\frac{1}{\lambda_D} - t_\Lambda)]^k}{k!} \sum_{n=k}^{\infty} \left(1 - \frac{1}{\mu_p}\right)^n \\
&= \left(1 - \frac{1}{\mu_p}\right) e^{-\lambda_{ip}(\frac{1}{\lambda_D} - t_\Lambda)} e^{\lambda_{ip}(\frac{1}{\lambda_D} - t_\Lambda)(1 - \frac{1}{\mu_p})} \\
&= \left(1 - \frac{1}{\mu_p}\right) e^{-\frac{\lambda_{ip}}{\mu_p}(\frac{1}{\lambda_D} - t_\Lambda)}.
\end{aligned} \tag{24}$$

From (23) and (24), we have

$$\theta_3 = 1 - \theta_1 - \theta_2 = \left(1 - \frac{1}{\mu_p}\right)\left[1 - e^{-\frac{\lambda_{ip}}{\mu_p}\left(\frac{1}{\lambda_D} - t_\Lambda\right)}\right]. \quad (25)$$

Next, we compute $E[N'_{1,3|i}]$ and $E[N''_{1,3|i}]$, and the associated mean total waiting time $E[W'_{1,3|i}]$ and $E[W''_{1,3|i}]$ for each of the three cases in Figure 6, where $1 \leq i \leq 3$. In case 1, the only packet of the packet call arrives at time t_Λ of the t_D period, and its mean waiting time is $1/\lambda_D - t_\Lambda$ in state S_3. Therefore,

$$E[N'_{1,3|1}] = 1,\ E[N''_{1,3|1}] = 0,\ E[W'_{1,3|1}] = \frac{1}{\lambda_D} - t_\Lambda, \text{ and } E[W''_{1,3|1}] = 0. \quad (26)$$

In case 2, according to the decomposition property of Poisson processes [17, Prop. 2.3.2], $N'_{1,3|2}$ is a shifted Poisson random variable (i.e., including the packet at t_Λ) with mean and the second moment

$$\begin{aligned}
E[N'_{1,3|2}] &= 1 + \lambda_{ip}\left(1 - \frac{1}{\mu_p}\right)\left(\frac{1}{\lambda_D} - t_\Lambda\right), \\
E[N'^2_{1,3|2}] &= Var[N'_{1,3|2}] + (E[N'_{1,3|2}])^2 \\
&= \lambda_{ip}\left(1 - \frac{1}{\mu_p}\right)\left(\frac{1}{\lambda_D} - t_\Lambda\right) + (E[N'_{1,3|2}])^2. \quad (27)
\end{aligned}$$

The mean total waiting time of these $N'_{1,3|2}$ packets is

$$E[W'_{1,3|2}] = \left(\frac{1}{\lambda_D} - t_\Lambda\right) + (E[N'_{1,3|2}] - 1)\frac{1}{2}\left(\frac{1}{\lambda_D} - t_\Lambda\right) + E\left[\sum_{k=1}^{N'_{1,3|2}} (k-1)\frac{1}{\lambda_x}\right]. \quad (28)$$

In (28), the first term represents the waiting time in state S_3 of the first packet arriving at t_Λ; the second term reflects the mean total waiting time in state S_3 of the other $N'_{1,3|2} - 1$ packets (note that these packet arrivals are uniformly distributed on the interval (t_Λ, t_D) [17, Th. 2.3.1], and the expected waiting time in state S_3 is thus $\frac{1}{2}(\frac{1}{\lambda_D} - t_\Lambda)$ for each of these packets); the third term corresponds to the mean total waiting time in state S_1 of the $N'_{1,3|2}$ packets. Substituting (27) into (28), we derive the conditional expectation $E[W'_{1,3|2}]$ for given t_Λ (see Appendix B for the details). Now, consider the mean total waiting time $E[W''_{1,3|2}]$ of the $N''_{1,3|2}$ packets that arrive in state S_1. Clearly, $N''_{1,3|2}$ is a geometric random variable with mean and the second moment

$$E[N''_{1,3|2}] = \mu_p,\ E[N''^2_{1,3|2}] = 2\mu_p^2 - \mu_p. \quad (29)$$

Then

$$E[W''_{1,3|2}] = E\left[\sum_{k=1}^{N''_{1,3|2}} \left\{\frac{N'_{1,3|2}}{\lambda_x} + \frac{k-1}{\lambda_x} - \frac{k}{\lambda_{ip}}\right\}\right]. \tag{30}$$

In (30), the first term represents the mean total service time of the $N'_{1,3|2}$ packet arrivals in state S_3; the second term corresponds to the mean total service time of the first $k-1$ packet arrivals in state S_1. Substituting (27) and (29) into (30), we derive $E[W''_{1,3|2}]$ for given t_Λ (see Appendix C for the details).

In case 3, all packets of the packet call arrive in state S_3, and thus

$$E[N''_{1,3|3}] = 0,\ E[W''_{1,3|3}] = 0. \tag{31}$$

Assume that the last packet of the packet call arrives at t_Υ within the t_D period (see Figure 6). Given t_Λ, from the decomposition property of Poisson processes, $t_H = t_\Upsilon - t_\Lambda$ is an exponential random variable with rate $\frac{\lambda_{ip}}{\mu_p}$ (see also the derivation in (24)). Therefore, t_Υ has the following conditional probability density function

$$f_{\Upsilon|\Lambda}(r|t) = \left[\frac{1}{1 - e^{-\frac{\lambda_{ip}}{\mu_p}\left(\frac{1}{\lambda_D} - t\right)}}\right]\left(\frac{\lambda_{ip}}{\mu_p}\right) e^{-\frac{\lambda_{ip}}{\mu_p}(r-t)}, \tag{32}$$

where $t \leq r \leq 1/\lambda_D$. The conditional mean $E[t_\Upsilon|t_\Lambda]$ and the conditional second moment $E[t_\Upsilon^2|t_\Lambda]$ of t_Υ can then be derived from (32), and are provided in Appendix D. We proceed to derive $E[N'_{1,3|3}]$ and $E[W'_{1,3|3}]$. Given t_Λ, similar to $N'_{1,3|2}$, $N'_{1,3|3}$ is a shifted Poisson random variable (i.e., including the packets at t_Λ and t_Υ) with mean and the second moment

$$\begin{aligned}
E[N'_{1,3|3}] &= 2 + \lambda_{ip}\left(1 - \frac{1}{\mu_p}\right)\left(E[t_\Upsilon|t_\Lambda] - t_\Lambda\right), \\
E[N'^2_{1,3|3}] &= Var[N'_{1,3|3}] + (E[N'_{1,3|3}])^2 \\
&= \lambda_{ip}\left(1 - \frac{1}{\mu_p}\right)\left(E[t_\Upsilon|t_\Lambda] - t_\Lambda\right) + (E[N'_{1,3|3}])^2.
\end{aligned} \tag{33}$$

Similar to the derivation in (28), the mean total waiting time of these $N'_{1,3|3}$ packets is expressed as

$$E[W'_{1,3|3}] = E[N'_{1,3|3}]\left[\frac{1}{\lambda_D} - \frac{1}{2}\left(E[t_\Upsilon|t_\Lambda] + t_\Lambda\right)\right] + E\left[\sum_{k=1}^{N'_{1,3|3}} (k-1)\frac{1}{\lambda_x}\right]. \tag{34}$$

Combining (22)-(31), (33) and (34), $E[N_{1,3}]$ and $E[W_{1,3}]$ are therefore

$$\begin{aligned} E[N_{1,3}] &= \int_{t=0}^{\frac{1}{\lambda_D}} \left\{ \sum_{i=1}^{3} \theta_i (E[N'_{1,3|i}] + E[N''_{1,3|i}]) \right\} f_\Lambda(t) dt, \\ E[W_{1,3}] &= \int_{t=0}^{\frac{1}{\lambda_D}} \left\{ \sum_{i=1}^{3} \theta_i (E[W'_{1,3|i}] + E[W''_{1,3|i}]) \right\} f_\Lambda(t) dt. \end{aligned} \quad (35)$$

Note that since the $N_{1,3}$ packets constitute a packet call, $E[N_{1,3}]$ in (35) is further simplified to have

$$E[N_{1,3}] = \mu_p.$$

The mean total waiting time $E[W_{1,3}]$ in (35) has no close-form solution, and can be easily computed by using numerical computing software, e.g., MATLAB [19].

- $E[N_{1,4}]$ and $E[W_{1,4}]$: The derivations of $E[N_{1,4}]$ and $E[W_{1,4}]$ are identical to that of $E[N_{1,3}]$ and $E[W_{1,3}]$ except that λ_{ipc} should be replaced by λ_{is}. This is because in state S_4, the MS would experience an inter-session idle time t_{is} with mean $1/\lambda_{is}$ rather than an inter-packet call idle time t_{ipc} with mean $1/\lambda_{ipc}$.

- $E[N_{2,j}]$ and $E[W_{2,j}]$ ($j \in \{1, 2, 3, 4\}$): Note that the only difference between state S_1 and state S_2 is that the busy period in state S_1 is followed by an inter-packet call inactivity period while the busy period in state S_2 is followed by an inter-session inactivity period. Therefore, the packets processed in state S_2 have the same statistical properties as those processed in state S_1, and we have

$$E[N_{2,j}] = E[N_{1,j}] \text{ and } E[W_{2,j}] = E[W_{1,j}],$$

where $j \in \{1, 2, 3, 4\}$.

Substitute the above $E[N_{i,j}]$ and $E[W_{i,j}]$ values ($i \in \{1, 2\}$ and $j \in \{1, 2, 3, 4\}$) into (17) to yield $E[N_t]$ and $E[W_t]$. Finally, substituting the obtained $E[N_t]$ and $E[W_t]$ into (16), we derive the mean packet waiting time $E[t_w]$.

Table 1: The comparison between the analytic and simulation results ($\lambda_{ip} = 5\lambda_x$, $\lambda_{ipc} = \lambda_x/800$, $\lambda_{is} = \lambda_x/16000$, $t_D = 20E[t_x]$, $\tau = E[t_x]$, $\mu_{pc} = 5$, and $\mu_p = 25$)

$t_I/E[t_x]$	200	280	360	440	520	600
$P_s(Analytic)$	0.899396	0.883885	0.869498	0.85613	0.843685	0.832079
$P_s(Simulation)$	0.899394	0.88391	0.869425	0.856189	0.843596	0.832103
Error rate	0.0002%	0.0028%	0.0084%	0.0069%	0.0105%	0.0029%
$t_I/E[t_x]$	200	280	360	440	520	600
$E[t_w]/E[t_x](Analytic)$	27.4320	26.8268	26.2782	25.7810	25.3302	24.9214
$E[t_w]/E[t_x](Simulation)$	27.4104	26.8274	26.2474	25.7858	25.3262	24.9462
Error rate	0.0787%	0.0022%	0.1172%	0.0186%	0.0158%	0.0995%

5 Numerical Results

The analytic model has been validated against simulation experiments. These simulation experiments are based on a discrete-event simulation model (including Packet arrival, Packet departure, Sleep, Reading, and Wakeup events) which simulates the MS power saving behaviors according to the UMTS DRX mechanism. The interested reader is referred to [22] for the details of the simulation model. Table 1 compares the analytic and simulation results, where $\lambda_{ip} = 5\lambda_x$, $\lambda_{ipc} = \lambda_x/800$, $\lambda_{is} = \lambda_x/16000$, $t_D = 20E[t_x]$, $\tau = E[t_x]$, $\mu_{pc} = 5$, and $\mu_p = 25$. The table indicates that for the power saving factor P_s, the discrepancies between analytic analysis and simulation are less than 0.01% in most cases; for the mean packet waiting time $E[t_w]$, the discrepancies are less than 0.1% in most cases. It is clear that the analytic analysis is consistent with the simulation results. Based on the analytic model, we investigate the DRX performance. Specifically, we consider the bursty packet data traffic. Figures 7-10 plot the P_s and $E[t_w]$ curves. In these figures, t_x has the cut-off Pareto distribution with shape parameter $\beta = 1.1$ and mean $E[t_x] = 0.5$ seconds, and t_D and t_I are of fixed values. The parameter settings are described in the captions of the figures.

Effects of t_{ipc}. Figure 7(a) indicates that the power saving factor P_s curves decrease and then increase as the inter-packet call idle time t_{ipc} increases. This phenomenon is explained as follows. For $t_{ipc} < 4000E[t_x]$, when t_{ipc} approaches zero, the packet arrivals in a packet ser-

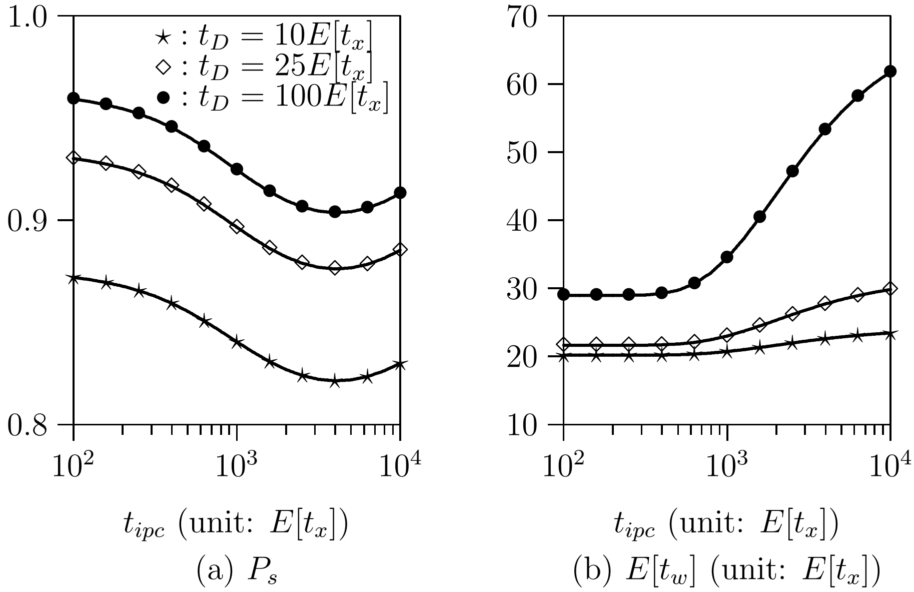

Figure 7: Effects of t_{ipc} ($\lambda_{ip} = 5\lambda_x$, $\lambda_{is} = \lambda_x/80000$, $t_I = 2000E[t_x]$, $\tau = E[t_x]$, $\mu_{pc} = 5$, and $\mu_p = 25$)

vice session degenerate into a single packet train. After a burst of these packet transmissions, the MS will immediately experience a long inter-session idle time t_{is}, and will eventually be switched into the sleep mode to reduce the power consumption. Therefore, we have high power saving factor P_s in this case. As t_{ipc} increases, P_s is affected by the operation of the RNC inactivity timer. Specifically, the MS is more likely to be found in the inactivity period when the next packet call arrives. Consequently, P_s decreases as t_{ipc} increases. On the other hand, if $t_{ipc} > 4000E[t_x]$, more of the inter-packet call idle times will be longer than the RNC inactivity timer threshold t_I as t_{ipc} increases. As a result, the MS is more likely to be in the power saving mode when subsequent packet calls arrive. Therefore, P_s increases as t_{ipc} increases. Figure 7(b) illustrates that the mean packet waiting time $E[t_w]$ is an increasing function of t_{ipc}. When t_{ipc} is sufficiently small (e.g., $t_{ipc} < 300E[t_x]$ in this experiment), the value of $E[t_w]$ is mainly dominated by the inter-session idle time t_{is}. Therefore, decreasing t_{ipc} will insignificantly affect the $E[t_w]$ performance. We also note that $E[t_w]$ is more sensitive to t_{ipc} for a large t_D than a small t_D.

Effects of t_{is}. Figure 8(a) shows the intuitive result that P_s is an increasing function of the inter-

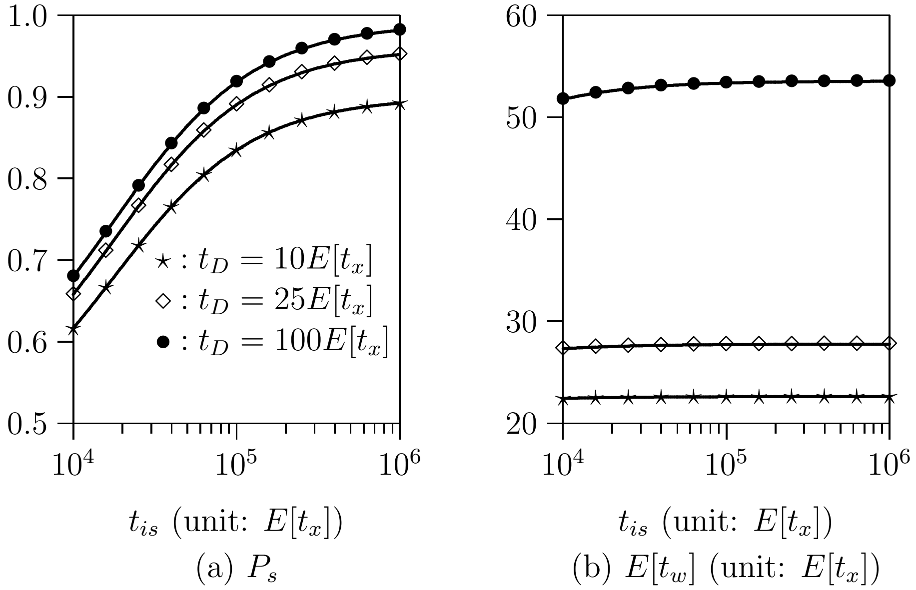

Figure 8: Effects of t_{is} ($\lambda_{ip} = 5\lambda_x$, $\lambda_{ipc} = \lambda_x/4000$, $t_I = 2000E[t_x]$, $\tau = E[t_x]$, $\mu_{pc} = 5$, and $\mu_p = 25$)

session idle time t_{is}. We observe that increasing t_{is} will not significantly improve the $E[t_w]$ performance in Figure 8(b). In our traffic model, if the current packet call is not served completely, the next packet call will not be generated. Therefore, no more than one packet calls will wait in the MS sleep period, and increasing t_{is} will not change the $E[t_w]$ value.

Effects of t_I. Figure 9 indicates that by increasing the RNC inactivity timer threshold t_I, P_s and $E[t_w]$ decrease. When t_I is small (e.g., $t_I < 200E[t_x]$ in this experiment), it is likely that the MS is found in the power saving mode as the next packet call arrives. Consequently, we observe high power saving factor P_s. However, the mean packet waiting time $E[t_w]$ is unacceptably high in this case. As $t_I \to \infty$, it is more likely that the MS will never enter the sleep mode, and the $E[t_w]$ decreases. Due to the characteristic of the packet burstiness within the packet call, the server state is likely to be busy when the next packets arrive. As a result, $E[t_w]$ is bounded at $20E[t_x]$ in this example, and increasing t_I will not enhance $E[t_w]$. We also note that $E[t_w]$ is more sensitive to t_I for a large t_D than a small t_D.

Effects of t_D. Figure 10 shows that P_s and $E[t_w]$ are increasing functions of the DRX cycle t_D. We observe that when t_D is large (e.g., $t_D > 100E[t_x]$ in Figure 10(a)), increasing t_D will

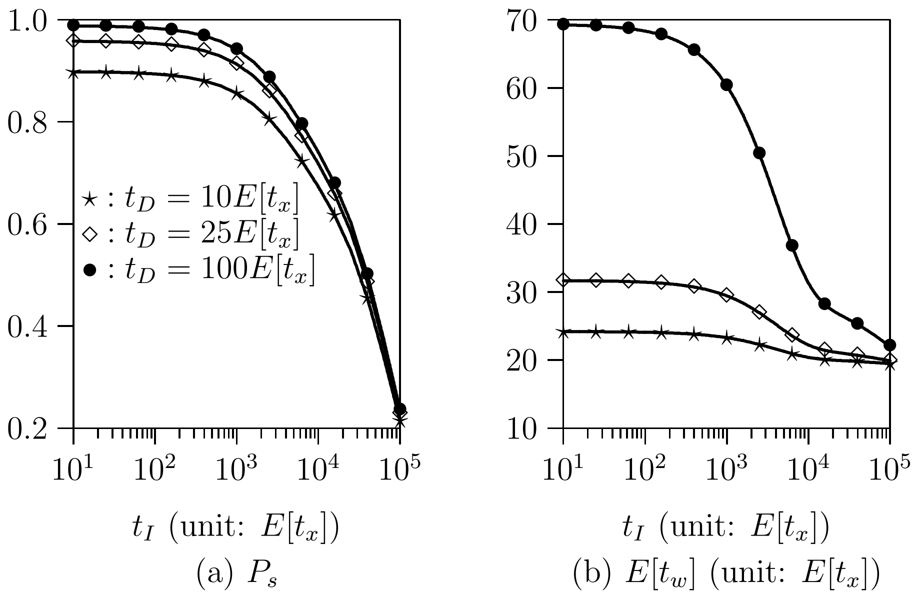

Figure 9: Effects of t_I ($\lambda_{ip} = 5\lambda_x$, $\lambda_{ipc} = \lambda_x/4000$, $\lambda_{is} = \lambda_x/80000$, $\tau = E[t_x]$, $\mu_{pc} = 5$, and $\mu_p = 25$)

not improve the P_s performance. On the other hand, when t_D is small (e.g., $t_D < 10E[t_x]$ in Figure 10(b)), decreasing t_D insignificantly improves the $E[t_w]$ performance. Therefore, for $t_I = 2000E[t_x]$, t_D should be selected in the range $[10E[t_x], 100E[t_x]]$.

Effects of τ. Figure 10(a) illustrates the impacts of the wakeup cost τ on P_s. When t_D is large (e.g., $t_D > 100E[t_x]$), τ is a small portion of a DRX cycle, and thus only has insignificant impact on P_s. When t_D is small (e.g., $t_D < 20E[t_x]$), P_s increases as τ decreases. Figure 10(b) demonstrates an intuitive result that $E[t_w]$ is not affected by τ.

6 Conclusion

The UMTS utilizes the DRX mechanism to reduce the power consumption of MSs. DRX permits an idle MS to power off the radio receiver for a predefined sleep period, and then wake up to receive the next paging message. The sleep/wake-up scheduling of each MS is determined by two DRX parameters: the inactivity timer threshold t_I and the DRX cycle t_D. Analytic and simulation models have been developed in the literature to study the DRX performance mainly for Poisson

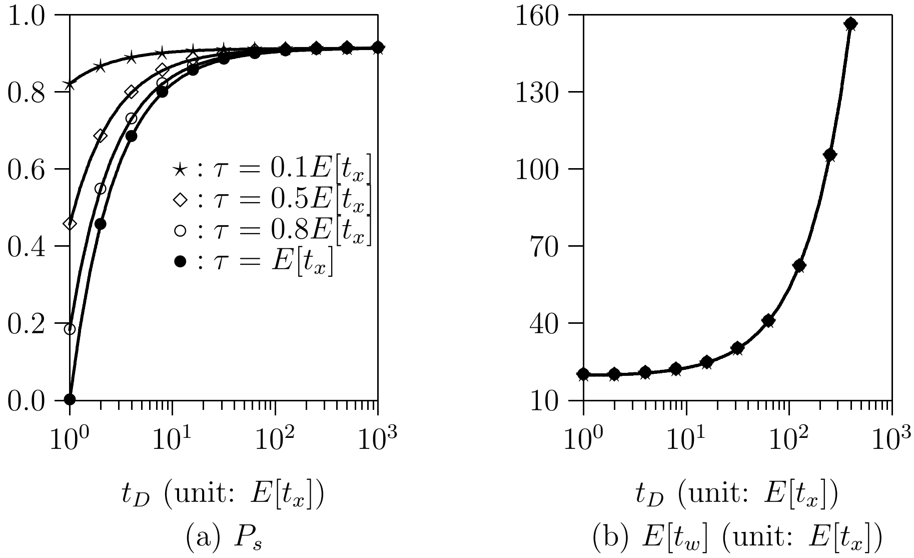

Figure 10: Effects of t_D ($\lambda_{ip} = 5\lambda_x$, $\lambda_{ipc} = \lambda_x/4000$, $\lambda_{is} = \lambda_x/80000$, $t_I = 2000E[t_x]$, $\mu_{pc} = 5$, and $\mu_p = 25$)

traffic. In this paper, we proposed a novel semi-Markov process to model the UMTS DRX with bursty packet data traffic. The analytic results were validated against simulation experiments. We investigated the effects of the two DRX parameters on output measures including the power saving factor P_s and the mean packet waiting time $E[t_w]$. Our study indicated the following.

- The power saving factor P_s curves decrease and then increase as the inter-packet call idle time t_{ipc} increases.
- The mean packet waiting time $E[t_w]$ is an increasing function of t_{ipc}.
- When t_{ipc} is sufficiently small, decreasing t_{ipc} will insignificantly affect the $E[t_w]$ performance.
- $E[t_w]$ is more sensitive to t_{ipc} for a large t_D than a small t_D.
- P_s is an increasing function of the inter-session idle time t_{is}.
- By increasing the RNC inactivity timer threshold t_I, P_s and $E[t_w]$ decrease.
- $E[t_w]$ is more sensitive to t_I for a large t_D than a small t_D.

- For the parameter settings considered in this paper, t_D should be selected in the range $[10E[t_x], 100E[t_x]]$ for better P_s and $E[t_w]$ performance.

- When t_D is large, τ is a small portion of a DRX cycle, and only has insignificant impact on P_s.

Acknowledgment

We would like to thank the anonymous reviewers. Their valuable comments have significantly enhanced the quality of this paper.

A Notation List

- $E[H'_3]$: the mean "effective" sleep period in state S_3
- $E[H'_4]$: the mean "effective" sleep period in state S_4
- $E[N_{i,j}]$: the mean number of packets delivered to the MS in state S_i, given that the previous state transition is from state S_j
- $E[N_t]$: the expected total number of packets that are processed in states S_1 and S_2
- $E[W_{i,j}]$: the mean total waiting time of the associated $N_{i,j}$ packet arrivals
- $E[W_t]$: the expected total waiting time of the N_t packet arrivals in states S_1 and S_2
- $1/\lambda_D$: the length of each DRX cycle t_D in a sleep period
- $1/\lambda_I$: the length of the RNC inactivity timer threshold t_I
- $1/\lambda_{ip}$: the expected value for the t_{ip} distribution
- $1/\lambda_{ipc}$: the expected value for the t_{ipc} distribution
- $1/\lambda_{is}$: the expected value for the t_{is} distribution

- $1/\lambda_x$: the expected value for the t_x distribution
- μ_p: the expected value for the N_p distribution
- μ_{pc}: the expected value for the N_{pc} distribution
- N_p: the number of packets within a packet call
- N_{pc}: the number of packet calls within a packet service session
- P_s: the power saving factor
- t_D: the DRX cycles in a sleep period
- t_{ip}: the inter-packet arrival time within a packet call
- t_{ipc}: the time interval between the end of a packet call transmission and the beginning of the next packet call transmission (i.e., the inter-packet call idle time)
- t_{is}: the time interval between the end of a packet session transmission and the beginning of the next packet session transmission (i.e., the inter-session idle time)
- t_I: the threshold of the RNC inactivity timer
- t_w: the packet waiting time in the RNC buffer
- t_x: the time interval between when the packet is transmitted by the RNC processor and when the corresponding positive acknowledgment is received by the RNC processor

B $E[W'_{1,3|2}]$ for Given t_Λ

$$
\begin{aligned}
E[W'_{1,3|2}] &= \left(\frac{1}{\lambda_D} - t_\Lambda\right) + (E[N'_{1,3|2}] - 1)\frac{1}{2}\left(\frac{1}{\lambda_D} - t_\Lambda\right) + E\left[\sum_{k=1}^{N'_{1,3|2}} (k-1)\frac{1}{\lambda_x}\right] \\
&= \left(\frac{1}{\lambda_D} - t_\Lambda\right) + (E[N'_{1,3|2}] - 1)\frac{1}{2}\left(\frac{1}{\lambda_D} - t_\Lambda\right) + \left(\frac{1}{\lambda_x}\right) E\left[\frac{N'_{1,3|2}(N'_{1,3|2} - 1)}{2}\right]
\end{aligned}
$$

$$= \left(\frac{1}{\lambda_D} - t_\Lambda\right) + (E[N'_{1,3|2}] - 1)\frac{1}{2}\left(\frac{1}{\lambda_D} - t_\Lambda\right) + \left(\frac{1}{2\lambda_x}\right)(E[N'^2_{1,3|2}] - E[N'_{1,3|2}]),$$

where $E[N'_{1,3|2}]$ and $E[N'^2_{1,3|2}]$ are given in (27).

C $E[W''_{1,3|2}]$ for Given t_Λ

$$\begin{aligned}
E[W''_{1,3|2}] &= E\left[\sum_{k=1}^{N''_{1,3|2}} \left\{\frac{N'_{1,3|2}}{\lambda_x} + \frac{k-1}{\lambda_x} - \frac{k}{\lambda_{ip}}\right\}\right] \\
&= \left(\frac{1}{\lambda_x}\right) E[N''_{1,3|2}]E[N'_{1,3|2}] - \left(\frac{1}{\lambda_{ip}}\right) E[N''_{1,3|2}] + \left(\frac{1}{\lambda_x} - \frac{1}{\lambda_{ip}}\right) E\left[\frac{N''_{1,3|2}(N''_{1,3|2} - 1)}{2}\right] \\
&= \left(\frac{1}{\lambda_x}\right) E[N''_{1,3|2}]E[N'_{1,3|2}] - \left(\frac{1}{\lambda_{ip}}\right) E[N''_{1,3|2}] + \frac{1}{2}\left(\frac{1}{\lambda_x} - \frac{1}{\lambda_{ip}}\right)(E[N''^2_{1,3|2}] - E[N''_{1,3|2}]),
\end{aligned}$$

where $E[N'_{1,3|2}]$, $E[N''_{1,3|2}]$ and $E[N''^2_{1,3|2}]$ are given in (27) and (29).

D $E[t_\Upsilon|t_\Lambda]$ and $E[t^2_\Upsilon|t_\Lambda]$

$$\begin{aligned}
E[t_\Upsilon|t_\Lambda] &= t_\Lambda + \left[\frac{1}{1 - e^{-\frac{\lambda_{ip}}{\mu_p}(\frac{1}{\lambda_D} - t_\Lambda)}}\right]\left[\frac{\mu_p}{\lambda_{ip}} - \left(\frac{1}{\lambda_D} - t_\Lambda + \frac{\mu_p}{\lambda_{ip}}\right) e^{-\frac{\lambda_{ip}}{\mu_p}(\frac{1}{\lambda_D} - t_\Lambda)}\right] \\
E[t^2_\Upsilon|t_\Lambda] &= t^2_\Lambda + \left[\frac{1}{1 - e^{-\frac{\lambda_{ip}}{\mu_p}(\frac{1}{\lambda_D} - t_\Lambda)}}\right]\left\{2t_\Lambda\left(\frac{\mu_p}{\lambda_{ip}}\right) + 2\left(\frac{\mu_p}{\lambda_{ip}}\right)^2 - \left[2t_\Lambda\left(\frac{1}{\lambda_D} - t_\Lambda + \frac{\mu_p}{\lambda_{ip}}\right)\right.\right. \\
&\quad \left.\left. + 2\left(\frac{\mu_p}{\lambda_{ip}}\right)^2 + 2\left(\frac{\mu_p}{\lambda_{ip}}\right)\left(\frac{1}{\lambda_D} - t_\Lambda\right) + \left(\frac{1}{\lambda_D} - t_\Lambda\right)^2\right] e^{-\frac{\lambda_{ip}}{\mu_p}(\frac{1}{\lambda_D} - t_\Lambda)}\right\}
\end{aligned}$$

References

[1] 3GPP. 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; RRC Protocol Specification for Release 1999. Technical Specification 3G TS 25.331 version 3.5.0 (2000-12), 2000.

[2] 3GPP. 3rd Generation Partnership Project; Technical Specification Group Services and Systems Aspects; General Packet Radio Service (GPRS); Service Description; Stage 2. Technical Specification 3G TS 23.060 version 3.6.0 (2001-01), 2000.

[3] 3GPP. 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; UTRA High Speed Downlink Packet Access. Technical Specification 3G TR 25.950 version 4.0.0 (2001-03), 2001.

[4] 3GPP. 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; UE Procedures in Idle Mode and Procedures for Cell Reselection in Connected Mode. Technical Specification 3G TS 25.304 version 5.1.0 (2002-06), 2002.

[5] CDPD Forum. Cellular Digital Packet Data System Specification: Release 1.1. Technical report, CDPD Forum, Inc., January 1995.

[6] ETSI. Universal Mobile Telecommunications System (UMTS); Selection Procedures for the Choice of Radio Transmission Technologies of the UMTS. Technical Report UMTS 30.03 version 3.2.0, April 1998.

[7] Holma, H. and Toskala, A. *WCDMA for UMTS*. John Wiley & Sons, Inc., 2000.

[8] IEEE. Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications. Draft Standard 802.11 D3.1, April 1996.

[9] Johnson, N.L. *Continuous Univariate Distributions-1*. John Wiley& Sons, 1970.

[10] Kwon, S.J., Chung, Y.W., and Sung, D.K. Queueing Model of Sleep-Mode Operation in Cellular Digital Packet Data. *IEEE Transactions on Vehicular Technology*, 52(4):1158–1162, July 2003.

[11] Lee, C.-C., Yeh, J.-H., and Chen J.-C. Impact of Inactivity Timer on Energy Consumption in WCDMA and cdma2000. *Wireless Telecommunications Symposium*, May 2004.

[12] Lin, Y.-B. and Chuang, Y.-M. Modeling the Sleep Mode for Cellular Digital Packet Data. *IEEE Communications Letters*, 3(3):63–65, March 1999.

[13] Nelson, R. *Probability, Stochastic Processes, and Queueing Theory*. Springer-Verlag., 1995.

[14] Paxson, V. and Floyd, S. Wide Area Traffic: The Failure of Poisson Modeling. *IEEE/ACM Transactions on Networking*, 3(3):226–244, June 1995.

[15] RAM Mobile Data. Mobitex Interface Specification. Technical report, RAM Mobile Data, 1994.

[16] Ross, S.M. *Introduction to Probability Models, 5th ed.* Academic Press, 1993.

[17] Ross, S.M. *Stochastic Processes, 2nd ed.* John Wiley, New York, 1996.

[18] Sarvagya, M. and Raja Kumar, R.V. Performance Analysis of the UMTS System for Web Traffic over Dedicated Channels. *The 3rd International Conference on Information Technology: Research and Education (ITRE), Taiwan*, June 2005.

[19] The MathWorks. *MATLAB User's Guide*. The MathWorks, Inc., Natick, MA, 1993.

[20] Willinger, W., Taqqu, M.S., Sherman, R., and Wilson, D.V. Self-Similarity through High-Variability: Statistical Analysis of Ethernet LAN Traffic at the Source Level. *IEEE/ACM Trans. Networking*, 5(1):71–86, February 1997.

[21] Yang, S.-R. and Lin, Y.-B. Modeling UMTS Discontinuous Reception Mechanism. *IEEE Transactions on Wireless Communications*, 4(1):312–319, January 2005.

[22] Yang, S.-R. and Yan, S.-Y. The Supplement to "*Modeling UMTS Power Saving with Bursty Packet Data Traffic*" (http://www.cs.nthu.edu.tw/∼sryang/submission/MUPSwBPDT.pdf). Technical report, National Tsing Hua University, 2006.

[23] Yeh, J.-H., Lee, C.-C., and Chen J.-C. Performance Analysis of Energy Consumption in 3GPP Networks. *Wireless Telecommunications Symposium*, May 2004.