Analysis of power saving with continuous connectivity

Vincenzo Mancuso, Sara Alouf

Analysis of power saving with continuous connectivity

Vincenzo Mancuso[1] and Sara Alouf[2]
[1] Institute IMDEA Networks, Madrid, Spain
[2] INRIA, Sophia Antipolis, France

Abstract

Always-on mobile users need high bandwidth channels with negligible access delay and limited power consumption. Such a *continuous connectivity* mode requires the management of high-speed channels, which can turn into substantial operational costs (i.e., power consumption rate) even in presence of low traffic, unless a power saving mechanism is enforced. In this paper, we analyze the impact of 3GPP-defined power saving mechanisms on the performance of users with continuous connectivity. We develop a model for packet transmission and operational costs. We model each downlink mobile user's traffic by means of an $M/G/1$ queue, and the base station's downlink traffic as an $M/G/1\,PS$ queue with multiple classes and inhomogeneous vacations. The model is validated through packet-level simulations. Our results show that consistent power saving can be achieved in the wireless access network, as high as 75% for mobiles and 55% for base stations.

keyword — green IT, continuous connectivity, power saving, analysis

1 Introduction

Thanks to technologies like WiMAX, HSPA, and LTE, today's mobile users can have a network performance experience similar to that provided by short range wireless LANs and even by wired DSL lines. The cost of providing such a service has been reported to be quite high for the network operator, e.g., of the order of tens of millions of dollars for a medium-small network with twenty thousand base stations [15]. However, most of the transmission cost might be dramatically reduced by using efficient power saving strategies in hardware, software and radio resource management domains.

We consider the case of users generating large volumes of traffic. These users browse the web, exchange email, share data on social networks, and access audio and video streaming applications. To shorten the delay to access the network as soon as new packets have to be exchanged, users need the continuous availability of a dedicated wideband data channel. This *continuous connectivity* requires frequent exchange of control packets, even in absence of data to be exchanged. So, unless power saving is enforced, a large amount of energy is required to control the high-speed connection.

The observation of current trends in the evolution of cellular standards, e.g., the evolution of 3GPP specifications, reveals that power saving is targeted via *sleep mode* operation, which will be mandatory in continuous connectivity at both user equipment (UE) and base station (evolved node B, namely eNB). However, sleep mode affects packet delay, thereby some constraints have to be considered when switching to power saving operations.

The literature presents various analytical and experimental studies on sleep mode in cellular networks, in particular on UE performance figures. The power saving mechanism for the UMTS UE has been evaluated in [18]. The performance of IEEE 802.16e power saving has been analytically evaluated in [10], where the authors use a semi-Markov chain approach. Other authors have used queueing theory to analyze the power saving. For instance, Seo *et al.* proposed an embedded Markov chain to model the system vacations in IEEE 802.16e, where the base station queue is seen as an $M/GI/1/N$ system [16]. An $M/G/1$ queue with repeated vacations has been proposed to model an 802.16e-like sleep mode and to compute the service cost for a single user download [4]. In a companion paper, we have analyzed the impact, on web traffic, of power saving mechanisms in continuous connectivity using a $G/G/1\ PS$ queue system [14]. In this paper, we adapt and extend the methodology of [4] and [14] to analyze multiple queues with a shared processor, without the restriction to web traffic.

Xiao proposed an analytical model, supported by simulations, for evaluating the performance of the UE in terms of energy consumption and access delay in both downlink and uplink [17]. Almhana *et al.* provide an adaptive algorithm that minimizes energy subject to QoS requirements for delay [3].

The work available in the literature does not tackle the base station viewpoint nor analytically capture the relation between cell load and service rate offered to the users. Conversely, we use an $M/G/1$ model to evaluate the behavior of each UE, and we compose the behavior of multiple $M/G/1$ queues into a single $M/G/1$ PS that models the eNB behavior. Then, we are able to analytically compute the cost reduction achievable thanks to sleep mode operations, and maximize this cost reduction both at the UE and the eNB under QoS constraints. In particular we refer to the 3GPP mechanism for downlink power saving in *Continuous Packet Connectivity* (CPC), namely the discontinuous reception (DRX) [1].

The importance of DRX in LTE and UMTS has been previously recognized in [19], where the authors model the DRX operation via a semi-Markov model for bursty packet data traffic. DRX advantages have been presented from the user viewpoint in [6], which proposes a very simple cost model over a detailed transmission model. Last, in [13], the authors use heuristics and simulation to show the importance of DRX for the UE.

The contribution of this paper is threefold: (*i*) we are the first to provide a complete queueing model for the behavior of users (UEs) and base stations (eNBs) in continuous connectivity, (*ii*) we provide a cost model that incorporates the different causes of power consumption, and (*iii*) we show how to use the model to minimize the power consumption rate under QoS constraints. Our model has been validated through packet-level simulations, and optimization results confirm that a dramatic economy of energy can be attained by correctly tuning the power saving parameters. UE costs can be reduced by a 75%, while eNB cost be lowered by more than 50%.

The paper is organized as follows: Section 2 reviews the concept of continuous connectivity. In Section 3 we derive a model for UE transmission activity and its cost. Section 4 extends the model to eNB. We validate the model in Section 5, and use it to compute the cost-QoS tradeoff at UE and eNB. Section 6 summarizes and concludes the paper.

2 Continuous connectivity

Consider a scenario in which user transmission activity is scheduled by the base station. Thereby the UE cannot transmit data unless the eNB grants a transmission opportunity. When using continuous connectivity, the UE should check the control channel continuously, and use it (in both uplink and downlink) for synchronization, power control, and traffic announcements. For instance, CPC has been defined by 3GPP for the next generation of high-speed mobile users, in which users register to the data packet service of their wireless operator and then remain online even when they do not transmit nor receive data for long periods [8]. A highly efficient sleep mode operation is thus strongly required, to allow disabling both transmission and reception of frames during idle periods. The UE, however, still has to transmit and receive control frames at regular pace, so that synchronization to the base station and power control loop can be maintained. Therefore, idle periods are limited by the mandatory control activity that involves the UE. To save energy, when there is no traffic for the user, the UE can enter a sleep mode in which it checks and reports on the control channels according to a fixed pattern, namely, only once every m time units (e.g., it listens to the control channel only one subframe out of m). This way, the energy consumption reduction at the mobile equipment is relevant, especially in case the transmitter is completely shut down during sleep mode operations. In change, the UE can transmit/receive new data only every m subframes.

However, in order to keep synchronization, the UE is always requested to listen to control channels every few tens of milliseconds, at most. Hence, the continuous connectivity cost can be sensibly higher than the cost incurred in WiMAX networks for instance, where no control channels are defined, and decoding the resource allocation table at the beginning of the downlink frame is not mandatory.

In 3GPP, DRX characterizes the downlink transmission behavior with sleep mode operations enabled. DRX allows the UE to save energy while monitoring the control information transmitted by the eNB over the High Speed Shared Control Channel. DRX affects data delivery, since no data can be dependably received without an associated control frame. In particular, 3GPP specifications define a DRX *long-cycle*, that is the total number of subframes in a listening/sleeping window out of which only one subframe is used for control reception. Valid values for this long-cycle are 4, 5, 8, 10, 16, and 20 subframes (i.e., using

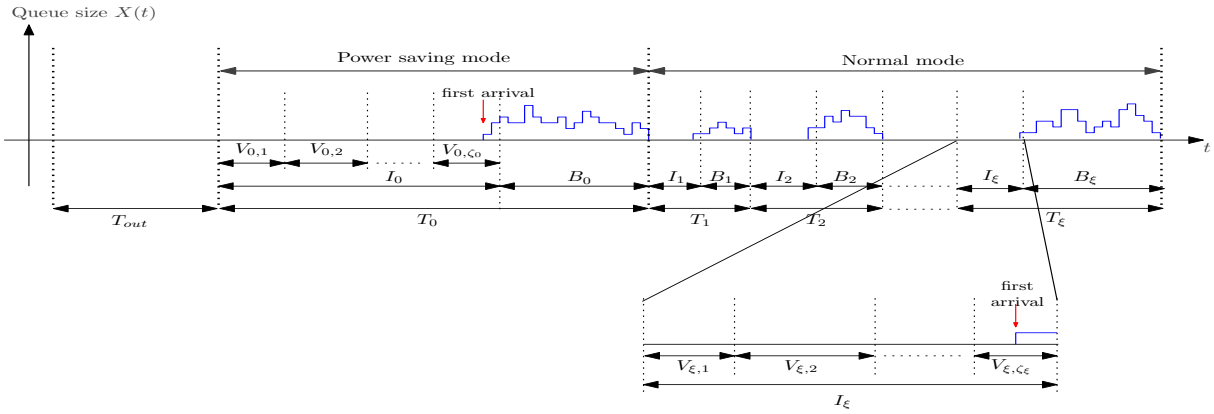


Figure 1: System cycle T_c.

a 2 ms subframe in HSPA yields cycles of 8, 10, 16, 20, 32, and 40 ms). Note that the DRX long-cycle is activated only upon a timeout after the last downlink transmission. The timeout threshold specified in the standard can be $M = 2^{a+1}$ subframes, $a \in \{0..8\}$.

3 Power saving at the UE

Power saving at the UE is composed of the saving done in downlink and that done in uplink. In downlink, the UE decodes the control channel following the DRX pattern, and receives packets accordingly [8]. Uplink control transmissions follow a scheme similar to DRX, namely DTX (Discontinuous transmission). However, the DTX behavior depends on the activity of multiple physical channels [1]. Therefore, for sake of clarity and simplicity, we only focus on the DRX in the downlink, and leave the analysis of DTX in uplink for the future work. In particular, our model can be used for the downlink of systems using slotted operations, and specifically for HSPA [8] and LTE/LTE-Advanced [8, 7]. We will show in Section 4 how to extend the model in order to compute the power saving for eNBs serving multiple users.

3.1 Framework

Our model describes the behavior of downlink sleep mode under DRX constraints. The power saving that can be obtained is expressed as a function of the subframe length T_{sub} and the DRX parameters, namely the timeout duration T_{out} and the DRX long-cycle duration mT_{sub}. The impact of sleep mode on the quality of service is evaluated in terms of the sojourn time experienced by downlink traffic, assuming that all the traffic is served.

In this framework, the eNB can deliver data to the UE according to the following pattern: the transmission time is slotted, each slot corresponding to a downlink *subframe*, and the eNB schedules data transmissions in any subframe unless an inactivity timer expires; after timeout expiration the eNB can transmit only every m subframes. The data to be delivered to the UE can result from the composition of many traffic patterns generated by multiple applications running on the same UE, e.g., streaming, web browsing, instant messaging, and so on. This allows us to model the per-user downlink buffer of the eNB as a queue of data packets arriving according to a Poisson process with rate λ. The Poisson assumption is also supported by the findings of [11] where the traffic crossing the Internet is shown to tend to behave approximately as Poisson over small to medium time scales (fraction of second to hours), and as a non-stationary Poisson over larger time intervals. Queued packets are served according to the scheduling discipline implemented at the eNB. In this paper, we assume the utilization of a GPS scheduler which closely approximates the operation of a weighted fair queueing scheduler, which is widely adopted in real devices. Therefore, if we assume that each downlink packet has a fixed size and fits in a subframe, then when a single user is in the system we obtain an $M/D/1$ queue with arrival rate λ, serving time T_{sub}, and timeout-triggered fixed-length repeated vacations lasting mT_{sub}, corresponding to DRX long-cycles.

Despite of the simplicity of the proposed optimization framework, in Section 5.1 we will show that our approach is robust to non-Poisson traffic behaviors.

The behavior of the downlink queue for each UE can be analyzed with the technique used in [4], namely, by exploiting the property of an $M/G/1$ queue during busy and idle periods separately, since results for $M/G/1$ queues apply to $M/D/1$ queues as well. In particular, busy and idle periods alternate in our model according to a bi-modal behavior, as follows: as long as an inactivity timer does not expire, a UE with continuous connectivity will be in a "normal mode", alternating idle and busy periods $\{I_k, B_k\}_{k\geq 1}$. When the inactivity timeout expires, the UE switches to a "power saving mode", and will stay in such a state for an interval I_0 depending on the next frame arrival at the download transmission queue, after which the UE becomes active for an interval B_0 (till the queue empties again). This bi-modal behavior is depicted in Figure 1 for generic duration of the vacation intervals. To be more precise, in the **normal mode**, there is a sub-cycle T_k repeated ξ times (with ξ being a random variable), consisting of:

1. An idle interval $I_k, 1 \leq k \leq \xi$, during which the UE monitors the queue at every subframe for ζ_k subframes while the queue is empty, and for no longer than a timeout T_{out} (i.e., in our case, this is equivalent to a queue server taking repeated vacations of fixed length $V_{k,i} = T_{sub}$, $k = 1, .., \xi$, $i = 1, .., \zeta_k$);

2. A busy period B_k during which all packets in the queue are served.

In the **power saving mode**, there is a single sub-cycle T_0 consisting of:

1. An idle interval I_0 in which the UE monitors the queue every m subframes until a packet arrives (i.e., the queue server takes repeated vacations of fixed length $V_{0,i} = mT_{sub}$, $i \geq 1$);

2. A busy period B_0 during which the queue is emptied.

The timeout interval consists of $M-1$ subframes after the last busy subframe, so that a timeout occurs when the UE counts M consecutive subframes without receiving data. Namely the timeout duration is $T_{out} = (M-1)T_{sub} = \left(2^{a+1} - 1\right) T_{sub}$, and we count the Mth idle subframe after a busy period as the beginning of the first vacation after the timeout expires.

Note that, according to [5], constant vacations are optimal for the minimization of the transmission cost in a system with Poisson arrivals. Hence constant DRX patterns are the optimal choice for minimizing the power consumption rate in power saving mode.

Figure 1 illustrates the composition of the system cycle. In the figure, $T_0 = I_0 + B_0$ is the interval in which the system is in power saving mode. T_0 always follows a timeout. For $1 \leq k \leq \xi$, interval $T_k = I_k + B_k$ is the kth of a set of consecutive intervals during which the system is in normal mode. Intervals T_k are optionally present in the system cycle after T_0. Since the arrival process is Poisson and the service process is uncorrelated with the arrivals, (i) the duration of each idle period I_k followed by a busy period B_k, $1 \leq k \leq \xi$, $\xi \geq 0$, is independent, and (ii) the duration of each busy period B_k, as for an $M/G/1$ queue, only depends on the number of packets queued at the beginning of the busy period, which, in turn, only depends on the arrivals in I_k. Given that a timeout occurs, the idle interval I_0 has a duration which only depends on the arrival process, and the following busy period B_0 only depends on the number of arrivals in I_0 and the serving discipline. In conclusion, each interval T_k, $k \neq 0$, is independent, and the interval $T_{out} + T_0$ is also independent from all other intervals. The sum of T_{out} and the sub-cycles T_i, $0 \leq i \leq \xi$, represents the duration of an overall system cycle T_c, which is a regeneration cycle. The timeout is a regeneration point for the system. Therefore, the system cycle duration is defined as the time between two consecutive timeouts.

3.2 Queueing model

Our objective in this section is to compute the expected sojourn time. For this, we derive the expectations of the initial backlogs, the idle/busy periods, the system cycle length, and the queue size, as detailed next.

Idle periods and initial backlogs in normal mode. To simplify the notation throughout the paper, we define the quantity p as:

$$p \triangleq e^{-\lambda T_{sub}}. \tag{1}$$

Consequently, the timeout probability is $P(T_{out}) = p^{M-1}$. When the inactivity timer does not expire, the system remains in normal mode for $\xi \geq 0$ intervals following T_0. Given there is no timeout, the (inactive) system is in an idle interval I_k consisting of $1 \leq \zeta_k < M$ subframes. We can write the joint distribution of the number of subframes ζ_k in I_k, and the backlog Z_k at the beginning of the busy period B_k, $1 \leq k \leq \xi$, as follows:

$$P(\zeta_k = i,\, Z_k = j) = \frac{e^{-\lambda T_{sub}(i-1)} \left[\frac{(\lambda T_{sub})^j}{j!} e^{-\lambda T_{sub}}\right]}{1 - p^{M-1}}; \tag{2}$$

where the denominator expresses the conditioning to the event that the inactivity timer does not expire, and $1 \leq i < M$ and $j \geq 1$. It follows that ζ_k, the number of idle subframes in each interval I_k, is distributed between 1 and $M-1$:

$$P(\zeta_k = i) = \frac{(1-p)p^{i-1}}{1 - p^{M-1}}, \qquad 1 \leq i < M. \tag{3}$$

The expectations of the number of subframes ζ_k, the idle interval I_k, and the initial backlog Z_k, for $1 \leq k \leq \xi$, can be readily written:

$$E[\zeta_k] = \sum_{i=1}^{M-1} iP(\zeta_k = i) = \frac{1 - Mp^{M-1} + (M-1)p^M}{(1-p)(1-p^{M-1})}; \tag{4}$$

$$E[I_k] = T_{sub} E[\zeta_k]; \tag{5}$$

$$E[Z_k] = \sum_{j=1}^{\infty} j \sum_{i=1}^{M-1} P(\zeta_k = i,\, Z_k = j) = \frac{\lambda T_{sub}}{1-p}. \tag{6}$$

Note that (4) to (6) do not depend on k.

Idle period and initial backlog in power saving mode. The computation of the expectation of the idle period I_0 and that of the initial backlog Z_0 at the beginning of the busy period B_0 is similar to the one found in [4]. Although a mandatory timeout is included here right before I_0, it does not impact the computation thanks to the memoryless property of Poisson flows.

The Laplace-Stieltjes transform of a generic random interval $V_{k,i}$ is denoted as $L_{k,i}(s)$ which, for $s = \lambda$, gives the probability of no arrivals in $V_{k,i}$:

$$P(\text{no arrivals in } V_{k,i}) \quad = \quad E\left[e^{-\lambda V_{k,i}}\right] \triangleq L_{k,i}(\lambda).$$

$L_{k,i}(s)$ is useful to compute the distribution of the number of consecutive system vacations ζ_0 and the joint distribution of Z_0 and ζ_0. We can write:

$$P(\zeta_0 = i) = [1 - L_{0,i}(\lambda)] \prod_{k=1}^{i-1} L_{0,k}(\lambda); \tag{7}$$

$$P(\zeta_0 \geq i) = \prod_{k=1}^{i-1} L_{0,k}(\lambda); \tag{8}$$

$$P(\zeta_0 = i,\, Z_0 = j) = E\left[\frac{(\lambda V_{0,i})^j}{j!} e^{-\lambda V_{0,i}}\right] \prod_{k=1}^{i-1} L_{0,k}(\lambda); \tag{9}$$

where $i \geq 1$ and $j \geq 1$. This yields the following expectations:

$$E[\zeta_0] = \sum_{i=1}^{\infty} P(\zeta_0 \geq i) = \sum_{i=1}^{\infty} \prod_{k=1}^{i-1} L_{0,k}(\lambda); \tag{10}$$

$$E[I_0] = \sum_{j=1}^{\infty} P(\zeta_0 = j) E\left[\sum_{i=1}^{j} V_{0,i}\right] = \sum_{i=1}^{\infty} E[V_{0,i}] \prod_{k=1}^{i-1} L_{0,k}(\lambda); \tag{11}$$

$$E[Z_0] = \sum_{j=1}^{\infty} j \sum_{i=1}^{\infty} P(Z_0 = j,\, \zeta_0 = i) = \lambda E[I_0]. \tag{12}$$

Second moment of initial backlogs. For later use, we need to derive $E[Z_k^2]$. For $1 \leq k \leq \xi$, it is straightforward to compute from (2) and (6):

$$E[Z_k^2] \quad = \quad (1 + \lambda T_{sub}) \cdot E[Z_k]. \tag{13}$$

For $k = 0$, we first define the following quantity (as in [4]):

$$I_a \triangleq \sum_{i=1}^{\zeta_0} V_{0,i}^2 = \sum_{i=1}^{\infty} V_{0,i}^2 \mathbf{1}_{\{\zeta_0 \geq i\}}, \tag{14}$$

whose expectation is:

$$E[I_a] = \sum_{i=1}^{\infty} E[V_{0,i}^2] \prod_{l=1}^{i-1} L_{0,l}(\lambda), \tag{15}$$

yielding, after some calculus:

$$E[Z_0^2] \quad = \quad E[Z_0] + \lambda^2 E[I_a]. \tag{16}$$

Busy periods. The expected busy time $E[B_k]$ is derived after the expected queue initial backlog $E[Z_k]$ by using the results for $M/G/1$ queues with no vacations [12]:

$$E[B_k] \quad = \quad E[Z_k] \frac{E[\sigma]}{1 - \rho}, \qquad 0 \leq k \leq \xi; \tag{17}$$

where $\rho = \lambda E[\sigma]$ is the offered load for an $M/G/1$ queue. In particular, for $k \geq 1$ we have $E[B_k] = \frac{\lambda T_{sub} E[\sigma]}{(1-\rho)(1-p)}$, which does not depend on k. The average busy period duration depends on λ, T_{sub}, m and M, i.e., the timeout. Observe that for 3GPP the service is deterministic, i.e., we have an $M/D/1$ queue and $E[\sigma] = T_{sub}$.

System cycle duration. A generic system cycle consists of a timeout, an interval $T_0 = I_0 + B_0$, and zero or more i.i.d. intervals $T_k = I_k + B_k$:

$$T_c = (M-1)T_{sub} + T_0 + \mathbf{1}_{\{\xi > 0\}} \sum_{k=1}^{\xi} T_k. \tag{18}$$

In (18), ξ is the number of times the inactivity timer does not expire in a row. The r.v. ξ is then distributed between 0 and infinite, and, due to the fact that intervals T_k are i.i.d., it behaves like the number of trials before a success for a Bernoulli process. The event of success is the timer expiration. Hence, the expected number of idle/busy periods before a timeout is:

$$E[\xi] = \sum_{k=0}^{\infty} k \left[1 - P(T_{out})\right]^k P(T_{out}) = \frac{1 - p^{M-1}}{p^{M-1}}. \tag{19}$$

The expected system cycle duration is then as follows:

$$\begin{aligned} E[T_c] \quad &= \quad (M-1)T_{sub} + E[I_0] + E[B_0] + P(\xi > 0) E\left[\sum_{k=1}^{\xi} (I_k + B_k) \middle| \xi > 0\right] \\ &= \quad \frac{1}{1-\rho} \frac{E[Z_0] + E[\xi]E[Z_1]}{\lambda} \quad \overset{\text{using (17)}}{=} \quad \frac{E[B_0] + E[\xi]E[B_1]}{\rho}. \end{aligned} \tag{20}$$

With $V_{0,i} = mT_{sub}$ and $V_{k,i} = T_{sub}$, (20) becomes

$$E[T_c] = \frac{T_{sub}}{1-\rho} \left(\frac{m}{1-p^m} + \frac{1-p^{M-1}}{(1-p)p^{M-1}} \right). \tag{21}$$

The system cycle duration depends on the timeout, the subframe length, the arrival rate, and the first moment of the service time (through ρ). In case the timeout is 0 ($M = 1$), i.e., the system never exits

the power saving mode, the expression for the system cycle reduces to the one obtained with the model in [4], without the warm-up period.

The system cycle grows with the vacation duration and with the timeout. For the case of constant vacations, we take the partial derivatives of $E[T_c]$ with respect to m and M:

$$\frac{\partial}{\partial m}E[T_c] = \frac{T_{sub}}{1-\rho}\left[\frac{1-(1-m\ln p)\,p^m}{(1-p^m)^2}\right]; \tag{22}$$

$$\frac{\partial}{\partial M}E[T_c] = \frac{\lambda T_{sub}^2}{(1-\rho)\,(1-p)\,p^{M-1}}. \tag{23}$$

From the definition of p in (1), and since m is positive, it follows that $(1-m\ln p)p^m = (1+m\lambda T_{sub})/e^{m\lambda T_{sub}} \leq 1$. Therefore, (22) is non-negative, and the system cycle is a non-decreasing function of m. Similarly, (23) is non-negative for all non-negative λ and null for $\lambda = 0$. Therefore the system cycle grows with M (i.e., with the timeout) for positive λ.

Number of packets served. The number of packets served in a cycle, N_P, on average is equal to the Poisson arrivals, i.e., $E[N_P] = \lambda E[T_c]$. In particular, the initial backlog Z_0 is, on average, λ times the duration of I_0 (cf. (12)), and the sum of the initial backlogs Z_k, $k \geq 1$, equates, on average, λ times the sum of the idle intervals I_k plus a timeout: $E[\xi]E[Z_1] = \lambda\{(M-1)T_{sub} + E[\xi]E[I_1]\}$. Hence, as expected, the total number of arrivals out of the busy intervals is λ times the duration of the non-busy intervals.

Queue size and sojourn time. Applying the methodology of [4] on each interval $T_k = I_k + B_k$ ($0 \leq k \leq \xi$), we can compute the expected queue size $E[X]$ and the expected sojourn time $E[T]$. In the interval T_k, the area under the curve $X(t)$ is given by $A_k + Q_{Z_k}$, with $A_k = A(V_{k,\zeta_k}) \triangleq \int_{V_{k,\zeta_k}} X(t)dt$ and $Q_{Z_k} \triangleq \int_{B_k} X(t)dt$. Here, the subscript Z_k expresses the fact that the initial backlog at the beginning of the busy interval B_k is Z_k.

The function $A(x)$ can be computed as follows, using the Poisson arrival process $N_\lambda(t)$,

$$A(x) = \frac{E\left[\int_0^x N_\lambda(t)dt\right]}{P(\text{at least one arrival in } x)} = \frac{\lambda x^2}{2}\frac{1}{1-e^{-\lambda x}}. \tag{24}$$

We can then write

$$E[A_k] = E[A(V_{k,\zeta_k})] = \begin{cases} \sum_{i=1}^{M-1} P(\zeta_k = i)E[A(V_{k,i})], & 1 \leq k \leq \xi; \\ \sum_{i=1}^{\infty} P(\zeta_0 = i)E[A(V_{0,i})], & k = 0. \end{cases} \tag{25}$$

The distributions of ζ_k and ζ_0 are given in (3) and (7) respectively. Since $V_{k,i} = T_{sub}$, we readily obtain $E[A_k] = \lambda T_{sub}^2/[2(1-p)]$, which does not depend on k. Similarly, if $V_{0,i} = mT_{sub}$, then $E[A_0] = \lambda(mT_{sub})^2/[2(1-p^m)]$.

The average of Q_{Z_k} depends on λ and the first two moments of Z_k and σ, for any k; cf. [4]:

$$E[Q_{Z_k}] = \frac{1}{2}\frac{E[Z_k]}{1-\rho}\left[\left(1+\frac{E[Z_k^2]}{E[Z_k]}\right)E[\sigma] + \frac{\lambda}{1-\rho}E[\sigma^2]\right]. \tag{26}$$

Specific expressions of $E[Q_{Z_k}]_{k\geq 1}$ and $E[Q_{Z_0}]$ can be found in Tables 1 and 2, respectively. These tables report the expressions of the quantities derived throughout this section, for the particular case of constant vacations $V_{0,i} = mT_{sub}$. Note that, for sake of generality, we did not replace in the formulas $E[\sigma]$ with T_{sub} as we have kept distinguishing $\rho = \lambda E[\sigma]$ from λT_{sub}.

A_0 and Q_{Z_0} are always present in a system cycle, as they always appear after a timeout, while A_k and Q_{Z_k}, $k > 0$, are in the cycle only if $\xi > 0$. Therefore, the expected queue size is:

$$E[X] \;=\; \frac{E[A_0] + E[Q_{Z_0}] + E[\xi]\,(E[A_1] + E[Q_{Z_1}])}{E[T_c]}. \tag{27}$$

Last, the expected sojourn time for a packet is computed via Little's formula as $E[T] = E[X]/\lambda$.

Table 1: Results for the normal mode ($V_{k,i} = T_{sub}$)

$E[\zeta_1]$	$\frac{1-Mp^{M-1}+(M-1)p^M}{(1-p)(1-p^{M-1})}$	$E[I_1]$	$T_{sub}E[\zeta_1]$
$E[Z_1]$	$\frac{\lambda T_{sub}}{1-p}$	$E[Z_1^2]$	$\frac{\lambda T_{sub}(1+\lambda T_{sub})}{1-p}$
$E[B_1]$	$\frac{\lambda T_{sub}E[\sigma]}{(1-p)(1-\rho)}$	$E[\xi]$	$\frac{1-p^{M-1}}{p^{M-1}}$
$E[A_1]$	$\frac{\lambda T_{sub}^2}{2(1-p)}$	$E[Q_{Z_1}]$	$\frac{\lambda T_{sub}\left[(1-\rho)(2+\lambda T_{sub})E[\sigma]+\lambda E[\sigma^2]\right]}{2(1-p)(1-\rho)^2}$

Table 2: Results for the power saving mode with $V_{0,i} = mT_{sub}$

$E[\zeta_0]$	$\frac{1}{1-p^m}$	$E[I_0]$	$mT_{sub}E[\zeta_0] = \frac{mT_{sub}}{1-p^m}$
$E[Z_0]$	$\frac{\lambda mT_{sub}}{1-p^m}$	$E[I_a]$	$\frac{m^2T_{sub}^2}{1-p^m}$
$E[B_0]$	$\frac{\lambda mT_{sub}E[\sigma]}{(1-p^m)(1-\rho)}$	$E[Z_0^2]$	$\frac{\lambda mT_{sub}(1+\lambda mT_{sub})}{1-p^m}$
$E[A_0]$	$\frac{\lambda(mT_{sub})^2}{2(1-p^m)}$	$E[Q_{Z_0}]$	$\frac{\lambda mT_{sub}\left[(1-\rho)(2+\lambda mT_{sub})E[\sigma]+\lambda E[\sigma^2]\right]}{2(1-p^m)(1-\rho)^2}$

3.3 Cost and power saving

The UE's receiver remains continuously active during the cycle, with a basic consumption rate c_{on}, except for the sleeping periods within I_0, during which the consumption rate is $c_{sl} < c_{on}$. Receiving a packet *increases* the basic consumption rate by c_{rx}. Listening to the control channels, i.e., receiving a control packet, also *increases* the basic consumption rate by c_{ln}. Since I_0 consists of ζ_0 sub-intervals, and since each of such intervals begins with a fixed-length listening window of T_{ln} seconds, then during the power saving periods, the receiver listens to the control channel for only $\zeta_0 T_{ln}$ seconds out of I_0. Hence, the average cost for receiving packets per time unit is a combination of the cost to receive packets, the cost to listen to the control channel, the energy spent in sleep mode, and the cost of being on, i.e.:

$$\begin{aligned} C_{UE} &= \frac{T_{sub}E[N_P]}{E[T_c]}c_{rx} + \frac{\frac{E[T_c]-E[I_0]}{T_{sub}} + E[\zeta_0]}{E[T_c]}T_{ln}c_{ln} + \\ &\quad + \frac{E[I_0] - E[\zeta_0]T_{ln}}{E[T_c]}c_{sl} + \frac{E[T_c] - E[I_0] + E[\zeta_0]T_{ln}}{E[T_c]}c_{on}. \end{aligned} \tag{28}$$

Considering the case of fixed vacations, when $E[I_0] = mT_{sub}E[\zeta_0]$, the total cost rate can be rewritten as:

$$C_{UE} = C_{UE}^{nps}(\lambda) - \alpha(m) \cdot \frac{E[I_0]}{E[T_c]}, \tag{29}$$

with

$$C_{UE}^{nps}(\lambda) = T_{sub}\lambda c_{rx} + \frac{T_{ln}}{T_{sub}}c_{ln} + c_{on}; \tag{30}$$

$$\alpha(m) = \left(1 - \frac{T_{ln}}{mT_{sub}}\right)(c_{on} - c_{sl}) + \left(1 - \frac{1}{m}\right)\frac{T_{ln}}{T_{sub}}c_{ln}. \tag{31}$$

$C_{UE}^{nps}(\lambda)$ is the cost with no power saving and depends on λ only. The second term in (29) is the cost reduction due to power saving, $\alpha(m)$ being a cost reduction factor which depends on the length of the power saving sub-cycle.

Impact of λ, m and M on the cost reduction. In case of constant vacations, the ratio $\frac{E[T_c]}{E[I_0]}$ can be expressed as follows:

$$\left.\frac{E[T_c]}{E[I_0]}\right|_{V_{0,i}=mT_{sub}} = \frac{1}{1-\rho}\left(1 + \frac{1}{1-p}\cdot\frac{1-p^{M-1}}{p^{M-1}}\cdot\frac{1-p^m}{m}\right). \tag{32}$$

It is easy to show that this ratio is: (i) null for $\lambda = 0$, and otherwise positive; (ii) insensitive to m if $M = 1$ and decreasing with m increasing if $M > 1$; (iii) increasing with both M and λ (recall $\rho = \lambda E[\sigma]$ and $p = e^{-\lambda T_{sub}}$).

Since $\alpha(m)$ increases with m and is insensitive to λ and M, the cost reduction $\alpha(m)E[I_0]/E[T_c]$ decreases with λ (the arrival rate), increases with m (i.e., with the vacation size) and decreases with M (i.e., with the timeout).

Power saving gain. The normalized cost reduction, or *power saving gain* G_{UE}, is the average rate of energy saved by using the power saving mode. It is then formally defined as follows:

$$G_{UE} \triangleq \frac{C_{UE}^{nps}(\lambda) - C_{UE}}{C_{UE}^{nps}(\lambda)} \overset{\text{using (29)}}{=} \frac{\alpha(m)}{C_{UE}^{nps}(\lambda)} \frac{E[I_0]}{E[T_c]}. \tag{33}$$

The second equality holds in the case of constant vacations. It can be shown that G_{UE} is a decreasing function of the arrival rate λ, an increasing function of the vacation size (through m), and a decreasing function of the timeout (through M).

4 Power saving at the base station

Similarly to the UE case of Section 3, here we show the power saving that can be achieved at the eNB when it transmits to a pool of N_u users (UEs). Here, we only consider the eNB downlink transmissions.

4.1 Queueing model for eNB

In order to compute the power saving at the eNB, we extend the model presented in Section 3 as follows: (i) up to N_u UEs can be active simultaneously; (ii) a separate $M/G/1$ queue is available for each UE, with independent arrivals; (iii) all queues share the same processor, i.e., the eNB scheduler, which has a fixed serving rate $\mu = 1/T_{sub}$; (iv) each queue behaves as analyzed in Section 3, hence it alternates a normal mode, during which packets are served as soon as they reach the head of the queue, and a power saving mode of duration T_0, during which head-of-line packets may not be served if the queue is on vacation; (v) the shared processor, representing a GPS scheduler, serves all head-of-line packets for all queues in parallel (generalized processor sharing model with variable number of queues and no priority); (vi) normal/power saving periods of different queues are considered as independent. This last assumption is not met in reality: queues are correlated given that they share the same processor. However, we will show in Section 5 that the approximation is good in the case of: (i) homogeneous arrival rates and (ii) heterogeneous arrival rates with low to medium traffic loads. Observe that in this model, the eNB is always operational (not sleeping) and ready to transmit packets to any UE that is operational. When a UE is sleeping, its corresponding $M/G/1$ queue will be in vacation, so the eNB cannot transmit any packet from this queue.

4.1.1 Homogeneous arrival rates

In case of homogeneous arrivals, the aggregate arrival rate in the system is $N_u\lambda$. Each queue is analyzed as in Section 3. The expected sleep period of each queue is $E[I_0]$, and the expected awake period is $E[T_c] - E[I_0]$. All expressions derived in Section 3 are valid for each queue, provided the arrival rate is the per-queue rate λ. The only (important) point that is different concerns the service time σ. We no longer have that the service time is an input parameter of the model (deterministic, equal to T_{sub} in 3GPP). Instead, σ depends on the number of active queues at each system slot, given that the total service rate is $\mu = 1/T_{sub}$. Besides the arrival rate λ and the power saving parameters m and M, the metrics derived in Section 3 depend also on the first and the second moments of the service time σ. To complete the analysis of the model, we need to derive the first and second moments of σ for the multiple queue case with single shared processor. This is done next.

We assume that the load of each queue is such that all the queues are stable. We can then interpret ρ as the fraction of time during which a queue is under service, or, equivalently, as the probability of the latter event.

From the point of view of a generic queue having a packet to be served, the service time at any instant is proportional to the number of queues being served simultaneously. Namely, $\sigma = T_{sub}N_a$ where N_a is a random variable taking values in the interval $[1, N_u]$, given that there are 1 to N_u queues to serve in parallel. Considering all queues as independent, we can write $N_a = 1 + \nu$, with ν a binomial random

variable having success probability ρ and number of trials $N_u - 1$. Therefore, the expected service time is:

$$E[\sigma] = T_{sub}E[1+\nu] = T_{sub}[1+(N_u-1)\rho]. \tag{34}$$

Hence, considering that $\rho = \lambda E[\sigma]$, we have a system of two equations in two variables, whose solution is:

$$E[\sigma] = \frac{T_{sub}}{1-\lambda T_{sub}(N_u-1)}; \qquad \rho = \frac{\lambda T_{sub}}{1-\lambda T_{sub}(N_u-1)}. \tag{35}$$

Since all arrivals are served, the expected service time only depends on the number of users and on the arrival rate, i.e., it does not depend on the power saving parameters m and M. Observe that in reality, queues are correlated and the degree of correlations increases with the load on the shared processor. The service time of a queue in case of correlations is actually smaller than the same in the absence of correlations, as will be observed in the validation section (cf. Section 5.1).

Similarly, the second moment of the service time is computed as follows:

$$E[\sigma^2] = T_{sub}^2\left[1+3(N_u-1)\rho+(N_u-1)(N_u-2)\rho^2\right]. \tag{36}$$

Note that for $N_u \to 1$, σ and ρ behave as described in Section 3 for the single queue case. The maximum allowable arrival rate is such that the aggregate rate equates the server rate, i.e., $N_u\lambda = 1/T_{sub}$. For very high traffic the system behaves as a regular $M/G/1$ PS queue with N_u equal classes (i.e., user's queues), each receiving $\frac{1}{N_u}$ of the overall service. In fact, for $\lambda \to 1/(T_{sub}N_u)$, we have $\rho \to 1$, $E[\sigma] \to T_{sub}N_u$, and $E[\sigma^2] \to T_{sub}^2N_u^2$.

4.1.2 Non-homogeneous arrival rates

We assume now that each packet arrival is independent and Poisson, but with a different rate λ_i per each user. The model for each user's queue is formally the same as for the case of homogeneous arrival rates, but the service time is no longer homogeneous. The utilization ρ_i of each user's queue is now $\rho_i = \lambda_i E[\sigma_i]$, where σ_i is the service time experienced at queue i. Assuming that all queues are independent (this assumption holds as long as the offered load is low to medium), when the ith queue is under service each other queue k can be under service with a probability ρ_k. The number of packets under service, i.e., the number of transmissions occurring when the ith queue has a packet under service, is a random variable $N_a^{(i)} = 1+\nu^{(i)}$. Here, $\nu^{(i)}$ is a sum of N_u-1 *independent* Bernoulli random variables Y_r, $r \in \{1,2,\ldots,N_u\}\setminus\{i\}$, where the success probability of r.v. Y_r is $\rho_r = \lambda_r E[\sigma_r]$. The service time for the ith queue is $\sigma_i = T_{sub}N_a^{(i)}$. Therefore, we can express the first and second moments of the service time for each queue i as a function of the utilization coefficients ρ_k, for $i = 1,2,\ldots,N_u$:

$$E[\sigma_i] = T_{sub}\Big(1+\sum_{k\neq i}\rho_k\Big); \tag{37}$$

$$E[\sigma_i^2] = T_{sub}^2\Big(1+3\sum_{k\neq i}\rho_k + 2\sum_{\substack{r<s,\\ r,s\neq i}}\rho_r\rho_s\Big). \tag{38}$$

In particular, (37) yields a system of N_u equations in N_u variables $E[\sigma_i]$, whose solution can be used to compute all ρ_k, and hence to solve (38). An explicit expression for $E[\sigma_i]$ is given by:

$$E[\sigma_i] = T_{sub}\left[1+\sum_{j=1}^{N_u-1}T_{sub}^j\sum_{\substack{k_1<..<k_j\\ k_1,..,k_j\neq i}}\prod_{a=1}^{j}\lambda_{k_a}\right]\Big/\left[1-\sum_{j=2}^{N_u}(j-1)T_{sub}^j\sum_{k_1<..<k_j}\prod_{a=1}^{j}\lambda_{k_a}\right]. \tag{39}$$

Observe that not all combinations of λ_i can be used, since we want $\max_i(\rho_i) < 1$, so that all queues are stable. From the expression of $E[\sigma_i]$, it is easy to see that $\lambda_i \leq \lambda_j$ implies $E[\sigma_i] \geq E[\sigma_j]$ and $\lambda_i E[\sigma_i] \leq \lambda_j E[\sigma_j]$. Hence, the system is stable if and only if the most loaded queue is stable.

4.2 Cost at the eNB

The power consumption rate at the eNB is the sum of a fixed component, c_f, that does not depend on the transceiver activity, and a variable component that depends on the activity of UEs in the cell. Namely, the power consumption rate at the eNB can be written

$$C_{BS} = c_f + \sum_{i=1}^{N_u} C_{tx}(\lambda_i, m, M) \overset{\text{homogeneous case}}{=} c_f + N_u C_{tx}(\lambda, m, M), \tag{40}$$

where $C_{tx}(\lambda, m, M)$ is the cost per time unit to transmit to a single UE having data rate λ. The fixed cost c_f is independent of user activity and relates to site control and management, power consumption of downlink pilots, etc. Recent studies show that the fixed cost c_f can be as much as 10 times the average cost for transmitting packets over the air interface [9].

Concerning the per-user transmission cost C_{tx}, it is the power consumption rate incurred by the transmission of data to a single UE with continuous connectivity. Each UE in the cell enables the DRX/DTX mode as soon as the inactivity timer expires, as discussed earlier. The cost C_{tx} at the eNB can be computed much in the same way as the reception cost C_{UE} at a UE. In the case of constant vacations, it can be written as a function of C_{tx}^{nps}, the cost with no power saving, and the transmission cost reduction factor α_{tx} as follows:

$$C_{tx}(\lambda_i, m, M) = C_{tx}^{nps}(\lambda_i) - \alpha_{tx}(m) \cdot \frac{E[I_0]}{E[T_c]}, \tag{41}$$

with

$$C_{tx}^{nps}(\lambda_i) = T_{sub}\lambda_i c_{tx} + \frac{T_{ln}}{T_{sub}} c_{sg} + c_{on}; \tag{42}$$

$$\alpha_{tx}(m) = \left(1 - \frac{T_{ln}}{mT_{sub}}\right)(c_{on} - c_{sl}) + \left(1 - \frac{1}{m}\right)\frac{T_{ln}}{T_{sub}} c_{sg}. \tag{43}$$

Comparing (42)-(43) with (30)-(31), a transmission cost c_{tx} now replaces the reception cost c_{rx} and a signaling cost c_{sg} replaces the listening cost c_{ln}. Observe that a reduction in the power consumption rate at the UE translates into a reduction in the power consumption rate at the eNB.

It is worth mentioning that the per-packet transmission cost, c_{tx}, is defined as the cost to transmit over the full bandwidth for a time unit T_{sub}. Therefore, the cost to transmit a packet (that fits in a subframe T_{sub}) over a generic bandwidth and an arbitrarily long transmission interval only depends on the packet size and equals $T_{sub}\, c_{tx}$. Thus, the total transmission cost is not affected by the per-packet serving time σ, and depends only on the number of packets to be served, hence the first summand in (42).

Power saving gain. It is simply the normalized cost reduction at the eNB and is denoted G_{BS}. Formally, we can write:

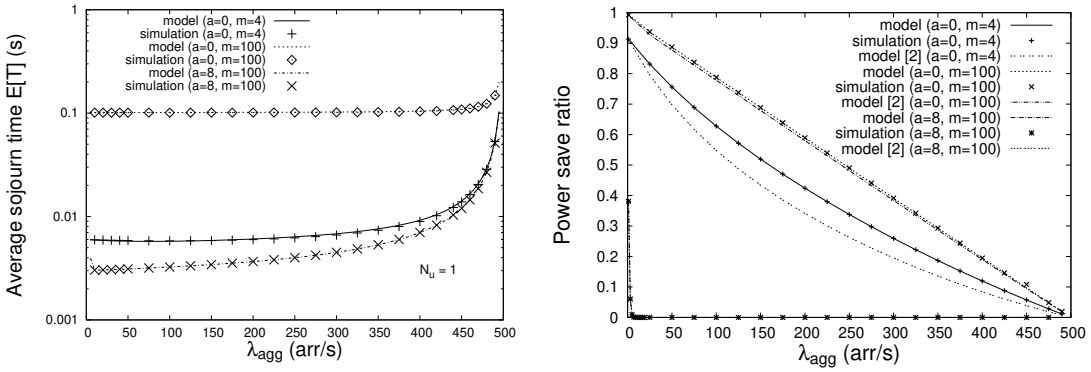
$$G_{BS} \triangleq \frac{C_{BS}^{nps} - C_{BS}}{C_{BS}^{nps}} \overset{\text{using (40)}}{=} \frac{\sum_{i=1}^{N_u} \left(C_{tx}^{nps}(\lambda_i) - C_{tx}(\lambda_i, m, M)\right)}{c_f + \sum_{i=1}^{N_u} C_{tx}^{nps}(\lambda_i)} \tag{44}$$

$$\overset{\text{using (41)}}{=} \frac{\alpha_{tx}(m) \sum_{i=1}^{N_u} \frac{E[I_0]}{E[T_c]}\Big|_{\lambda=\lambda_i}}{c_f + \sum_{i=1}^{N_u} C_{tx}^{nps}(\lambda_i)} \overset{\text{homogeneous case}}{=} \frac{\alpha_{tx}(m)\frac{E[I_0]}{E[T_c]}}{\frac{c_f}{N_u} + C_{tx}^{nps}(\lambda)}. \tag{45}$$

Equation (45) holds in the case of constant vacations. When arrivals are homogeneous, the power saving gain increases with the number of users N_u. Observe that the cost reduction at user i is $\alpha(m)\, \frac{E[I_0]}{E[T_c]}\Big|_{\lambda=\lambda_i}$ (cf. Eq. (29)) while that at the eNB is $\alpha_{tx}(m) \sum_{i=1}^{N_u} \frac{E[I_0]}{E[T_c]}\Big|_{\lambda=\lambda_i}$ (numerator of (45)). Therefore, the cost reduction at the eNB is a factor $\frac{\alpha_{tx}}{\alpha}$ of the cost reductions at all users combined.

5 Validation and evaluation of the model

In this section we validate the model using simulations. Then we use the model to compute the power saving parameters which maximize the cost reduction at the UE and the eNB, subject to an upper bound on the packet sojourn time. Throughout this section, $T_{sub} = 2$ ms and arr/s stands for arrivals per second.



(a) Average sojourn time.

(b) Power save ratio (with $T_{ln} = T_{sub}/3$).

Figure 2: Model vs. simulation with one user.

5.1 Validation the model through simulation

In order to evaluate the model, we developed a C++ event-driven simulator that reproduces the behavior of a *time slotted* $M/G/1$ PS queue with $N_u \geq 1$ users. In the simulator, the shared processor resources are allocated fairly between users at the *beginning* of each time slot of duration T_{sub}. Each user represents a downlink user, which can be in normal mode or in power saving mode. Packet interarrivals are exponentially distributed with rate λ_i, and arrivals for different users are independent. Simulated packets have the same size, and each requires one slot of service time. If only one user is under service, a packet is served completely in one slot. Otherwise, since the processor is shared, all backlogged active users have a fraction of packet served in that slot. In the simulator, the fair per-user share is computed as one over the number of backlogged active users. However, if an active user has not enough backlog to use all its processor share during a slot, unused resources are redistributed amongst other users. The service of a packet can last one or more time slots, and the service is considered complete at the end of the last service slot. Observe that queues states are correlated in the simulator. This will allows us to test the robustness of the eNB model to violation of the independence assumption.

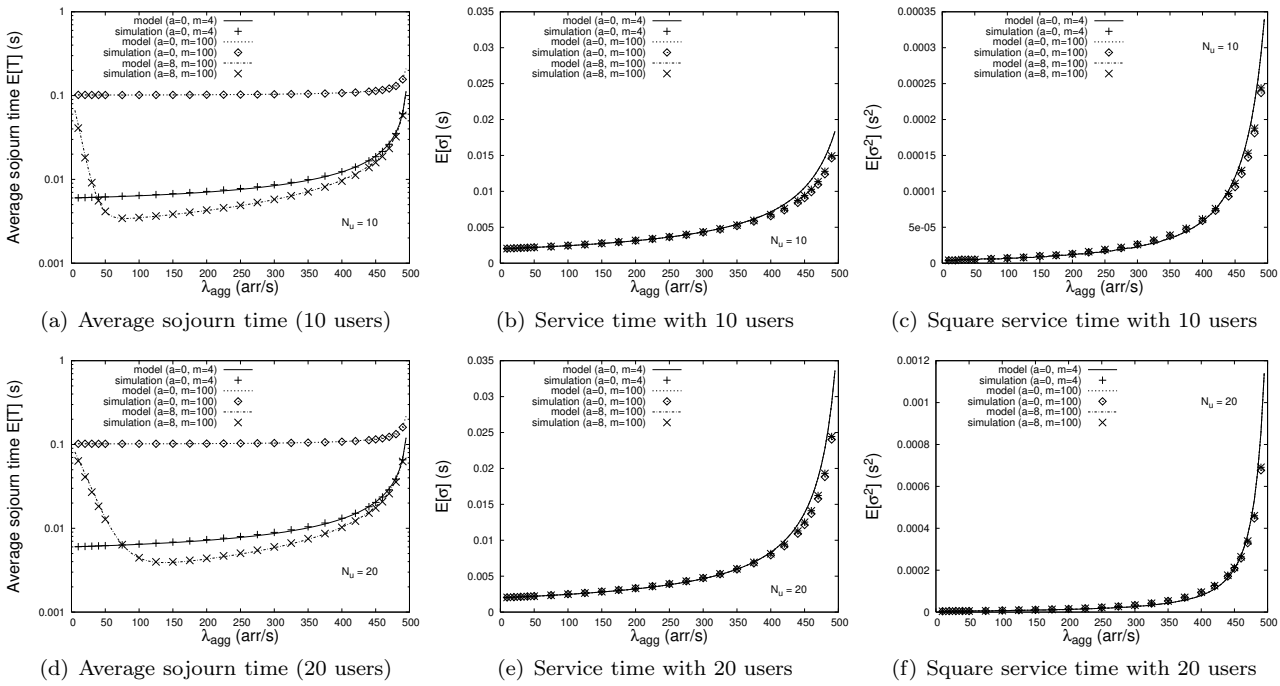
We simulate different values of number of users N_u and arrival rates λ_i. Also, we simulate three different settings for the power saving by changing the timeout duration (through a, using the relation $T_{out} = (2^{a+1} - 1)T_{sub}$), and the length of power saving cycles m. The three power saving settings are: (i) configuration "$a = 0, m = 4$" which shows the results for short timeouts ($T_{out} = T_{sub}$) and short power saving sub-cycles ($4T_{sub}$); (ii) configuration "$a = 0, m = 100$" which shows the results for short timeouts and long vacations (i.e., yielding high savings); and (iii) configuration "$a = 8, m = 100$" which shows the system performance for long timeouts (and hence low power saving) and long vacations.

Each simulation consists of a warm-up period lasting 10,000 seconds (5,000,000 slots), followed by 20 runs, each lasting 10,000 seconds. In each run, statistics are collected separately. At the end of simulation, all statistics are averaged over the 20 runs and 99% confidence intervals are computed.

We are interested in three performance parameters: the average sojourn time $E[T]$ and the first two moments of the packet service time $E[\sigma]$ and $E[\sigma^2]$. These are computed using the analytical models and collected from simulations as explained earlier.

Single user case. The service time is constant ($\sigma = T_{sub}$) for both model and simulation. The expected sojourn time is computed according to the model in Section 3.2. Figure 2(a) shows that the sojourn time is correctly evaluated through the model, for all sustainable values of the aggregate arrival rate λ_{agg}. Figure 2(b), reports the *power save ratio*, i.e., the fraction of time during which the UE sleeps. With our model, the power save ratio can be computed as the time spent in I_0 during a cycle T_c, excluding listening intervals, i.e., $\left(1 - \frac{T_{ln}}{mT_{sub}}\right) E[I_0]/E[T_c]$. Figure 2(b) shows that our model matches with high accuracy simulation results. Furthermore, the figure includes the power save ratio computed with the analytical model proposed in [18] for DRX in UMTS systems. In that model, a continuous-time approach is adopted, in contrast with the more realistic slotted time assumption of our model. Notwithstanding the different modeling assumptions, the two models yield pretty similar results in all cases.

Multiple users, homogeneous arrivals. Analytical results are those of the model in Section 4.1.1.



(a) Average sojourn time (10 users) (b) Service time with 10 users (c) Square service time with 10 users

(d) Average sojourn time (20 users) (e) Service time with 20 users (f) Square service time with 20 users

Figure 3: Analytic/simulation results for 10 and 20 users (homogeneous arrivals).

Figure 3 depicts the results of simulation and model for the case of 10 and 20 users. The figure shows a good match between the model and the simulations, in all cases. As predicted by the model, the service time only depends on the arrival rate and the number of users, but not on the power saving parameters a and m. Observe that the model slightly overestimates the moments of the service time at very high traffic rates (cf. Figs. 3(b), 3(c), 3(e), 3(f)). This is a consequence of the independence assumption that is less good at high traffic. However, this overestimation does not affect the sojourn time as analytic and simulation results perfectly match at all traffic rates (cf. Figs. 3(a) and 3(d)).

Multiple users, heterogeneous arrival rates. Analytical results are those of the model in Section 4.1.2. Three users are considered. Users 1 and 2 have rates $\lambda_1 = 50$ arr/s and $\lambda_2 = 100$ arr/s. Different values are simulated for the arrival rate of user 3, as reported in Figure 4. The figure depicts the average sojourn time of user 3 as a function of its arrival rate λ_3. Model and simulations yield similar results until $\lambda_3 < 250$ arr/s, which turns in serving about 400 packets/s. Hence the model is accurate in the heterogeneous case for low to medium arrival rates. Recall that the independence assumption is not met in the simulator.

From the comparison of simulation and model, we can conclude that the assumptions we used in order to compute the service time's moments are not impairing the quality of estimation for both the average service time and the average sojourn time. Consequently, we can use the model to optimize the power saving parameters when the admissible sojourn time is upper-bounded.

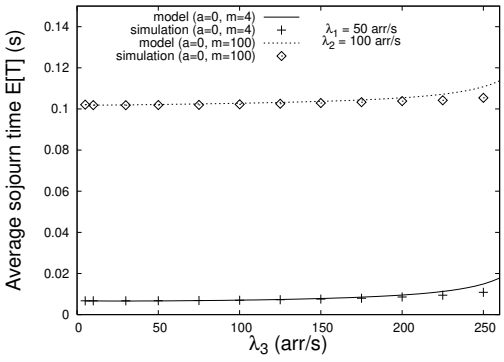


Figure 4: Results for the heterogeneous case.

Poisson vs. web traffic. Packet arrivals with real traffic might be far from Poisson. To evaluate the impact of the Poisson assumption adopted throughput the paper, we simulated web traffic according to the web traffic evaluation model proposed by 3GPP2 in [2]. With the 3GPP2 traffic generation model, each user generates a web request after the previous request has been com-

pletely served. All web pages are generated according to the same distribution, so that, to change the offered load we can only change the number of users. Therefore, in this set of experiments, for each fixed number of users in the system, we first simulate web traffic according to the 3GPP2 model, and then we repeat the simulation using the same number of users generating Poisson traffic with the same average as in the web-based experiment. Figure 5(a) shows that generating traffic with Poisson arrivals or according to the 3GPP2 evaluation model does not significantly affect the average service time σ. Similar results, not shown here for lack of space, hold for the second moment of the service time. In fact, the first two service time's moments depend on the average load of the various users, as enlightened by Eqs. (37) and (38). However, the time spent in power saving state can be radically different with Poisson or web traffic. To illustrate this point, Figures 5(b) and 5(c) show the power save ratio achieved with Poisson and with web traffic. In particular, those figures show that when the timeout is high (e.g., $a = 8$), Poisson arrivals yield very few power saving opportunities, while the web traffic, being more bursty, would still allow for several power saving opportunities. Note also that using Poisson traffic the power save ratio is always smaller than using the 3GPP2 model.

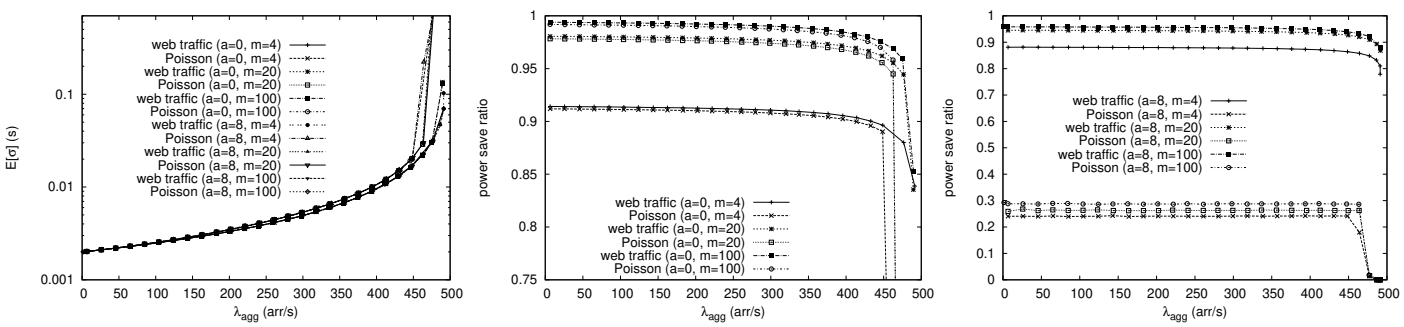
Considering that (*i*) power saving opportunities decrease with a, (*ii*) Poisson traffic yields pessimistic power saving ratios, and (*iii*) Poisson traffic results are close to web traffic results for small values of a, we conclude that Poisson traffic can be reasonably used to estimate the *optimal* power saving under realistic traffic conditions.

5.2 Maximization of UE cost reduction

Here we want to find the parameters that maximize the energy saving at the UE ($N_u = 1$), using constant vacations and keeping the packet sojourn time bounded. The system parameters are: (*i*) the timeout duration, through the parameter a; (*ii*) the length of the power saving cycle, m, in subframes; and (*iii*) the arrival rate λ. In particular, we look for the optimal values of a and m for a given value (or for a range of values) of the arrival rate λ. The function to be optimized is the relative gain G_{UE} averaged over a selected range of λ. The constraint to the optimization is represented by the sojourn time $E[T]$, after averaging over the selected range for λ:

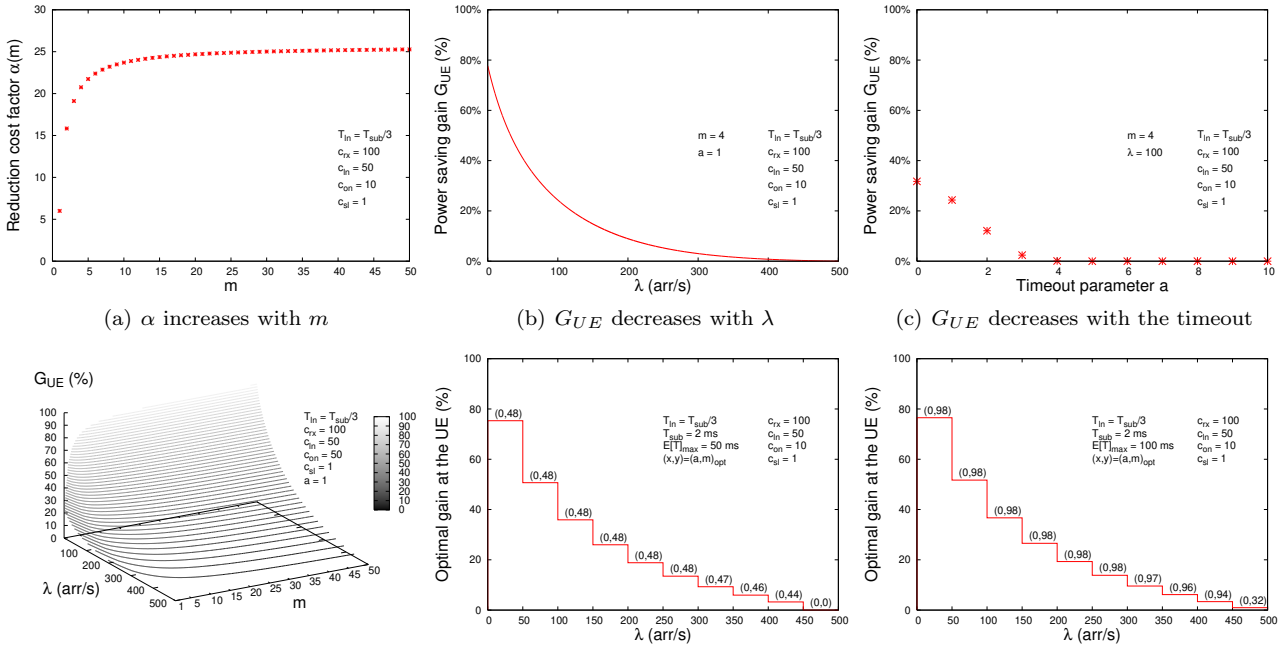
$$\begin{cases} \max\limits_{m\geq 1, a\geq 0} \frac{1}{\lambda_{max}-\lambda_{min}} \int\limits_{\lambda_{min}}^{\lambda_{max}} G_{UE}(\lambda)d\lambda; \\ \text{subject to} \quad \frac{1}{\lambda_{max}-\lambda_{min}} \int\limits_{\lambda_{min}}^{\lambda_{max}} E[T](\lambda)d\lambda \leq D_x. \end{cases} \tag{46}$$

Reasonably, the cost for receiving a packet is larger than the cost for receiving a control packet (i.e., for "listening"), which is usually shorter and transmitted at low rate. Both receiving and listening costs are much higher then the cost to stay on, which, in turn, is at least one order of magnitude greater than the cost to stay in sleep mode. As an example, we use the following values: $c_{rx} = 100$, $c_{ln} = 50$, $c_{on} = 10$, and $c_{sl} = 1$. Furthermore we assume that control packets have a duration $T_{ln} = \frac{T_{sub}}{3}$, e.g., the UE has to listen to the control channel only during the first of the three slots composing an HSPA subframe.



(a) Average service time (b) Power save ratio ($a = 0$). (c) Power save ratio ($a = 8$).

Figure 5: Performance comparison with Poisson and web traffic.



(a) α increases with m

(b) G_{UE} decreases with λ

(c) G_{UE} decreases with the timeout

(d) G_{UE} for fixed timeout ($a = 1$)

(e) Optimal (a, m) for $E[T] \leq 50$ ms

(f) Optimal (a, m) for $E[T] \leq 100$ ms

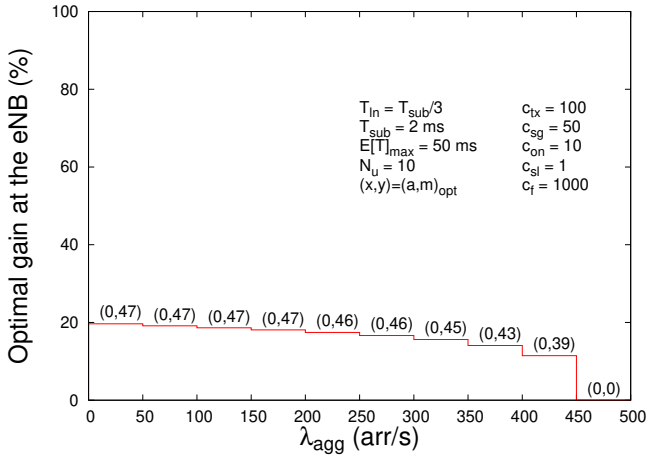
Figure 6: Relative gain G_{UE} at the UE, parameter optimization over intervals of 50 arr/s.

The value of $\alpha(m)$, the cost reduction factor, is depicted in Figure 6(a): it grows very fast for small m, but quickly saturates. In practice, values bigger than 20 do not give substantial gain advantages with respect to $m = 20$, which is the maximum value suggested by 3GPP. Figure 6(b) shows a dramatic cost reduction if the network is underloaded. With $m = 4$ and $a = 1$, the gain can be as high as 75% for negligible arrival rates, and 20% if λ is one forth of the maximum server capacity. Higher values of m and $a = 0$ would give even higher gains, but also higher delays. The impact of the timeout is shown in Figure 6(c), where we fix $m = 1$, $\lambda = 100$ arr/s (yielding $\rho = 0.2$), and plot the relative gain G_{UE} as a function of a. Only small values of a enable a considerable gain. Figure 6(d) shows the combined effect of varying λ and m when the timeout is fixed and small. Remarkably, the gain can be as high as 90% with low arrival rates, and remains above 20% for medium loads (up to 300 arr/s, i.e., $\rho = 0.6$).

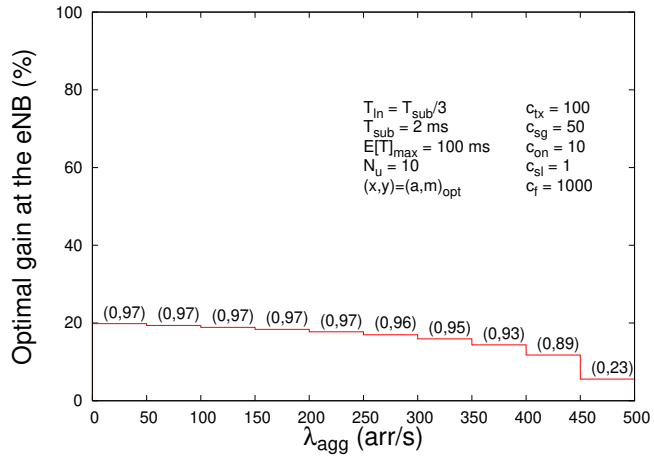
Figures 6(e) and 6(f) illustrate the gain that can be achieved at the UE through the optimization of power saving parameters a and m, subject to keeping the average sojourn time not greater than 50 ms and 100 ms, respectively. Optimal values of the parameters are reported in the figure, above the average gain level. The optimization described in (46) has been performed by splitting the total arrival rate range (0 to 500 arr/s) into 10 intervals, and by numerically optimizing the average gain in each interval, subject to an expected sojourn time whose average over the considered arrival rate interval is not greater than 50 ms (100 ms). In both cases, the gain is consistent as far as the arrival rate is below 250 arr/s (i.e., $\rho \leq 0.5$), and it can be as high as 75%.

5.3 Maximization of eNB cost reduction

At the eNB, the cost is also a function of the number of users. Hence the optimization problem has the form of (46) with G_{BS} replacing G_{UE}. Following the same rationale as for the UE case, we use the following cost parameters for illustrative purposes: $c_{tx} = 100$, $c_{sg} = 50$, $c_{on} = 10$, and $c_{sl} = 1$. Additionally, as suggested by experimental measurements [9], we consider a huge fixed base station cost $c_f = 1000$. Figure 7(a) illustrates the gain in a system with 10 users and homogeneous arrival rates. The optimization of a and m is performed for intervals of $\lambda_{\max} - \lambda_{\min} = 50$ arr/s, and subject to keeping the average sojourn time not greater than 50 ms. Figure 7(b) refers to the case that the maximum tolerable average sojourn time is 100 ms. Here, the gain is not high (<20%), but, unlike the UE case, it does not degrade fast with λ_{agg}. A much higher gain can be obtained if the number of users grows. In particular,



(a) Optimal (a, m) for $E[T] \leq 50$ ms



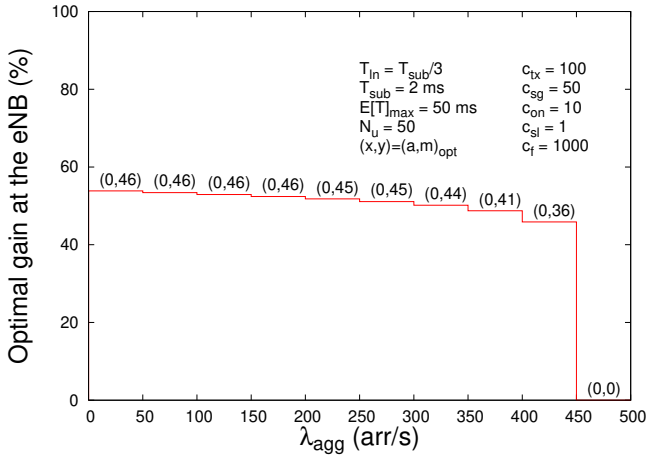
(b) Optimal (a, m) for $E[T] \leq 100$ ms

Figure 7: Optimization of the eNB gain with 10 homogeneous users.

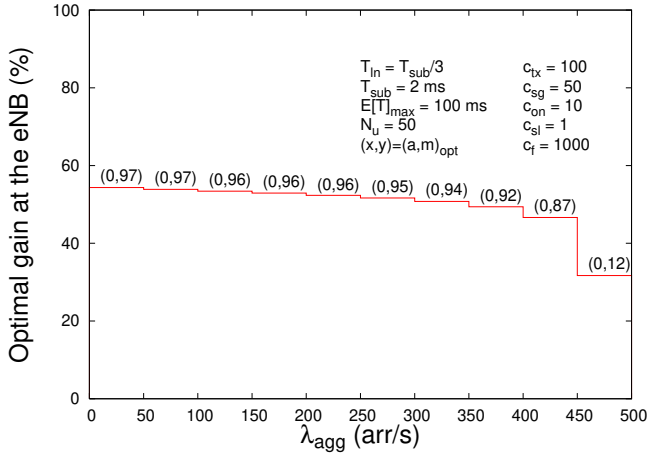
Figures 8(a) and 8(b) show the optimal gain with 50 users, subject to an average sojourn time not greater than 50 ms and 100 ms, respectively. The gain can be as high as 50% for a large range of arrival rates. Thereby, the use of power saving strategies at eNB is attractive only if the number of users is not low.

6 Conclusions

In this paper, we first have shown how to use the properties of an $M/G/1$ system to model the transmission of a user adopting the continuous connectivity model. We have derived the quantities that characterize the regeneration cycle of the system, allowing us to compute the packet performance figures. Second, and most important, we have shown how to extend the model to the case of multiple users sharing the same base station. We have modeled the per-user activity in order to evaluate the service share that the base station processor can grant to each user. After that, we have modeled the base station behavior with an $M/G/1$ PS system in which a user is excluded from the services when it is in power saving mode. The model has been validated through simulations. Finally, we have proposed a cost model and shown how to optimize the power saving parameters to minimize the cost with a bounded queueing delay. Remarkably, we have shown that up to 75% of the user cost, and 55% of the base station one, can be saved while preserving the quality of the packet flow in the downlink.



(a) Optimal (a, m) for $E[T] \leq 50$ ms



(b) Optimal (a, m) for $E[T] \leq 100$ ms

Figure 8: Optimization of the eNB gain with 50 homogeneous users.

References

[1] 3GPP TS 25.214. Physical layer procedures (FDD), release 8, v8.9.0, March 2010.

[2] 3GPP2 C.R1002-B v1.0. CDMA2000 evaluation methodology - Revision B, December 2009.

[3] J. Almhana, Z. Liu, C. Li, and R. McGorman. Traffic estimation and power saving mechanism optimization of IEEE 802.16e networks. In *Proc. of IEEE ICC 2008*, pages 322–326, Beijing, China, May 2008.

[4] S. Alouf, E. Altman, and A.P.. Azad. M/G/1 queue with repeated inhomogeneous vacations applied to IEEE 802.16e power saving. In *Proc. of ACM SIGMETRICS 2008*, volume 36 of *Performance Evaluation Review*, pages 451–452, Annapolis, Maryland, USA, June 2008.

[5] Amar Prakash Azad, Sara Alouf, Eitan Altman, Vivek Borkar, and Georgios Stavrou Paschos. Optimal control of sleep periods for wireless terminals. *IEEE Journal on Selected Areas in Communications*, 29(8):1605–1617, September 2011. special Issue on Energy-Efficient Wireless Communications.

[6] C. Bontu and E. Illidge. DRX mechanism for power saving in LTE. *IEEE Communications Magazine*, 47(6):48–55, June 2009.

[7] E. Dahlman, S. Parkvall, and J. Skold. *4G: LTE/LTE-Advanced for Mobile Broadband*. Academic Press, San Diego, CA, USA, 2011.

[8] E. Dahlman, S. Parkvall, J. Skold, and P. Beming. *3G Evolution: HSPA and LTE for Mobile Broadband*. Academic Press, Oxford, UK, Second edition, 2008.

[9] F. Corrêa Alegria and F.A. Martins Travassos. Implementation details of an automatic monitoring system used on a Vodafone radiocommunication base station. *IAENG Engineering Letters*, 16(4):529–536, November 2008.

[10] K. Han and S. Choi. Performance analysis of sleep mode operation in IEEE 802.16e mobile broadband wireless access systems. In *Proc. of IEEE VTC 2006-Spring*, volume 3, pages 1141–1145, Melbourne, Australia, May 2006.

[11] T. Karagiannis, M. Molle, M. Faloutsos, and A. Broido. A nonstationary Poisson view of Internet traffic. In *Proc. of IEEE INFOCOM 2004*, volume 3, pages 1558–1569, Hong Kong, mar 2004.

[12] L. Kleinrock. *Queueing Systems: Theory*, volume 1. 1975.

[13] T. Kolding, J. Wigard, and L. Dalsgaard. Balancing power saving and single user experience with discontinuous reception in LTE. In *Proc. of IEEE ISWCS 2008*, pages 713–717, Reykjavik, Iceland, 2008.

[14] Vincenzo Mancuso and Sara Alouf. Power save analysis of cellular networks with continuous connectivity. In *Proc. of IEEE WoWMoM 2011*, Lucca, Italy, June 2011.

[15] Nujira Ltd. State of the art RF power technology for defense systems. white paper, February 2009. `http://www.nujira.com/_uploads/whitepapers/State_of_the_Art_RF_Power_Technology_for_Defence_Systems_EU.pdf`.

[16] J.B. Seo, S.Q. Lee, N.H. Park, H.W. Lee, and C.H. Cho. Performance analysis of sleep mode operation in IEEE 802.16e. In *Proc. of IEEE VTC 2004-Fall*, volume 2, pages 1169–1173, Los Angeles, CA, USA, September 2004.

[17] Y. Xiao. Performance analysis of an energy saving mechanism in the IEEE 802.16e wireless MAN. In *Proc. of IEEE CCNC 2006*, volume 1, pages 406–410, Las Vegas, Nevada, USA, January 2006.

[18] S.R. Yang and Y.B. Lin. Modeling UMTS discontinuous reception mechanism. *IEEE Transactions on Wireless Communications*, 4(1):312–319, January 2005.

[19] L. Zhou, H. Xu, H. Tian, Y. Gao, L. Du, and L. Chen. Performance analysis of power saving mechanism with adjustable DRX cycles in 3GPP LTE. In *IEEE VTC 2008-Fall*, Calgary, Alberta, Canada, September 2008.