Power save analysis of cellular networks with continuous connectivity

Vincenzo Mancuso
Institute IMDEA Networks
Madrid, Spain
Email: vincenzo.mancuso@imdea.org

Sara Alouf
INRIA Sophia Antipolis Méditerranée
Sophia Antipolis, France
Email: sara.alouf@sophia.inria.fr

***Abstract*—In this paper, we analyze the power save and its impact on web traffic performance when customers adopt the continuous connectivity paradigm. To this aim, we provide a model for packet transmission and cost. We model each mobile user's traffic with a realistic web traffic profile, and study the aggregate behavior of the users attached to a base station by means of a processor-shared queueing system. In particular, we evaluate user access delay, download time and expected economy of energy in the cell. The model is validated through packet-level simulations. Our model shows that dramatic energy save can be achieved by both mobile users and base stations, e.g., as much as 70% of the energy cost due to packet transmission at the base station.**

***Keywords*-power saving; cellular network; analytical model;**

I. Introduction

The total operating cost for a cellular network is of the order of tens of millions of dollars for a medium-small network with twenty thousand base stations [1]. A relevant portion of this cost is due to power consumption, which can be dramatically reduced by using efficient power save strategies. Power save can be achieved in cellular networks operating WiMAX, HSPA, or LTE protocols by optimizing the hardware, the coverage and the distribution of the signal, or also by implementing energy-aware radio resource management mechanisms. In particular, we focus on power save in wireless transmissions, which would enable the deployment of compact (e.g., air conditioning free) and green (e.g., solar power operated) base stations, thus requiring less operational and management costs.

An interesting case study is offered by the behavioral analysis of users that remain online for long periods. These users request a continuous availability of a dedicated wide-band data channel, in order to shorten the delay to access the network as soon as new packets have to be exchanged. This *continuous connectivity* requires frequent exchange of control packets, even when no data are awaiting for transmission. Therefore, in case of continuous connectivity, a huge amount of energy might be spent just to control the high-speed connection, unless power save is enforced. However, since power save mode affects packet delay, some constraints have to be considered when turning to the power save operational mode.

Power save and sleep mode in cellular networks have been analytically and experimentally investigated in the literature, mainly from the user equipment (UE) viewpoint. E.g., power save in the UMTS UE has been evaluated in [2] and [3] by means of a semi-Markov chain model. The authors of [4] 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. In [5], the authors use an $M/G/1$ queue with repeated vacations to model an 802.16e-like sleep mode and to compute the service cost for a single user download. Analytical models, supported by simulations, were proposed by Xiao for evaluating the performance of the UE in terms of energy consumption and access delay in both downlink and uplink [6]. The authors of [7] provide an adaptive algorithm that minimizes energy subject to QoS requirements for delay.

The existing work does not tackles the base station (or evolved node B, namely eNB) viewpoint nor analytically captures the relation between cell load and service rate statistics. Furthermore, for sake of tractability, many of those studies assume that packet arrivals follow a Poisson model. Instead, in real networks, the user traffic can be very bursty and follow long tail distributions [8]. In contrast, we use a $G/G/1$ queue with vacations to model the behavior of each UE, and we compose the behavior of multiple users into a single $G/G/1\,PS$ queue that models the eNB traffic. We analytically compute the cost reduction achievable thanks to power save mode operations, and show how to minimize the system cost under QoS constraints. In particular we refer to the mechanisms made available by 3GPP for *Continuous Packet Connectivity* (CPC), i.e., the discontinuous transmission (DTX) and discontinuous reception (DRX) [9].

The importance of DRX has been addressed in [10], where the authors model a procedure for adapting the DRX parameters based on the traffic demand, in LTE and UMTS, via a semi-Markov model for bursty packet data traffic. A description of DRX advantages in LTE from the user viewpoint is given in [11] by means of a simple cost model. In [12], 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 model for the behavior of users (UEs) and base stations (eNBs) in continuous connectivity

This work was supported in part by the WiNEM project (funded by ANR RNRT 2006), the ICT FLAVIA Project (funded by the EU Framework 7 Programme), and the MEDIANET Project (grant S2009/TIC-1468) from the General Directorate of Universities and Research of the Regional Government of Madrid.

978-1-4577-0351-5/11/$26.00

and with non-Poisson traffic, (ii) we provide a cost model that incorporates the different causes of operational costs, and (iii) we show how to use the model to minimize operational costs under QoS constraints. Our model has been validated through packet-level simulations, and our results confirm that a tremendous cost reduction can be attained by correctly tuning the power save parameters. In particular, eNB transmission costs can be lowered by more than 70%.

The paper is organized as follows: Section II presents power save operations in continuous connectivity mode; Section III describes a model for cellular users generating web traffic. Section IV illustrates a model for downlink transmissions, and Section V describes how to evaluate flow performance and transmission costs. In Section VI we validate the model through simulation, and show the achievable power saving. Section VII concludes the paper.

II. Continuous connectivity

Cellular packet networks, in which the base station schedules the user activity, require the online UEs to check a control channel continuously, namely for T_{ln} seconds per system slot (i.e., per subframe T_{sub}). 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 or receive any data for long periods [13]. A highly efficient power save mode operation is then strongly required, which would allow disabling both transmission and reception of frames during the idle periods. The UE, however, has to transmit and receive control frames at regular rhythm, every few tens of milliseconds, so that synchronization with 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 power save mode in which it checks and reports on the control channels according to a fixed pattern, i.e., only once every m time slots. Relevant energy economy can be achieved. In change, the queued packets have to wait for the mth subframe before being served.

Discontinuous transmission. DTX has been first defined by 3GPP release 7. It is a UE operational mode for discontinuous uplink transmission over the Dedicated Physical Control Channel (DPCCH). With DTX, UEs transmit control information according to a cycle. There are actually two possible DTX cycles. The first cycle is used when some data activity is present in the uplink (normal operation), and it is a short cycle (one or very few subframes). The second cycle is longer (up to tens of subframes), and is triggered after an inactivity timeout in the uplink data channel expires (power save mode operation). The threshold M for inactivity period is a power of 2 subframes. Since transmissions on uplink data channel can only start in parallel with DPCCH transmissions, DTX also regulates data transmissions.

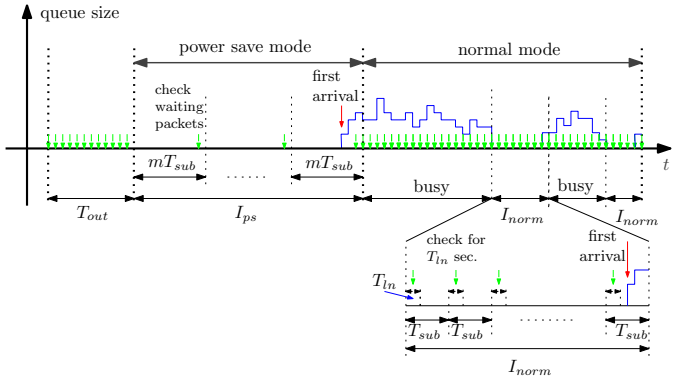

Figure 1. Downlink queue activity with power save and normal operation.

Discontinuous reception. DRX is an operational mode defined by 3GPP release 6 for the UE to save energy while monitoring the control information transmitted by the eNB. It also affects data delivery, since no data can be dependably received without an associated control frame. 3GPP specifications define a 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 cycle are 4 to 20 subframes (i.e., using a 2 ms subframe in HSPA yields a cycle of 8 to 40 ms). DRX is activated only upon a timeout expiry after the last downlink transmission, and like DTX, the timeout threshold specified in the standard is M subframes, with M being a power of 2.

III. Power save model

We focus on the power consumption due to wireless activity on the air interface of mobile users (UEs) and base station (eNB). On the one hand, we assume that uplink control transmission follows the DTX pattern. On the other hand, the UE has to decode the downlink control channel according to the DRX pattern, and receive packets accordingly [13]. Thus, uplink power save can be enabled by means of a long DTX cycle, with a timeout whose duration can be of the same order of the subframe size. Downlink power save is similarly enforced by setting the DRX cycle and timeout.

Thereby, power save issues in uplink and downlink can be modeled in a similar way, and there is little difference between the cost computation of a single UE and the one of a base station. In fact, the evaluation of the costs at the eNB, can be seen as the collection of costs over the control and data channels towards the various UEs, plus a fixed per-cell operational cost that the eNB has to pay to notify its presence and maintain the users synchronized. Therefore, here we focus on the downlink only, and begin our analysis with the behavior of a UE receiving a data stream.

Power save in downlink. As illustrated in Figure 1, downlink power save can be obtained by alternating between two possible DRX cycles: after any downlink data activity there is a short cycle in which the UE continuously checks the control channel at each subframe (normal operation

mode); instead, upon the expiration of an inactivity timeout T_{out}, consisting of M subframes, there is a longer cycle in which the UE checks the control channel periodically, with period m subframes (power save mode).[1] In power save mode, the UE samples the downlink control channel every m subframes, and returns to normal mode as soon as the channel sampling detects a control message indicating that the downlink queue is no longer empty. Note that UEs do not receive any service during: (i) I_{norm}, i.e., idle intervals in normal operation, (ii) timeout intervals, and (iii) I_{ps}, i.e., idle intervals in power save mode.

To quantify the power save that can be achieved at the UE, in Section IV we model the behavior of downlink transmissions with DRX operations enabled and users generating web traffic. Then, in Section V we discuss the tradeoff between per-packet performance and per-UE cost. Our model can be used for systems using slotted operations, and in particular LTE and HSPA [13]. The model can be applied to both uplink and downlink. However, for sake of clarity, we explicitly deal with the downlink case.

Achievable cost saving and performance metrics will be expressed as a function of the subframe length T_{sub} and the DRX parameters, namely the timeout duration, through the parameter M, and the DRX power save cycle duration, through the parameter m. We assume fixed-length packets, and the server capacity is exactly one packet per subframe. However, no packet is served for UEs in power save mode, and the server capacity is shared, in each subframe, between the UEs operating in normal mode. Therefore, we model a system which behaves as a $G/G/1\,PS$ queue with repeated fixed-length vacations of mT_{sub} seconds.

Before proceeding with the model derivation, we introduce the traffic model adopted in this study.

Traffic model. We assume that downlink traffic is the composition of users' web browsing sessions. Traffic profile is the same for all users and is as follows. The size of each web request is modeled as suggested by 3GPP2 in [14]: a web page consists of one main object, whose size is a random variable with truncated lognormal distribution, and zero or more embedded objects, each with random, truncated lognormal distributed size. The number of embedded objects is a random variable derived from a truncated Pareto distribution. Each web page request triggers the download of the packets carrying the main object only. Then a *parsing time* is needed for the user application to parse the main object and request the embedded objects, if any. The parsing time distribution is exponential with rate λ_p. After having received the last packet of the last object, the customer *reads* the web page for an exponentially distributed

[1]The actual system timeout is M-subframe long. However, since the UE checks for new traffic at the beginning of a subframe, the UE switches to power save mode if it does not receive any traffic alert at the beginning of the Mth idle subframe. Therefore, it is enough to have no arrivals for $M-1$ subframes and the UE will not receive any packet for M subframes.

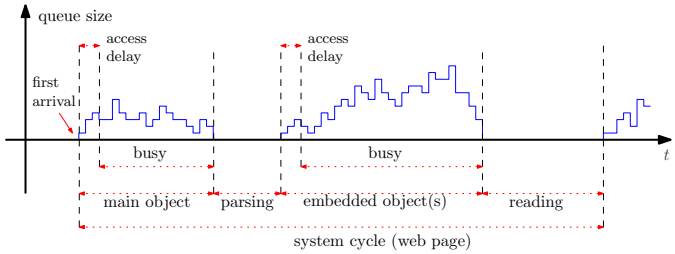

Figure 2. System cycle with web traffic as defined in [14].

reading time, whose rate is λ_r. If no object is embedded, the reading time includes the parsing time. Finally he/she requests another web page. Figure 2 represents the UE's downlink queue size at the eNB during a generic web page request and download. Table I summarizes the parameters used for the generation of web browsing sessions. Note that the probability ψ_0 to have no embedded objects in a web page can be computed through the distribution of the truncated Pareto random variable Y described in Table I: $\psi_0 = P(y_{\min} \leq Y < y_{\min}+1) = 1 - \left(\frac{y_{\min}}{y_{\min}+1}\right)^{\alpha}$. Note also that the downlink of the web page experiences a small access delay due to the completion of the current DRX cycle before the first packet of the new burst could be served.

In our model, we assume that the time to request a web object with an http GET command is negligible in comparison with the time needed to parse the main object, and therefore also in comparison with the time needed for a customer to read the web page. Hence we incorporate this request delay in the parsing time and in the reading time. In this way, we clearly focus our study on the sole impact of the wireless technology on the system performance and costs. Furthermore, packet arrivals are supposed to be bursty after each GET request, so that no power save mode can be triggered after an object download begins, i.e., all power save intervals are contained in either parsing or reading times. With these assumptions, we study the system performance through the analysis of a generic web page download and its fruition. More precisely, we study the system cycle defined as the time in between two consecutive web page requests. Therefore, the system cycle can be decomposed in four phases, as depicted in Figure 2: (i) download of the main object of the web page, (ii) parsing of the main object, (iii) download of embedded objects, and (iv) web page reading. The first three phases represent the web page download time, from the first packet arrival in the eNB queue to the last packet delivery to the UE. Access delay and download time characterize the service experienced by the customer.

IV. Model Derivation

Here we derive the time spent by the system in the various cycle phases. For ease of notation, we define $\beta_p = e^{-\lambda_p T_{sub}}$ and $\beta_r = e^{-\lambda_r T_{sub}}$ as the probabilities that, respectively, the exponentially distributed parsing time and reading time are longer than one subframe. Hence the timeout probability is β_r^{M-1} in reading time, and β_p^{M-1} in parsing time.

Table I
PARAMETERS SUGGESTED BY 3GPP2 FOR THE EVALUATION OF WEB TRAFFIC

Quantity	Derivation	Probability distribution	Parameters
Main object size	$S_{mo} = \lceil X \rceil$	$f_X(x) = \frac{\left(2\pi\sigma_X^2\right)^{-\frac{1}{2}} e^{-\frac{(\ln x - \mu_X)^2}{2\sigma_X^2}}}{\int_{x_{\min}}^{x_{\max}} \left(2\pi\sigma_X^2\right)^{-\frac{1}{2}} e^{-\frac{(\ln t - \mu_X)^2}{2\sigma_X^2}} dt}$ $x \in [x_{\min}, x_{\max}]$	$\mu_X = 8.35$, $\sigma_X = 1.37$, $x_{\min} = 100$ bytes, $x_{\max} = 2 \cdot 10^6$ bytes
Number of embedded objects	$N_{eo} = \lfloor Y \rfloor - y_{\min}$	$f_Y(y) = \alpha \frac{y_{\min}^{\alpha}}{y^{\alpha+1}}, \quad y \in [y_{\min}, y_{\max}[$ $f_Y(y_{\max}) = \left[1 - \left(\frac{y_{\min}}{y_{\max}}\right)^{\alpha}\right] \delta(y - y_{\max})$	$y_{\min} = 2$, $y_{\max} = 55$ $\alpha = 1.1$
Embedded object size	$S_{eo} = \lceil Z \rceil$	$f_Z(z) = \frac{\left(2\pi\sigma_Z^2\right)^{-\frac{1}{2}} e^{-\frac{(\ln z - \mu_Z)^2}{2\sigma_Z^2}}}{\int_{z_{\min}}^{z_{\max}} \left(2\pi\sigma_Z^2\right)^{-\frac{1}{2}} e^{-\frac{(\ln t - \mu_Z)^2}{2\sigma_Z^2}} dt}$ $z \in [z_{\min}, z_{\max}]$	$\mu_Z = 6.17$, $\sigma_Z = 2.36$, $z_{\min} = 50$ bytes, $z_{\max} = 2 \cdot 10^6$ bytes
Reading time	Λ_r	$f_{\Lambda_r}(t) = \lambda_r e^{-\lambda_r t}$, $t \geq 0$	$\lambda_r = 0.0\overline{3}$
Parsing time	Λ_p	$f_{\Lambda_p}(t) = \lambda_p e^{-\lambda_p t}$, $t \geq 0$	$\lambda_p = 7.69$

Timeouts in a cycle. Each cycle always includes one reading time, while the parsing time is present with probability $1-\psi_0$, i.e., only if there are embedded objects. Therefore, the average number of timeouts in a system cycle is:

$$E[N_{to}] = \beta_r^{M-1} + (1-\psi_0)\,\beta_p^{M-1}. \tag{1}$$

Hence each cycle includes, on average, $E[N_{to}](M-1)T_{sub}$ seconds due to timeout occurrences.

Idle time in power save mode. The average time per cycle during which the system is in power save mode, denoted as I_0, is computed by summing up the time spent in power save mode (the intervals I_{ps} as in Figure 1) occurring in the reading time and in the parsing time, if any is present in the cycle: $I_0 = I_{ps|\text{reading}} + I_{ps|\text{parsing}}$. Thanks to the memoryless property of exponential arrivals, the interval between the timeout expiration and the arrival of the next data packet is exponential too, and has the same exponential rate. In particular, the power save interval that begins in the reading time lasts a multiple number of checking intervals mT_{sub}, with the following distribution and average:

$$\begin{aligned}
&P\left(I_0 = jmT_{sub}|\text{ reading timeout}\right)\\
&= P(0 \text{ arrivals in } (j-1)mT_{sub})[1 - P(0 \text{ arrivals in } mT_{sub})]\\
&= (\beta_r^m)^{j-1}\left(1-\beta_r^m\right), \qquad j \geq 1;\\
&E[I_0|\text{reading}] = \beta_r^{M-1}\frac{mT_{sub}}{1-\beta_r^m};
\end{aligned} \tag{2}$$

where we also removed the conditioning on the timeout occurrence. Similarly, for the parsing time:

$$E[I_0|\text{parsing}] = \beta_p^{M-1}\frac{mT_{sub}}{1-\beta_p^m}. \tag{3}$$

Therefore, the expected value of the time spent in power save mode in a system cycle is given by the following average:

$$E[I_0] = \beta_r^{M-1}\frac{mT_{sub}}{1-\beta_r^m} + (1-\psi_0)\,\beta_p^{M-1}\frac{mT_{sub}}{1-\beta_p^m}. \tag{4}$$

Note that $E[I_0]$ is a function of m and M, the web traffic parameters being fixed. It is easy to find that $\frac{\partial}{\partial m}E[I_0] > 0$, and $\frac{\partial}{\partial M}E[I_0] < 0$, hence the power save interval I_0 monotonically grows with the duration of the DRX cycle, and decreases with the duration of the timeout.

Idle time in normal mode. The amount of time spent in normal mode without serving any traffic is the sum of the normal mode idle intervals due to parsing and reading times. Since we counted apart the time spent in timeouts by means of (1), here we only count the intervals I_{norm}, whose sum over a system cycle is denoted by $I_1 = I_{norm|\text{reading}} + I_{norm|\text{parsing}}$. Considering that I_{norm} is always a multiple of T_{sub} but smaller than a timeout, and since the component of I_1 in reading time is $I_{norm|\text{reading}}$, the conditional distribution of I_1 in reading time is as follows:

$$\begin{aligned}
&P(I_1 = jT_{sub}|\text{reading})\\
&= P(I_{norm} = jT_{sub}|\text{ exp. arrivals with rate } \lambda_r)\\
&= \begin{cases} \beta_r^{M-1} & j = 0; \\ \beta_r^{j-1}(1-\beta_r) & 1 \leq j \leq M-1. \end{cases}
\end{aligned}$$

Hence the conditional expected value of this interval I_1 is:

$$E[I_1|\text{reading}] = T_{sub}\frac{1 - M\beta_r^{M-1} + (M-1)\beta_r^M}{1-\beta_r}. \tag{5}$$

Similarly, the expected value for the time spent in normal mode with no traffic to be served during parsing, without counting the timeout, is given by

$$E[I_1|\text{parsing}] = T_{sub}\frac{1 - M\beta_p^{M-1} + (M-1)\beta_p^M}{1-\beta_p}. \tag{6}$$

Therefore, on average, the time spent in normal mode without serving any traffic during a system cycle is given by the timeout intervals plus $E[I_1|\text{reading}]$, plus $1-\psi_0$ times $E[I_1|\text{parsing}]$. So, the expected value of I_1 increases with the timeout duration, through M.

Cumulative idle time. The cumulative amount of idle time I in a cycle is the sum of timeouts, I_0, and I_1. Its expected value is then as follows:

$$E[I] = \beta_r^{M-1}\frac{mT_{sub}}{1-\beta_r^m} + T_{sub}\frac{1-\beta_r^{M-1}}{1-\beta_r} + (1-\psi_0)\left(\beta_p^{M-1}\frac{mT_{sub}}{1-\beta_p^m} + T_{sub}\frac{1-\beta_p^{M-1}}{1-\beta_p}\right). \quad (7)$$

$E[I]$ is a decreasing function of M, and increases with m. However, with our model assumptions, $E[I]$ is slightly larger than the sum of reading and parsing times. More precisely, its value is bounded as follows:

$$\frac{1}{\lambda_r}+\frac{1-\psi_0}{\lambda_p} < E[I] < \frac{1}{\lambda_r}+mT_{sub}+(1-\psi_0)\left(\frac{1}{\lambda_p}+mT_{sub}\right).$$

Given that m can be as high as few tens, and T_{sub} is only few milliseconds, the product mT_{sub} is negligible in comparison with the average parsing and reading times. Hence, for all realistic values of m, the per-cycle idle time can be considered constant and equal to its lower bound.

Busy time. The expected time spent to serve the packets of a web page, i.e., the busy time in a cycle, is given by the expected number of packets $E[N_p]$ per web page times the expected service time $E[\sigma]$. The number of packets depends on the distribution of the web page objects, and it is 39.47 with the 3GPP2 traffic model reported in Table I.[2] The service time depends on the number of active UEs and on the server capacity, as we show later in this section.

System cycle duration. Putting together the results for the time spent in timeouts, idle intervals, and busy periods, the expected duration of a cycle is given by:

$$E[T_c] = E[N_{to}](M-1)T_{sub}+E[I_0]+E[I_1]+E[N_p]E[\sigma] = E[I] + E[N_p]E[\sigma]. \quad (8)$$

The relation between $E[T_c]$ and $E[\sigma]$ is linear with a coefficient that is determined by the web page object distribution. Since $E[\sigma]$ too will be shown to grow with m and decrease with M (see next paragraph), the entire expected system cycle increases with m and decreases with M. Furthermore, as the expected service time increases with the number N_u of UEs attached to the eNB, the system cycle behaves likewise. However, both $E[I]$ and $E[\sigma]$ are barely affected by m and M, thereby $E[T_c]$ is mainly affected by N_u only.

Service time. We assume that there are N_u homogeneous UEs in the cell. The activity factor of each UE is:

$$\rho = \frac{E[N_p]E[\sigma]}{E[T_c]} = \frac{E[N_p]E[\sigma]}{E[N_p]E[\sigma]+E[I]} < 1. \quad (9)$$

Equivalently, we can interpret ρ as the probability that a UE is under service. Note that $E[\sigma]$, $E[N_p]$, and $E[I]$ assume always positive values, and thus $E[T_c]>0$ and $0<\rho<1$.

[2]We use 1500-byte packets and consider each object as an integer number of packets. Hence, after having computed the number of bytes N_b in an object, we consider that object as consisting of $\left\lceil \frac{N_b}{1500} \right\rceil$ packets.

From the point of view of a generic queue, the service time in the lth subframe only depends on the number $N_a(l)$ of queues which transmit in that specific subframe. In fact, the downlink bandwidth is shared between the active and backlogged queues, the total serving capacity being fixed to one packet per subframe. Thus, given that the ith queue has a packet under service in the lth system subframe, the service time for the ith queue is $T_{sub}N_a(l)$. Since we are interested in the service time for the ith queue, we condition the observation of the service time to the transmission of a packet queued in the ith queue. Hence, considering all queues as i.i.d., the number of active queues is a random variable $N_a = 1+\nu$, with ν being a random variable exhibiting a binomial distribution between 0 and N_u-1 with success probability ρ. Thereby, the average service time is:

$$E[\sigma] = T_{sub}E[1+\nu] = T_{sub}[1+(N_u-1)\rho]. \quad (10)$$

Hence, considering the expression (9) of ρ as a function of $E[\sigma]$, we have a system of two equations in two variables, from which we can compute $E[\sigma]$.

Proposition 1: The expected packet service $E[\sigma]$ is the unique positive solution of the following quadratic equation:

$$E[N_p]\,E^2[\sigma]+(E[I]-E[N_p]\,N_uT_{sub})\,E[\sigma]-E[I]T_{sub}=0.$$

Proof: the equation is obtained by combining (9) and (10). Since $E[N_p]$ and $E[I]$ are positive numbers, the quadratic coefficient in the equation is always positive, whilst the constant term is negative: this is necessary and sufficient to have one positive solution and one negative solution. However, the negative solution has no physical meaning. Thus, the positive solution is the only acceptable solution candidate. ■

Corollary 1: The expected packet service is

$$E[\sigma] = \frac{(E[N_p]N_uT_{sub}-E[I])+\sqrt{(E[I]-E[N_p]\,N_uT_{sub})^2+4E[I]\,E[N_p]\,T_{sub}}}{2E[N_p]}.$$

As we stressed before, the term $E[I]$ increases with m and decreases with M, but its variations are quite limited. So, thanks to the corollary, we can conclude that $E[\sigma]$ behaves as $E[I]$, i.e., it is barely affected by m and M. Furthermore, $E[\sigma]$ grows with N_u, i.e., with the number of UEs in the cell. Notably, the impact of N_u on $E[\sigma]$ is amplified by a factor equal to the average page size $E[N_p]$.

Since a new web page is requested only after the reading time of the previous request, the number of customers has no theoretical upper bound. In fact, service time and system cycle just keep growing with the number of UEs, and the average cumulative traffic generated and served per subframe is $N_u\frac{E[N_p]}{E[N_p]\,E[\sigma]+E[I]} \leq \frac{1}{T_{sub}}$. Thus, as the system approaches saturation, $E[\sigma]$ tends to N_uT_{sub}, since in saturation the N_u users are always active and receive a fraction $1/N_u$ of the server capacity. The asymptotic distribution of the system cycle duration is constant and

equal to $T_c^{up} = E[N_p]\, N_u T_{sub} + E[I]$, which scales linearly with the number of users and loosely depends on the power save parameters m and M. T_c^{up} is an upper bound for the evaluation of the system cycle, and can be used to limit the maximum number of customers, thus guaranteeing a maximum web page processing time to any customer.

V. Performance and cost metrics

The impact of power save mode on web traffic can be evaluated in terms of access delay and page download time, assuming that all the traffic is served. Costs due to wireless transmission and reception of packets are to be traded off with such indicators. Therefore, we first derive an expression for performance metrics and show how to compute the fraction of time during which power save can be realistically obtained. Then we derive the parametric expressions for cost and power save at both UE and eNB.

A. Performance metrics and power save opportunities

Page download time. The time W needed to download a web page includes the time to download each and every page's packet, the time to parse the main object of the page, and the access delay. Hence, we can derive $E[W]$ as the difference between $E[T_c]$ and the expected reading time:

$$E[W] = E[T_c] - \frac{1}{\lambda_r}. \tag{11}$$

Access delay. The access delay is the delay experienced after any download request. In our model we consider only that part of the access delay which is due to the wireless access protocol. In particular, we have two epochs within each cycle at which a request can experience access delay: at the end of the reading time, corresponding to a new page request, and at the end of the parsing time, corresponding to the request for the embedded objects. We name D the total access delay experienced within a web page download, thus accounting for the delay accumulated in both reading and parsing times. $E[D]$ can be easily computed by subtracting the parsing time, the reading time and the busy time from the expected system cycle duration (see Figure 2), i.e.:

$$E[D] \;=\; E[I] - \left(\frac{1}{\lambda_r} + \frac{1-\psi_0}{\lambda_p}\right). \tag{12}$$

The expected access delay is a function of the power save parameters used in the DRX configuration, plus the traffic profile parameters, through λ_r, λ_p, $E[N_p]$, and ψ_0. However, using the upper bound for $E[I]$, one can conclude that the access delay is upper bounded to $(2-\psi_0)mT_{sub}$.

Power save time ratio. Economy of energy can be achieved by reducing the radio activity, including the possibility to turn off the radio transceiver, according to the DTX/DRX pattern. Therefore, power save opportunities can be measured through the fraction of cycle during which the transceiver can be deactivated. In practice, UE and eNB can save power during I_0, which is a multiple of mT_{sub}, but for the intervals in which the UE has to check the control channel, i.e., exactly T_{ln} seconds out of m subframes. The power save time ratio is then defined as follows:

$$R \doteq \left(1 - \frac{T_{ln}}{mT_{sub}}\right) \frac{E[I_0]}{E[T_c]}. \tag{13}$$

Considering that $E[T_c]$ is almost insensible to m and M, but increases with N_u, and recalling that $E[I_0]$ increases with m and decreases with M, we conclude that R is an increasing function of m, and it decreases with M and N_u.

B. Cost analysis

Cost at the UE. Whenever the UE receiver is active, its consumption rate is c_{on}, and $c_{ps} < c_{on}$ otherwise. Decoding a packet has an *additional* consumption rate c_{rx}, while listening to the control channel has an *additional* consumption rate c_{ln}. The average consumption is a combination of these four consumption terms. Using definitions (9) and (13), recalling that control channel listening is performed in each subframe in normal mode, but only in one out of m subframes in power save mode, and taking the average over a system cycle, we obtain the following cost per UE:

$$\begin{aligned} C_{UE}(m, M, N_u) \;=\; & (1-R)\, c_{on} + R\, c_{ps} + \rho\, c_{rx} \\ & + \left(1 - \frac{m-1}{m}\frac{E[I_0]}{E[T_c]}\right)\frac{T_{ln}}{T_{sub}} c_{ln}. \end{aligned} \tag{14}$$

Considering a fixed web traffic profile, the cost is a function of the power save parameters m and M affecting R, ρ, $E[I_0]$, and $E[T_c]$, and of the number of users N_u which appears in $E[T_c]$ and hence in R. The cost with no power save mode is computed by plugging $E[I_0] = 0$, which is equivalent to setting $m = 1$ and $M \to \infty$, in (14):

$$C_{UE}(1, \infty, N_u) \;=\; c_{on} + \rho\, c_{rx} + \frac{T_{ln}}{T_{sub}} c_{ln}. \tag{15}$$

Finally, the relative power save gain that can be attained is:

$$\begin{aligned} G_{UE}(m, M, N_u) \;&\doteq\; \frac{C_{UE}(1, \infty, N_u) - C_{UE}(m, M, N_u)}{C_{UE}(1, \infty, N_u)} \\ &=\; \frac{\gamma(m)}{C_{UE}(1, \infty, N_u)} \frac{E[I_0]}{E[T_c]}, \end{aligned} \tag{16}$$

where the quantity $\gamma(m)$ is a cost reduction factor which increases with the DRX power save cycle length, that is:

$$\gamma(m) \doteq \left(1 - \frac{T_{ln}}{mT_{sub}}\right)(c_{on} - c_{ps}) + \left(1 - \frac{1}{m}\right)\frac{T_{ln}}{T_{sub}} c_{ln}. \tag{17}$$

We can conclude that the relative gain is a function that increases with the duration of the DRX power save cycle (i.e., with m), and decreases with the timeout (i.e., with M) and with the number N_u of users in the cell.

Cost at the eNB. The discontinuous reception and transmission is defined on a per-UE basis, and hence the eNB power save can be expressed as the sum of power save over all users. However, the eNB experiences some additional

cost for cell management (synchronization, pilots, etc.). Hence, the eNB cost per associated UE, namely C'_{UE} is expressed similarly to the UE cost computed earlier in this section, where the reception cost rate c_{rx} is replaced by a transmission cost rate c_{tx}, and the listening cost c_{ln} is replaced by the signaling cost c_{sg}. The additional per-eNB fixed cost c_f does not depend on the transceiver activity and it is normally huge. Recent works show that it can be as high as 10 times the average cost for transmitting data over the air interface [15]. In sum, the total base station cost rate with homogeneous users is:

$$C_{BS}(m, M, N_u) \quad = \quad N_u C'_{UE}(m, M, N_u) + c_f. \quad (18)$$

The relative power save gain is then as follows:

$$G_{BS}(m, M, N_u) = \frac{\gamma'(m)}{C'_{UE}(1, \infty, N_u) + \frac{c_f}{N_u}} \cdot \frac{E[I_0]}{E[T_c]}, \quad (19)$$

where γ' is obtained from (17) by replacing c_{ln} with c_{sg}. Note that with few users the main eNB cost is represented by the fixed cost c_f, thereby the gain increases with the number of users until the per-user cost becomes the predominant term in the denominator of (19).

VI. Evaluation

In this section we first evaluate the model using a packet-level simulator that reproduces the behavior of downlink transmissions. Second, we use the model to perform the optimization of power save parameters m and M in order to minimize the transmission/reception cost, subject to an upper bound for access delay $E[D]$ and download time $E[W]$.

A. Simulating the $G/G/1\,PS$ queue with web traffic

We developed a C++ event-driven simulator that reproduces the behavior of a time slotted $G/G/1\,PS$ queue with N_u homogeneous classes. In the simulator, each class can be in two different operational modes, namely normal mode and power save mode. The shared processor resources are allocated equally to all classes in normal mode at the beginning of each time slot of duration T_{sub}. The traffic is homogeneously generated, in accordance to the 3GPP2 evaluation methodology discussed in Section III. Furthermore, all simulated packets have the same size, i.e., 1500 bytes, and the processor capacity is 1500 bytes per slot. Hence, if only one class is under service, a packet is served completely in one slot. Otherwise, since the processor is shared, all classes in normal mode have a fraction of packet served in that slot. The fair per-class share is computed as one over the number of classes in normal mode. However, if a class has not enough backlog to use all its processor share, unused resources are redistributed amongst the remaining classes. The service process can last one or more time slots per packet, and packet service is considered complete at the end of its last service slot.

Simulations are performed for different numbers of classes N_u, duration of the timeout (through M), and duration of DRX power save cycle (through m). Each simulation consists of a warm-up period lasting 10,000 seconds (5,000,000 slots), followed by 100 runs, each lasting 10,000 seconds. Statistics are separately collected in each run. At the end of a simulation, all statistics are averaged over the 100 runs and 99% confidence intervals are computed for each average result. Simulations have to be run for such a long time to have statistics with relatively small confidence intervals. In fact, due to heavy tailed distributions involved in the generation of web traffic, the number of packets per cycle has a huge variance. Furthermore, simulations with a high number of users require very long CPU time (in our specific case, a single simulation point requires up to 12 hours of a 3 GHz Intel Core™2 Duo E6850 CPU), which makes it prohibitive to explore in detail all possible values of the input parameters. As a reference, our model can be run with the Maple software in as few as 30 seconds on the same machine used for simulations. The model, however, neglects the correlation between the activity of different users, e.g., in the computation of $E[\sigma]$.

However, the comparison between model and simulation shows that the model approximates the system performance with a good accuracy. In particular, here we compare three performance indicators: system cycle duration $E[T_c]$, power save time ratio R, and service time $E[\sigma]$. $E[W]$ could be easily computed from $E[T_c]$. For clarity of presentation, we show only a subset of the results obtained. In particular we selected some extreme cases that well depict the variability of performance with the parameters m, M, and N_u.

Figure 3 compares the estimates of $E[T_c]$ obtained with the model (lines with marks) and with the simulator (marked points) for two very different values of m (4, which is the minimum in the 3GPP recommendations, and 100). The lower part of the figure contains the results obtained with one user, and the upper part reports the results with $N_u = 400$ users. The results of the simulation are highly variable due to the heavy tailed distribution in web page size statistics, hence 99%-confidence intervals appear large over the zoomed y-scale used in the figure. Though the average values show some small difference, both simulations and model behave similarly. The maximum relative difference between model and simulation with one user is within 1%, and it is below 2% with $N_u = 400$. However, model estimates are within the 99%-confidence intervals of simulation estimates. The main cause of the difference between the results of the model and the ones obtained via simulation is in the estimation of the service time, which linearly affects the cycle duration. In fact, by observing Figure 4, it is clear that the model slightly overestimates the service time for high values of N_u, i.e., when the correlation between multiple users, in terms of probability to share the same transmission slot, becomes relevant. As predicted, m and M do not significantly affect

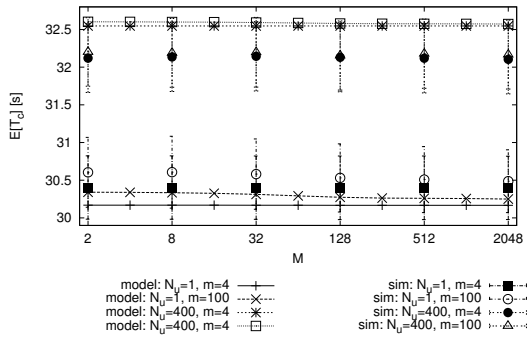

Figure 3. System cycle duration is affected by the number of users. It slightly grows with m and is almost insensitive to M.

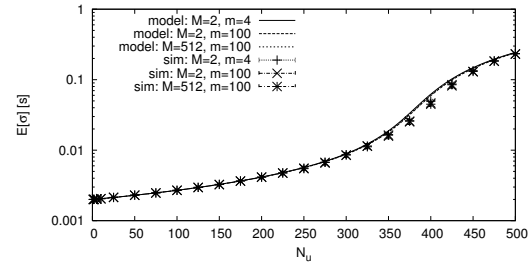

Figure 4. The service time grows with the number of users and is almost not affected by the timeout and the DRX power save cycle durations.

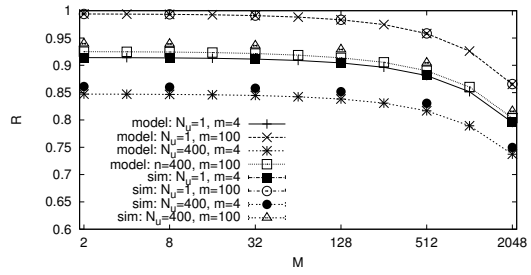

Figure 5. The power save time ratio computed for $T_{ln} = T_{sub}/3$.

$E[\sigma]$. Figure 5 illustrates the power save time ratio R. Model's and simulation's results are very close in all cases, and confidence intervals are very small, so we omitted them in the figure. The results are sensitive to m and N_u, while the effect of M is almost negligible for short timeouts.

In conclusion, simulations suggest that we can safely use the model to estimate the system performance and evaluate its potentialities for power save with good accuracy.

B. Model-based parameter optimization

Here we want to compute the optimal values of m and M that yield the highest gain while keeping low the access delay and the download time. We consider the eNB cost only, but the results can be easily extended to the UE.

Reasonably, the cost for transmitting a data packet is larger than the cost for transmitting a control packet, which usually takes less bandwidth. Both transmitting and signaling costs are much higher than the cost to stay on, which, in turn, is at least one order of magnitude greater than the cost to stay in power save mode. As an example, we use

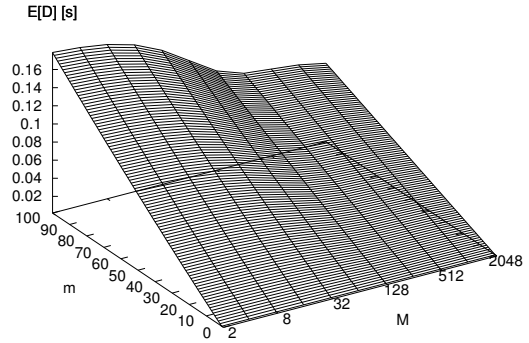

Figure 6. Access delay (independent on the number of users).

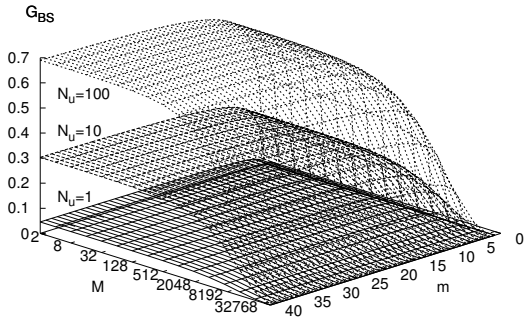

Figure 7. Relevant power save gain can be obtained with small timeouts, even for power save intervals lasting few subframes.

the following values: $c_{tx} = 100$, $c_{sg} = 50$, $c_{on} = 10$, and $c_{ps} = 1$. Additionally, as suggested by experimental measurements [15], we consider a base station cost one order of magnitude higher than the transmission cost: $c_f = 1000$. 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.

The access delay experienced in the network is reported in Figure 6. $E[D]$ is sensitive to m, especially with low timeout values. However, reasonable values of m, e.g., below 20, yield access delay times not higher than 40 ms.

With the chosen cost parameters, the function $\gamma'(m)$—not depicted here for lack of space—grows very fast for small m, but it quickly saturates. In practice, values of m larger than 20 do not give substantial gain advantages with respect to $m = 20$, that is the maximum value suggested by 3GPP for CPC. The relative gain at the eNB is reported in Figure 7 for a few values of N_u. One can notice that low to medium values of the timeout, jointly with moderately high values of m, allow to obtain a relevant gain as soon as the number of users reaches 10. In fact, when few users are attached to the eNB, the main cost figure becomes c_f, which is fixed. However, as shown in Figure 8, if the number of users grows above 350, the gain recedes. In fact, with too many users, the system saturates and the power save opportunities diminish.

Last, Figure 9 shows some particular cases of system optimization. In the figure, D_x and W_x denote the maximum allowable access delay and download time, respectively. Each optimization is performed over m and M, given a fixed number of users N_u. Each optimized value of the gain

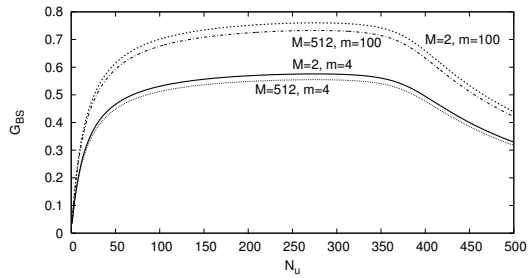

Figure 8. A large gain can be obtained over a wide spectrum of number of users as soon as m grows to few tens.

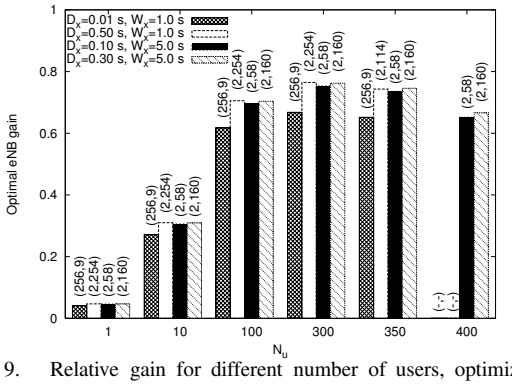

Figure 9. Relative gain for different number of users, optimized over bounded download time and access delay.

is labeled with the pair (M, m) which corresponds to the optimum. The figure shows that the gain can exceed 70% while keeping the access delay bounded to less than half second, and the total web page download time below one second. However, with 400 users, the minimum download time grows above one second and the system cannot be optimized unless W_x was raised to a few seconds. Note also that the optimization with very small values of the access delay can only be obtained by setting a long timeout and short power save intervals (e.g., $M = 256$ and $m = 9$ with 100 users yields a $\sim 60\%$ gain with no more than 10 ms of access delay). With higher access delay bounds, e.g., as high as 100 ms, the optimal timeout is the shortest possible, i.e., $M = 2$. Almost in all cases, the optimization suggests to use very large values for m. However, observing Figure 7, it is clear that near-optimal gain can be obtained with values of m as low as 20.

VII. Conclusions

The paper shows how to model and simulate a $G/G/1\,PS$ system representing the download transmission queues of cellular users adopting the continuous connectivity model. The model, which has been validated through simulation, is based on two basic assumptions: (i) users can receive traffic according to the DRX paradigm, and (ii) the user generated traffic is a realistic sequence of web page requests. We model the per-user activity and evaluate the service share that the base station processor can grant to each user. Furthermore, we propose a cost model and show how to optimize the power save parameters to minimize the cost under bounded access delay and page download time. Remarkably, we show that up to 70% or more of the downlink transmission cost can be saved while preserving the quality of packet flows.

References

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

[2] S. Yang and Y. Lin, "Modeling UMTS discontinuous reception mechanism," *IEEE Transactions on Wireless Communications*, vol. 4, no. 1, pp. 312–319, Jan. 2005.

[3] 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*, vol. 3, Melbourne, Australia, May 2006, pp. 1141–1145.

[4] J. Seo, S. Lee, N. Park, H. Lee, and C. Cho, "Performance analysis of sleep mode operation in IEEE 802.16e," in *Proc. of IEEE VTC 2004-Fall*, vol. 2, Los Angeles, California, USA, Sep. 2004, pp. 1169–1173.

[5] 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*, Annapolis, Maryland, Jun. 2008.

[6] Y. Xiao, "Performance analysis of an energy saving mechanism in the IEEE 802.16e wireless MAN," in *Proc. of IEEE CCNC*, vol. 1, Jan. 2006, pp. 406–410.

[7] 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*, vol. 19-23, Beijin, China, May 2008, pp. 322–326.

[8] H. Choi and J. Limb, "A behavioral model of web traffic," in *ICNP'99*, Washington, DC, USA, Oct. 1999.

[9] 3GPP TS 25.214, "Physical layer procedures (FDD), rel. 8."

[10] 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, Cananda, Sep. 2008, pp. 1–5.

[11] C. Bontu and E. Illidge, "DRX mechanism for power saving in LTE," *IEEE Communications Magazine*, vol. 47, no. 6, pp. 48–55, Jun. 2009.

[12] 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, pp. 713–717.

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

[14] 3GPP2 C.R1002-B v1.0, "CDMA2000 evaluation methodology - Revision B," Dec. 2009.

[15] 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*, vol. 16, no. 4, Nov. 2008.