Analysis of power save and its impact on web traffic in cellular networks with continuous connectivity

Sara Alouf, Vincenzo Mancuso, Nicaise Choungmo Fofack

Analysis of power save and its impact on web traffic in cellular networks with continuous connectivity*

Sara Alouf,[1] Vincenzo Mancuso,[2] and Nicaise Choungmo Fofack[1,3]
[1] INRIA, Sophia Antipolis, France
[2] Institute IMDEA Networks, Madrid, Spain
[3] University of Nice Sophia Antipolis, France

Abstract

In this work, 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. Our study shows that dramatic energy save can be achieved by mobile devices and base stations, e.g., as much as 70%-90% of the energy cost in cells with realistic traffic load and the considered parameter settings.

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

1 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 [16]. 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 wideband 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 [20, 12] by means of a semi-Markov chain model. The authors of [18] propose 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. The authors of [4] 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. Using Laplace-Stieltjes Transform and Probability Generating Functions, [13] derives closed form expressions for the average power consumption (objective) and the average packet delay (constraint) for an UE. The authors of [13] also design a sleep mode mechanism based on traffic estimation and a solution of the optimization problem. 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

*This manuscript is an author version and an extension of [15]

[19]. Almhana *et al.* provide an adaptive algorithm that minimizes energy subject to QoS requirements for delay [3]. The works [5, 6] closely relate to our proposal and mainly focus on the analysis of the discontinuous reception mode in 3GPP LTE and IEEE 802.16m respectively. The authors consider both the uplink and downlink packets for the UE and show that uplink packets increase the power consumption and decrease the delay.

The existing work does not tackle 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) [1].

The importance of DRX has been addressed in [21], 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 [7] by means of a simple cost model. In [14], the authors use heuristics and simulation to show the importance of DRX for the UE.

The contributions of our work are as follows: (*i*) we are the first to provide a complete model for the behavior of users (UEs) and base stations (eNBs) in continuous connectivity and with non-Poisson traffic (namely web traffic), (*ii*) we provide a cost model that incorporates the different causes of operational costs, (*iii*) we validate our model using packet-level simulations, (*iv*) we study the importance of a variety of model parameters by means of a *sensitivity analysis*, and (*v*) we show how to use the model to minimize operational costs under QoS constraints. Our results confirm that a tremendous cost reduction can be attained by correctly tuning the power save parameters. In particular, transmission costs can be lowered by more than 90% with realistic traffic loads.

This article extends our work published in the proceedings of IEEE WoWMoM 2011 [15]. Compared with the conference paper, we implemented the following modifications and additions: (*i*) We have done a research review of recent works and have amended the related work section by adding three new references. (*ii*) The presentation of the analytical model has been improved as some equations/derivations have been explicitly written. The cost model has also been refined impacting all the numerical results which rely on it. (*iii*) We have performed simulations in which each user has p parallel browsing sessions; the aim is to evaluate whether our study can be used when each user's traffic consists of superposed arrival processes. (*iv*) A sensitivity analysis has been performed. We provide both first order and total sensitivity indices and comment on the implications and the interpretations of these indices. (*v*) The numerical analysis has been revisited and expanded with new numerical results. (*vi*) A "lessons learned" section has been added, summarizing our recommendations and suggesting a setup which achieves a good tradeoff between energy savings and QoS performance.

The rest of the manuscript is organized as follows: Section 2 presents power save operations in continuous connectivity mode. Section 3 describes a model for cellular users generating web traffic. Section 4 illustrates a model for downlink transmissions, and Section 5 describes how to evaluate flow performance and transmission costs. In Section 6 we validate the model through simulation. A sensitivity analysis is performed in Section 7, and a performance analysis and optimization is done in Section 8 showing the achievable power saving. Section 9 concludes the article.

2 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 [10]. A highly efficient power save mode operation is then strongly required, which would allow disabling both transmission and reception of frames during

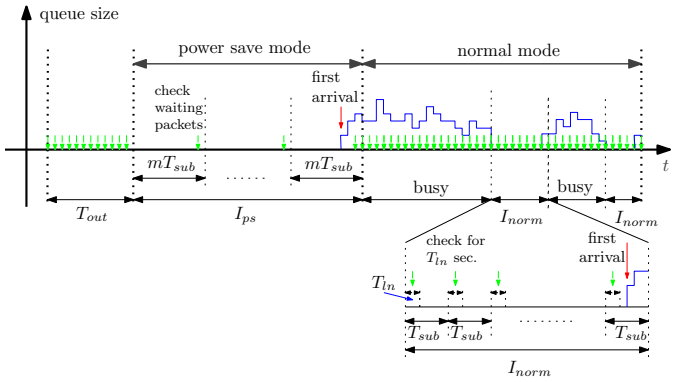


Figure 1: Downlink queue activity with power save and normal operation.

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 short (few subframes) and is used when some data activity is present in the uplink (normal operation). The second cycle is longer (up to tens of subframes) and is activated when an inactivity timer in the uplink data channel expires (power save mode operation). The threshold M for inactivity period is a power of 2 subframes (specified values are in $\{2^1, 2^2, \ldots, 2^9\}$).

Discontinuous reception. DRX is an operational mode defined by 3GPP release 6. It allows the UE to save energy while monitoring the control information transmitted by the eNB. DRX affects data delivery, since no data can be dependably received without an associated control frame. 3GPP specifications define a DRX 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 cycles of 8 to 40 ms). DRX is activated only upon a timeout after the last downlink transmission, and like DTX, the timeout threshold M specified in the standard is a power of 2 subframes.

3 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 [10]. 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. Indeed, the overall cost 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 checks the control channel at each subframe (normal operation mode); instead, upon the

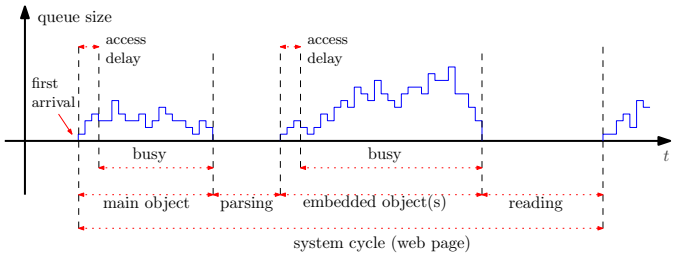


Figure 2: System cycle with web traffic as defined in [2].

expiration of T_{out} (inactivity timer 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 monitors 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 4 we model the behavior of downlink transmissions with DRX operations enabled and users generating web traffic. Then, in Section 5 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 [10]. 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 [2]: 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 *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 1 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 1: $\psi_0 = P(y_{\min} \leq Y < y_{\min} + 1) = 1 - \left(\frac{y_{\min}}{y_{\min}+1}\right)^{\alpha}$. 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 a 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

[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.

Table 1: Parameters suggested by 3GPP2 for the evaluation of web traffic

Quantity	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 $N_{eo} = \lfloor Y \rfloor - y_{\min}$	$f_Y(y) = \begin{cases} \alpha \frac{y_{\min}^{\alpha}}{y^{\alpha+1}}, & y \in [y_{\min}, y_{\max}[\\ 1 - \left[\frac{y_{\min}}{y_{\max}}\right]^{\alpha}, & y = y_{\max} \end{cases}$	$y_{\min} = 2$ $y_{\max} = 55$ $\alpha = 1.1$
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	$f_{\Lambda_r}(t) = \lambda_r e^{-\lambda_r t}$, $t \geq 0$	$\lambda_r = 0.03$ s
Parsing time Λ_p	$f_{\Lambda_p}(t) = \lambda_p e^{-\lambda_p t}$, $t \geq 0$	$\lambda_p = 7.69$ s

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.

4 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.

Timeouts in a cycle. Cycles always include one reading time, but the parsing time is present only if there are embedded objects (i.e., with probability $1 - \psi_0$). The average number of timeouts in a cycle is then:

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

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:

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

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

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

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

$$E[I_0] = \beta_r^{M-1} \frac{mT_{sub}}{1-\beta_r^m} + (1-\psi_0)\,\beta_p^{M-1} \frac{mT_{sub}}{1-\beta_p^m}. \tag{2}$$

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 and its expectation are as follows:

$$\begin{aligned} P\big(I_1 = jT_{sub}|\text{reading}\big) &= P\Big(I_{norm} = jT_{sub}\Big|\ \text{exp. arrivals with rate } \lambda_r\Big) \\ &= \begin{cases} \beta_r^{M-1}, & j=0; \\ \beta_r^{j-1}\,(1-\beta_r)\,, & 1 \le j \le M-1; \end{cases} \\ E[I_1|\text{reading}] &= T_{sub}\frac{1-M\beta_r^{M-1}+(M-1)\beta_r^M}{1-\beta_r}. \end{aligned} \tag{3}$$

Similarly, the expected value of the time spent in normal mode with no traffic to be served during parsing, not counting the timeout, is given by

$$E[I_1|\text{parsing}] = T_{sub}\frac{1-M\beta_p^{M-1}+(M-1)\beta_p^M}{1-\beta_p}. \tag{4}$$

Therefore, the average duration of I_1, attained by using (3) and (4), is an increasing function of the timeout duration, as expressed by the following formula:

$$E[I_1] = T_{sub}\left[\frac{1-M\beta_r^{M-1}+(M-1)\beta_r^M}{1-\beta_r} + (1-\psi_0)\frac{1-M\beta_p^{M-1}+(M-1)\beta_p^M}{1-\beta_p}\right]. \tag{5}$$

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:

$$E[I] = \frac{\beta_r^{M-1} mT_{sub}}{1-\beta_r^m} + T_{sub}\frac{1-\beta_r^{M-1}}{1-\beta_r} + (1-\psi_0)\left[\frac{\beta_p^{M-1} mT_{sub}}{1-\beta_p^m} + T_{sub}\frac{1-\beta_p^{M-1}}{1-\beta_p}\right]. \tag{6}$$

$E[I]$ 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:

$$\frac{1}{\lambda_r} + \frac{1-\psi_0}{\lambda_p} < E[I] < \frac{1}{\lambda_r} + mT_{sub} + (1-\psi_0)\left(\frac{1}{\lambda_p} + mT_{sub}\right). \tag{7}$$

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 in a cycle. It is the time spent to serve the packets of a web page. Its expectation is the expected number of packets per web page, $E[N_p]$, times the expected service time $E[\sigma]$. The number of

packets depends on the distribution of the web page objects. Assuming the 3GPP2 traffic model reported in Table 1 and 1500-byte long packets, we can compute:

$$E[N_p] = E\left[\left\lceil \frac{S_{mo}}{1500} \right\rceil\right] + E[N_{eo}] \cdot E\left[\left\lceil \frac{S_{eo}}{1500} \right\rceil\right] = 39.476.$$

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 cycle duration is:

$$E[T_c] = E[I] + E[N_p]E[\sigma]. \tag{8}$$

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:

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

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 $0 < \rho < 1$.

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 all backlogged active 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:

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

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

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

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. *The expected packet service time is*

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

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 $N_u T_{sub}$, 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 on $E[T_c]$, and can be used to limit the maximum number of customers, thus guaranteeing a maximum web page processing time to any customer.

5 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.

5.1 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:

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

Considering (7) and (8), a tight lower bound on the expected page download time is $E[N_p]E[\sigma] + (1 - \psi_0)/\lambda_p$.

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. Let D be the total access delay experienced within a web page download, accounting for the delay accumulated in both reading and parsing times. $E[D]$ can be easily computed by subtracting parsing, reading, and busy times from the expected system cycle duration (see Figure 2), i.e.:

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

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} during which no transmissions occur. However, in the interval I_0, the UE has to periodically be active to listen to the control channel for exactly $T_{ln} \leq T_{sub}$ seconds out of m subframes. The power save time ratio is then defined as follows:

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

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

5.2 Cost analysis

Cost at the UE. The basic consumption rate of the UE receiver is c_{on} if active and $c_{ps} < c_{on}$ otherwise. Receiving a packet *increases* the basic consumption rate by c_{rx}, while listening to the control channel *increases* it by c_{ln}. The average consumption is a combination of these four consumption terms. For sake of generality we assume that listening to the control channel can last differently, depending on whether data are associated to the control message or not. For instance, in HSPA systems, the user can switch from control to data channel after having decoded the initial part (one third) of the control frame indicating the arrival of a new data frame [10]. We denote by T_{ln} the listening time when no data are transmitted, and by T'_{ln} the listening time when data follow the control message. Therefore, using definitions (9) and (13), and 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, we can compute the cost per UE by taking the average over a system cycle while keeping separated the listening occurrences with and without associated data transmissions. Namely,

$$C_{UE}(m,M,N_u) = (1-R)c_{on} + Rc_{ps} + \rho c_{rx} + \left[\frac{\left[1-\rho-\frac{m-1}{m}\frac{E[I_0]}{E[T_c]}\right]T_{ln}}{T_{sub}} + \frac{\rho T'_{ln}}{T_{sub}} \right] c_{ln}. \tag{14}$$

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):

$$C_{UE}(1,\infty,N_u) = c_{on} + \rho\, c_{rx} + \left[(1-\rho)\frac{T_{ln}}{T_{sub}} + \rho\frac{T'_{ln}}{T_{sub}} \right] c_{ln}. \tag{15}$$

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

$$G_{UE}(m,M,N_u) \triangleq \frac{C_{UE}(1,\infty,N_u) - C_{UE}(m,M,N_u)}{C_{UE}(1,\infty,N_u)} = \frac{\gamma(m)E[I_0]/E[T_c]}{C_{UE}(1,\infty,N_u)}, \tag{16}$$

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

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

Note that T'_{ln} does not affect the cost reduction (numerator of (16)).

Summarizing, 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 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 is

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

where $C'_{UE}(m,M,N_u)$ is the cost per time unit to transmit to a single UE. It can be written as follows:

$$C'_{UE}(m,M,N_u) = C'_{UE}(1,\infty,N_u) - \gamma'(m) \cdot \frac{E[I_0]}{E[T_c]}, \tag{19}$$

with

$$C'_{UE}(1,\infty,N_u) = c_{on} + \rho\, c_{tx} + \frac{T'_{ln}}{T_{sub}} c_{sg}; \tag{20}$$

$$\gamma'(m) = \left(1 - \frac{T'_{ln}}{mT_{sub}}\right)(c_{on} - c_{ps}) + \left(1 - \frac{1}{m}\right)\frac{T'_{ln}}{T_{sub}} c_{sg}. \tag{21}$$

Here, c_{tx} is a transmission cost rate and c_{sg} is a signaling cost. Last, the relative power save gain is:

$$G_{BS}(m, M, N_u) = \frac{\gamma'(m)}{C'_{UE}(1, \infty, N_u) + \frac{c_f}{N_u}} \cdot \frac{E[I_0]}{E[T_c]}. \tag{22}$$

Note that with few users the main eNB cost is represented by the fixed cost c_f. Hence, the gain increases with the number of users until the per-user cost becomes the predominant term in the denominator of (22).

6 Validation through simulations

In this section we evaluate the robustness of the model by comparing the analytical results to simulations. The main assumption used in the model states that queues related to different active UEs are i.i.d.; however, queues are correlated in practice as they share the same processor. This assumption is not met in the simulations.

We developed a $C{+}{+}$ packet-level 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 suggested web traffic model of Table 1. 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. If a class has not enough backlog to use all its processor share, unused resources are redistributed among the remaining classes. 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 M, and length of DRX power save cycle m. Hereafter, we will use $\lambda_r = 1/30$ s, $\lambda_p = 1/0.13$, $\psi_0 = 1 - (2/3)^{1.1}$, and $T_{sub} = 2$ ms. 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.

We need to run simulations 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$^{\text{TM}}$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 $E[\sigma]$. Nevertheless, the comparison between model and simulation shows that the model approximates the system performance with a good accuracy. Numerical results for $E[\sigma]$, obtained from both the model and the simulations, are reported in Figure 3. It is clear from the figure 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 $E[\sigma]$.

We now compare two performance metrics: the system cycle duration $E[T_c]$ and the power save time ratio R. $E[W]$ can be easily computed from $E[T_c]$; cf. (11). 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 m, M, and N_u.

Figure 4(a) compares the estimates of $E[T_c]$ obtained with the model (lines) 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

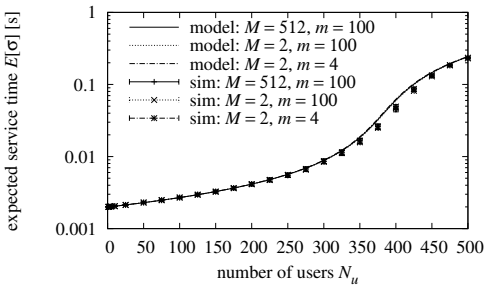


Figure 3: $E[\sigma]$ grows with N_u and is almost not affected by the timeout and the DRX power save cycle durations.

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 $N_u = 400$. Noticeably, model estimates are within the 99%-confidence intervals of simulation estimates.

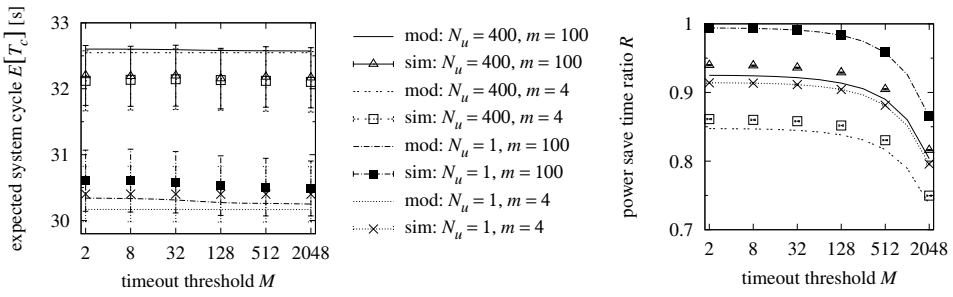
The main cause of the difference between the results of the model and the ones obtained via simulation is in the estimation of the service time, which linearly affects the cycle duration. Similar differences can be observed for the power save time ratio R with $N_u = 400$ in Figure 4(b). Analytic and simulation results remain however very close. 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.

6.1 Impact of parallel user's browsing sessions

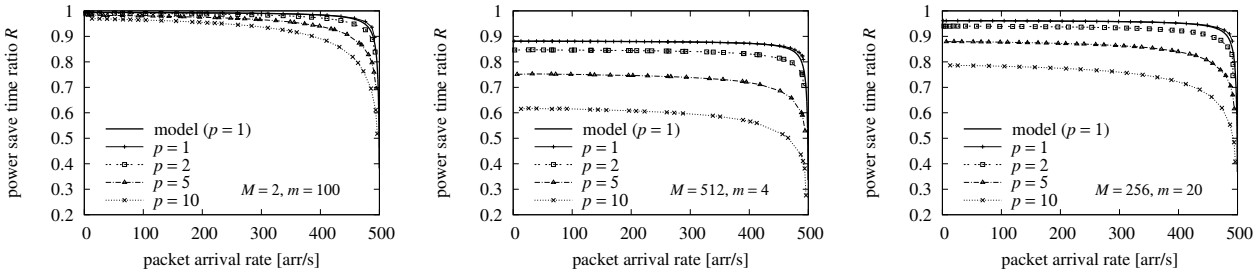
In real life, a user can activate more than one browsing window and switch from window to window while a page is being loaded. Thus, in practice it is not uncommon to have more than one browsing session active on the same device. Therefore, in that case, the arrival process at the user's download queue will result from the superposition of various per-browsing session arrival processes. Here we simulate the occurrence of multiple active http browsing sessions for each user, and we compare the performance with the case of single browsing session. Our model does not capture the effect of parallel http sessions, hence the experiments proposed in this subsection are aimed at evaluating whether our study can be used to approximate the network behavior in more generic and realistic traffic scenarios. Specifically, we focus on one particular metric, namely the power save time ratio, since it is representative of the system's power save performance.

In Figure 5, we plot the power saving time ratio R for three scenarios: a configuration for the DRX parameters (M, m) yielding high power save (Figure 5(a)), a configuration yielding the minimum power



(a) Expected system cycle duration $E[T_c]$ (b) Power save time ratio R

Figure 4: Comparison of analytic and simulation results: (a) $E[T_c]$, and (b) R.



(a) high save configuration (b) low save configuration (c) recommended setup

Figure 5: Impact of the number p of parallel user's browsing sessions on R, for $T_{ln} = \frac{T_{sub}}{3}$.

save for realistic values of (M, m) (Figure 5(b)), and the configuration that we recommend in light of our optimization analysis reported in Section 8, i.e., $(M, m) = (256, 4)$ (Figure 5(c)). The recommended configuration, yields a good tradeoff between power save and serving delay incurred by the packets due to DRX operations.

Note that, since a user generates a traffic volume which depends on the number p of parallel http browsing sessions, in Figure 5 we plot R as a function of the offered load, expressed in terms of packet arrivals per second. For each represented curve, we change the load by changing the number of users N_u, and report the corresponding arrival rate in the x-axis, and the power save time ratio R in the y-axis. As reference, we include in each figure the results obtained with the model by increasing N_u from 1 to 1000, then computing $E[\sigma]$ by solving the system consisting of Eqs. (9) and (10), or with the formula given in the Corollary in Section 4, for each given value of N_u, and eventually computing the load factor as $N_u E[N_p] T_{sub}/E[T_c]$. The latter formula represents the fraction of time spent in a cycle to serve the average aggregate volume of downlink packets $N_u E[N_p]$ generated in that cycle, when the volume of data corresponding to one packet is served in exactly one subframe T_{sub}.[2]

In the model, the load in packet arrivals per second is computed by scaling the load factor by $\frac{1}{T_{sub}}$, which is the maximum number of packets that can be served in a second, and is 500 in our case, corresponding to the capacity of a HSPA downlink with $2-$ms subframes. Clearly, a given arrival rate corresponds to a different number of users N_u when p changes, and the relation between the packet arrival rate, the number of browsing sessions p, and the number of users N_u cannot be predicted with our model. Therefore, for $p > 1$ we only show simulation results.

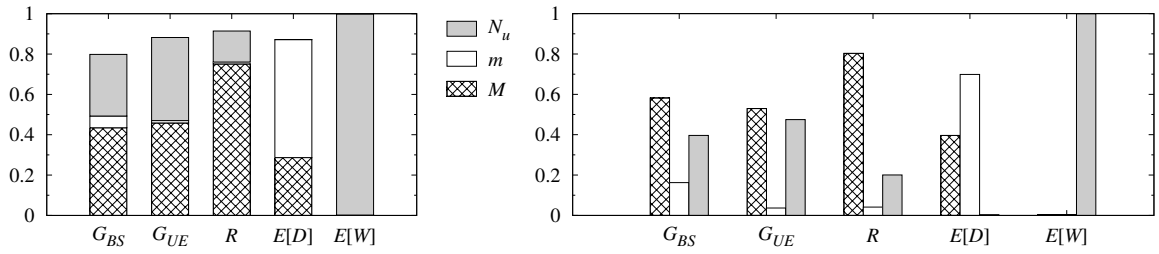
Observing Figure 5, one can notice that (i) the model accurately predicts the simulation for $p = 1$, and (ii) values of p as large as 10 can have a remarkable impact on the power save time ratio R. However the impact of p is important only for high loads and for large values of the DRX timeout M, causing up to a $\sim 25\%$ drop in power save opportunities. However, for reasonable values of p, e.g., 2 to 5, R remains always very high, and within a few percent from the value achieved with $p = 1$. In light of this result, we argue that using our model can suitably approximate the computation of the power save opportunities of a system with users browsing a few (up to 5) web pages in parallel.

We will now perform a sensitivity analysis on our model to evaluate which parameters mostly affect the performance metrics.

7 Sensitivity analysis

In the previous sections, we gave the expressions of the performance and cost metrics that enables us, by a partial derivation, to outline a preliminary behavior of our metrics according to the input parameters, namely M, m and N_u. Our objective now is to characterize qualitatively and quantitatively the impact of our input parameters on the variability of our metrics. We will further analyze the sensitivity of the metrics when the expected web page size $E[N_p]$, the expected reading rate λ_r, and the expected parsing rate λ_r are uncertain, in addition to the three input parameters.

[2]Equivalently, the load factor can be computed as the sum of N_u activity factors expressed as in Eq.(9), multiplied by a coefficient $T_{sub}/E[\sigma]$ which represents the fraction of resources allotted to a user when sharing the processor with other users.



(a) First order sensitivity indices S_i

(b) Total order sensitivity indices S_i^T

Figure 6: Sensitivity indices of M, m and N_u for the defined metrics.

Performing a *sensitivity analysis* is evaluating how variability in the output of a model can be apportioned to different input parameters. Variance-based techniques define sensitivity indices (i) to measure the main effect of a given input on the output, (ii) to measure the relative importance of any combination of input in the output variability, and (iii) to measure the total effect of a given input on the output. More precisely, assuming the inputs to be random variables $X_1, \ldots, X_n$, and the model output to be a random variable $Y = f(X_1, \ldots, X_n)$, the first order and total sensitivity indices for random variable X_i are defined as follows:

$$S_i = \frac{\mathrm{Var}\Big(E[Y|X_i]\Big)}{\mathrm{Var}(Y)}, \qquad S_i^T = \frac{E\Big[\mathrm{Var}(Y|X_{-i})\Big]}{\mathrm{Var}(Y)}, \tag{23}$$

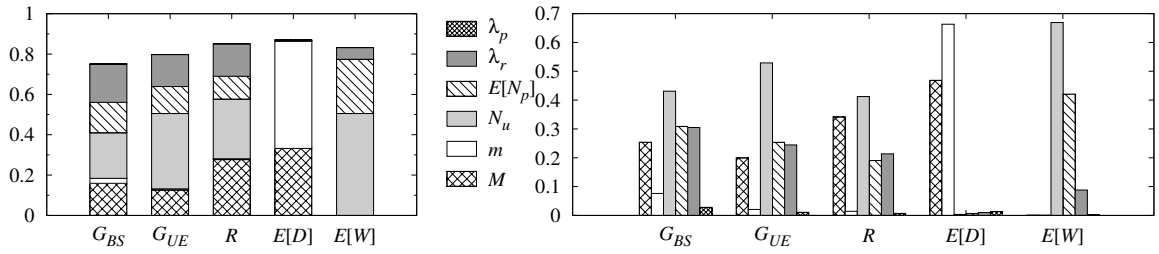
where X_{-i} denotes all input random variables except X_i. S_i is a quantitative measure of the main effect of X_i on output Y (through its variance) and S_i^T is a quantitative measure of the total effect of X_i, including the interactions with other input random variables. The difference $S_i^T - S_i$ measures the importance of interactions in the total effect of X_i. When there are no interactions between the input random variables, the sum $\sum_{i=1}^{n} S_i = 1$; otherwise, this sum is less than 1. If S_i^T is small, then this means that the value of X_i is not essential, and it can be considered as deterministic, taking any value within its range, without any significant impact on the model output. Note that an exhaustive sensitivity analysis requires to compute $2^n - 1$ sensitivity indices, including those accounting for interactions between any combination of input random variables. The sum of these $2^n - 1$ indices amounts to 1.

One method for estimating S_i and S_i^T for non-correlated variables is the Extended Fourier Amplitude Sensitivity Test (EFAST), introduced by Saltelli et al. in 1999 [17]. The EFAST method does not require any knowledge on the function $f(.)$, which can be seen as a black box. The advantages of EFAST are its robustness, especially at small sample size, and its computational efficiency. EFAST expands the output of the model by using the Fourier Series, then assigns an integer frequency to each input parameter, to finally compute the variance of output as well as the contribution of each input to this variance. Using a brute-force approach, computing S_i and S_i^T requires to evaluate a multidimensional variance integral. The main advantage of EFAST is to reduce the computation of this complex integral to a monodimensional integral over a curve exploring the n-dimensional space. For a detailed description of the method we refer to [17, 9].

We will now show the results of the sensitivity analysis (SA) for the five performance and cost metrics introduced in Sections 4 and 5. We have checked our results with two different softwares that implement SA.

7.1 SA results with three input parameters: M, m and N_u

We first apply the EFAST method to our model with the web traffic configuration specified by 3GPP2 (see Table 1). We consider the following ranges for the three input parameters: $M \in \{2, 2^2, \ldots, 2^{15}\}$; $m \in \{1, \ldots, 50\}$ and $N_u \in \{1, \ldots, 600\}$. All other parameters are constant in this analysis. We compute the first order (S_i) and the total (S_i^T) sensitivity indices of each of the parameters M, m and N_u for the five performance and cost metrics defined in Section 5. The results are displayed in Figure 6. It is clear that parameters with small impact on some metric may well be essential for other metrics. We can make the following observations:



(a) First order sensitivity indices S_i

(b) Total order sensitivity indices S_i^T

Figure 7: Sensitivity indices of $M, m, N_u, \lambda_r, \lambda_p$ and $E[N_p]$ for the defined metrics.

- The download time $E[W]$ is affected only by the number of cell users N_u; the DRX parameters M and m may take any value within their range without impacting $E[W]$.
- The cycle length m is essential for the access delay $E[D]$ and has a minor effect on the eNB's relative gain G_{BS}. Noticeably, about two thirds of the total effect of m on G_{BS} comes from interactions with other variables.
- The timeout threshold M is the most relevant parameter as concerns the power save time ratio R and the gains G_{UE} and G_{BS}. The second input parameter affecting mostly these metrics is N_u.
- Last, interactions between multiple variables are mostly relevant for the gain G_{BS}.

7.2 SA results with six input parameters: $M, m, N_u, \lambda_r, \lambda_p$ and $E[N_p]$

As the Internet (and so the web) is evolving very fast, it is easy to predict that the traffic parameters suggested by 3GPP2 (see Table 1) will have to be modified. Therefore, power save performance will change accordingly, and network optimization will require a different setup. In particular, the actual trend for mobile devices is to increase memory and data processing speed; meanwhile, the web page sizes tend to increase because of the embedded objects, some of which are large images/videos or heavy scripts. Furthermore, some websites offer light versions of web pages specifically for mobile clients. To give an insight on the relevance of these changes, we now present the results of our sensitivity analysis extended to the model parameters that characterize the user traffic behavior, namely the reading and parsing time, through λ_r and λ_p, and the web page average size $E[N_p]$.

In our sensitivity analysis, we consider the following ranges of variability for the three additional parameters: $E[N_p] \in \{20, \ldots, 100\}$, $\lambda_r \in [0.02, 0.1]$, and $\lambda_p \in [1, 50]$. The selected ranges include the original 3GPP2 parameters, and account for reasonable parameter modifications. For the resulting 6-parameter SA of our model, Figure 7 shows the first order and total sensitivity indices for cost and performance metrics. The following is observed.

- The parsing rate λ_p is definitely unessential (this is mainly due to the negligible value of the average parsing time compared to the other durations) and can be fixed to any value within its range.
- The observation on $E[D]$ remains unchanged: it is only impacted by the DRX parameters M and m (including their interactions).
- Interactions between multiple variables play a more important role than in the SA with three input parameters.
- $E[N_p]$ and λ_r are equally relevant as concerns G_{BS}, G_{UE} and R as they have almost the same total sensitivity index.
- The download time $E[W]$ is still mostly affected by N_u, but it is also impacted by the web page size $E[N_p]$ and to a lesser extent by the reading rate λ_r.

Our analysis reveals that λ_r and $E[N_p]$ are essential for our model. It is recommended to accurately estimate λ_r and $E[N_p]$ before using the model to optimize the power save configuration in the network.

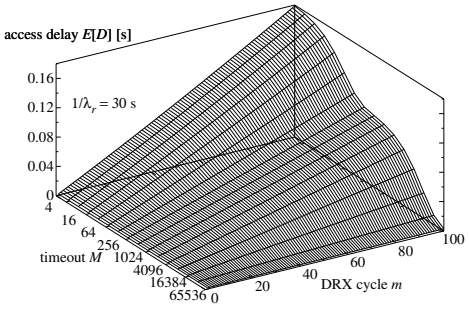
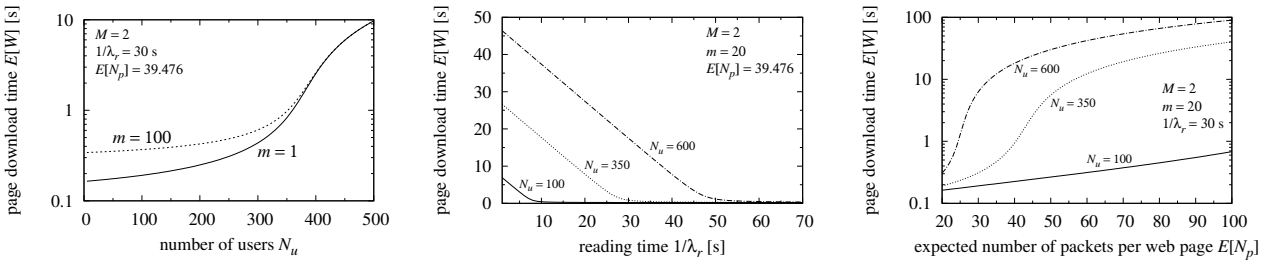


Figure 8: Access delay (independent of N_u).



(a) $E[W]$ grows with N_u (b) $E[W]$ decreases with reading time (c) $E[W]$ increases with $E[N_p]$

Figure 9: The expected page download time is insensible to DRX cycle length m and timeout M; it is roughly $E[\sigma]E[N_p]$ (tight lower bound).

8 Performance analysis and optimization

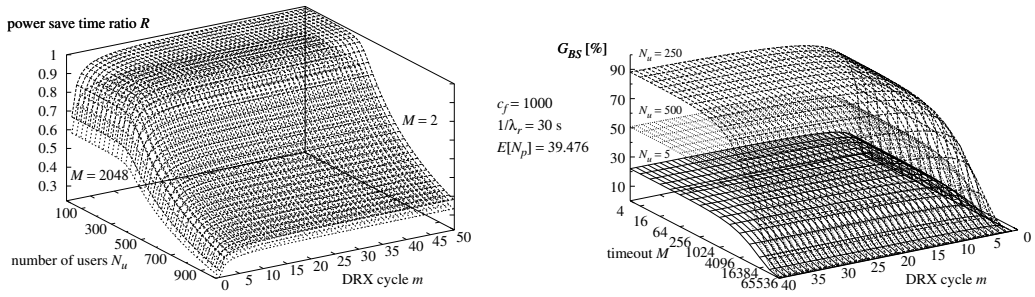
The section focuses on the analysis of the performance and on its optimization, using the model developed in Section 4 and validated in Section 6. Where not specified, we use the traffic parameters reported in Table 1.

Access delay. The access delay is the performance metric mostly impacted by the tunable parameters M (timeout threshold) and m (DRX cycle length), as confirmed by the sensitivity analysis. The access delay experienced in the network is reported in Figure 8 for the parameter set given in Table 1. $E[D]$ 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. As for the timeout threshold, an interesting value is $M = 256$ (see shape of $E[D]$ around $M = 256$ in Figure 8).

Page download time. Figure 9 depicts the behavior of the expected page download time when one of the following terms is varied: (a) the number of users in the cell N_u; (b) the expected reading time $1/\lambda_r$ (user's behavior); and (c) the expected web page size in packets $E[N_p]$. The following is observed. The page download time is small as long as $N_u < 350$. For a larger number of users, $E[W]$ increases abruptly (and linearly) with N_u. The value of m has only a negligible impact on the page download time: the latter is sensibly the same whatever the value of m. The same is observed concerning the parameter M (not reported here). Also, $E[W]$ increases with the reading rate λ_r as can be inferred from Figure 9(b). Indeed, longer reading times lower the load on the shared processor, thereby decreasing the expected service time and consequently the page download time. Last, longer web pages (this is a trend currently observed due mainly to large embedded objects and heavy scripts) yield longer download times.

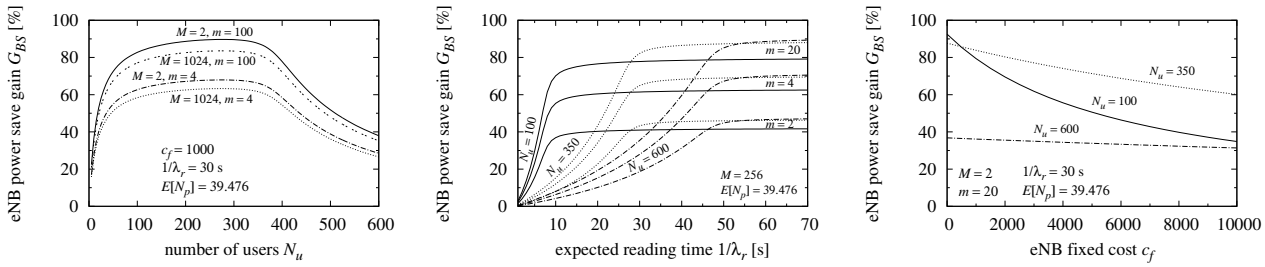
Power save time ratio. We now consider the power save time ratio R. The most impacting parameters are m and N_u. We report the analytical results in Figure 10(a). The power save time ratio R saturates quickly with m. It is sensibly the same for a large range of number of users values but decreases as soon as $N_u > 350$.

Relative power save gain at eNB. Reasonably, the cost for transmitting a data packet is larger



(a) Great save opportunities with $m \geq 4$ (b) Relevant gain with $M \leq 1024$ and $m \geq 4$

Figure 10: Power save time ratio vs. m and N_u (a) and eNB gain vs. m and M (b).



(a) Large gain for many N_u's if $m \geq 4$ (b) Gain saturates as λ_r decreases (c) Smaller c_f yields a larger gain

Figure 11: G_{BS} vs. the number of cell users N_u, the reading time $\frac{1}{\lambda_r}$ and eNB's fixed cost.

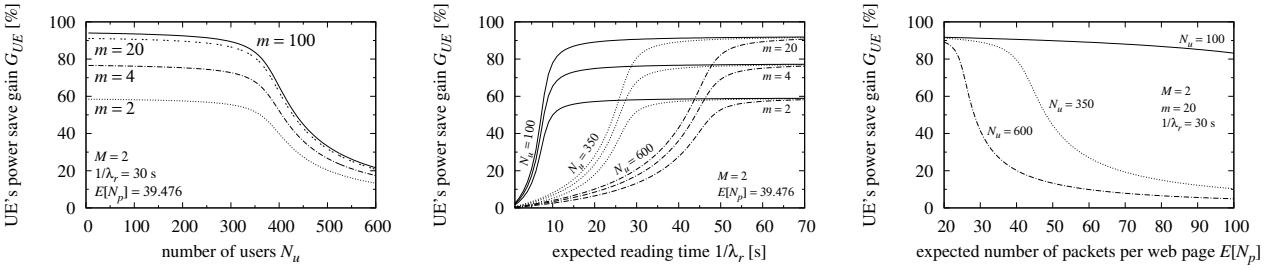
than the cost for transmitting a control packet, which usually takes less bandwidth. Both transmitting and signaling costs are much higher than the cost to stay on, which, in turn, is at least one order of magnitude greater than the cost to stay in power save mode. As an example, we use the following values: $c_{tx} = 100$, $c_{sg} = 50$, $c_{on} = 10$, and $c_{ps} = 1$. Additionally, as suggested by experimental measurements [11], we consider a base station cost one order of magnitude higher than the transmission cost: $c_f = 1000$. In the following $T_{ln} = T_{sub}/3$ and $T'_{ln} = T_{sub}$.

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 10(b) 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 most of the potential gain for the current value of N_u. Observe that when few users are attached to the eNB, the main cost figure is c_f, which is fixed. However, as shown in Figure 11(a), if the number of users grows beyond 350, the gain recedes. In fact, with too many users, the system saturates and the power save opportunities diminish (cf. Figure 10(a)).

We have investigated the effect of the expected reading time $1/\lambda_r$ on the eNB gain. The results are depicted in Figure 11(b) for various values of N_u and m, the timeout M being fixed to 256. We observe that the relative gain at the eNB saturates as soon as the reading time reaches some value (which depends on the number of users). The saturation level of G_{BS} depends on the cycle length m. It is clear that small variations around the current recommended simulation value, that is equal to 30 seconds, will not affect the gain in cells with a moderate number of users (say $N_u \leq 300$). Decreasing the expected reading time in very large cells will yield less gain.

We now vary c_f. It is expected to obtain smaller relative power save gain should the fixed cost at the eNB be larger, and vice-versa (larger gain if smaller fixed cost). This is observed in Figure 11(c).

Relative power save gain at UE. The last metric we analyzed is the user's relative power save gain.



(a) Lesser gain with larger capacity cell (b) More gain if load decreases (c) Larger web pages impair user gain

Figure 12: Analytical evaluation of the relative power save gain at UE.

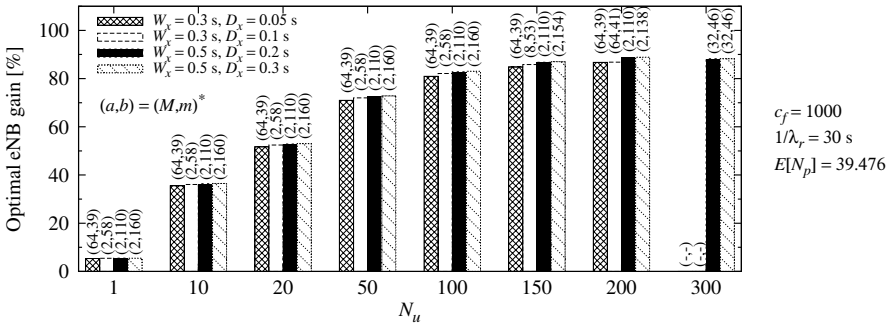


Figure 13: Relative gain for different number of users, optimized over bounded download time and access delay.

We have varied successively the number of users N_u, the reading time $1/\lambda_r$, and the web page size (in packets) $E[N_p]$. Results are reported graphically in Figure 12 for timeout threshold $M = 2$. Substantial user power save gain is possible if $m \geq 20$. Having timeout thresholds longer than 2 subframes slightly decreases the user gain (about 2% loss in G_{UE} if $M = 256, 1/\lambda_r = 30$ s, and $E[N_p] = 39.476$).

DRX protocol parameter optimization. We want to compute now 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.

Figure 13 shows some particular cases of system optimization. In the figure, W_x and D_x denote the maximum allowable download time and access delay, respectively. Each optimization is performed over M and m, given a fixed number of users N_u. Each optimized value of the gain is labeled with the pair (M, m) corresponding to the optimum. The figure shows that the gain exceeds 70% in cells with at least 50 users, while keeping the total web page download time bounded to less than 0.3 s, and the access delay below 0.05 s. However, with 300 users, the minimum download time grows above 0.3 s and the system cannot be optimized unless W_x was raised to at least 0.44 s. Note also that the optimization with very small values of the access delay (e.g., $D_x = 0.05$ s) can only be obtained by setting relatively long timeout and short DRX cycle values (e.g., $M = 64$ and $m = 9$). With higher access delay bounds (e.g., 0.1–0.3 s), and in cells with at most 100 users, the optimal timeout is the shortest possible, i.e., $M = 2$. In all cases reported in Figure 13, the optimization suggests to use very large values for m (larger than 39). However, observing Figure 10(b), it is clear that near-optimal gain can be obtained with values of m as low as 20.

Lessons learned. Our cost and sensitivity analysis shows that significant power save can be achieved while users are guaranteed to experience high performance. In particular, we have unveiled that the threshold timeout does not need to be excessively short in order to enable a remarkable power save, e.g., using $M = 256$ turns into reasonable access delay (tens of milliseconds). We also observed that using $m = 20$ is a very good tradeoff between power save and access delay. In order to limit the download time, it is crucial to limit the number of active users in the cell (to less than 350 users, which is reasonable

for 3GPP LTE, 802.16 and HSPA networks). What is also needed is to limit the web page size. In conclusion, we suggest that enforcing a *green attitude* for web designers, in terms of reducing the web page size and the number of embedded objects, would enable the cellular operator to use reasonable power save parameters (e.g., $m = 20$, $M = 256$) and so achieve a dramatic cost economy at both base station and mobile user sides, without any quality degradation.

9 Conclusions

In this article, we have shown how to model the activity of cellular users adopting the continuous connectivity model under realistic traffic conditions. To this aim, we used a $G/G/1\,PS$ queueing model, which has been validated through simulation. We first modeled the per-user activity and evaluated the service share that the base station processor can grant to each user. Thus, we have derived close-form expressions for busy and idle periods for each mobile user's connection. Second, we provided performance metrics and a cost model enlightening the impact of traffic and power save parameters on quality and cost of transmission. Third, we provided a sensitivity analysis to figure out the impact of each model parameter on performance and power consumption. Forth, we showed how to optimize the power save parameters to minimize the transmission cost under bounded access delay and page download time. Remarkably, we showed that with the considered parameter settings up to 90% or more of the transmission cost can be saved while preserving the quality of packet flows.

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] Sangkyu Baek and Bong Dae Choi. Analysis of discontinuous reception (DRX) with both downlink and uplink packet arrivals in 3GPP LTE. In *Proc. of ACM QTNA 2011*, pages 8–16, Seoul, Korea, August 2011.

[6] Sangkyu Baek and Bong Dae Choi. Performance analysis of sleep mode operation in IEEE 802.16m with both uplink and downlink packet arrivals. In *Proc. of IEEE CAMAD 2011*, pages 112–116, Kyoto, Japan, June 2011.

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

[8] H.K. Choi and J.O. Limb. A behavioral model of Web traffic. In *Proc. of ICNP 1999*, pages 327–334, Washington, DC, USA, October 1999.

[9] R. Cukier, J. Schaibly, and K. Shuler. Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. III. Analysis of the approximations. *Journal of Chemical Physics*, 63(3):1140–1149, 1975.

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

[11] 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.

[12] 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.

[13] Sunggeun Jin, Xi Chen, Daji Qiao, and Sunghyun Choi. Adaptive sleep mode management in IEEE 802.16m wireless metropolitan area networks. *Computer Networks*, 55(16):3774–3783, November 2011.

[14] 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.

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

[16] 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`.

[17] A. Saltelli, S. Tarantola, and K. Chan. A quantitative model-independent method for global sensitivity analysis of model output. *Technometrics*, 41(1):39–56, February 1999.

[18] 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.

[19] 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.

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

[21] 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.