Reducing the Energy Consumption of Small Cell Networks Subject to QoE Constraints

Nikolaos Sapountzis[1], Stylianos Sarantidis[1], Thrasyvoulos Spyropoulos[1], Navid Nikaein[1], and Umer Salim[2]

[1]Mobile Communications Department, EURECOM, 06410, Biot, France, firstname.lastname@eurecom.fr
[2]Intel Mobile Communications, Sophia Antipolis, 06560, France, umer.salim@intel.com

***Abstract*—Small cell networks (SCNs) are widely considered as a promising solution for future cellular deployments. Lately, the benefits of small cells to improve spectrum utilization and the user quality of experience (QoE) have been well documented. In addition, the power consumption of current deployments, for instance due to idle power and cooling equipment, is a major concern for operators. Small cells offer the opportunity for more dynamic power management of base stations, due to coverage overlaps and larger spatio-temporal load fluctuations. Yet, such power management decisions (e.g. turning off a base station) should not lead to excessive performance degradation for users associated with it or additional power consumption. This tradeoff becomes significantly more challenging to evaluate in future networks, due to the diversity of services offered to users beyond the traditional voice calls, as well as the complexity of traffic scheduling algorithms. The goal of this paper is to make a first step towards an analytical investigation of this tradeoff. To this end, we propose a number of QoE constraints that a power management decision should consider, and analytically relate them to key parameters such as user traffic mix, cell load, user density, etc. We then use this framework to perform a preliminary study of the potential energy savings an operator could achieve, while guaranteeing the satisfaction of these constraints. Our results provide some qualitative and quantitative insights on the interesting tradeoff between switch-off duration and number of small cells one can safely switch off.**

I. Introduction

The growing demand for Internet-enabled wireless devices, and bandwidth-hungry multimedia services from the increasing number of "heavy" users and smartphones create significant capacity problems. Thus, operators tend to build more dense deployments. Nevertheless, the higher the deployment density, the higher the chance that cellular nodes will carry no traffic or only a low traffic-load due to spatial and temporal traffic fluctuations. Currently, 15-20% of all sites carry about 50% of the total traffic [1]. Hence, a considerable number of sites waste energy (for staying ON, as well as for cooling), despite serving little or no traffic [2].

A large research effort has been initiated recently in the area of "green" networks. Among the earliest efforts, [3] addresses energy efficiency issues in fixed networks. As Base Stations (BS) are responsible for most of the energy consumed by a cellular network [4], several techniques that consider the BS utilization have been proposed. For example, optimizing the use of sleep modes according to daily traffic variations is explored in [5]. In addition, centralized and distributed cell zooming techniques [6] that adjust the cell size according to traffic load, user requirements, and channel conditions, have also been widely investigated.

Nevertheless, most past studies are performed in the context of large macrocells under homogeneous traffic profiles, and with large time-scales (e.g. turning off BSs during the night [7]). Furthermore, usually simple QoS requirements are considered when applying such techniques, e.g. signal quality as in [8], or traditional blocking probabilities as in [9]. In modern and future cellular networks, dealing with energy consumption issues becomes more challenging. Significantly more opportunities arise for switching off BSs in smaller time scales (e.g. in the order of some minutes), due to (a) coverage overlaps stemming from heterogeneous and/or independent deployment of cells, (b) larger spatio-temporal load variations due to the smaller number of users associated to each cell, and (c) power-proportional and load-dependent BSs. Yet, exploiting such opportunities must be done without violating agreed QoE performance for users. The evaluation of the latter is a rather daunting task, due to the diversity of user traffic (streaming, voice, web, file download, etc.) and service and performance requirements offered to users. As a result, a number of interesting questions arise: Which QoE metric(s) should be used in such future SCNs? Which types of users and BSs should one consider when making a power management decision? Should the duration of switching-off period, affect our decision, and if so, how?

Towards answering these questions, in this paper we identify three QoE constraints, related to different ways that the performance of a User Equipement (UE) could deteriorate. We then derive analytically the probability of violating each of them, as a function of user and network parameters and planned switch-off duration. Specifically, we consider:

- **Network coverage**, i.e. the probability that a random UE experiences poor signal quality when it needs to use the network (e.g. making a call, or sending a web request). (Section II-A).
- **Admission control and "blocking" probabilities**, i.e. the probability that a flow that requires a certain amount of (dedicated) bandwidth, is blocked due to the lack of the available resources (Section II-B).
- **Admission control and "service delay"** for regular "best-effort" flows, i.e. the ongoing delay for the flows that are multiplexed and have to compete for resources. (Section II-C).

Our general methodology is to, first, identify the key parameters for each QoE constraint, and then use analytical tools, mostly coming from queueing theory, to evaluate the probability of violating each one of them, if a BS is switched-off. Our goal in this direction is to strike a tradeoff between realistically capturing some features of new, data-centric cellular systems, while maintaining a certain analytical tractability to provide insights into the QoE vs. Energy savings. The novelty of

our methodology is that we can select even a small time-interval, for the sleeping period X, and evaluate the energy-QoE tradeoff by switching to transient analysis (rather than stationary analysis) of the stochastic model in hand. Based on these QoE constraints and the time duration X, we perform a preliminary study and show that significant energy savings can be achieved even for switching-off periods of the order of some minutes (Section III).

II. System Model and Problem Formulation

In this section, we present some general assumptions about the problem setting. Without loss of generality, we consider a cellular network that is overlaid with M_s small cells (with partially overlapping coverage) and M_m macro cells (macro cells are used to avoid disconnections for users that cannot be connected to a small cell). Our aim is to decrease energy consumption of this cellular network, by dynamically switching off one or more of small cells, during a defined time-duration of X minutes, referred to as *switch-off* period, hereafter[1].

Our first observation is that different users will affect the considered constraints differently. For example, ensuring good signal quality for a UE that has some *ongoing* traffic (e.g. doing a VoIP call, or streaming a video), is more important (and more challenging) than for a UE currently "on" the network but idle. If the former experiences poor signal quality, the communication session might be dropped immediately. We consider three different types of users:

- **Active users** (AU): users that are connected to a BS *and* have one or more on-going traffic flows currently. These users reside in EMM (EPS-Mobility Management) REGISTERED and ECM (EPS Connection Management) CONNECTED states.
- **Connected users** (CU): users associated with a BS but without any ongoing traffic sessions[2]. These users are in EMM REGISTERED and ECM IDLE states.
- **Disconnected users** (DU): users *in the vicinity* of the BS, but currently not ON (or in airplane mode); while their exact number and location cannot be known their impact should be estimated, especially when the switch-off duration increases, as one of them might decide to switch on the UE and use the network. These users are in the EMM DEREGISTERED and ECM IDLE states.

In addition to the above classification of users, we also need to make some assumptions about the classification of flows.

- **Dedicated** flows (e.g. voice and video), where *dedicated* bearers are coupled with the Guaranteed Bit Rate (GBR) to meet the required application bit rate under the latency constraint [10]. They are differentiated by their QoS class of identifier (QCI) ranging from 1 to 4 [11].
- **Best Effort** flows, where default or dedicated bearers are coupled with the non-GBR, which are differentiated by their QCI ranging from 5 to 9 [11].

While the above classification divides flows into two groups [10], it provides a service differentiation between flows that affect the proposed system model very differently.

The probability that the next flow generated by a user is a dedicated or best effort flow depends on the aggregate traffic mix (e.g. percentage of VoIP calls vs. video streaming vs. simple browsing, etc.). We will assume this to be an input parameter. Furthermore, we assume that each BS has a peak data rate R_{total} to allocate among all flows from all serving users, with R_{d}, and R_{b} being allocated to dedicated and best effort flows ($R_{total} = R_{\mathrm{d}} + R_{\mathrm{b}}$), respectively[3].

A. Coverage Constraint

When the decision to switch off a BS with users is made, those users will have to be handed-over to an available neighboring BS. This will often result into a weaker than average signal level. Hence, before a decision to switch off a target BS is made, we must ensure that it will not lead to a disconnection or unacceptable quality for one or more handed-over users. To this end, as our first QoE constraint we will consider the probability that a user, originally associated with a switched-off BS, will experience low-signal quality (e.g. a deep fade) *if* it needs to use the network during the switch-off period.

It turns out this probability changes for different types of users, namely AU, CU, and DU. Specifically, an AU with a current ongoing session will be immediately affected by a signal quality drop. In contrast, a CU or DU will be affected only if *both* the following events occur: (i) it becomes active (e.g. initiates a new call or data session) during the switch-off period X, and (ii) the signal quality is low. Consequently, we need to calculate the following quantities:

- the *outage probability*, which is the probability that the signal strength of user is not sufficient to maintain an ongoing service,
- the *activation probability*, which is the probability that a user covered by the BS in question (e.g. a CU or a DU) becomes active during the next X minutes, and
- the *coverage failure probability*, which depends on both the outage (AU, CU, DU) and activation probabilities (CU, DU), and is the quantity we are interested in.

Outage probability. For simplicity, we use the SNR to calculate the outage probability[4]. Thus, we assume that the SNR for the l^{th} UE associated with the j^{th} BS, is given by [8]:

$$\mathrm{SNR}_{lj} = \frac{G_{lj} R_{lj} p_j}{N_0}. \quad (1)$$

The noise power is denoted as N_0, and the transmission power of the j^{th} BS is p_j. G_{lj} represents the nonnegative path loss between the j^{th} BS and the l^{th} UE (it may also encompass antenna and coding gains) that is often modeled

[1] "Idle" power consumption (related to both electronics, but also cooling) is a major component one could thus save. The additional "load-dependent" power consumption would essentially be shifted over to a neighboring base station, not leading to significant further gains. We defer exploring more complex power management techniques (e.g. cell zooming [6]) to future work.

[2] For simplicity, we will ignore background traffic as it usually less delay-sensitive, and often "lightweight" (e.g. email client polling, social network notifications, etc.).

[3] The actual values are operator-specific, which is why in our analysis it is considered as an input parameter. Note also that, depending on the deployment, the available rate R_{total} might not be bounded by the radio access capacity, but rather by the backhaul capacity [12].

[4] The use of SINR could also be introduced in this constraint, but would make our analysis more complex.

TABLE I. NOTATION

Variable	Meaning
X	Duration of the switch-off period.
p_d, p_b	Probability that a random flow requires dedicated, or best-effort resources respectively.
p_f, p_{block}, D_{max}	Thresholds for Failure Probability (Proposition 1), Blocking Probability (Proposition 2), and ongoing service delay (Proposition 3)
$R_{\text{d}}, R_{\text{b}}, R_{\text{total}}$	Available peak bit rates for "dedicated'" flows, for "best effort" flows, and their sum (total available).
$\lambda_{\text{AU}}, \lambda_{\text{CU}}, \lambda_{\text{DU}}$	Data rates for active, connected and disconnected users.
$E[B_d]$, Y_b	Expected bit rate (in bps) of dedicated flows, and (average) length (in bits) for best effort flows.

as proportional to r_{lj}^{-n} (n is the power fall-off factor and r_{lj} denotes distance). R_{lj} corresponds to a Rayleigh fading component, and is exponentially distributed with unit mean. The distribution of the received power from the j^{th} BS at the l^{th} UE is then exponentially distributed with mean value $E[G_{lj}R_{lj}p_j] = G_{lj}p_j$.

Thus, the outage probability for the l^{th} AU or CU associated with the j^{th} BS is:

$$\text{P}_{\text{out}}(r_{lj}) = P(\text{SNR}_{lj} < \gamma) = 1 - e^{-\frac{\gamma N_0}{G_{lj}p_j}} = 1 - e^{-\frac{\gamma N_0}{r_{lj}^{-n}p_j}}. \quad (2)$$

The above formula is applicable for AUs and CUs, as their actual distance r_{lj} is known. In the case of DUs, their location and the total number is unknown. Assuming that there are ρ_{DU} DUs per m^2, and the transmission range of a base station is r_{max}, the expected number of DUs in the considered cell is:

$$N_{DU} = \rho_{DU}\pi r_{max}^2. \quad (3)$$

If we now consider a specific DU that becomes active, and whose "local" BS is switched off, it will try to connect to one of the neighboring cells. Let r_d denote the distance of the chosen BS from the local BS (its mean value is a function of deployment density). Thus, we can replace r_{lj} in Eq.(2) with r_d to get an estimate for the DU outage probability:

$$P_{out}^{DU} = 1 - e^{-\frac{\gamma N_0}{(r_d)^{-n}p_j}}. \quad (4)$$

Activation Probability. We now consider the probability that a CU or DU becomes active during the next X minutes. We denote these probabilities as $P_{act}^{CU}(X)$ and $P_{act}^{DU}(X)$, respectively. For simplicity, we assume that the time until a CU or a DU generates a new session (call, data session, etc.) is exponentially distributed with rate λ_{CU} and λ_{DU}, respectively (we assume $\lambda_{DU} \leq \lambda_{CU}$). Hence, we can calculate the activation probabilities as follows:

$$P_{act}^{CU}(X) = 1 - e^{-\lambda_{CU}X}, \quad (5)$$
$$P_{act}^{DU}(X) = 1 - e^{-\lambda_{DU}X}. \quad (6)$$

The above equations can be easily extended to general user session interarrival distributions. However, Poisson arrivals are often assumed for user-initiated sessions [13].

Coverage Failure Probability. Assume that the candidate BS serves N_{AU} active and N_{CU} connected users. We denote the set of active and connected users as $\mathcal{N}_{AU}$ and $\mathcal{N}_{CU}$, respectively, and we assume that some DUs are also in the covered region, whose number is given by Eq.(3). If the BS is switched off, then let $J(i)$ denote the BS that user i is handed-over to[5]. Finally, assume that the desired QoE is described by a maximum failure probability p_f, chosen by the operator or indicated in a Service Level Agreement (SLA). Then, the following Proposition captures the first system constraint:

Proposition 1: (Constraint I) A BS cannot be switched off if the average user associated with it will experience a coverage failure probability, during the switch-off period X, that exceeds a threshold p_f. This probability is given by[6]:

$$\frac{\sum_{i\in\mathcal{N}_{AU}} P_{out}(r_{iJ(i)}) + \sum_{i\in\mathcal{N}_{CU}} P_{act}^{CU}(X)P_{out}(r_{iJ(i)}) + N_{DU}P_{act}^{DU}(X)P_{out}^{DU}}{N_{AU}+N_{CU}+N_{DU}}. \quad (7)$$

Impact of switch-off duration X: The above analysis gives qualitative insight about the impact of the switch-off duration. If X is short, compared to the average inactivity time for CUs (DUs), one can more aggressively switch off BSs as a smaller percentage of node is affected. However, for large X Eq.(5) and (6) converge to 1. In that case, all users in the vicinity of a BS must be considered, and the decision only depends on the average outage probability.

B. Admission Control: Blocking Probability Constraint

In this subsection, we focus on the impact of switching off a BS on the admission control mechanism of the neighboring BSs, where users will have to be handed over. A given BS is allocated a finite set of k resources, where k could be frequencies, time slots, etc.. If a user initiates a new session (e.g. call) when the system is already using all its k resources, this session will be *blocked*. An M/M/k/k loss system, like the one shown in Fig. 1, can be used to calculate this *blocking probability*, which is given by the well-known Erlang-B formula [13].

This simple loss system has been the basis of most of the early works on "green" cellular networks, in the context of macrocells [9]: switching off a BS will save energy, but it will also increase the load, λ, in neighboring BSs, thus increasing the blocking probability. However, in modern networks with data flows (rather than voice) comprising the cell's load, a number of issues make the above approach not directly applicable:

1) Each user might generate different *dedicated* flows that require different bit rates;
2) A given dedicated bit rate might require a different amount of resources (e.g. bandwidth, power) to be scheduled by the BS, depending mainly on the distance (channel quality);
3) The switch-off duration X can be small, making the use of stationary probabilities for the Markov chain of Fig.1 (and thus the Erlang-B formula) incorrect;
4) Different dedicated flows might have different priorities and buffer occupancies.

[5]Note that in the real system, this is done using RSSI and RSRP measurements coupled with the received system information assuming that the terminal is eligible. In our analysis, we will assume that either maximum SNR or, simply, distance is used as the criterion.

[6]Instead of this weighted average approach, one could also consider a very conservative, worst-outage probability minimization approach, that has been considered in [14], using the Perron-Frobenius theorem. We omit this case due to space limitations.

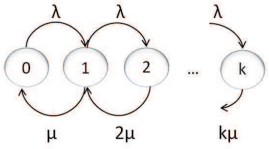


Fig. 1. k-server Loss system

We would like to still use the M/M/k/k system by making appropriate modifications in the model and respective analysis, to address the first three points. To keep our discussion simplified, we assume that there is only one class of dedicated flows (that has priority over non-dedicated ones). In the remaining discussion, when we consider a given BS, we use the term "handed over" or "remote" (sub-/superscript "HO") to refer to users that have been "transferred" to this BS from a neighboring BS that is switched off, and "local" (sub-/superscript "l") for existing users of this BS. Handed-over users are generally further away from the BS than local ones.

Arrival and service rate of dedicated flows: Consider a given BS being switched off, whose users are handed over to (different) neighboring BSs. Consider now one of this neighboring BSs, and let us denote as N_i^l, and N_i^{HO} the number of local and remote users, respectively, of type i ($i \in \{AU, CU, DU\}$), associated with this BS. Let further λ_i denote the flow arrival rate *per user* of type i. The total load for this BS is the sum of all flow rates across these users, and we'll assume that the actual arrival process is Poisson with the sum rate (This assumption is motivated by the Palm-Khintchine theorem, which states that the sum of many independent arrival processes becomes Poisson in the limit [13]). Finally, assume that each arriving flow requires dedicated resources with a probability p_d. Then, due to Poisson splitting, the arrival process remains Poisson with total rate λ_d, given by:

$$\lambda_d = p_d\Big(\lambda_{AU} \sum_{i\in\{AU,CU,DU\}} (N_i^l + N_i^{HO}) P_{act}^i(X)\Big), \quad (8)$$

where $P_{act}^i(X)$ are the activation probabilities defined in (5) and (6). Here, we also need to define the activation probability for AUs, since an AU might ask for additional dedicated flows: $P_{act}^{AU}(X) = 1 - e^{-\lambda_{AU} X}$, where λ_{AU} is the rate that AU ask for new flows . We will also assume that the flow sizes are exponentially-distributed with parameter μ_d, that is, approximately, the average one between dedicated flows. Thus, we can replace λ and μ in the Markov chain of Fig. 1, with λ_d and μ_d, for the case of dedicated flow admission control.

Resource constraint k: As explained earlier, k, the resource constraint in a loss system is a "hard" resource constraint, related to countable resources (e.g. servers, time slots, etc.). In the context considered, the available peak rate for dedicated flows R_d, is a flexible resource, whose allocation is a function of the number of flows, respective dedicated rate demand, and user channel quality. Thus, we apply a "softer", estimated value of k in our loss system.

First, we estimate the average bit rate demand per dedicated flow. Assume that there are different types of "dedicated flows", and that (i) a flow of type i requires a data rate of b_i bits-per-second (bps), and (ii) the ratio of flows with rate b_i is equal to p_i (where $\sum_i p_i = 1$). Thus, the average data rate for an incoming dedicated flow, denoted by $E[B_d]$, can be approximated as:

$$E[B_d] = \sum_i b_\mathrm{i} p_\mathrm{i} \quad (9)$$

If a peak rate R_d is available at the BS, the resource constraint k could be approximated in the M/M/k/k system as $\frac{R_d}{E[B_d]}$, since an "average" flow consumes a percentage $\frac{E[B_d]}{R_d}$ of the available rate. However, this nominal peak rate is only available when the SNR is ideal, or more simply, within a certain distance from the BS (assuming e.g. a simple log-distance path loss model). Hence, to better estimate the maximum number of "average" dedicated flows that can be served, we need to also consider the (potential) distances of different users generating these flows. For this purpose, we adopt a simple model associating peak rate to distance, proposed in [15], [16], stating that the peak rate available drops with distance r_{ij} from a BS j as:

$$c(r_{lj}) = \begin{cases} 1, & r_{lj} \le r_0 \\ (\frac{r_0}{r_{lj}})^n, & \text{otherwise} \end{cases} \quad (10)$$

where r_0 is some threshold range within which the maximal rate is obtained, and n, is the attenuation factor.

Hence, if all dedicated flows where requested from a distance r_{lj}, then the total rate available to them would be only $c(r_{lj})R_d$ ($\le R_d$), or stated differently, the effective rate requirement per average flow would be higher at a large distance r and given by:

$$B(r_{lj}) = \frac{E[B_d]}{c(r_{lj})}. \quad (11)$$

We can now approximate the peak rate drop factor $c(r)$ based on the combination of UEs and distances, e.g. using again a weighted average. Specifically, our estimated resource constraint k for dedicated flows is given by:

$$k = \frac{R_d}{\tilde{B}_d}, \quad (12)$$

where

$$\tilde{B}_d = \frac{\sum_{l=1}^{N_{AU}} B(r_{lj}) + \sum_{m=1}^{N_{CU}} B(r_{mj})}{N_{AU} + N_{CU}}, \quad (13)$$

and $N_i = N_i^l + N_i^{HO}$, $i \in \{AU, CU\}$, denotes the total number of users, local and remote, of type i. We can thus replace k with the approximated value of k in the loss system of Fig. 1[7]. Finally, note that we have assumed that DUs will not affect the peak data rate of the considered BSs (but only affect this constraint through Eq.(8), where we assume that DUs might also switch on and generate some flows during X).

Transient analysis of M/M/k/k: So far, we have shown how to calculate the necessary parameters for the Markov chain of Fig. 1. However, to calculate the probability that a newly arrived flow that needs dedicated resources will be blocked, it does not suffice to replace these parameters in the Erlang B formula. The latter gives the *stationary* blocking probability, that requires the respective chain to be converged, and thus corresponds to large values of X. Instead, we need

[7]We should stress that, as mentioned earlier, this is only an estimate. In practice, there will be a few times when the system is serving more than k dedicated flows (e.g. when all users are close-by or flows require lower rates than average), and times when a new flow might be blocked even if less than k flows are served.

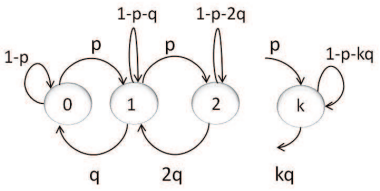


Fig. 2. DTMC k-Loss system

to apply *transient analysis* to this system, and estimate the blocking probability via the *occupation time* in state k during the intended switch-off duration X.

The initial state for the Markov chain, at time 0 (the beginning of the switch-off period), corresponds to the current number of active dedicated flows. Denoted as s,it is:

$$s = p_d \cdot (N_{AU}^{l} + N_{AU}^{HO}) \cdot \xi, \tag{14}$$

where we use ξ to denote the expected number of ongoing flows per AU (this is an input parameter). Starting from s, the occupation time in state i, denoted as $O_i(X)$ $(0 \leq i \leq k)$, is the time that the MC spends in state during the next X minutes (or time units). We are interested in deriving the quantity $\frac{E[O_k(X)]}{X}$. This corresponds to the percentage of time that the system is in state k (all resources are used), during the switch-off period X and starting from state s. Hence, due to the PASTA (Poisson Arrivals See Time Averages) property, this also corresponds to the probability that a newly arrived dedicated flow will be blocked due to non-available capacity. We can estimate this percentage of time either by *uniformization* in CTMC (Continuous Time Markov Chain) [17], or by converting it to DTMC (Discrete Time Markov Chain), as an approximation. To simplify our discussion, we follow the second approach. Let Δt be a small time inteval. The DTMC depicted in Fig. 2, is the discrete-time approximation of our system continuous-time M/M/k/k system, where a state transition occurs every Δt time units. If P_{ij} denotes the probability that the chain goes from state i to state j $(0 \leq i, j \leq k)$, then it follows from standard properties of the Poisson distribution [17] that (see Fig. 2):

$$p = \lambda_d \cdot \Delta t, \quad P_{i,i+1} = p, \;\; 0 \leq i < k$$
$$q = \mu_d \cdot \Delta t, \quad P_{i,i-1} = i \cdot q, \;\; 0 < i \leq k$$
$$P_{i,i} = 1 - P_{i,i+1} - P_{i,i-1}, \;\; 0 \leq i \leq k$$

Hence, if $\mathbf{P} = \{P_{i,j}\}$ denotes the probability transition matrix, and $\mathbf{P}^n = \{P_{i,j}^n\}$ the n-step transition matrix ($\mathbf{P}^n = (\mathbf{P})^n$), then the expected occupation time is given by:

$$E[O_k(X)] = \sum_{n=0}^{\frac{X}{\Delta t}} P_{\mathrm{s,k}}^n, \tag{15}$$

where s denotes the initial state (initial number of active dedicated flows) and $\frac{X}{\Delta t}$ the total switch-off duration (counted in discrete time steps of duration Δt).

Proposition 2: (Constraint II) Assume a desired maximum blocking probability is given for dedicated flows, defined as p_{block}. A given BS can be switched off only if the following inequality holds for all neighboring BSs to which users of the switched-off BS are handed-over:

$$\frac{\sum_{n=0}^{\frac{X}{\Delta t}} P_{\mathrm{s,k}}^n}{X/\Delta t} \leq p_{block}. \tag{16}$$

Impact of switch-off duration X. The computational complexity of Proposition 2 can be traded off with accuracy by increasing the time step Δt. In addition, as X becomes large (specifically, larger than the mixing time for the MC of Fig. 2), the condition of Eq.(16) converges to the Erlang-B formula.

Remark 1: When $X \to \infty$, the condition of Eq. (16), converges to

$$\frac{(\frac{\lambda_{\mathrm{d}}}{\mu_d})^k / k!}{\sum_{j=0}^{k} (\frac{\lambda_{\mathrm{d}}}{\mu_d})^j \frac{1}{j!}} \leq p_{block}. \tag{17}$$

Proof:

$$\begin{aligned} \lim_{\Delta t \to 0} \lim_{X \to \infty} \frac{E[O_k](X)}{X/\Delta t} &= \lim_{\Delta t \to 0} \lim_{X \to \infty} \frac{\sum_{n=0}^{\frac{X}{\Delta t}} P_{s,k}^n}{X/\Delta t} \\ = \lim_{\Delta t \to 0} \pi_k(\Delta t) &= \pi_k, \end{aligned}$$

where $\pi_k(\Delta t)$ is the stationary probability for state k in the DTMC approximation with unit step (Δt). As $\Delta t \to 0$ this quantity converges to π_k, the stationary probability of state k for the CTMC corresponding to the standard M/M/k/k loss system of Fig.1, which is the Erlang B formula [13]. ■

C. Admission Control: Service Delay Constraint

As our last constraint, we consider the delay for a best-effort flow, i.e. a flow that does not require dedicated resources. Consider again a given BS being switched-off, whose users are handed over to different neighboring BSs, and let's pick one of them and focus on it. As before, this BS will have some local users and some remote users, that were handed over from the switched-off BS, all of which might generate (new) best effort flows. While there are no guarantees for such flows, we might still want to keep their expected delay below a certain threshold. Our goal is to model and analytically bound this delay.

Unlike the case of dedicated flows, that are allocated separate resources each, best-effort flows are multiplexed and have to compete for resources. The more flows in parallel in the system, the larger the expected delay for each flow. A lot of effort has been devoted to the study of scheduling algorithms for such "elastic" types of traffic [13]. In this work, we will assume that best-effort flows are scheduled using Processor Sharing (PS) [18], as PS works better than First Come First Serve for high variability loads, and is also widely considered as a fair queueing discipline. For simplicity, we are only considering one priority-class of best-effort flows.

To analyze the delay of the PS flow scheduler for best effort flows we need to know λ_b, the arrival rate of best-effort flows, and μ_b the service rate for best-effort flows. Let p_b denote the probability that a new flow arrival is best effort (rather than dedicated), and thus $p_b = 1 - p_d$, it follows that incoming "best-effort" flows are Poisson distributed with total rate:

$$\lambda_b = p_b \Big(\lambda_{AU} \sum_{i \in \{AU, CU, DU\}} (N_i^l + N_i^{\mathrm{HO}}) P_{act}^i(X) \Big). \tag{18}$$

To find the service rate for best-effort flows, let R_{b} denote again the peak bit rate for best-effort flows. As explained before, if a *single* best effort flow exists in the system for a user at distance

r_{lj}, then the actual bit rate received is only $c(r_{lj})R_b$, where $c(r_{lj})$ is given by Eq(10). The actual average peak rate is:

$$\tilde{R}_b = R_b \cdot \frac{\sum_{l=1}^{N_{AU}} c(r_{lj}) + \sum_{m=1}^{N_{CU}} c(r_{mj})}{N_{AU} + N_{CU}}, \quad (19)$$

where $N_i = N_i^l + N_i^{HO}$, $i \in \{AU, CU\}$. The above estimated rate corresponds to a single flow. If there are n total best-effort flows currently in the system, then PS would split this rate equally, and each flow would be served with a bit rate $\frac{\tilde{R}_b}{n}$.

To find the actual service rate μ_b of the PS queue, the number of *flows served per time unit* (note that this is not equal to $\tilde{R}_b$, which is just the effective bit rate), we also need to know the average length of best effort flows. If we assume that the sizes of the best-effort flows are exponentially distributed with mean Y_b, then $\mu_b = \frac{\tilde{R}_b}{Y_b}$. When the system is stationary, i.e. when X is quite large, the expected delay for a newly arriving flow corresponds to the delay of an M/M/1/PS system:

$$E[D_b] = \frac{1}{\mu_b - \lambda_b} \quad (X \to \infty).$$

However, for general values of X, the Markov chain corresponding to the PS system is not stationary. Thus, we must again apply transient analysis, assuming an initial state. Let s again denote the initial number of best-effort flows in the BS, at the beginning of the switch-off period X, where $s = p_b \cdot (N_{AU}^l + N_{AU}^{HO})\xi$, similar to Eq. (14). Consider now a new flow of size Y_b arriving at some time $t \in [0, X]$. The number of active best effort flows in the system that has to share the PS capacity with, is a random variable, denoted as n. Our approach will be to find the expected delay conditional on this value of n, and then take the average.

If our flow of size Y_b finishes transmitting while in state n (i.e. no new flows arrive and no existing flows finish), the service rate remains fixed at $R_n = \tilde{R}_b/n$ and the expected delay for this flow is $\frac{Y_b \cdot n}{\tilde{R}_b}$. However, if a state transition occurs before all Y_b bits are transmitted, then the remaining bits will be transmitted at a lower ($\tilde{R}_b/(n+1)$) or higher rate ($\tilde{R}_b/(n-1)$), if a new flow arrived, or an existing finished, respectively. Let us denote as T_n the time spent in this state until the next transition. This time is exponentially distributed with rate $\lambda_b + \mu_b$, so $E[T_n] = \frac{1}{\lambda_b + \mu_b}$. Hence, putting everything together, we can define the following recursion to derive the (conditional) delay ($D^n(Y_b)$) of a flow of Y_b bits finding another n ongoing flows when it arrives. R_n denotes the transmission rate at state n.

$$D^n(Y_b) = \begin{cases} \frac{Y_b}{R_n}, & \text{if } \frac{Y_b}{R_n} \le E[T_n] \\ E[T_n] + D^{n+1}(Y_b - R_n \cdot E[T_n]), & \text{if } \frac{Y_b}{R_n} > E[T_n] \\ & \text{and } n \to n+1 \\ E[T_n] + D^{n-1}(Y_b - R_n \cdot E[T_n]), & \text{if } \frac{Y_b}{R_n} > E[T_n] \\ & \text{and } n \to n-1 \end{cases} \quad (20)$$

It is: $P(n \to n+1) = \frac{\lambda_b}{\lambda_b + \mu_b}$, and $P(n \to n-1) = \frac{\mu_b}{\lambda_b + \mu_b}$.

However, the actual number of initial active flows n at time t is also a random variable, which depends on the evolution of the system, starting at initial state s until time t. To find these probabilities, we will again use a DTMC approximation and n-step transitions as before. Due to space limitations, we omit here the details and give the final result which is:

$$E[D_b] = \sum_{t=0}^{X/\Delta t} \sum_{n=1}^{\infty} \frac{P_{s,n}^{t/\Delta t} D^n(Y_b)}{X/\Delta t}, \quad (21)$$

In practice, we can add up only a finite number of terms in the inner sum, to reduce the calculations.

Proposition 3: (Constraint III) Assume a desired maximum delay for best effort flows, D_{max}. A given BS can be switched off only if the following inequality holds for all neighboring BSs to which users of the switched-off BS are handed-over:

$$\sum_{t=0}^{X/\Delta t} \sum_{n=1}^{\infty} \frac{P_{s,n}^{t/\Delta t} D^n(Y_b)}{X/\Delta t} \le D_{max}.$$

Impact of switch-off duration X. The computational complexity of Prop. 3 can be traded off with accuracy by increasing the step Δt. Also, as X becomes large, the individual probabilities of (21) converge to their stationary distribution.

III. Simulation Results

To evaluate our QoE constraints, we consider a network composed of $M_s = 120$ small cells (μ-cell), and $M_m = 2$ macro cells (m-cell) that are uniformly distributed in an area of $45km^2$. Each μ-cell has radius 400m (meters), and each m-cell had radius 2.2 km. We assume that there are 500 AUs and CUs, plus 120 DUs. We also assume total peak rate R_{total}= 70 Mbps; average length and bit-rate for best-effort and dedicated flows Y_b = 20 Kbytes and $E[B_d]$ = 200 kbps, respectively; coverage threshold $\gamma = 50dB$[8]; probability for dedicated flows $p_d = 0.7$; and rates μ_d, $\lambda_{\text{AU}}, \lambda_{\text{CU}}, \lambda_{\text{DU}}$ 10, 2, 1 and 0.1 flows/hour, respectively. Finally, the maximum number of concurrent users that each μ-cell can handle is set to[9] 11.

We are interested in investigating how the different values of the predefined thresholds p_f (failure probability), p_{block} (blocking probability) and D_{max} (service delay) affect the portion of energy savings[10]. In Fig. 3(a), 3(b) and 4(a), we assume switching-off duration $X = 10min$. Each figure contains two curves; the "top" curve corresponds to the portion of energy saved when we consider only a certain constraint active, while the "bottom" curve considers all constraints to be active, at fixed thresholds (when not explicitly mentioned, we assume them to be $p_f = 0.3$, $p_{block} = 10^{-3}$ and $D_{max} = 50$msec).

In the "top" curve of Fig. 3(a), on the x-axis we increase the p_f and plot the savings. It can be seen that, increasing the threshold (making the constraint less strict) increases savings, as it allows for more BSs to be switched off. For instance, we can save up to 68% for $p_f = 0.4$. As for the "bottom" curve, savings increase too, but less sharply, as the other two constraints can overrule the switch-off decision, especially for large p_f. For example, with $p_f = 0.4$ and the other two thresholds fixed, the energy savings can be up to 30%.

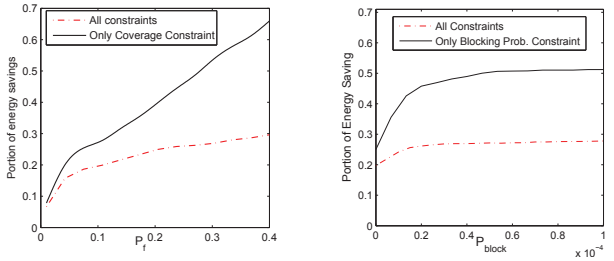
Similarly, Fig. 3(b) and 4(a) depict the portion of the energy saved, by taking into account the blocking probability and service delay constraints. For example the top (bottom) curve of Fig. 3(b), shows that the portion of energy savings can be up to 50% (28%), by considering only the blocking probability constraint (plus the other two with fixed). Finally, Fig. 4(a)

[8]The threshold γ is an input parameter and is chosen to ensure the coverage constraint with a relatively good signal quality.

[9]This number can vary, depending on the type of the small cell [4], and does not affect the blocking probability.

[10]This portion is equal to the energy we can save, divided by the energy needed for all BSs switched-on during X i.e. $\frac{E_{ALL} - E_{part}}{E_{ALL}}$, where E_{ALL} is the energy needed if all BSs are switched-on, and E_{part} is the (decreased) energy needed if we safely switch-off some BSs based on our policies.

shows that the portion of energy savings for the delay constraint can be 70% by maintaining only the $D_{\max}$ in 100msec, and 30% by holding the other two fixed.



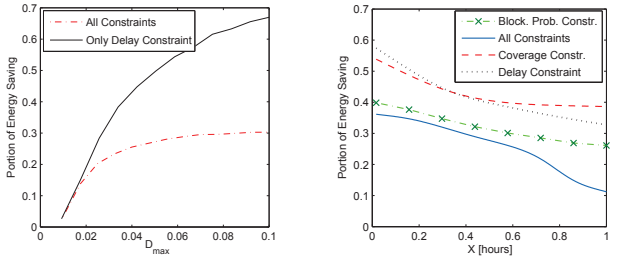
(a) Portion of Energy Saving versus the Failure probability.

(b) Portion of Energy Saving versus the Blocking probability.

Fig. 3. Portion of Energy Saving versus the Failure and Blocking probability

Another interesting parameter is X, the switch-off duration. Fig. 4(b) depicts the portion of energy saved for different values of X with fixed constraint thresholds ($p_f = 0.4$, $p_{block} = 10^{-3}$, $D_{max} = 200$msec). To be more precise, energy savings are maximum when X is relatively small, but start decreasing and eventually flatten out, as X increases. The reason is that, for small X, one needs to only consider the impact of AUs when evaluating the constraint and the impact of hand overs to neighboring BSs. However, as X increases, there is a higher chance that CUs and DUs will add traffic to the total transferred load (see Eqs .(8) and (18)), which might prevent us from switching off a BS. Finally, the plot corresponding to each constraint is not always linear, as some additional phenomena, such as convergence to stationarity for the stochastic systems we use in constraints 2 and 3, also affect systems' behavior.

Thus, a small switch-off duration X promises larger energy savings. However, it also implies that the system will (a) have to re-evaluate the state of the system and repeat its decision quite frequently (computation complexity) and (b) it might lead to some additional energy wastage (and performance degradation) due to the fixed power (and delay) needed to switch-off and back-on a BS. This suggests an interesting trade-off that we plan to explore in future.



(a) Portion of Energy Saving versus the Service delay (sec).

(b) Portion of Energy Saving versus the Switching-off period.

Fig. 4. Portion of Energy Saving versus the Delay and switching-off period.

IV. Conclusion

In this paper, we consider the problem of energy saving in future SCN by switching off underloaded BSs. Specifically, we have shown how the potential degradation of user QoE could be analytically captured and bounded along different dimensions, namely coverage probability, blocking probability (for dedicated flows), and delay (for best effort flows). Based on the proposed framework, we have also showed how a significant amount of energy in SCNs could be saved while maintaining some desired QoE levels. Finally, complementary to research works focusing on large switch-off periods (e.g. night hours) we show that savings can be achieved for short periods as well, and that the switch-off duration presents an interesting tradeoff.

V. Acknowledgments

The research leading to these results has received funding from the European Research Council under the European Community Seventh Framework Programme (FP7/2012- 2015) under the ICT theme of DG-CONNECT n^{o} 317941 (iJOIN).

References

[1] T. K. H. Guan and P. Merz, "Discovery of cloud-RAN," *in Cloud-RAN Workshop*, April 2010.

[2] G. Auer, V. Giannini, C. Desset, I. Godor, P. Skillermark, M. Olsson, M. Imran, D. Sabella, M. Gonzalez, O. Blume, and A. Fehske, "How much energy is needed to run a wireless network?" *in IEEE Wireless Communications*, 2011.

[3] M. Gupta and S. Singh, "Greening of the internet," in *Conf. on Applications, Technologies, Architectures, and Protocols for Computer Communications*, ser. SIGCOMM, 2003.

[4] N. S. Networks, "Improving 4G coverage and capacity indoors at hotspots with lte femtocells," *White Paper*, 2011.

[5] C. Peng, S.-B. Lee, S. Lu, H. Luo, and H. Li, "Traffic-driven power saving in operational 3G cellular networks," in *Proceedings of the 17th Annual International Conf. on Mobile Computing and Networking*, ser. MobiCom, 2011.

[6] Z. Niu, Y. Wu, J. Gong, and Z. Yang, "Cell zooming for cost-efficient green cellular networks," *in IEEE Communications Magazine*, November 2010.

[7] M. A. Marsan, L. Chiaraviglio, D. Ciullo, and M. Meo, "On the effectiveness of single and multiple base station sleep modes in cellular networks," *in Comp. Networks*, 2013.

[8] S. E. Nai, T. Quek, and M. Debbah, "Shadowing time-scale admission and power control for small cell networks," in *Wireless Personal Multimedia Communications (WPMC)*, Sept 2012.

[9] L. Chiaraviglio, D. Ciullo, M. Meo, and M. Marsan, "Energy-efficient management of UMTS access networks," in *Teletraffic Congress*, 2009.

[10] F. Capozzi, G. Piro, L. Grieco, G. Boggia, and P. Camarda, "Downlink packet scheduling in LTE cellular networks: Key design issues and a survey," *in IEEE Comm. Surveys*, 2013.

[11] Y. Q. Jie Wang, "A new call admission control strategy for LTE femtocell networks," *in 2nd International Conference on Advances in Computer Science and Engineering*, 2013.

[12] O. Simeone, O. Somekh, H. Poor, and S. Shamai (Shitz), "Downlink multicell processing with limited-backhaul capacity," *in EURASIP Journal on Advances in Signal Processing*, 2009.

[13] M. Harchol-Balter, *Performance Modeling and Design of Computer Systems*. Cambridge University Press, 2010.

[14] C. wei Tan, "Optimal power control in rayleigh-fading heterogeneous networks," in *IEEE INFOCOM*, April 2011.

[15] S. Aalto and L. Pasi, *Impact of Size-Based Scheduling on Flow Level Performance in Wireless Downlink Data Channels*. Springer, 2007.

[16] T. Bonald and A. Proutiere, "Wireless downlink data channels: User performance and cell dimensioning," in *Proceedings of the 9th Annual International Conference on Mobile Computing and Networking*, ser. MobiCom, 2003.

[17] S. M. Ross, *Introduction to Probability Models*. Academic Press, 10th edition, 2009.

[18] A. Elwalid and D. Mitro, "Design of generalized processor sharing schedulers which statistically multiplex heterogeneous QoS classes," in *IEEE INFOCOM*, Mar 1999.