Energy and Cost Implications of a Traffic Aware and QoS Constrained Sleep Mode Mechanism

Siyi Wang§, Weisi Guo
Institute for Telecommunications Research, University of South Australia, Australia§
School of Engineering, University of Warwick, United Kingdom
Email: siyi.wang@mymail.unisa.edu.au§, weisi.guo@warwick.ac.uk

Abstract

This paper considers the theoretical and realistic performance of a traffic aware sleep mode mechanism in the cellular network. The paper devises a sleep mode mechanism that maximises the energy saving of the network, whilst maintaining both the user throughput and reliability performance. The analysis combines recent developments in stochastic geometry and packet data modelling, as well as a simulation of a real cellular network. The main contribution is to show that accurate theoretical modelling of realistic network data can allow deterministic sleep mode triggers to be devised. This is shown to be highly applicable to a real network, reducing energy consumption by 15 to 60% throughout a day and the annual operational expenditure by 4%.

I. INTRODUCTION

A. ICT Energy Consumption

There is a growing focus, both in the research community and in industry on improving the energy- and cost-efficiency of cellular networks [1]–[4]. Approximately 70% of global information and communication technology (ICT)’s energy is consumed by the outdoor cells of a cellular network. The global cellular networks consume 60 TWh of electricity and the utility bill for the network operators stands at over $10 billion in 2010–11. In the face of increasing data demand and uncertain average revenue per user (ARPU) trends, operators are looking at ways to reduce the costs in order to improve their competitiveness. In terms of the environmental impact, over 40 million metric tons of CO₂ produced is directly attributed to the electricity consumption of the cellular network, and up to 10 times more is indirectly attributed.

For the aforementioned reasons, there is a genuine desire from operators to significantly reduce their energy consumption. One of the key challenges facing the cellular operators is the inefficiency of network deployment. Over 50% of the data traffic is carried by less than 10% of the cells. This is primarily due to the dynamic spatial- and temporal-variations in traffic loads. The consequence is that the network wide
energy and cost expenditure is higher than perhaps necessary. This body of work examines how a sleep mode mechanism can both satisfy the traffic offered and reduce energy and cost expenditures.

B. Review

Existing literature has demonstrated that different network deployment configurations can result to different spatial capacity distributions [4], and consequently incur different energy consumption levels. For a single link, the relationship between spectral- and energy-efficiency was investigated for noise-limited channels in [1] and for interference-limited channels in [3].

In terms of sleep mode research, most research has considered isolated energy savings for individual cells or small clusters of cells [2], [5], operating under simple traffic models [6]. The investigation in [7] proposed the expansion of traditional cell coverage after a certain set of cells enter sleep mode. However, this analysis was performed for a regular hexagonal cell geometry without interference modelling. Numerical investigations for realistic networks with interference has been considered in [6]. More recently, our work considered the application of stochastic geometry [8], [9] to assist in predicting the performance of cell coverage expansion with uniform traffic loads [10].

A shortfall in current research is an analysis of sleep mode behaviour for interference-limited networks that are subject to realistic traffic models. Furthermore, whilst the coverage of the network (signal strength target) has been used as a constraint [6], the minimum achievable throughput is less well considered.

C. Contribution

This investigation evaluates the potential operational cost and energy savings that can arise from implementing a sleep mode mechanism. Let us consider a network that consists of multiple co-frequency and potentially mutually interfering cells. In order to implement a sleep mode mechanism that can maintain performance and reduce energy consumption, we must address the following questions:

1) If a cell is switched off (sleep), what is the new coverage and capacity pattern of the network?
2) When does a cell decide to sleep or stay active?

The analysis employs recent theoretical methods in stochastic geometry, as well as simulation results. The main novelty and benefit of the investigation is to devise a sleep mode mechanism that can address the aforementioned questions. The authors demonstrate how theoretical models can be adapted to a realistic network, and be used to aid the solution.
The paper presents the investigation in the following logical order:

1) Derive the theoretical multi-cell interference-limited capacity of a radio-access-network (RAN), and verify the theory against simulation results.

2) Present the probability distribution of packet-switched traffic model based on real traffic data, and the energy-cost consumption models.

3) Devise a sleep mode mechanism that adapts the operational state of cell based on the traffic load and achievable capacity, and present the savings on energy and cost.

4) Adapt theoretical sleep mode thresholds to a realistic network and consider the impact of heterogeneous network elements.

In summary, the objective is to maximise the energy savings, whilst satisfying the performance both in terms of the achievable throughput and reliability. The paper also extends the results and discussion onto a heterogeneous network, consisting of femto-cells and macro-cells in co-existence.
II. Achievable Capacity Model

A. Stochastic Geometry Network Model

The multiple-access network considered is the Orthogonal Frequency-Division Multiple Access (OFDMA) based 4G Long-Term-Evolution (LTE). In this section, the authors present the theoretical achievable capacity of the multi-cell network, which is effectively the averaged sum-rate of all user locations. Generally speaking, the average achievable capacity of a single arbitrary cell $n$ is the aggregate of the capacity of each user $m$ served in the coverage area of the cell:

$$C_n = \sum_{m=1}^{M_n} B \frac{1}{M_n} \mathbb{E}(C_{m,n}(\gamma_{m,n})),$$

where $\mathbb{E}(C_{m,n}(\gamma_{m,n}))$ is the expectation of a single user’s capacity over multiple fading realisations, $M_n$ is the number of active users at different locations in the cell and $B$ is the bandwidth available in the cell. $\Phi$ represents the homogeneous Spatial Poisson Point Process (SPPP). The parameter $\gamma_{m,n}$ is the received signal-to-interference-noise-ratio (SINR). As shown in Figure 1, the SINR for user $m$ attached to cell $n$ is:

$$\gamma_{m,n} = \frac{H_{m,n} P_n \lambda r_{m,n}^{-\alpha}}{W + \sum_{\substack{k \neq n \in \Phi}} H_{m,k} P_k \lambda r_{m,k}^{-\alpha}},$$

for: $g_{m,n} = H_{m,n} P_n \lambda$,

where $P_n$ is the transmit power, $r_{m,n}$ and $r_{m,k}$ are the distance between the user and the serving cell and interference cells, respectively. The pathloss follows the WINNER urban macro pathloss model, and the parameter definitions are given in Table I. The antenna gain is assumed to be unity (dipole) and log-normal shadowing is neglected for this framework, and the impact will be considered later in this section. The interference power is written as $I$, which is a function of the distance and fading. The combined gain $g$ follows an exponential distribution $\sim \exp(\beta)$, where $\beta = 1/P_n \lambda$.

The key assumptions of this paper are that: the cells in the network are homogeneous and that the users are uniformly randomly distributed.

For multiple user positions, each user is attached to its closest cell at a distance of $r_{m,n}$. Given the location of a user (represented by $r_{m,n}$), it has been shown that one can model the distribution of serving
distance as a 2-D Poisson process with random variable $r$ [11]:

$$P(D, k) = \frac{\Lambda |D|^k \exp(-\Lambda |D|)}{k!},$$

$$= \exp(-\Lambda \pi r^2) \quad \text{for: } D = \pi r^2,$$

and $k = 0$ as there are no interference cells closer to the user than the attached serving-cell. Given that (3) is the complementary cumulative distribution function (CCDF), the probability density function (PDF) is the differential of the CDF:

$$f_R(r) = \frac{dF_R(r)}{dr} = 2\Lambda \pi r \exp(-\Lambda \pi r^2),$$

where $\Lambda$ is the density of cells in the RAN.
Given the spatial distribution of users relative to cells, the definition of the average downlink capacity in a cell is given by:

\[
\overline{C}_n = \int_0^{+\infty} \mathbb{E}(C_{m,n}) f_R(r) \, dr, \\
= 2\Lambda \pi \int_0^{+\infty} \mathbb{E} \left[ B \log_2 (1 + \gamma_{m,n}) \right] r \exp(-\Lambda \pi r^2) \, dr,
\]  
(5)

where the instantaneous SINR \( \gamma_{m,n} \) is a function of \( r \).

Given that the capacity is a continuous random variable with non-negative values, the expectation of the capacity is defined as:

\[
\mathbb{E}(C_{m,n}) = \int_0^{+\infty} \mathbb{P} \left\{ B \log_2 \left[ 1 + \frac{g_{r,n} r^{-\alpha}}{W + I(r,H)} \right] > \zeta \right\} \, d\zeta, \\
= \int_0^{+\infty} \exp \left[ -\beta r^\alpha W (2\zeta/B - 1) - \Lambda \pi r^2 Q(\zeta,\alpha) \right] \, d\zeta,
\]  
(6)

where:

\[
Q(\zeta,\alpha) = \int_{(2\zeta/B - 1)^{-2/\alpha}}^{+\infty} \frac{(2\zeta/B - 1)^{2/\alpha}}{1 + u^{\alpha/2}} \, du, \\
= \sqrt{2\zeta/B - 1} \arctan \left( \sqrt{2\zeta/B - 1} \right) \quad \text{for: } \alpha = 4.
\]  
(7)

Whilst, in reality the distance between a user and a cell is lower bounded by the cell height and upper bounded by the receiver sensitivity, in theory we have bounded the distance between a user to a cell from 0 to infinity. The full proof is given in Appendix A.

Evaluating all the possible user locations with respect to the distribution expression (4), the average achievable capacity (spectral-efficiency) of users in a cell is given by (full proof is given in Appendix B):

\[
\overline{C}_n = \int_0^{+\infty} \frac{1}{1 + Q(\zeta,\alpha)} \, d\zeta = 2.14 B \text{ bits/s},
\]  
(8)

for an interference-limited network \( I \gg W \). The integral in (8) can be solved using the Gauss-Kronrod technique. The average capacity achieved by all user locations is the equivalent to the average capacity achieved by a single cell with a single user.

Furthermore, two useful capacity metrics can be derived from the above framework (Proof in Appendix C):
• CDF of Average Cell Throughput:
\[
P(\overline{C}_n > \zeta) = \frac{1}{1 + Q(\zeta, \alpha)};
\]  
(9)

• CDF of Average Cell-edge Throughput:
\[
P(\overline{C}_{n, R} > \zeta) = \exp(-\kappa Q(\zeta, \alpha)),
\]  
(10)

where \(\kappa\) is a geometry-related constant (see Appendix C). The cell-edge in this paper is defined as the region where the distance to the serving and nearest interference cell is the same.

The paper will now compare the average cell throughput performance with that of different Monte-Carlo simulation results.

**B. Monte-Carlo Simulation**

**TABLE II**  
**BENCHMARK SYSTEM CONFIGURATIONS**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Monte-Carlo Hexagon</th>
<th>Monte-Carlo London</th>
<th>Stochastic Geometry</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of Cell-sites</td>
<td>19</td>
<td>96</td>
<td>unlimited</td>
</tr>
<tr>
<td>No. of Interference Cells</td>
<td>wrap-around</td>
<td>2 km range</td>
<td>infinite</td>
</tr>
<tr>
<td>Pathloss Model</td>
<td>statistical [12]</td>
<td>deterministic (PACE 3G)</td>
<td>statistical</td>
</tr>
<tr>
<td>Antenna Patterns</td>
<td>realistic</td>
<td>realistic</td>
<td>none</td>
</tr>
<tr>
<td>Terrain Effects</td>
<td>none</td>
<td>realistic</td>
<td>none</td>
</tr>
<tr>
<td>Indoor Coverage</td>
<td>none</td>
<td>realistic</td>
<td>none</td>
</tr>
<tr>
<td>Fading Effects</td>
<td>statistical</td>
<td>deterministic (PACE 3G)</td>
<td>statistical</td>
</tr>
<tr>
<td>Shadowing Effects</td>
<td>statistical</td>
<td>deterministic (PACE 3G)</td>
<td>none</td>
</tr>
<tr>
<td>Data Sample Area</td>
<td>unlimited</td>
<td>5 km × 2km area</td>
<td>unlimited</td>
</tr>
<tr>
<td>Software Tool</td>
<td>VCEsim</td>
<td>Forsk Atoll</td>
<td>none</td>
</tr>
<tr>
<td>Simulation Speed</td>
<td>fast</td>
<td>medium</td>
<td>instant</td>
</tr>
</tbody>
</table>

In current literature, different performance modelling approaches have been taken to characterise the network performance. The paper compares the capacity modelling results of the following techniques:

• **Monte-Carlo Simulation of Hexagonal Cells with Wrap Around**: yields network upper-bound performance and can include realistic antenna patterns;

• **Monte-Carlo Simulation of a Real London Area**: realistic performance with ray-traced pathloss models, antenna patterns and terrain-clutter effects;

• **Theoretical Framework - Stochastic Geometry**: yields closed-form mathematical expressions of multicell performance with full interference modelling.
A summary of the system configuration parameters and capabilities is given in Table II. The software package used for pathloss calculations and performances are industrially bench-marked. Figure. 2 shows the CDF of the multi-cell RAN capacity for: hexagonal cell RAN with wrap-around, real London RAN, and theoretical network. The London RAN is a 3G HSPA network that consists of approximately 100 macro-cells deployed over a 9 km × 6 km urban area. The paper assumes that a potential 4G network would share the same cell locations. The effects of deterministic pathloss, antenna gain, shadow fading and terrain clutter are modeled. The results show that a close match between theocratical framework and London network is achieved (80% CDF correlation) [13]. The theoretical performs worse at high capacity values due to the lack of antenna gain benefits. The hexagonal model yields an upper-bound network performance, as intended. Therefore, from Figure. 2, the stochastic model of network capacity appears significantly accurate.

![Figure 2](image-url)

Fig. 2. CDF of the multi-cell RAN capacity for: hexagonal cells with wrap-around (simulation), a real London cellular network (simulation), and a stochastic geometry network (theory).
III. TRAFFIC AND CONSUMPTION MODELS

A. Packet Switched Traffic

The paper employs real 3G HSPA traffic data, which can be statistically characterised. Data from a European city’s 3G network [14] show that the number of active users in a cell at time $t$ is: $M_{n,t} \sim \text{Pois}(\Upsilon)$. The PDF for number of active users $x$ is given as:

$$f_{M_{n,t}}(x; \Upsilon) = \frac{\Upsilon^x \exp(-\Upsilon)}{x!},$$

for $\Upsilon = 0.003 \exp(1.1 \log_{10} C_n)$. This means the number of active users in a cell is correlated strongly with the achievable capacity of the cell.

The mean traffic throughput demanded per user session (bits/s) is related to the log-normal distribution: $R_{m,t} = 10^\Omega$, where $\Omega \sim \text{Log-N}(\mu, \sigma^2)$. The PDF for traffic demanded per user $y$ is given as [14]:

$$f_{R_{m,t}}(y; \mu, \sigma) = \frac{\exp \left( - \frac{(\ln(\log_{10} y) - \mu)^2}{2\sigma^2} \right)}{y \log_{10}(y) \sigma \sqrt{2\pi \ln 10}},$$

for $y > 1$. The statistical parameter values are: $\mu = 1.35$ and $\sigma^2 = 0.2$.

The traffic throughput demanded for all users in a cell is defined as a random variable $R_{n,t} = M_{n,t} R_{m,t}$. Provided that they are independent (traffic demand per user is independent to the number of users), $R_{n,t}$ is proved to be a continuous R.V. as well and the PDF for traffic demanded per cell can be derived as follows (Proof in Appendix D):

$$f_{R_{n,t}}(z) = \sum_{x=1}^{z} \left\{ \frac{\Upsilon^x \exp(-\Upsilon)}{x!} \exp \left[ - \frac{(\ln(\log_{10} \frac{z}{x}) - \mu)^2}{2\sigma^2} \right] \right\} \frac{z \log_{10} \left( \frac{z}{x} \right) \sigma \sqrt{2\pi \ln 10}}{\sigma},$$

for $\frac{z}{x} > 1$.

The expected value of the traffic rate per cell is approximated by (Proof in Appendix E):

$$E(R_{n,t}) \simeq 0.003 \times 10^{\exp(\mu)} \exp(1.1 \log_{10} C_n),$$

where in reality the value of $E(R_{n,t})$ varies slowly with the time of the day ($\tau \gg t$). This is explored in more detail later.

Given the achievable cell capacity in (8), the load on the cell at time $t$ and the expected load are defined
as:

\[ L_{n,t} = \frac{R_{n,t}}{C_n}, \quad \text{and} \quad \mathbb{E}(L_{n,t}) = \frac{21.6 \exp(1.1 \log_{10} C_n)}{C_n}, \]  

(15) respectively. The paper assumes that the cell is never over-loaded \((L = [0, 1])\).

### B. Energy and Cost Consumption

The total operational energy consumption of a cell over time period \(T\) is modelled as a binary partition of the power consumption [15]:

\[ E_n = \sum_{t=1}^{T} \left( \frac{P_n}{\nu} L_{n,t} + O \right), \]  

(16)

where \(P_n\) is the transmit power, \(\nu\) is the radio-head efficiency, and \(O\) is the overhead power consumption (independent of load). Typical values for a macro-cell with a single transmitting element is: \(P_n = 40\) W, \(\nu = 0.3\), and \(O = 200\) W (including backhaul power consumption). Therefore, even when a cell has no active users and traffic load, the power consumption is \(O\). In sleep mode, the power consumption of the cell is approximately \(\rho O\), where typically \(\rho = 0.5\) [3].

The traffic demand distribution in (13) has a mean of \(\mathbb{E}(R_n) = 21.6 \exp(1.1 \log_{10} C_n)\) bits/s. Therefore, the mean energy consumed by a cell is:

\[ \mathbb{E}(E_n) = T \left( \frac{P_n}{\nu} \mathbb{E}(L_{n,t}) + O \right), \]

\[ = T \left\{ \frac{P_n}{\nu} \left[ \frac{21.6 \exp(1.1 \log_{10} C_n)}{C_n} \right] + O \right\}. \]  

(17)

The energy consumption gain (ECG) is defined as the ratio between energy consumed in the original network and the energy consumed in the proposed network. The energy reduction gain (ERG) is defined as the ratio between energy saved in proposed network and the energy consumed in the original network.

The operational cost expenditure (OPEX) of a cell is modelled as a binary partition [16]:

\[ K_n = \psi E_n + \Psi, \]  

(18)

where \(\psi\) is the price of electricity and \(\Psi\) is the pedestal costs associated with cell maintenance and rentals. Typical values for a macro-cell is: \(\psi = 0.1\) $/kWh, and \(\Psi = 10,000\) $ (not including spectrum
and marketing costs) [3].

IV. SLEEP MODE MECHANISM

A. Definition

The paper proposes a sleep mode mechanism, whereby a cell can enter sleep mode, even when there are active users in its coverage area. These users are handed-off to neighbouring cells by triggering a hand-over procedure. Figure 3 shows an illustration of a cell $n$ in sleep mode, where user $m$ that was previously served by cell $n$ is now served by a neighbouring cell $n'$. The prolonged distance to the new serving cell is $r_{m,n'}$, which is longer than a conventional cell coverage radius [7]. In order to avoid

![Illustration of cellular RAN setup with a cell $n$ in sleep mode. The user $m$ that was previously served by cell $n$ is now served by a neighbouring cell $n'$. That distance of the new serving cell $n'$ to the user is $r_{m,n'}$.](image-url)
comprising the cell-edge downlink quality-of-service (QoS), the operational mechanism is defined as a cell decides to go to sleep if:

1) the traffic load in the cell falls below a certain threshold ($\Xi$);

2) whilst the probability that the data-rate delivered to the cell-edge user maintained above a threshold ($\Theta$) is above a percentage value ($\theta$).

Therefore, each operation mode of cell at time $t$ is:

$$\text{Mode}_{n,t} = \begin{cases} 
\text{Sleep} & L_{n,t} < \Xi \text{ and } \mathbb{P}(C_{m,n',t} > \Theta) > \theta \\
\text{Active} & \text{otherwise}
\end{cases}.$$  \hspace{1cm} (19)

A high sleep-mode trigger threshold $\Xi$ will send more cells into sleep mode and yield a greater energy saving. Therefore, the **objective** is to find: for a given cell-edge performance requirement ($\Theta, \theta$), what is the highest sleep-mode load threshold ($\Xi$) and the resulting energy-cost saving. The paper now examines the two necessary conditions for triggering sleep mode in a cell and their respective probabilities of occurrence.

### B. Trigger 1: Low Traffic Load Probability

The probability that the traffic load ($L_{n,t}$) in a cell is below a certain threshold ($\Xi$) is given by the ratio of CDF of (13) to $C_n$ (Proof in Appendix F): (noting that $\Xi = \frac{R}{C_n}$)

$$\mathbb{P}(L_{n,t} < \Xi) = \int_{L_0 C_n}^{\Xi C_n} f_{R_{n,t}}(z) \, dz,$$

$$= \frac{1}{2} \sum_{M_{n',t}} \left[ \frac{\Upsilon M_{n',t} \exp(-\Upsilon)}{(M_{n',t})!} \left[ \Gamma(L_0 C_n) - \Gamma(\Xi C_n) \right] \right],$$  \hspace{1cm} (20)

where:

$$\Gamma(z) = \text{erf} \left( \frac{\mu + \ln \lfloor \ln(10) \rfloor - \ln \left( \frac{z}{M_{n',t}} \right)}{\sqrt{2} \sigma} \right),$$  \hspace{1cm} (21)

and $\lfloor L_0 \rfloor$ is the largest integer not greater than $L_0$ which is the smallest traffic load in a cell.

Overall, the active cell density $\Lambda_{\text{active}}$ in the RAN is the total cell density multiplied by the proportion
of cells not in sleep mode:

$$\Lambda_{\text{active}} = [1 - \mathbb{P}(L_{n,t} < \Xi | \Theta, \theta)] \Lambda,$$

(22)

where $$\Lambda = \frac{\kappa}{\pi r^2}$$ is the deployed cell density with an average coverage radius of $$r$$. The parameter $$\kappa$$ is a geometric parameter and its physical meaning is explained in Appendix C.

The active cell $$n'$$ has also taken onboard extra users from the neighbouring sleeping cell $$n$$, so the average number of users per cell is modified to:

$$\mathbb{E}(\hat{M}_{n',t}) = \frac{\Upsilon \Lambda}{\Lambda_{\text{active}}} = \frac{\Upsilon}{1 - \mathbb{P}(L_{n,t} < \Xi)},$$

(23)

for a sleep mode threshold of $$\Xi$$. The higher the threshold $$\Xi$$ is, the higher probability that a cell will enter in sleep mode.

C. Trigger 2: Cell-Edge QoS Constraint

Given that the observed cell $$n$$ is in sleep mode. A user that is previous attached to cell $$n$$ has to be now attached to a neighbouring cell $$n' \neq n$$. The user that is the furthermost from cell $$n'$$ is defined as the cell-edge user with a serving signal transmission distance of $$R_{m,n'}$$. The probability that its achievable capacity is above a threshold $$\Theta$$ is a function of the active cell density (22) and bandwidth per user (23) is formed from (10):

$$\mathbb{P}(C_{R_{m,n'},n} > \Theta) = \exp \left\{ -[1 - \mathbb{P}(L_{n,t} < \Xi)] \kappa \left( \frac{R_{m,n'}}{R} \right)^2 Q \right\},$$

for: $$Q(\Theta, \alpha = 4) = \Delta \arctan(\Delta),$$

and: $$\Delta(\Theta, \Xi) = \sqrt{2 \mathbb{P}(L_{n,t} < \Xi) - 1},$$

(24)

where $$R_{m,n'}$$ is the distance from an alternative serving cell to the furthermost user (cell-edge). Given that a cell is in sleep mode, the paper assumes that the ratio between the extended cell coverage range [7] and traditional cell coverage range cannot exceed a ratio of $$\frac{R_{m,n'}}{R} = 2$$, as shown in Figure. 3.

V. Sleep Mode Performance

For LTE standards [12] and referring to (26), the paper considers an example deployment. The specifications are given in Table II, where each cell has $$B = 5$$ MHz bandwidth with an average of $$\Upsilon$$ users,
each with \( \frac{B}{T} \) bandwidth per user on average. The QoS requirement is to be able to achieve a cell-edge capacity of \( \Theta \leq 0.5 \) Mbits/s per user with a reliability probability of \( \theta \leq 0.95 \).

For a QoS constraint at the cell-edge, the probability that the cell-edge capacity is satisfied at a reliability level greater than \( \theta \) is given by:

\[
P(C_{R_{m,n'}} > \Theta) > \theta,
\]

where \( P(C_{R_{m,n'}} > \Theta) \) is defined by (24). For a certain set of edge QoS requirements \((\Theta, \theta)\), the load threshold \((\Xi)\) that maximises the energy saving can be found using the Gauss-Kronrod quadrature technique.

### A. Snap-Shot Mechanism Performance

![Plot of the maximum sleep mode traffic load threshold](image)

Fig. 4. Plot of the maximum sleep mode traffic load threshold \((\Xi)\) that satisfies different cell-edge capacity constraints \((\Theta, \theta)\), given in (25).

Figure 4 shows the maximum sleep mode threshold \(\Xi\), that satisfies the set of edge QoS requirements \((\Theta, \theta)\). The higher the sleep mode threshold, the more likely a cell will enter sleep mode. The results show that the greater the edge QoS requirements \((\Theta, \theta)\) are, the less likely cells can enter sleep mode, which is represented by a low \(\Xi\) value. The results also show that the value of \(\Xi\) is mostly either 0 or approximately 0.4, meaning that all of the cells in the RAN are either in active, or some fixed percentage
of cells are in sleep mode.

The probability of triggering sleep mode is given by the probability of the traffic load falling below $\Xi$, where $\Xi$ is conditioned on a certain edge QoS requirement $(\Theta, \theta)$:

$$
\mathbb{P}_{n, \text{sleep}}(\Theta, \theta) = \mathbb{P}(L_{n,t} < \Xi | \Theta, \theta),
$$

which can be solved by employing the closed-form expression (20). Figure 5 shows the probability of a cell in sleep mode for different bandwidths $B$ and sleep mode thresholds $\Xi$, conditioned on the cell edge QoS given in (25). For the same traffic profile, the intuitive results show that the higher the bandwidth per cell, or sleep mode threshold, the greater the probability of a cell entering sleep mode.

![Plot](image.jpg)

**Fig. 5.** Plot of the probability of a cell in sleep mode for different bandwidths $B$ and sleep mode thresholds $\Xi$, conditioned on the cell edge QoS given in (25).

### B. Time-Varying Network Performance

The energy consumed actually varies slowly with time (hours) across the day $\tau$, which is much greater than the stochastic time effects per hour ($\tau \gg t$) [17]. The energy consumed across a day, for a set of
QoS constraints \((\Theta, \theta)\) is:

\[
E_{n,t,\text{sleep}}(\Theta, \theta) = \sum_{\tau=1}^{24} \left[ P_{n,\tau,\text{sleep}\rho} + (1 - P_{n,\tau,\text{sleep}}) \mathbb{E}(E_{n,\tau,\text{ref}}) \right],
\]

for: \( \mathbb{E}(E_{n,\tau,\text{ref}}) = \frac{P_n}{\nu} \left[ \frac{A(\tau) 10^{\exp(\mu)}}{C_n} \right] + O. \) \hspace{1cm} (27)

The reference (ref.) energy consumption refers to the case where sleep mode is not implemented and is given by (16). The average traffic rate’s mean varies as a function of the hourly time \( \tau \) [17]:

\[
A(\tau) = \mathbb{E}(\hat{M}_{n',t}) \left[ 0.6 - 0.4 \sin \left( \frac{\tau - 2}{12} \right) \right],
\]

where \( \mathbb{E}(\hat{M}_{n',t}) \) is given in (23).

Figure 6 shows the variation in power consumption over time (energy), for a reference RAN and a RAN with the proposed sleep-mode activated. The cell edge QoS requirements are \( \Theta = 0.1 \) Mbits/s, and \( \theta = 0.95 \). The achieved energy reduction gain (ERG) are highest (60\%) during low load conditions (morning, \( 3 < \tau < 11 \) am) and lowest (15\%) during the high load conditions (night, \( 7 < \tau < 9 \) pm).

![Figure 6](image_url)

Fig. 6. Theoretical power consumption results over the course of an average day for a RAN with a cell density of \( \Lambda = 2.54 \) cells per km\(^2\) and a cell edge QoS requirement of \( \Theta = 0.1 \) Mbits/s, \( \theta = 0.95 \). The sleep mode cell power consumption is assumed to be \( \rho = 0.5 \) overhead.

In terms of cost savings, using (18) shows that the maximum OPEX cost saving is approximately 4\% ($400 million per year). That is to say, the energy price at the current moment is not a big fraction of
operator expenditures, but with growing traffic demand, it is set to grow rapidly as a proportion of total OPEX.

C. London Network Performance

The paper now considers the sleep mode threshold mechanism implemented across a simulation of a real London cellular network. The parameters of the simulation are given in Table II. It is worth mentioning that because the data was taken for a particular day in the year, the traffic rate at the beginning and at the end of the day may not be exactly the same. This is valued at a fraction of the mean traffic across all samples in the week.

In the results shown in Figure 7, a number of values for different sleep mode threshold factors $\epsilon$, whereby the threshold for sleep mode is:

$$\xi_{\epsilon,\text{London}} = \epsilon \tau.$$  \hspace{1cm} (29)

The factor $\epsilon$ is added to tailor the mechanism to the realistic London environment. It can be seen that a more aggressive sleep mode can increase the ERG, but in order to satisfy the cell-edge QoS requirements only a set $\epsilon$ can be implemented for the specific London environment.

Fig. 7. Simulated Energy Reduction Gain (ERG) for a London cellular RAN, with different sleep mode threshold factors $\epsilon$ given in brackets in the legend.
In order to check this, Figure 8 shows the reliability probability $\theta$ for cell-edge capacity ($\Theta > 0.15$ Mbits/s) in London Area, for different sleep mode threshold factors $\epsilon$. It can be seen that for a reliability of 0.95 or over, the adjustment factor $\epsilon$ is below 1. The achievable energy savings are between 15 to 45% for $\epsilon = 0.5$. In terms of cost savings, using (18) shows that the maximum OPEX cost saving is approximately 3%. These results closely resembles the energy savings achieved in the theoretical framework, shown in Figure 6. The back-off factor $\epsilon$ is primarily needed due to the effects of antenna gains and terrain clutter, which were not included in the theoretical model.

![Fig. 8. Reliability probability $\theta$ for cell-edge capacity ($\Theta > 0.15$ Mbits/s) in London Area, for different sleep mode threshold factors $\epsilon$.]

That is to say, the stochastic geometry model is a sufficiently good modelling framework for dynamic network adjustments and that mechanisms can be devised that can maximise a certain objective (energy saved), whilst being constrained by QoS requirements (cell-edge capacity and reliability).

VI. DISCUSSION

A. Extension to Heterogeneous Network

In this section, the paper briefly analyses the impact of the sleep mode threshold approach for a heterogeneous network. The section considers a heterogeneous network that consists of $S = 2$ tiers of co-frequency and open access [8], [9]:
1) Macro-cells with deployment density $\Lambda_{\text{macro}}$ cells per square metre;
2) Femto-cells with deployment density $\Lambda_{\text{femto}}$ cells per square metre.

By co-frequency, the paper means that all cells are full-buffer and that they are constantly transmitting data on all radio resource blocks. By open access, we mean that a user will always attach to the cell with the strongest mean signal strength, irrespective of the tier (cell type). The following analysis considers how the sleep mode mechanism’s governing expressions (25), (26) change.

The CCDF of the cell-edge capacity is given by:

$$P(C_{n,\text{RT}} > \zeta) = \sum_{s}^{S} \exp \left[ - \sum_{s' = 1}^{S} \left( \frac{P_{s'}}{P_{s}} \right)^{2} \kappa Q(\zeta, \alpha) \right],$$

which equates to the 1-tier case (10) for $S = 1$. From (30), it can be shown that a $S$ tier heterogenous network has the same mean capacity (bits/s/Hz) per cell coverage area:

$$C_n = \int_{0}^{+\infty} \frac{1}{1 + Q(\zeta, \alpha)} \, d\zeta,$$

which is the same as the 1-tier case in (8). That is to say, a heterogeneous network only provides more bandwidth, but no spectral efficiency gain across the whole network coverage area.

Recall the previous definitions for triggering sleep mode: the traffic load in the cell must be below a threshold and that the cell-edge performance must be above a threshold. For a given traffic profile (13), the load incurred upon the serving cell is a factor of the capacity of cell and the bandwidth available. Given that the spectral efficiency profile across the network is unchanged, only the bandwidth increase in heterogeneous networks will reduce the traffic load experienced. The cell-edge performance is however affected by the transmit powers of all heterogeneous tiers, as well as the bandwidth available in each tier.

**B. Mean and Median Traffic Load**

Recall that the sleep mode mechanism requires estimating the mean traffic rate demanded by users, which was given in (14) from Appendix E. This is in fact a median average, because the traffic random variable considered is unbounded (i.e., from 0 to infinite bits/s). This was used to provide a generic framework to the analysis without setting a numerical upper-limit. If one is set, then the mean traffic rate can be found and the analysis remains the same in this paper.
VII. Conclusions

This paper has presented a sleep mode mechanism that allows cells in a network to switch-off and save energy, whilst maintaining two cell-edge requirements: minimum throughput and reliability. This is performed for real 3G packet data. The analysis employs a double verification technique that consists of a mathematical framework, and a Monte-Carlo simulation model in a real city.

The primary contribution is to show that the theoretical framework can be used to accurately derive trigger thresholds for a realistic network. The capacity and energy results between the real London network and the theoretical model are very similar, and the investigation found that the proposed solution can reduce energy consumption between 15 to 60% throughout the day and reduce annual OPEX by 4% ($400 million per year). Future work will focus on the analysis of sleep mode operations in heterogeneous networks, but preliminary analysis shows that the performance of sleep mode is primarily affected by the additional bandwidth.

APPENDIX A

EXPECTATION OF SINGLE USER LOCATION CAPACITY

Given that the capacity is a R.V. that is strictly non-negative. The expectation of the capacity of a single user location is by definition the integral of the CCDF:

\[
\mathbb{E}(C_{r,n}) = \int_0^{+\infty} \mathbb{P} \left\{ B \log_2 \left[ 1 + \frac{g_{r,n}r^{-\alpha}}{W + I(r,H)} \right] > \frac{\xi}{B} \right\} \, d\xi
\]

\[
= \int_0^{+\infty} \mathbb{P} \left\{ g_{r,n} > \xi \right\} d\xi
\]

\[
= \int_0^{+\infty} \int_0^{+\infty} \int_0^{+\infty} f_G(g) f_R(I|r) \, dg \, dI \, d\xi,
\]

where \( \xi = r^\alpha(W + I(r,H))(2^{\xi/B} - 1) \). The multipath fading has a PDF of \( f_G(g) \sim \exp(\beta) \), and \( f_R(I|r) \) is the joint interference distribution, which includes the distribution of the fading and distance R.V.s.

Inserting the fading distribution yields:

\[
\mathbb{E}(C_{r,n}) = \int_0^{+\infty} \exp(-\beta r^\alpha W(2^{\xi/B} - 1)) \int_0^{+\infty} \exp(-\beta r^\alpha I(2^{\xi/B} - 1)) f_R(I|r) \, dI \, d\xi
\]

\[
= \int_0^{+\infty} \exp(-\beta r^\alpha W(2^{\xi/B} - 1)) \mathcal{L}[\beta r^\alpha(2^{\xi/B} - 1)] \, d\xi,
\]

(33)
where \( L[x] \) is the Laplace transform of the interference term \( I \), evaluated at \( x \):

\[
L[x] = \int_{0}^{+\infty} \exp(-xI) f_R(I|r) \, dI,
\]

for \( x = \beta r^\alpha (2^{\xi/B} - 1) \).

The Laplace transform of the interference signal power is:

\[
L[x] = \int_{r}^{+\infty} \int_{0}^{+\infty} \exp(-\sum_{k \neq n} x g_k r_k^{-\alpha}) f_G(g_k) f_R(r_k) \, dg_k \, dr_k
\]
\[
= \int_{r}^{+\infty} \int_{0}^{+\infty} \prod_{k \in \Phi} \exp(-x g_k r_k^{-\alpha}) f_G(g_k) f_R(r_k) \, dg_k \, dr_k
\]
\[
= \int_{r}^{+\infty} \prod_{k \in \Phi} \frac{1}{1 + (2^{\xi/B} - 1)(r/r_k)^\alpha} f_R(r_k) \, dr_k
\]
\[
= \exp[-\Lambda \pi r^2 Q(\zeta, \alpha)],
\]

where the distance function is Poisson distributed and the PDF is given in (4). The full proof of (35) can be found in [9]. The \( Q(\zeta, \alpha) \) function is given by:

\[
Q(\zeta, \alpha) = \int_{(2^{\xi/B} - 1)^{2/\alpha}}^{+\infty} \frac{(2^{\xi/B} - 1)^{2/\alpha}}{1 + u^{\alpha/2}} \, du,
\]
\[
= \sqrt{2^{\xi/B} - 1} \arctan(\sqrt{2^{\xi/B} - 1}), \quad \text{for: } \alpha = 4
\]

Therefore, substituting \( \beta r^\alpha W(2^{\xi/B} - 1) \) in for \( x \) in the third step of Eq. (35):

\[
L[\beta r^\alpha W(2^{\xi/B} - 1)] = \exp[-\Lambda \pi r^2 Q(\zeta, \alpha)].
\]

The resulting expectation of the capacity of a single user location is:

\[
\mathbb{E}(C_{r,n}) = \int_{0}^{+\infty} \exp(-\beta r^\alpha W(2^{\xi/B} - 1) - \Lambda \pi r^2 Q(\zeta, \alpha)) \, d\zeta,
\]

where \( W \) is the AWGN power defined in (2).
APPENDIX B

EXPECTATION OF MULTI-USER LOCATION CAPACITY

The single user location capacity, which is averaged over all fading and interference terms is given in (38). The network-wide multi-user capacity is defined as a function of the single user location capacity:

\[
\mathbb{C} = \int_0^{+\infty} \mathbb{E} \left[ \log_2(1 + \gamma_{r,n}) \right] f_R(r) \, dr,
\]

\[
= \int_0^{+\infty} \int_0^{+\infty} 2\Lambda \pi r \exp \left\{ -\beta r^\alpha W(2^{\frac{\zeta}{B}} - 1) - \Lambda \pi r^2 \left[ 1 + Q(\zeta, \alpha) \right] \right\} \, d\zeta \, dr. \tag{39}
\]

When the interference signal power is much greater than the AWGN power (interference-limited), the network-wide multi-user capacity is:

\[
\mathbb{C} = \int_0^{+\infty} \frac{1}{1 + Q(\zeta, \alpha)} \, d\zeta \quad \text{for: } W \approx 0 \tag{40}
\]

The integral in (40) can be solved using the Gauss-Kronrod quadrature technique. The average capacity achieved by all user locations is the equivalent to the average capacity achieved by a single cell with a single user, in a multi-cell RAN.

APPENDIX C

CELL-EDGE CAPACITY

Two useful capacity metrics can be derived from the stochastic geometry framework. First of all, the complementary CDF of the average cell capacity is defined as:

\[
\mathbb{P}(C_n > \zeta) = \frac{1}{1 + Q(\zeta, \alpha)}, \tag{41}
\]

as proven previously in (40).

The cell-edge capacity is defined as the capacity achieved by a single user location that is at a distance \( R \) from the serving cell. A multi-cell RAN covers an area \( A \), which upper-bounds the total coverage area of all \( N \) cells with a circular coverage area of radius \( \mathfrak{R} \), such that \( \kappa A = N \pi \mathfrak{R}^2 \), for a geometry constant \( \kappa \leq 1 \). For a cell density of \( \Lambda = N/A = \kappa/\pi \mathfrak{R}^2 \), the CCDF of cell-edge capacity is simply from (38):

\[
\mathbb{P}(C_{\mathfrak{R},n} > \zeta) = \exp[-\beta \mathfrak{R}^\alpha W(2^{\frac{\zeta}{B}} - 1) - \Lambda \pi \mathfrak{R}^2 Q(\zeta, \alpha)],
\]

\[
= \exp[-\Lambda \pi \mathfrak{R}^2 Q(\zeta, \alpha)], \quad \text{for: } W \to 0 \tag{42}
\]

\[
= \exp[-\kappa Q(\zeta, \alpha)] \quad \text{for: } \kappa = \Lambda \pi \mathfrak{R}^2.
\]
APPENDIX D

JOINT PROBABILITY DENSITY DISTRIBUTION OF TRAFFIC DATA RATE

Define the R.V. of $R_{n,t}$ as the traffic rate per cell. The CDF of R.V. $R_{n,t}$ can be expressed as the joint distribution of the number of users in that cell ($M_{n,t}$) and the traffic rate per user ($R_{m,t}$):

$$f_{R_{n,t}}(z) = \mathbb{P}(R_{n,t} \leq z),$$

$$= \mathbb{P}(M_{n,t}R_{m,t} \leq z),$$

$$= \mathbb{P} \left( \bigcup_{m=0}^{+\infty} \left( M_{n,t} = m, R_{m,t} \leq \frac{z}{m} \right) \right),$$

$$= \sum_{m=1}^{+\infty} \mathbb{P} \left( M_{n,t} = m, R_{m,t} \leq \frac{z}{m} \right),$$

$$= \sum_{m=1}^{+\infty} \mathbb{P}(M_{n,t} = m) \mathbb{P} \left( R_{m,t} \leq \frac{z}{m} | M_{n,t} = m \right),$$

$$= \sum_{m=1}^{+\infty} \mathbb{P}(M_{n,t} = m) \int_{\frac{z}{m}}^{+\infty} f_{R_{m,t}}(r_{m,t}; \mu, \sigma) \, dr_{m,t},$$

for $M_{n,t}$ and $R_{m,t}$ are independent R.V.s.

The PDF of $R_{n,t}$ can be obtained by differentiating the above CDF expression with respect to $z$:

$$f_{R_{n,t}}(z) = F'_{R_{n,t}}(z),$$

$$= \sum_{m=0}^{+\infty} \mathbb{P}(M_{n,t} = m) \frac{1}{m} f_{R_{m,t}} \left( \frac{z}{m}; \mu, \sigma \right),$$

$$= \sum_{m=1}^{[z]} \frac{\Upsilon^m \exp(-\Upsilon)}{m!} \exp \left\{ -\frac{\left( \ln \left[\log_{10}(\frac{z}{m})\right] - \mu \right)^2}{2\sigma^2} \right\} \left( z > m \right),$$

where the operation $[z]$ is the largest integer not greater than $z$.

APPENDIX E

EXPECTED VALUE OF TRAFFIC DATA RATE

The expected value of the traffic rate per cell is given by: $\mathbb{E}(R_{n,t}) = \mathbb{E}(M_{n,t}) \mathbb{E}(R_{m,t})$, for independent R.V.s. The value of $\mathbb{E}(M_{n,t}) = \Upsilon$, but $\mathbb{E}(R_{m,t})$ is undefined.

For the traffic rate per user, the paper substitutes the mean with the median. The median ($\tilde{R}_{m,t}$) is
defined as the root to the equation \( \int_{1}^{\tilde{R}_{m,t}} f_{R_{m,t}}(t; \mu, \sigma) \, dt = \frac{1}{2} \), which is:

\[
\begin{align*}
\text{erf} \left( \frac{\mu - \ln \left[ \ln \left( \frac{\tilde{R}_{m,t}}{\ln(10)} \right) \right]}{\sqrt{2} \sigma} \right) &= 0 \\
\mu - \ln \left[ \frac{\ln \left( \frac{\tilde{R}_{m,t}}{\ln(10)} \right)}{\ln(10)} \right] &= 0 \\
\ln \left( \frac{\tilde{R}_{m,t}}{\ln(10)} \right) &= \exp(\mu) \\
\ln(\tilde{R}_{m,t}) &= \ln[10^{\exp(\mu)}] \\
\tilde{R}_{m,t} &= 10^{\exp(\mu)}.
\end{align*}
\]

(45)

For a \( \mu = 1.35 \), this yields a median of data traffic rate of 7.9 kbits/s per user. A discussion on the impact of using the media for this paper can be found in the main body of the paper.

Therefore, the mean of the traffic rate per cell is approximated by:

\[
E(R_{n,t}) \approx E(M_{n,t}) \tilde{R}_{m,t} = 21.6 \exp(1.1 \log_{10} C_n).
\]

(46)

### APPENDIX F

**Sleep Mode Trigger Probability for Traffic Threshold \( R \)**

Given that the traffic per cell is distributed as a PDF \( f_{R_{n,t}}(z) \). The probability that \( z \) falls below some threshold \( R \) is:

\[
F_{R_{n,t}}(z < R) = \int_{z_0}^{R} f_{R_{n,t}}(z) \, dz,
\]

\[
= \sum_{M=1}^{\lfloor z_0 \rfloor} \frac{\Gamma^M}{M!} \exp(-\Upsilon) \times \int_{z_0}^{R} \frac{\exp \left\{ - \frac{\left( \ln \left( \log_{10} \left( \frac{z}{M} \right) \right) - \mu \right)^2}{2\sigma^2} \right\}}{z \log_{10} \left( \frac{z}{M} \right) \sigma \sqrt{2\pi \ln 10}} \, dz,
\]

\[
= \frac{1}{2} \sum_{M=1}^{\lfloor z_0 \rfloor} \frac{\Gamma^M}{M!} \exp(-\Upsilon) \left\{ \text{erf} \left[ \frac{\mu + \ln \left( \ln \left( \frac{z}{M} \right) \right)}{\sqrt{2}\sigma} \right] - \text{erf} \left[ \frac{\mu + \ln \left( \ln \left( \frac{R}{M} \right) \right)}{\sqrt{2}\sigma} \right] \right\}
\]

(47)

for \( z_0 \geq M > \Upsilon \).
REFERENCES


