Article

On Optimal Cell Flashing for Reducing Delay and Saving Energy in Wireless Networks

Jaeik Jeong and Hongseok Kim *

Department of Electronic Engineering, Sogang University, 35 Baekbeom-ro, Mapo-gu, Seoul 121-742, Korea; jaeik1213@sogang.ac.kr
* Correspondence: hongseok@sogang.ac.kr; Tel.: +82-2-705-7989

Abstract: To save energy from cellular networks or to increase user-perceived performance, studying base station (BS) switching on–off is actively ongoing. However, many studies focus on the tradeoff between energy efficiency and user-perceived performance. In this paper, we propose a simple technique called cell flashing. Cell flashing means that base stations are turned on and off periodically and rapidly so that, when one base station is turned on, the adjacent base stations which make interferences are always off. Thus, both energy efficiency and cell edge user performances can be improved. In general, switching off base stations to save energy can lead to longer file download time (or delay) to customers. Using flow-level dynamics, we analyze average delay and energy consumption of cellular networks when cell flashing is used. We show that both of total energy consumption and average flow-level delay decrease in the case of small cells. Extensive simulations confirm that cell flashing can significantly save the energy of the base stations, e.g., by up to 25% and, at the same time, reduce the average delay by up to 75%.

1. Introduction

The main objectives for next generation communications, although they may be different for various organizations, are high data rate, low latency and low energy consumption [1]. Data rate means not only average data rate but also cell edge data rate, which is the worst data rate, because of inter-cell interference. With the deployment of broadband data services, demands of mobile traffic are growing rapidly [2], and thus the 4th generation wireless networks may reach their limits soon. This implies that the above-mentioned three objectives may not be achieved at the same time, and thus to resolve these challenges, research towards the 5th generation of mobile cellular networks is actively being conducted.

To improve user-perceived performance specifically focused on energy efficiency and cell edge users, many cell switch-off techniques have been proposed. The works in [3,4] proposed interference management schemes by switching off the nearest cells. The authors in [5,6] proposed inter-cell interference (ICI) management by exploiting adaptive fractional time reuse (FTR) instead of fractional frequency reuse (FFR). However, these works focused on improving edge user performance and neglected considering energy efficiency. A series of papers [7–13] studied about energy-efficient cellular systems by switching off base stations and/or transmit antennas. The works such as [14,15] proposed energy harvesting wireless communication systems that exploit renewable energy such as solar power, wind power, etc. However, these works did not consider cell edge user performances. The work such as [16] proposed increasing energy efficiency and cell edge user performances simultaneously by cell switch-off with coordinated multi-point (CoMP). The work such as [17]

proposed energy conservation and interference management in a multi-cell time division multiple access (TDMA) network. However, these works consider only the static system where the number of users are fixed, and thus cannot capture real systems where the number of users (or flows, equivalently) changes over time. The work such as [18] studied energy and delay in the downlink of a proportional-fair cellular system with inter-cell interferences and fading. However, this work focused on the trade-off between energy and delay. To the best of our knowledge, increasing energy efficiency and cell edge user performances (or even improving the average performance of all users) simultaneously by cell switch-off, specifically in a dynamic system, has not been studied in the literature.

In this paper, we investigate a mechanism that improves three objectives at the same time, i.e., energy efficiency, cell edge user performances, and the average user performances by managing cell activation in a timely orthogonal fashion. We call it cell flashing because base stations are turned on and off periodically and rapidly so that, when one base station is turned on, the adjacent base stations that make interferences are always off; switching off adjacent cells significantly improves energy efficiency, and cell edge users obviously experience less inter-cell interference. Cell center users who were not experiencing much inter-cell interference, however, might be slightly penalized because the associated base station is not always on. Then, it becomes not clear, even though energy efficiency would be mostly improved with cell flashing, how overall system performance would be. In this regard, we are interested in under which circumstance cell flashing becomes beneficial for both energy efficiency as well as user-perceived performance (overall and/or in the cell edge). Note that, in general, switching off base stations to save energy leads to longer delay, i.e., there is a tradeoff between energy saving and user-perceived performance. Thus, a proper switching on–off mechanism and its theoretical analysis are essential.

One may think that cell flashing may not be practical, at least at this moment, because turning base stations on and off may take significant time, either base station (BS) goes to sleep mode or returns back to active mode, or uplink signals cannot be continuously received. When BS goes to deep-sleep mode, which consumes very low power, it takes a long time and uplink signals cannot be continuously received. However, when fast discontinuous transmission (DTX) is used [19], the transition period between sleep mode and active mode is similar to the case of state on/off change in time division duplexing (TDD) system, and subframes for uplink signals can be continuously received at the expense of energy consumption [20,21]. According to [22], it is possible to switch on/off states in the order of 17 μs. This is a much shorter time than long term evolution advanced (LTE-A) scheduling period of 0.5 ms, and thus cell flashing can be viable. Of course, energy saving of DTX cannot be as much as that of deep-sleep mode.

To better understand energy saving and the user-perceived performance, we consider stochastic traffic loads where new file transfers, or equivalently flows, are initiated at random and leave the system after being served. This is usually referred to as flow-level dynamics [23–25]. Unlike the previous works of FTR [5,6] where energy efficiency was ignored, and static population (or infinitely backlogged users) was assumed, our work is unique in the sense that it captures the impact of dynamic user population on energy efficiency and the long term average performance, which is closer to user-perceived performance, e.g., file download time (or called delay hereafter). Note that user experience is based on the time-averaged performance rather than instantaneous capacity. Analyzing such systems is, however, more challenging because of the dynamic nature of the system, and its analysis is hardly tractable in general [24].

When cell flashing is employed in a dynamic system, we consider the user-perceived performance, i.e., the average delay measured by file download time and the energy consumption of base stations and formulate an optimization problem. Based on flow-level dynamics, we prove that cell flashing can improve both total energy saving and the average delay in the interference-limited regime, which is typically the case of small cells. This can be achieved because large average delay is mainly incurred by cell edge users' poor performances, and thus improved cell edge user

performances make the average delay lower. Thus, cell flashing has three advantages: higher energy efficiency, improved cell edge user performances, and lower average delay in small cells. These benefits are promising because cell densification is essential to achieve higher network-level capacity [2] and lower energy consumption [26]. However, as the cell radius increases, there is a tradeoff between energy saving and reducing delay. Our extensive simulations show that when noise power is negligibly small compared to transmit power, energy consumption of the base stations can be reduced by 25%, and at the same time, the overall average delay is reduced by up to 75% compared to the case of non-flashing. Furthermore, the performance improvement for cell edge users is also noticeable, e.g., the conditional average delay at cell edge is reduced by 80%.

We would like to emphasize that previous papers mainly focused on deep-sleep mode on a long-term scale, but our work analyzes the impact of turning BSs on and off rapidly while receiving uplink signals continuously. Thus, the time scales of energy saving are different, and cell flashing can be implemented independently of deep-sleep mode; even if the BS decides to remain on because traffic volume is moderate, the BS still can be turned on/off rapidly for extra energy saving. In addition, cell flashing can reduce both energy and delay simultaneously in small cells (see Proposition 1).

The rest of paper is organized as follows. In Section 2, we describe the hexagonal cellular model used in our study. In Section 3, we analyze the optimal delay and energy efficiency when cell flashing is used. In Section 4, numerical results are given followed by conclusion in Section 5.
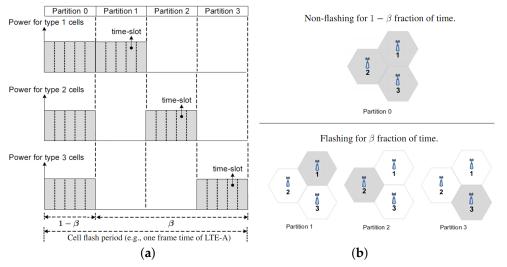
2. System Model

We assume a basic small cell wireless network where each base station is equipped with an omnidirectional antenna. For analytical tractability, we consider small cell hexagonal networks, but our model can be further extended to randomly deployed small cells using stochastic geometry, which is beyond the scope of this paper. We consider a region $\mathcal{L} \subset \mathbb{R}^2$ that is served by a set of base stations $\mathcal{B}$. Let $x \in \mathcal{L}$ denote a location and $i \in \mathcal{B}$ be a BS index. Note that we use the BS index i for cell index because BS and cell refer to the same object with slightly different purposes, which can be understood in the context. Let $k \subset \mathcal{B}$ be a *cluster*, which consists of a set of adjacent cells that may orthogonally use a time-domain resource. Even though the number of the BSs in a cluster can be two or more, *three* is meaningful in a practical system because too many BSs in a cluster sacrifice the benefit of frequency reuse. Thus, we consider three BSs in a cluster hereafter. We assume that file transfer or web page download requests, i.e., flow arrivals follow a Poisson point process with arrival rate per unit area $\lambda(x)$ and file sizes that are independently distributed with mean $1/\mu(x)$ at location $x \in \mathcal{L}$, so the traffic load density is defined by $\gamma(x) := \frac{\lambda(x)}{\mu(x)}$; we assume $\gamma(x) < \infty$ for $x \in \mathcal{L}$. This captures spatial traffic variability. For example, a hot spot can be characterized by a high arrival rate and/or possibly large file sizes [23–25].

Figure 1 depicts the hexagonal cellular networks operated by cell flashing. Figure 1a shows the case when time-domain resources are divided into four partitions. In partition 0, all three BSs transmit while in partition i. Only BS i transmits and other BSs remain silent. Figure 1b shows the coverage based on signal-to-interference-plus-noise ratio (SINR) in each partition. For example, during partition 0, each cell serves its own coverage, while during partition i, only BS i is activated and serves its own coverage. Let β denote a fraction of time that cells flash, i.e., the fraction of time for partitions 1, 2, and 3. Thus, with a ratio of $1 - \beta$, all BSs are on (partition 0), and with a ratio of β, only one BS $i \in \{1, 2, 3\}$ is on as shown in Figure 1b. Thus, the inter-cell interference can be decreased just like frequency reuse 3, and capacity seen by cell edge users can be improved.

One may want to consider load balancing where radio resource, either time or frequency, is allocated depending on traffic variation in space. For example, our previous work of [25] proposed adaptively changing the spectrum bandwidth of each BS in accordance with the loads of BSs. In this work, however, we do not consider load balancing because when partition time is not orthogonalized among adjacent clusters, inter-cluster interferences from adjacent clusters may

become severe. This makes complicated dynamic inter-cell interference unavoidable, and the analysis of flow-level dynamics is not applicable [24,25,27]. In this regard, we consider the case where all cells have the equal partition length. The study on cell flashing with load balancing remains for further study.

Figure 1. Example of cell flashing with four partitions. (**a**) Power configuration in each partition; and (**b**) cell coverage based on SINR. The gray cells represent base stations that are active and the white cells represent base stations that are inactive.

This paper focuses on scenarios where users see (a roughly) static interference from neighboring cells, which was also assumed in [23–25]. Static interference assumption makes sense when interfering cells independently operate with the interfered cell and also interferences are not severe by using, for example, FFR or enhanced inter-cell interference coordination (eICIC). Then, the summation of interference can be considered as static Gaussian-like noise [23–25].

We use Shannon capacity to model the transmission rate to a user at location x, i.e.,

$$c_i(x) = W \log_2(1 + SINR_i(x)), \tag{1}$$

where W denotes the channel bandwidth, and $SINR_i(x)$ is the receive signal-to-interference-plus-noise ratio at location x for the signal from BS i. For analytical tractability, we assume that $c_i(x)$ does not change over time, i.e., we do not consider fast fading or dynamic inter-cell interferences. Instead, $c_i(x)$ can be considered as a time-averaged transmission rate. This assumption is reasonable when the time scale for measuring channel gain $g_i(t)$ is assumed to be much larger than the time scale of fast fading [23,24]. $SINR_i(x)$ is then given by:

$$SINR_i(x) = \frac{P_i g_i(x)}{\sigma^2 + I_i(x)}, \tag{2}$$

where P_i denotes the transmission power spectral density (W/Hz) of BS i, $g_i(x)$ denotes the total channel gain from the BS i to the user equipment (UE) at location x, including path loss, shadowing, and other factors, if any. In addition, σ^2 denotes noise power spectral density (W/Hz) and $I_i(x)$ denotes the average interference power spectral density seen by the UE at location x.

Assumption 1 (fast flashing)**.** We assume that the cell flashing is much faster than flow arrival process, which is reasonable because one frame time in LTE-A is 10 ms, which is much shorter than average file download time that spans seconds to minutes. In addition, we consider that cell flashing period is one frame time of LTE-A or less.

Then, transition time between active and sleep mode becomes negligible because transition time in DTX is 17 µs according to [22]. We can operate cells with non-flashing or flashing mode. If β is 0, all base stations are always on, i.e., non-flashing, and this is same to the case of full time reuse. If β is 1, cellular networks always operate with cell flashing, and each base station is on one third of time in a timely orthogonal fashion. Let $c_i^0(x)$ denote the instantaneous capacity at location x when cells are non-flashing and $c_i^1(x)$ denote the instantaneous capacity when one of three cells are on. Thus, the time-averaged capacity $c_i(x)$ at location x with flashing ratio β is given by:

$$c_i(x, \beta) = (1 - \beta)c_i^0(x) + \frac{\beta}{3} c_i^1(x). \tag{3}$$

It should be noted that $c_i(x, \beta)$ is location-dependent but not necessarily determined by the distance from the BS i. For example, $c_i(x, \beta)$ can be very small in a shadowed area where $g_i(x)$ is very small. Hence, $c_i(x, \beta)$ can capture shadowing as well [25]. According to Assumption 1, we can use the time-averaged capacity Equation (3) to define the system-load density as below:

$$\rho_i(x, \beta) := \frac{\gamma(x)}{c_i(x, \beta)} = \frac{\gamma(x)}{(1 - \beta)c_i^0(x) + \frac{\beta}{3} c_i^1(x)}, \tag{4}$$

which denotes the fraction of time required to deliver traffic load $\gamma(x)$ from BS i to location x. We assume that $\min_i \rho_i(x, \beta)$ is finite, i.e., at least one BS has physical capacity to location $x \in \mathcal{L}$ that is not arbitrarily close to zero [25]. Then, the load of BS i (or utilization) is given by:

$$\rho_i(\beta) = \int_{\mathcal{L}_i} \frac{\gamma(x)}{(1 - \beta)c_i^0(x) + \frac{\beta}{3} c_i^1(x)} \, dx, \tag{5}$$

where $\mathcal{L}_i$ is the coverage of BS i. Note that the meaning of $\rho_i(\beta)$ is the busy fractional time of BS i. We do not consider the unstable system, so $\rho_i(\beta)$ will be smaller than 1 [25].

The system can be modeled by an $M/GI/1$ multi-class processor sharing system, which reflects the fact that users see different service rates and file sizes based on their locations [24]. We also consider infinitely many classes because we address this problem in a continuous space $\mathcal{L}$. The average number of flows at BS i is then simply given by $\frac{\rho_i(\beta)}{1 - \rho_i(\beta)}$ and total number of flows in all base stations is $\sum_{i \in \mathcal{B}} \frac{\rho_i(\beta)}{1 - \rho_i(\beta)}$ [28]. From Little's formula, minimizing the average number of flows is equivalent to minimizing the average delay experienced by a typical flow. Minimizing $\sum_i \frac{\rho_i(\beta)}{1 - \rho_i(\beta)}$ is equivalent to minimizing $\sum_i \frac{1}{1 - \rho_i(\beta)}$ because $\sum_i \left(\frac{\rho_i(\beta)}{1 - \rho_i(\beta)} + 1 \right) = \sum_i \frac{1}{1 - \rho_i(\beta)}$, which serves as one part of our objective functions in Equations (6) and (8).

Our notations are summarized in Table 1.

Table 1. Notation summary (BS: base station).

x	location in continuous space $\mathcal{L}$
$i \in \mathcal{B}$	BS index
β	cell flashing ratio
$\lambda(x)$	flow arrival rate per unit area
$1/\mu(x)$	average file size at x
$\gamma(x)$	$:= \frac{\lambda(x)}{\mu(x)}$, inhomogeneous traffic load density
$c_i^0(x)$	the physical capacity at x from BS i during Partition 0
$c_i^1(x)$	the physical capacity at x from BS i during Partition 1, 2 or 3
$c_i(x, \beta)$	$:= (1 - \beta)c_i^0(x) + \frac{\beta}{3} c_i^1(x)$, reference capacity
$\rho_i(x, \beta)$	$:= \frac{\gamma(x)}{c_i(x, \beta)}$ system-load density (fractional time)
$\mathcal{L}_i$	the coverage of BS i
$\rho_i(\beta)$	$:= \int_{\mathcal{L}_i} \rho_i(x, \beta) dx$, load of BS i having β

3. Optimal Flashing Ratio

Our problem is to find an optimal flashing ratio β so as to minimize the system cost function given below. Let the average delay be $\phi(\beta)$, which is given by:

$$\phi(\beta) := \sum_{i \in \mathcal{B}} \frac{1}{1 - \rho_i(\beta)}. \tag{6}$$

Let the average total power consumption be $\psi(\beta)$. From [23], we know that power consumed by operating BS i is $\rho_i(\beta)P_t + P_o$, where P_t denotes the transmit power, and P_o denotes static power when BS operates. Power consumed by sleeping BS is P_s. Note that BS is on with the fraction of $1 - \frac{2}{3}\beta$, so $\psi(\beta)$ is given by:

$$\psi(\beta) := \sum_{i \in \mathcal{B}} \left\{ \left(1 - \frac{2}{3}\beta \right) (\rho_i(\beta)P_t + P_o) + \frac{2}{3}\beta P_s \right\}. \tag{7}$$

We then formulate our problem as optimization as follows.

Problem:

$$\min_{\beta} \quad f(\beta) = \phi(\beta) + \eta \, \psi(\beta), \tag{8}$$

$$s.t. \quad 0 \le \beta \le 1,$$

$$0 \le \rho_i(\beta) < 1, \forall i \in \mathcal{B}, \tag{9}$$

where $\eta \ge 0$ is the parameter that balances the tradeoff between delay and energy consumption. We analyze $\phi(\beta)$ and $\psi(\beta)$ first to investigate under which circumstance the optimal β^* can be found. Since all base stations are symmetric, we consider BS i in analysis hereafter.

Lemma 1. $\phi(\beta)$ *is a convex function of* β.

Proof. To check the convexity, we obtain the second order derivatives of $\phi(\beta)$, which is given by:

$$\phi''(\beta) = \sum_{i \in \mathcal{B}} \frac{\rho_i''(\beta)(1 - \rho_i(\beta)) + 2\rho_i'(\beta)^2}{(1 - \rho_i(\beta))^3}. \tag{10}$$

We then obtain the second order derivatives of $\rho_i(\beta)$, which is given by:

$$\rho_i''(\beta) = \int_{\mathcal{L}} \frac{2 \left(\frac{1}{3}c_i^1(x) - c_i^0(x) \right)^2 \gamma(x)}{\left((1 - \beta)c_i^0(x) + \frac{\beta}{3} c_i^1(x) \right)^3} \, dx. \tag{11}$$

Then, since $\gamma(x) > 0$ and $(1 - \beta)c_i^0(x) + \frac{\beta}{3} c_i^1(x) > 0$ for $\forall \beta$, $0 \le \beta \le 1$, we have $\rho_i''(\beta) > 0$. Thus, it is obvious that $\phi''(\beta) \ge 0$, which completes the proof. $\square$

Lemma 2. $\psi(\beta)$ *is a monotone decreasing function of* β, *and we have*:

$$\{ \tilde{\beta} = 1 \mid \tilde{\beta} = \operatorname*{argmin}_{\beta} \psi(\beta) \}. \tag{12}$$

Proof. The derivative function of $\psi(\beta)$ is:

$$\psi'(\beta) = \sum_{i \in \mathcal{B}} \frac{2}{3} \left(\left(-\rho_i(\beta) + \left(\frac{3}{2} - \beta \right) \rho_i'(\beta) \right) P_t + P_s - P_o \right). \tag{13}$$

From Equation (5), we can obtain the coefficient of P_t in Equation (13):

$$-\rho_i(\beta) + \left(\frac{3}{2} - \beta\right)\rho_i'(\beta) = -\frac{1}{2}\int_{\mathcal{L}} \frac{\gamma(x)\left(c_i^1(x) - c_i^0(x)\right)}{\left((1-\beta)c_i^0(x) + \frac{\beta}{3}\,c_i^1(x)\right)^2}\,dx. \tag{14}$$

Note that $\gamma(x) > 0$ and $c_i^0(x) < c_i^1(x)$ because of the reduced interference power. Then, it is obvious that the coefficient of P_t in Equation (13) is negative and because operating power is larger than sleep power, the derivative function of $\psi(\beta)$ is negative, which completes the proof. $\square$

From Lemmas 1 and 2 we now know that $\phi(\beta)$ is a convex function of β and $\psi(\beta)$ is a monotone decreasing function of β. Hereafter, we are interested in under which circumstance the optimal solution can be obtained. As one special case, we are interested in when the optimal β^* is 1, i.e., cell flashing all the time is optimal. The following proposition is about cell flashing in interference-limited regime.

Proposition 1 (Cell flashing in interference-limited regime)**.** Suppose that the traffic load density is uniform, i.e., $\gamma(x) = \gamma$ for $\forall x = \mathcal{L}$. Then, in the interference limited regime, $f(\beta)$ is a monotone decreasing function, i.e.,

$$\{\beta^* = 1 \mid \beta^* = \underset{\beta}{\operatorname{argmin}}\ f(\beta)\}. \tag{15}$$

Proof. Since all base stations are symmetric, we focus on BS i in the analysis hereafter. Let the path loss factor be α, $y_i \in \mathcal{L}$ be the location of BS i and $\mathcal{B}_f$ be the set of turned-on base stations when cells flash. Then, for any $x \in \mathcal{L}_i$, $c_i^0(x)$ and $c_i^1(x)$ are given by:

$$c_i^0(x) = W\log_2\left(1 + \frac{|x - y_i|^{-\alpha}}{\Sigma_{j\in\mathcal{B}, j\neq i}\,|x - y_j|^{-\alpha}}\right), \tag{16}$$

$$c_i^1(x) = W\log_2\left(1 + \frac{|x - y_i|^{-\alpha}}{\Sigma_{j\in\mathcal{B}_f, j\neq i}\,|x - y_j|^{-\alpha}}\right). \tag{17}$$

Note that in the interference limited regime, transmit power of the transmitter and the interferers are canceled out. Then, we obtain $\rho_i(\beta)$ as follows:

$$\rho_i(\beta) = \int_{\mathcal{L}_i} \frac{\gamma}{(1-\beta)c_i^0(x) + \frac{\beta}{3}\,c_i^1(x)}\,dx. \tag{18}$$

From Equation (18), we can obtain the derivative function of $\rho_i(\beta)$ as follows:

$$\rho_i'(\beta) = \int_{\mathcal{L}_i} \frac{-\gamma\left(\frac{1}{3}c_i^1(x) - c_i^0(x)\right)}{\left((1-\beta)c_i^0(x) + \frac{\beta}{3}\,c_i^1(x)\right)^2}\,dx. \tag{19}$$

Figure 2 shows $\rho_i'(\beta)$ is negative for all β where the path loss factor is $2 \leq \alpha \leq 7$ [29]. We are only interested in whether $\rho_i'(\beta)$ is negative or not, so we set $W = 1$ and $\gamma = 1$. Thus, $\phi'(\beta)$ is also negative for all β and we can know that $\phi(\beta)$ is a monotone decreasing function in β. It should be noted that we numerically check that $\rho_i'(\beta) < 0$ even when α is unrealistically high, e.g., 20. Since we verify all practically meaningful α, we do not attempt to prove that $\rho_i'(\beta) < 0$ holds for all α. $\square$
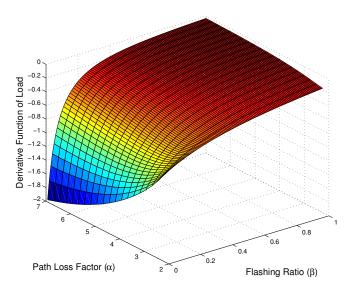
Figure 2. The derivative function of load in interference limited regime.

Remark 3.1. *Note that by cell flashing, we can minimize average delay and average energy consumption simultaneously in the interference limited regime. Interference limited regime is usually the case with small cells. Furthermore, in 5G communication systems, cell coverage will be smaller than now for high data rates and low energy consumption, and thus cell flashing achieves both of them as well as improves the edge user performances.*

Remark 3.2. *When the effect of the noise is high enough, the problem becomes non-trivial because $\phi(\beta)$ can be an increasing function. In Section 4, we consider the impact of noise throughout all simulations.*

Remark 3.3. *In general, $f(\beta)$ is not a convex function. However, if P_t is negligible in Equation (7), which means P_o is much larger than $\rho_i(\beta)P_t$, $\psi(\beta)$ can be considered as a decreasing linear function [30]. Thus, $f(\beta)$ becomes a convex function, and we can easily find the optimal β^*. It is a reasonable assumption because P_o is much larger than P_t in small cells (see Table 2), and $\rho_i(\beta)$ lies between 0 and 1. However, when P_t is not negligible, the optimal β^* can be found by exhaustive search; this may not be technically challenging because there is only one optimization variable.*

Table 2. Power model parameters for different omnidirectional BS types from [22].

BS Type	P_t (W)	P_o (W)	P_s (W)
RRH	5.0	103.0	69.0
	1.0	96.2	62.0
	0.25	13.6	8.6
Pico	5.0	103.0	69.0
	1.0	96.2	62.0
	0.25	13.6	8.6
Femto	0.1	9.6	5.8

4. Simulation Results

In this section, we verify cell flashing by extensive simulations. To evaluate the performance of the proposed scheme, we consider two-tier multi-cell networks composed of 19 hexagonal cells. Each BS has an omnidirectional antenna. The simulation parameters are summarized in Table 3 where power related parameters are from [22]. Since we do not consider fast fading, the average interference power spectral density is computed by the path loss from interfering cells.

Table 3. Simulation parameters.

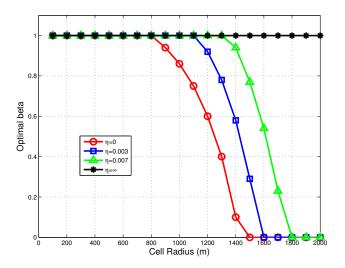| Parameter	Value
Transmission power (P_t)	5 W
Operation power (P_o)	103 W
Sleep power (P_s)	69 W
Average File Size	10 Mbit
Channel bandwidth	20 MHz
Noise power spectral density	−174 dBm/Hz
Path loss factor	3.5

We analyze the optimal β^* with numerical simulations and compare four different cases depending on the use of cell flashing and the cell size as follows:

- Case 1: Cell flashing is not applied for small cells with radius 500 m.
- Case 2: Cell flashing is applied for small cells with radius 500 m.
- Case 3: Cell flashing is not applied for macro cells with radius 1500 m.
- Case 4: Cell flashing is applied for macro cells with radius 1500 m.
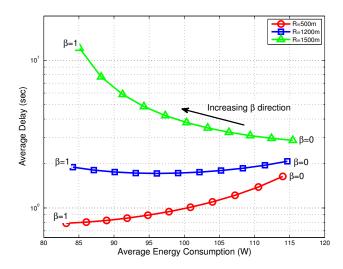
4.1. Optimal β^*

Figure 3 shows the optimal β^* that minimizes $f(\beta)$ as cell radius changes. According to Proposition 1, when cell radius is small, i.e., in the interference limited regime, β^* is 1 regardless of η. This is because, when cell radius is small, the effect of interference power is dominant, so cell flashing reduces both the average delay and energy consumption in the interference limited regime. However, as the cell radius increases, the system operates more and more in a noise-limited regime, and thus β^* decreases, and eventually β^* becomes zero (unless η is infinite). This implies that cell flashing is not suitable for conventional macro cells albeit it can improve edge user performances. For a given cell radius, as η increases, more emphasis is put on energy conservation, and thus β^* increases, i.e., cell flashing becomes more beneficial as expected.

Figure 3. Optimal β^* for minimizing $f(\beta)$ ($\lambda = 1.7$).

Figure 4 shows the relations between the average delay and the average energy consumption as cell flashing ratio β varies when cell radius is 500 m, 1200 m, 1500 m, respectively. When cell radius is 500 m (i.e., β^* is 1), we can see that the average energy consumption decreases when the average delay decreased. This means that cell flashing reduces delay and energy consumption simultaneously. When cell radius is 1200 m (i.e., $0 < \beta^* < 1$), average delay is minimized at some β. When cell radius

is 1500 m, there is a tradeoff between minimizing delay and energy consumption, i.e., to save energy, the delay needs to get longer.

Figure 4. Relationship between average delay and average energy consumption depending on cell radius ($\lambda = 1.7$).
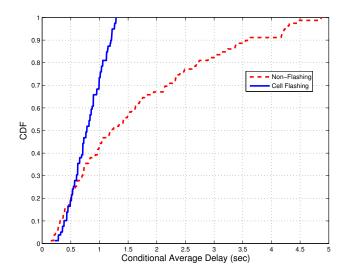
4.2. Cell Flashing vs. Non-Flashing

Figures 5 and 6 show a cumulative distribution function (CDF) of the conditional average delay when cell radius is 500 m and 1500 m, respectively. Let T_i be a random variable denoting the flow-level delay of BS i. In the case of an $M/GI/1$ multi-class processor sharing system model, the conditional average delay experienced by the flow at location x when served by BS i is given by [25]:
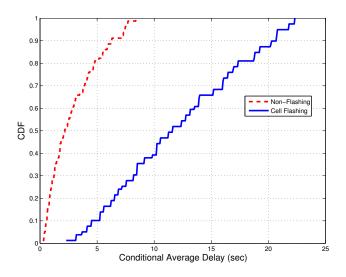
$$
\begin{aligned}
E[T_i|X=x,\beta] &= \frac{1}{\lambda(x)}\frac{\rho_i(x,\beta)}{\rho_i(\beta)}\frac{\rho_i(\beta)}{1-\rho_i(\beta)} \\
&= \frac{1}{\mu(x)c_i(x,\beta)(1-\rho_i(\beta))}.
\end{aligned}
$$

We plot the CDF of the conditional flow delay in Figures 5 and 6 for Cases 1–4. As can be seen in Figure 5, by slightly sacrificing cell center user performances, the overall delay as well as cell edge user performances are significantly improved when cell flashing is used for cell radius 500 m. This is because the impact of inter-cell interference is lessened by cell flashing. However, when the cell radius is 1500, non-flashing is better in terms of average delay.

To see the average delay under various flow arrival rates, we compare Cases 1–4 in Figures 7 and 8 by increasing flow arrival rates when cell radius is 500 m and 1500 m, respectively. When cell radius is 500 m, average delay is reduced by up to 75%. This means the effect of decreased delay of cell edge users is higher than increased delay of cell center users. When cell radius is 1500 m, however, the average delay increases when cell flashing is used. From Figures 5–8, we see that cell flashing is very suitable to small cell networks where lower delay and less energy consumption are achievable.

Figure 5. Cumulative distribution function (CDF) of conditional delay (cell radius = 500 m, $\lambda = 1.7$).

Figure 6. Cumulative distribution function (CDF) of conditional delay (cell radius = 1500 m, $\lambda = 1.7$).
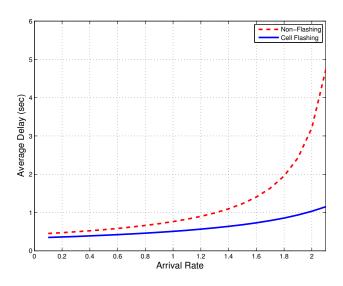

Figure 7. The comparison of average delay vs. arrival rate (cell radius = 500 m).
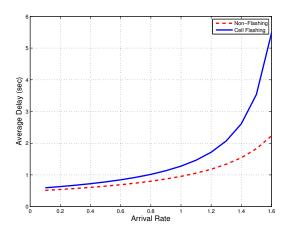
Figure 8. The comparison of average delay vs. arrival rate (cell radius = 1500 m).

5. Conclusions

In this paper, we investigated a technique that turns the BS on and off periodically and rapidly, which is called cell flashing. Because adjacent cells are always off, cell edge user performances and energy efficiency are increased simultaneously by slightly sacrificing cell center user performances. We formulated an optimization problem with flow-level delay and energy consumption. In general, switching off BSs to save energy leads to longer delay but when cell flashing is exploited in small cells, both overall average delay and energy consumption are significantly improved. This implies that cell flashing in small cells has three advantages: higher energy efficiency, lower average delay and also improved cell edge user performances. Our extensive simulations show that when small cells are considered, energy consumption of the BSs can be saved by 25%, and the average delay can be reduced by up to 75% by cell flashing. Furthermore, conditional average delay for cell edge users is reduced by 80%. In our future work, cell flashing can be further extended to randomly deployed small cells using stochastic geometry or also to the case of considering load balancing.

Author Contributions: Jaeik Jeong designed the algorithm, performed the simulations, and prepared the manuscript as the first author. Hongseok Kim led the project and research. Both of the authors discussed the simulation results and approved the publication.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Andrews, J.G.; Buzzi, S.; Choi, W. What will 5G be? *IEEE J. Sel. Areas Commun.* **2014**, *32*, 1065–1082.
2. Yunas, S.F.; Valkama, M.; Niemela, J. Spectral and energy efficiency of ultra-dense networks under different deployment strategies. *IEEE Commun. Mag.* **2015**, *53*, 90–100.
3. Jafar, S.A. Topological interference management through index coding. *IEEE Trans. Inf. Theory* **2014**, *60*, 529–568.
4. Gao, Y.; Wang, G.; Jafar, S.A. Topological Interference Management for Hexagonal Cellular Networks. *IEEE Trans. Wirel. Commun.* **2015**, *14*, 2368–2376.
5. Cho, Y.H.; Seo, S.; Song, J.; Lee, S.; Lee, H.S. Adaptive fractional time reuse for multi-cell OFDMA networks. *IEEE Commun. Lett.* **2013**, *17*, 1798–1801.
6. He, C.; Liu, F.; Yang, H.; Chen, C.; Sun, H.; May, W.; Zhang, J. Co-channel interference mitigation in MIMO-OFDM system. In Proceedings of the International Conference on Wireless Communications, Networking and Mobile Computing (WiCom 2007), Shanghai, China, 21–25 September 2007; pp. 204–208.
7. Niu, Z.; Wu, Y.; Gong, J.; Yang, Z. Cell zooming for cost-efficient green cellular networks. *IEEE Commun. Mag.* **2010**, *48*, 74–79.
8. Niu, Z. TANGO: Traffic-aware network planning and green operation. *IEEE Wirel. Commun.* **2011**, *18*, 25–29.

9. Gong, J.; Zhou, S.; Niu, Z.; Yang, P. Traffic-aware base station sleeping in dense cellular networks. In Proceedings of the 2010 18th International Workshop on Quality of Service (IWQoS), Beijing, China, 16–18 June 2010.

10. Wu, J.; Zhou, S.; Niu, Z. Traffic-aware base station sleeping control and power matching for energy-delay tradeoffs in green cellular networks. *IEEE Trans. Wirel. Commun.* **2013**, *12*, 4196–4209.

11. Niu, Z.; Guo, X.; Zhou, S.; Kumar, P.R. Characterizing energy–delay tradeoff in hyper-cellular networks with base station sleeping control. *IEEE J. Sel. Areas Commun.* **2015**, *33*, 641–650.

12. Vu, Q.D.; Tran, L.N.; Farrell, R.; Hong, E.K. Energy-efficient zero-forcing precoding design for small-cell networks. *IEEE Trans. Commun.* **2016**, *64*, 790–804.

13. Tervo, O.; Tran, L.N.; Juntti, M. Optimal energy-efficient transmit beamforming for multi-user MISO downlink. *IEEE Trans. Signal Process.* **2015**, *63*, 5574–5588.

14. Zhang, S.; Zhang, N.; Zhou, S.; Gong, J.; Niu, Z.; Shen, X. Energy-aware traffic offloading for green heterogeneous networks. *IEEE J. Sel. Areas Commun.* **2016**, *34*, 1116–1129.

15. Xu, J.; Zhang, R. Cooperative energy trading in CoMP systems powered by smart grids. *IEEE Trans. Veh. Technol.* **2016**, *65*, 2142–2153.

16. Cili, G.; Yanikomeroglu, H.; Yu, F.R. Cell switch off technique combined with coordinated multi-point (CoMP) transmission for energy efficiency in beyond-LTE cellular networks. In Proceedings of the 2012 IEEE International Conference on Communications (ICC), Ottawa, ON, Canada, 10–15 June 2012; pp. 5931–5935.

17. Fu, L.; Kim, H.; Huang, J.; Liew, S.C.; Chiang, M. Energy conservation and interference mitigation: From decoupling property to win-win strategy. *IEEE Trans. Wirel. Commun.* **2011**, *10*, 3943–3955.

18. Hossain, M.M.A.; Koufos, K.; Jantti, R. Minimum-energy power and rate control for fair scheduling in the cellular downlink under flow level delay constraint. *IEEE Trans. Wirel. Commun.* **2013**, *12*, 3253–3263.

19. Frenger, P.; Moberg, P.; Malmodin, J.; Jading, Y.; Gódor, I. Reducing energy consumption in LTE with cell DTX. In Proceedings of the 2011 IEEE 73rd Vehicular Technology Conference (VTC Spring), Budapest, Hungary, 15–18 May 2011.

20. Hiltunen, K. Utilizing eNodeB sleep mode to improve the energy-efficiency of dense LTE networks. In Proceedings of the 2013 IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), London, UK, 8–11 September 2013; pp. 3249–3253.

21. Wang, R.; Thompson, J.S.; Haas, H.; Grant, P.M. Sleep mode design for green base stations. *IET Commun.* **2011**, *5*, 2606–2616.

22. NTT DOCOMO, Alcatel-Lucent, Alcatel-Lucent Shanghai Bell, Ericsson, Telecom Italia. 3GPP [R1-113495] Base station power model, 2011. Available online: http://www.3gpp.org/ftp/tsg_ran/WG1_RL1/TSGR1_66b/Docs/R1-113495.zip (accessed on 20 September 2016).

23. Son, K.; Kim, H.; Yi, Y.; Krishnamachari, B. Base station operation and user association mechanisms for energy-delay tradeoffs in green cellular networks. *IEEE J. Sel. Areas Commun.* **2011**, *29*, 1525–1536.

24. Kim, H.; De Veciana, G.; Yang, X.; Venkatachalam, M. Distributed α-optimal user association and cell load balancing in wireless networks. *IEEE/ACM Trans. Netw.* **2012**, *20*, 177–190.

25. Kim, H.Y.; Kim, H.; Cho, Y.H.; Lee, S.H. Self-Organizing Spectrum Breathing and User Association for Load Balancing in Wireless Networks. *IEEE Trans. Wirel. Commun.* **2016**, *15*, 3409–3421.

26. Leem, H.; Baek, S.Y.; Sung, D.K. The effects of cell size on energy saving, system capacity, and per-energy capacity. In Proceedings of the 2010 IEEE Wireless Communications and Networking Conference (WCNC), Sydney, NSW, Australia, 18–21 April 2010.

27. Rengarajan, B.; De Veciana, G. Practical adaptive user association policies for wireless systems with dynamic interference. *IEEE/ACM Trans. Netw. (TON)* **2011**, *19*, 1690–1703.

28. Walrand, J. *An Introduction to Queueing Networks*; Prentice Hall: Upper Saddle River, NJ, USA, 1998.

29. Goldsmith, A. *Wireless Communications*; Cambridge University Press: Cambridge, MA, USA, 2005; pp. 46–48.

30. Lee, G.; Kim, H.; Kim, Y.T.; Kim, B.H. Delaunay triangulation based green base station operation for self organizing network. In Proceedings of the 2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing, Beijing, China, 20–23 August 2013.