# Traffic-Driven Power Saving in Operational 3G Cellular Networks

Chunyi Peng, Suk-Bok Lee, Songwu Lu, Haiyun Luo*, Hewu Li+
UCLA Computer Science, Los Angeles, CA 90095, USA; Tsinghua University, Beijing 100084, China+
{chunyip,sblee, slu}@cs.ucla.edu, hluo@live.com*, lihewu@cernet.edu.cn+

## ABSTRACT

Base stations (BSes) in the 3G cellular network are not energy proportional with respect to their carried traffic load. Our measurements show that 3G traffic exhibits high fluctuations both in time and over space, thus incurring energy waste. In this paper, we propose a profile-based approach to green cellular infrastructure. We profile BS traffic and approximate network-wide energy proportionality using non-load-adaptive BSes. The instrument is to leverage temporal-spatial traffic diversity and node deployment heterogeneity, and power off under-utilized BSes under light traffic. Our evaluation on four regional 3G networks shows that this simple scheme yields up to $53\%$ energy savings in a dense large city and $23\%$ in a sparse, mid-sized city.

## Categories and Subject Descriptors

C.2.1 [**Computer Systems Organization**]: Computer-Communication Networks—*Network Architecture and Design*; C.4 [**Computer Systems Organization**]: Performance of Systems

## General Terms

Design, Measurement, Performance

## Keywords

Energy Efficiency, Cellular Networks, 3G Network Traffic

## 1. INTRODUCTION

We are currently experiencing surging energy consumptions on the wireless cellular infrastructure. Recent reports show energy consumption of mobile networks would reach 124.4B KWh in 2011 [3], and the power bill is expected to double in five years for one Chinese mobile operator [19]. To build a green cellular network, we need to first improve the most critical subsystem that is the dominant contributing factor to overall energy. In the 3G context, it is the base station (BS) subsystem. BSes consume about 80% of overall infrastructure energy, while the user clients typically take around 1% [15].

In this paper, we seek to make the 3G infrastructure more energy efficient. We use real traffic traces, actual BS deployment map and measured BS power consumption, collected from four regional 3G networks, each of which has 45 to 177 BSes and is operated by a largest mobile operator in the world. Our analysis reveals that, 3G traffic load exhibits wide-range fluctuations both in time and over space. However, energy consumption of current networks is not load adaptive. The used energy is unproportionally large under light traffic. The root cause is that each BS is not energy proportional, with more than 50% spent on cooling, idle-mode signaling and processing, which are not related to the runtime traffic load.

We design a solution that approximates an energy-proportional (EP) 3G system using non-EP BS components, in order to cope with temporal-spatial traffic dynamics. The main instrument of our proposal is to completely power off under-utilized BSes when their traffic load is light and power them on when the traffic load becomes heavy. The challenge is to devise a distributed solution that uses a small set of active BSes, while satisfying three requirements of traffic capacity, communication coverage, and minimal on/off switching of each BS. To this end, we take a location-dependent profile-based approach. We divide the network into grids, so that BSes in each local cell can replace each other when serving user clients. We then perform location-dependent profiling to estimate the aggregate traffic among BSes in the grid. Based on the peak/idle of the traffic profile, we decide the corresponding set of active BSes for each duration. It turns out that, if we select the active sets appropriately, we only need to power on a sleep BS and shut down an active BS at most only once during each 24-hour period.

Our evaluation using real traces shows that our scheme leads to average daily energy saving of 52.7%, 46.6%, 30.8% and 23.4% in the four regional 3G networks. The savings are more significant during midnight and weekends and in dense deployment areas, while the miss rate to deny client requests is kept lower than $0.1\%$ in the worst case. While our scheme saves energy on cellular infrastructure, it does negatively increase client power for *uplink* transmission during *idle* hours (e.g., late nights and weekends).

The rest of the paper is organized as follows. Section 2 introduces 3G background and Section 3 explores the operating energy-load curve in 3G networks and models BS power consumption. Section 4 analyzes 3G traffic, and Section 5 describes the proposed solution as well as its implementation within the 3G standard. Section 6 evaluates the performance and Section 7 discusses the related work. Section 8 concludes the paper.

## 2. BACKGROUND

The 3G network infrastructure has two main parts of radio access

*This work was done when he collaborated with UCLA as an independent contributor not associated with his current affiliation.


*MobiCom'11,* September 19–23, 2011, Las Vegas, Nevada, USA.

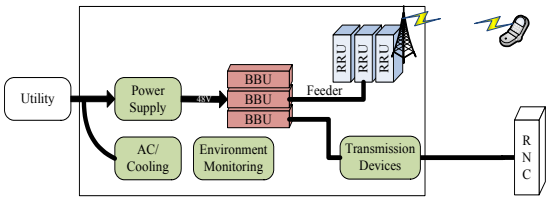



**Figure 1: A typical BS in 3G networks.**

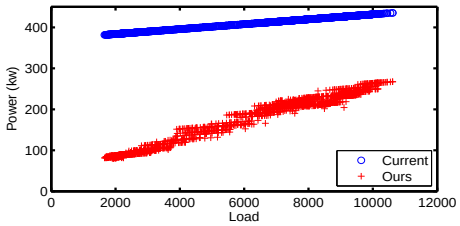



**Figure 2: Energy-Load Curve for Region 1 Network.**

network (RAN) and core network (CN). Its RAN is composed of the User Equipment (UE), the Base Station (BS)[1], and the Radio Network Controller (RNC). Each RNC manages tens of BSes, each of which provides network access services to mobile users via its air interface to the UE.

Figure 1 shows a typical BS in 3G UMTS networks. It has the communication subsystem and the supporting subsystem. The communication subsystem includes Remote Radio Unit (RRU), Base Band Unit (BBU), and Feeder. RRU is the radio specific hardware for each sector. Each BS may install several RRUs near the antennas to provide different coverage and capacity. BBU, as the main unit, provides all other communication functions, including control, base band, switching and Iub interfaces to RNC. Each BS may have several BBUs. Feeder is the optical-fiber pair cable that connects RRUs to BBUs. The supporting subsystem includes the cooling subsystem and other auxiliary devices. The cooling subsystem, including air conditioning and fans, maintains an appropriate operation temperature at the BS. The auxiliary devices include power supply and environment monitoring modules.

From the energy efficiency perspective, the cooling subsystem and some transmission modules consume a significant portion of overall power at each BS, regardless of the traffic load intensity. Our measurement shows that it reaches $50\%$ or more in an operational BS. This is a main factor that leads to energy inefficiency for the 3G infrastructure as we show next.

# 3. TOWARDS TRAFFIC LOAD-ADAPTIVE ENERGY CONSUMPTION

We now describe the problem with current 3G networks from the energy consumption perspective, present the BS power models based on measurements, and identify the roadmap to the solution.

## 3.1 Energy-Load Curve in 3G networks

Our study on real traces of 3G networks shows that the current network operation is not energy proportional to its carried traffic load. The used energy is unproportionally large under zero or light traffic load. Figure 2 shows an illustrative example based on our trace analysis on Region 1 network (see Table 1 for more details). The Region 1 network is an operational 3G network in a big city with 177 BSes shown in Figure 6(a). From the plot of the total consumed power[2] versus the aggregate traffic load in Figure 2, we see that even with light traffic (say, 2000 or below), the consumed power is still quite significant, about 380Kw in total, approximately $95\%$ of the peak power. In contrast, the desired energy proportional operation (also shown in the figure) will consume much less power, about 100Kw in total, under light traffic.

We have also digged into the trace and discovered why. It turns out that the traffic load at each BS varies significantly over time (see Figure 7(a) for a snapshot of traffic at four BSes in different regions). There is a large fraction of time (more than 10 hours over each 24-hour period) that the BS carries very light traffic. This implies that each BS system is not energy proportional to the traffic load. The root cause is that the large fraction of cooling power and fixed radio transmission overhead are invariant of traffic load (as we show next), further contributing to non-energy-proportionality feature at each BS. Therefore, without energy-proportional operations, the 3G network suffers from large energy waste.

## 3.2 Understanding BS Power Consumption

We now model the overall BS power consumption, including both the radio communication and the auxiliary parts (e.g., cooling). We use real measurement data taken on both transmission and cooling systems at BSes. Early models only consider radio transmission but ignore power for cooling and other auxiliary devices [9, 11], or over/under-estimated power consumption coefficients [5, 26].

The total power consumption $P$ at a BS is given by

$$P = P_{tx} + P_{misc},$$

where the first part $P_{tx}$ accounts for power used to provide network access to mobile clients. It includes power consumed by RRUs, BBUs, feeder and RNC transmission. The second part $P_{misc}$ records the auxiliary power for cooling, power supply and monitoring. We next show that $P_{tx}$ mainly changes with carried load while $P_{misc}$ typically remains constant given a fixed operating environment.

**Modeling $P_{tx}$** Using real measurement data on transmission power, we find out that linear models can offer reasonably good approximation for a variety of BSes; This model has also been widely adopted in the literature [5, 9, 11, 26]. Figure 3 gives the scatter plot of power and load at three BSes. The figure clearly shows that a linear model can approximate the transmission power with respect to the carried traffic load, i.e., $P_{tx}(L) = P_\alpha \cdot L + P_\beta$, where $L$ is the utilization level, i.e., the traffic load factor.

The above empirical model can be also explained by the actual BS operations. The two dominant components in $P_{tx}$ are the power consumed by RRUs and BBUs. When the traffic load is heavy, RRU has to spend more power to support more active links. Therefore, it increases in proportion to the traffic volume. On the other hand, BBU does baseband processing for all frequency carriers used by the BS. No matter how many links are active, its power consumption is mainly determined by the number of frequency carriers unless it is in sleep mode. Moreover, signaling over control channels, even during idle modes, also incurs energy overhead on transmission modules.

Note that the power coefficients (i.e., slope and offset) may vary over BSes. This is caused by different vendor products and the changing number of installed BBUs and RRUs at each BS. Product data sheets show that $P_{tx}$ varies from 600w to 3000w [13, 18, 23]. In our model, transmission power also increases when the operational range expands. Specifically, when the BS reaches its maximum transmission range via cell breathing or duplicate long-range

[1]It is also called Node B in the 3G context.

[2]The power is averaged over a time window (e.g., 15 minutes). We do not differentiate between power and energy hereafter.

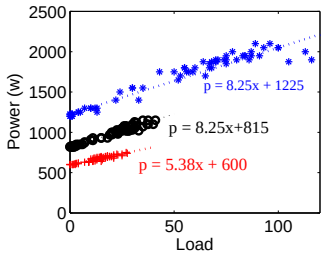



**Figure 3: $P_{tx}$vs. load at 3 BSes.**

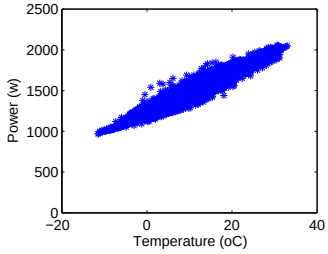



**Figure 4: $P_{misc}$ vs. temperature.**

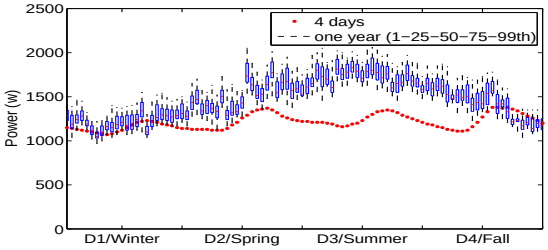



**Figure 5: $P_{misc}$ on 4-day and 1-year time scales.**

radios (see Section 5.3 for details), we model that the power also grows in proportion to the traffic load but uses a larger coefficient $P_a$, say, $P_a$ doubles at its maximum range. Our design and evaluation consider such diversity factors.

**Modeling $P_{misc}$** We focus on modeling the cooling (i.e., air conditioner) power consumption since it is the dominant factor in $P_{misc}$ based on real measurement. It depends on the amount of the extracted heat and the desired operating temperature. It also varies with chillers that use a variety of compressors and drivers. Previous work does not model this part, though it is known that cooling may consume about 50% power at BSes [15].

Figure 4 shows the scatter plot of the cooling power and temperature at a BS in 2010. It can be seen that the cooling power mainly depends on the temperature. It increases approximately linearly from 1000w to 2000w when the environment temperature varies from $-10^oC$ to $30^oC$ (i.e., from winter to summer). We also check daily and yearly pattern in Figure 5. The upper line is the yearly pattern (1-25-50-75-99th percentile in 96 bins) that varies with four seasons. The lower line is a 4-day measurement in early winter. It shows that though the cooling power fluctuates slightly at different hours of a day (e.g., BBUs and RRUs tend to raise the air temperature and the chiller workload). It can still be approximated as a constant within a short period of time (say, a day), here in [1200w, 1400w]. Over a larger time window (say, a year), it varies with the external environment temperature. For simplicity, we assume $p_{misc}$ remains constant on a daily basis but changes with seasons.

In summary, our analysis shows that each BS is not energy proportional to its carried load, mainly due to the residual factors of $P_{misc}$ and $P_\beta$. Recent efforts [12, 14] have been made to reduce them to some extent, but cannot eliminate them.

## 3.3 Roadmap to the Solution

Given that 3G network is not energy proportional to traffic load, our ultimate goal is to build a load-adaptive solution to energy savings in operational 3G networks. To this end, we need to address three issues: (1) What are the characteristics of traffic load in operational 3G networks? We use real traffic traces and BS deployment map to conduct detailed analysis on traffic dynamics both in time and over space (Section 4); (2) Given the traffic dynamics, how can we achieve network-wide energy proportionality (EP) using non-EP components shown in Section 3.2? We need a solution that can achieve load-adaptive energy operation using the current non-EP BS (Sections 5.1 and 5.2). (3) How can the proposed solution work with the current 3G standard? The solution needs to be standard compliant (Section 5.3). We next elaborate on these aspects.

# 4. 3G DATA TRAFFIC: DIVERSITY IN TIME AND SPACE

In this section, we present our measurement results on 3G traffic diversity in both time and space, and show the design insights on how to improve the current 3G network's non-energy-proportionality. We use traces collected from the operational 3G network in four regions to study their temporal-spatial traffic patterns. All four regional networks are managed by one of the largest operators in the world. Figure 6 shows the BS locations in these four regions; we hide the detailed deployment map for privacy concerns. They have different geographic scales and represent diverse city types: Region 1 is a large, populous city, Region 2 is a medium-size city, and Regions 3 and 4 are large cities in a large metropolitan area. All regions have diverse residential and downtown areas. The coverage area and the number of BSes in each region are given in Table 1. Our data sets contain 15min-bin traffic volume records for two months from August 2010 to October 2010. For proprietary reasons, the presented traffic volume is normalized by an arbitrary constant, but normalization does not change the dynamic range in the figures.

| | Region 1 | Region 2 | Region 3 | Region 4 |
|---|---|---|---|---|
| Area (km) | 11x11 | 8x4 | 16x28 | 30x45 |
| # BS | 177 | 45 | 154 | 164 |
| BS density | dense | dense/normal | normal/sparse | sparse |

**Table 1: Basic statistics of 4-region traces.**

## 4.1 Temporal Diversity

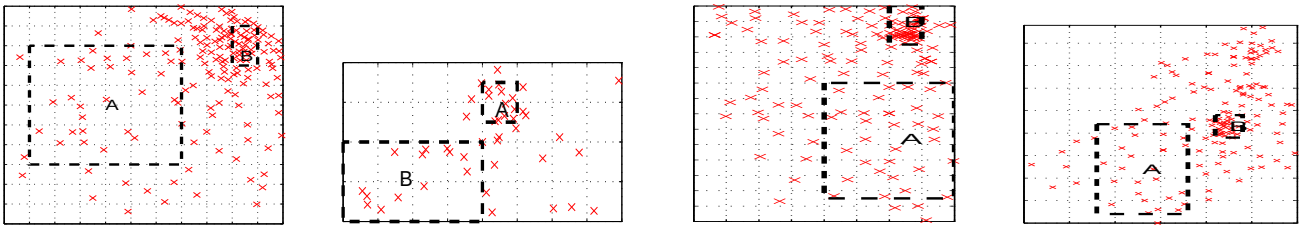
**Temporal traffic dynamics** We first find that each BS exhibits high traffic dynamics over time. Figure 7(a) plots the traffic load at four individual BSes in different regions. We observe strong diurnal patterns on both daily and weekly basis, alternating between peak and idle durations[3]. We separate the weekday and weekend data here, and only present the weekday case unless explicitly stated; the result for weekend is similar.

To quantify the degree of temporal traffic dynamics, we compute the ratio of peak-to-idle traffic load at each BS in four regions. We define the peak (/idle) duration of each BS as the hour $h$, when it has the maximum (/minimum) traffic load (typically between 10AM-18PM for peak, or 1AM-5AM for idle), plus two adjacent hours, i.e., $h-1$ and $h+1$. Figure 7(b) presents CCDF of peak-to-idle traffic-load ratios in four regions. We see that the peak-to-idle traffic ratio is larger than 4 in most (70-90%) BSes, and the smaller ratio (say, 2–4 in Region 1) is only due to relative small traffic volume at BSes. We also study the effect of time window size (here, 3hr) and find that large peak-to-idle ratios still exist when the window is smaller than 8 hours.

**Design insight 1:** *This result shows that the traffic distribution of each BS is quite diverse over time everywhere. Such strong temporal diversity indicates the under-utilization of each BS in the time domain, resulting in system-wide energy inefficiency at BSes.*

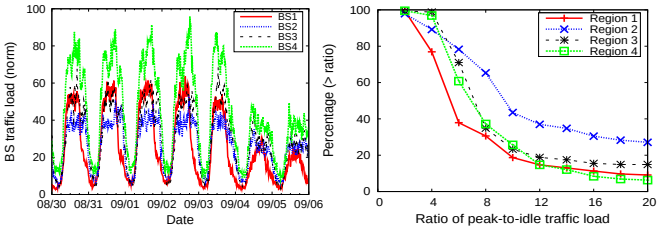
**Near-term traffic stability** We also observe that the traffic volume is stable over short term (e.g., the same time of consecutive days), while it may slowly evolve over a long term (e.g., 26% global increase in 2010 [10]). Although the traffic load fluctuates

[3]We use the term "idle" duration for light traffic cases in our work.




(a) Region 1 (b) Region 2 (c) Region 3 (d) Region 4

**Figure 6: Base station locations in four regions. Dotted rectangles A and B indicate residential and downtown areas.**




(a) One-week traffic pattern (b) Peak-to-idle traffic ratio

**Figure 7: Temporal traffic diversity.**

| | Traffic load variation in consecutive days | | | | |
|---|---|---|---|---|---|
| **Location** | **10th** | **30th** | **50th** | **70th** | **90th** |
| Region 1 | 2.1% | 6.7% | 12.1% | 20.1% | 38.7% |
| Region 2 | 1.9% | 6.4% | 11.8% | 20.6% | 45.0% |
| Region 3 | 2.1% | 6.8% | 12.3% | 20.4% | 42.0% |
| Region 4 | 2.0% | 6.3% | 11.4% | 18.8% | 36.6% |

**Table 2: Near-term stability in four regions. The values indicate the traffic difference between consecutive days.**

over time, the time-of-the-day traffic at each BS is quite stable over consecutive days (see Figure 7(a)). For example, BS1 has similar traffic load at 5 pm in Days 1 and 2, Days 2 and 3, and so on.

We first assess the similarity of near-term traffic by computing their autocorrelation at each BS with the time-lag factor being 24 hours. Our results show that, in all four regions, the autocorrelation values are higher than 0.963 for 70% BSes – confirming strong correlation between traffic load during two consecutive days. To further measure the near-term stability, we also compute the near-term traffic variation $V(i,t)$ at time $t$ at BS $i$:

$$V(i,t) = |R(i,t_{cur}) - R(i,t_{prev})|/R(i,t_{prev}),$$

where $R(i,t_{cur})$ and $R(i,t_{prev})$ denote the traffic load of BS $i$ at time $t$ on the current day and on the previous day. Table 2 shows the near-term variation statistics using our two-month data in four regions. We see that, at any time, the traffic load difference in two consecutive days is less than 20% for 70% BSes. We also note that high variation values are mostly caused by the low traffic volume at the idle time and their absolute values of traffic difference are, in fact, quite small. We further examine the impact of different aggregation granularity (e.g., from 15-min bins to several hours). The near-term variation increases as the aggregation granularity grows, but remains highly stable when the window is smaller than 2 hours.

**Design insight 2:** *The near-term stability result makes a case for traffic profiling to estimate the next day's traffic trend and motivates us to develop power-saving schemes using traffic profiles. The measurement also indicates that hourly traffic aggregation achieves good balance between estimation accuracy and simplicity.*

**Time-domain multiplexing diversity** We find that the aggregate traffic load in a region hardly reaches the aggregate BS capacity in the region. To verify such a trend, we define time-domain "multiplexing" gain $M(t)$ as the ratio of the sum of the peak traffic at each BS (i.e., lower bound of BS capacity) to the aggregate traffic load at time $t$ in the region: $M(t) = \sum_i R(i,t_{max}) / \sum_i R(i,t)$, where $R(i,t)$ is the traffic load of BS $i$ at time $t$; $t_{max}$ is the peak traffic time. Figure 8 plots the multiplexing gains $M(t)$ in four regions. We see that the multiplexing gain is around 2 even during daytime in all regions. Note that the gain can be even larger in reality because the operators often deploy BSes with much larger capacity than the actual traffic demand to account for forthcoming market growth. The root cause for large multiplexing gain is that not all BSes reach their peak load simultaneously. We study the peak hour distribution in subregions A (residence area) and B (business area) and find that the peak hour spans from 10 AM to 6 PM in A, and from 4 PM to 8 PM in B (see Figure 9). The operator has to deploy the infrastructure that can accommodate the peak traffic at each location, even though the peak load may only last two or three hours a day. As the peak hour varies with each location, the deployed capacity (i.e., the sum of each BS's capacity) is much larger than the actual traffic volume at the time. Note that, our observation also explains why current operators tend to be overly conservative in BS deployment density, since they largely ignore the multiplexing effect of traffic load.

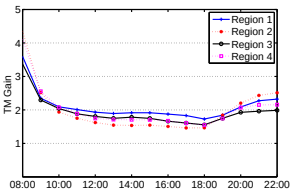



**Figure 8: Time multiplexing gain.**

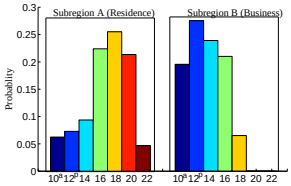



**Figure 9: Peak hour distribution.**

**Design insight 3:** *This multiplexing gain shows that the aggregate BS capacity is highly under-utilized in each region. It also explains why current BS deployment tends to be overly conservative in operational networks. The inherent temporal-spatial diversity opens venue for energy savings via aggregating traffic load.*

## 4.2 Spatial Diversity

**Diverse BS deployment density** The BS deployment density varies across locations (see Figure 6 for location distributions). In the hot spots of a city (e.g., subregion B), more BSes are provisioned, thus creating location-dependent diversity. Figure 10 depicts the distribution of the number of neighbors per BS (within 1 Km, which is the typical communication range of many BS products [22]), representing the BS deployment density in four regions. We see that the deployment density is quite diverse across different regions, as well as in the same region. We also see that a large

number of BSes have multiple neighbors, especially in Regions 1 and 2. For example, for more than half of BSes in Region 1, each has at least 10 neighbors within its 1Km range. In contrast, Region 4 has the most sparse deployment; only 40% BSes have multiple neighbors. The dense BS deployment is partly due to the current practice that operators mostly ignore the traffic multiplexing effect we have discovered before.

**Design insight 4:** *This BS deployment practice provides us an opportunity to exploit such topological "redundancy" for energy-proportional power savings, and the expected gain tends to vary across regions.*

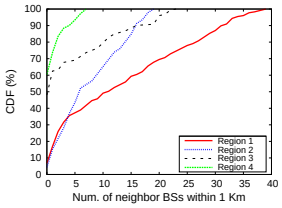



**Figure 10: BS deployment density.**

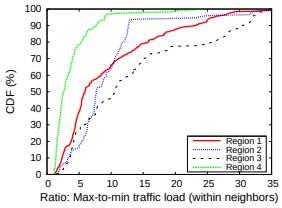



**Figure 11: Neighbor-scale traffic diversity.**

| | Peak time | | | Idle time | | |
|---|---|---|---|---|---|---|
| **Location** | **20th** | **50th** | **80th** | **20th** | **50th** | **80th** |
| Region 1 | 4.0 | 6.5 | 17.1 | 2.3 | 3.9 | 6.0 |
| Region 2 | 6.2 | 9.1 | 12.3 | 3.7 | 6.2 | 8.7 |
| Region 3 | 7.5 | 13.4 | 28.2 | 3.6 | 4.9 | 10.6 |
| Region 4 | 1.7 | 3.2 | 6.0 | 1.5 | 1.9 | 3.1 |

**Table 3: Max-to-min traffic ratio in neighboorhood.**

**Spatial traffic diversity** Another key observation is that traffic load intensity is quite diverse even in each local neighborhood (i.e., traffic loads among the closely located BSes). Figure 11 shows the spatial traffic diversity among neighboring BSes. Each point represents, at any given time of the day, the traffic-volume ratio of the maximum-traffic BS and the minimum-traffic BS within 1 Km range of each BS in four regions. We see that the max-to-min traffic ratio is larger than 5 in 50% cases, and larger than 10 in 30% cases (in Regions 1, 2, and 3)[4]. We also observe that such neighborhood-scale spatial traffic diversity is more evident during the peak time. Table 3 presents the max-to-min BS traffic ratio at peak and idle times. Note that, for example, the spatial diversity at the peak time becomes a factor of 13.4 in 50% cases in Region 3.

**Design insight 5:** *Such strong neighborhood-scale traffic diversity indicates the under-utilization of a group of BSes in the spatial domain, and sheds lights on energy savings in each local area.*

# 5. DESIGN

Using the gained insights on traffic analysis, we seek to achieve load-adaptive energy consumption in the 3G infrastructure. If traffic varies over time (Insight 1) and in space (Insight 5), consuming energy adaptive to the traffic variation becomes critical. The near-term traffic stability (Insight 2) makes a case for *profile*-based approach to estimating traffic envelope at any time. Since the multiplexing gain is high (Insight 3), the profile-based scheme on aggregate traffic, rather than individual BS traffic, will deem more effective. To leverage diversity in BS deployment (Insight 4), we can power off under-utilized BSes when their traffic is light and power them on under heavy traffic. This way, we devise a distributed solution that uses a small set of active BSes based on traffic estimate (defined in terms of active user clients). The solution has to satisfy three requirements of traffic capacity (i.e., traffic does not exceed BS capacity), communication coverage (i.e., each location is covered by at least an active BS), and minimal on/off switching of each BS (i.e., we avoid powering on/off each BS frequently).

Our overall design takes a grid-based, location-dependent profiling approach. We divide the entire network into grids, so that BSes in each local grid cell can replace each other when serving user clients. Once the grid is established, we perform location-dependent profiling, which estimates the traffic envelope for the aggregate BS traffic in the grid. Given the peak and idle hours of the traffic profile, we decide the corresponding set of active BSes for each duration. It turns out that, if we select the sets appropriately, we only need to power on a sleep BS and shut down an active BS only once during each 24-hour period. This minimal on/off switching works well with the cooling subsystem, which needs 10s of minutes when adjusting to the desired operating temperature inside each BS upon power-on.

Our design also eliminates several limitations of the popular optimization-driven approach [9, 11, 21]. These drawbacks include a centralized rather than distributed scheme, approximation to the optimal solution, excessive on/off switching, unrealistic BS power consumption model, one-time optimization targeting instantaneous traffic load, and difficulty in addressing deployment diversity and node heterogeneity. Finally, the related work does not exploit multiplexing gain to minimize active BSes.

## 5.1 Grid-based location-dependent profiling

The grid-based profiling approach estimates traffic in a given area. It addresses two issues: (1) How to determine the grid to partition the network and facilitate powering off under-utilized BSes in a given area? The proposed solution has to accommodate diversity in node deployment and communications. (2) How to perform location-dependent traffic profiling to exploit the multiplexing gain over time and among local BSes? We now elaborate our solution to these two issues.

**Grid size** We partition the grid so that BSes in each grid cell are equivalent. BSes are equivalent if they can replace each other when communicating with user clients. We use location information and transmission range of each BS to decide whether BSes in spatial proximity are equivalent or not. Location coordinates can be obtained by GPS or other location systems when operators plan and deploy their infrastructure. Transmission range of a BS may vary from 200m to 1km in cities and from 1km to 5km in rural areas [22]. It can be different among BSes due to antenna configuration and placement, transmit power and environment.

Specifically, let the distance between two BSes $i$ and $j$ be $d(i,j)$, then BSes $i$ and $j$ are equivalent if

$$r_i + d(i,j) \leq R_j, \qquad r_j + d(i,j) \leq R_i,$$

where $r_i$ and $r_j$ are the normal communication ranges, and $R_i$ and $R_j$ are the maximum possible communication range of $i$ and $j$, respectively. Note that the above procedure can handle the parameter diversity across BSes. Moreover, deployment density can also vary, reflected by changing distance $d_{i,j}$ between any pair of nodes $i, j$. In the example of Figure 12, BS 1 is equivalent to BSes 2 and 3, but is not equivalent to BS 4.

A virtual grid cell is formed when all BSes in it are equivalent. Once a BS is not equivalent to every BS in the current grid, we create a new grid cell. Since grid formation can be nonunique, we use a simple heuristic "northwest rule" to decide our grid construction. We start from the northwest corner in the BS deployment map (i.e., top-left corner), cluster all equivalent BSes from top to down

[4]Region 4 gives the lowest max-to-min traffic ratio due to its sparse BS deployment.

and from left to right, and generate a new grid-cell when a BS is found to not be equivalent to at least one BS in the current cell. We repeat the process until we reach the southeast corner and exhaust all the BSes in the 3G network. In the illustrative example of Figure 12, three grid cells are thus formed following this rule. We note that formation along other directions may generate different virtual grids, but would not much affect energy savings. No matter what formation is created, it does not change the inherent proximity. Close nodes belong to the same grid with high probability. For example, if we form the grid in "northeast" rule (i.e., top-right first), we will get three grids: 6 and 5, 4 and 3, 2 and 1. Each virtual grid still has similar redundancy (the average density is 2 here) and offers local capacity at slightly different spots.

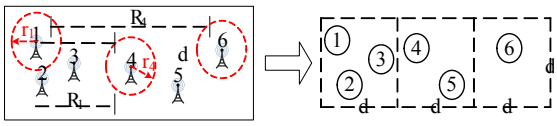



**Figure 12: Example of virtual grid. Left: geo location. Right: virtual grids.**

**Location-dependent traffic profiling** We now devise a profiling scheme that estimates the envelope of aggregate traffic demand in a local grid.

We divide each day into 24 hourly intervals, compute the statistics of each hourly interval, and derive the traffic envelope for the given hour[5]. We differentiate a weekday from a weekend day, but treat all weekdays or weekend-days similarly. Specifically, for the $i$-th hour of $k$-th day that we stack together consecutive weeks[6], we compute the moving average $\bar{S}(i,k)$ and standard deviation $\bar{D}(i,k)$ as follows:

$$\bar{S}(i,k) = (1-\alpha)\cdot\bar{S}(i,k-1) + \alpha\cdot S(i,k),$$

$$\bar{D}(i,k) = (1-\beta)\cdot\bar{D}(i,k-1) + \beta\cdot|S(i,k) - \bar{S}(i,k)|,$$

where $S(i,k)$ is the hourly sample value of the aggregate traffic in the grid for $i$-th hour during the $k$-th day, and $\alpha, \beta$ are the smoothing parameters, chosen as $\alpha = \frac{1}{8}$ and $\beta = \frac{1}{4}$ in our prototype. Consequently, we estimate the hourly traffic envelope as $EV(i,k) = \bar{S}(i,k) + \gamma\cdot\bar{D}(i,k)$ where $\gamma$ is a design parameter that offers a tuning knob to balance between tight estimate and miss ratio. We evaluate its impact on the performance in Section 6.2.

An alternative approach is to first profile each individual BS and then sum up all as the grid profile. It estimates each individual traffic envelope first without extracting the multiplexing effect of traffic among local BSes. In contrast, our group-based profiling can improve energy efficiency when traffic load is heavy. Figure 13 shows an example of 15 BSes in one grid with several micro grids. The peak hours in two micro grids (marked by "+" and "x") vary slightly and exhibit different patterns even within the same grid. As a result, it leads to about 5-8% energy-saving gain at peak hours when using the group profiling scheme (see Section 6.1 for details).

## 5.2 Graceful Selection of Active BSes

Given the traffic profile in each grid, we next select the right set of active BSes and power off under-utilized BSes. The design has to reach two goals of minimizing the number of on/off operations and satisfying both coverage and capacity requirements. To this end, our solution has three components: (1) selection of active BSes for the peak hour(s), (2) selection of active BSes for the idle hour(s), and (3) smooth transition between the idle and the peak.

**Selection of active BSes for peak hour** Given the 24-hour traffic profile at a given grid, we first find the hour(s) with heaviest traffic. For this peak duration, we need to select the set of active BSes in the grid, denoted by $S_{max}$. Based on the fact that the residual energy ($P_{misc}+P_\beta$) of Section 3.2 contributes a large percentage, we reduce the number of active BSes to save energy. On the opposite side, the *local*, aggregate capacity of all active BSes has to be large enough to accommodate *local* traffic. Our algorithm thus prefers the BSes with larger capacity. We rank all the BSes in the grid in decreasing order of their capacity values $C(BS_i)$, say, $C(BS_1) \geq C(BS_2), \ldots, \geq C(BS_n)$. Then we select the largest number $m$ of active BSes so that $\sum_{k=1}^{m} C(BS_k) \geq EV_{max}$. Then, the set of active BSes for the peak hour $S_{max}$ is given by $S_{max} = \{BS_1, \ldots, BS_m\}$. This heuristic ensures the minimum number of active BSes in the grid. Assume that all local BSes use same power models, we can easily prove that the algorithm is optimal to ensuring minimum total energy in the grid. When BSes have heterogeneous power models (i.e., different $P_\alpha, P_\beta, P_{misc}$), we will select the high-energy-efficiency BSes with high priority if their capacity exceeds the traffic demand.

We repeat the above algorithmic procedure for each grid in the network, thus obtaining the active BSes for each grid during its peak hour. Note that the peak hours in different grids may be different. Once active BSes are selected for each grid, our algorithm can meet requirements for both coverage and capacity. Note that two nodes in adjacent grids may cover each other. This offers new opportunity to further save power by merging active BSes in adjacent grids. However, our study shows that this option would add much higher complexity to the design; we trade marginal power savings for design simplicity in this work.

**Comparing with the optimization scheme** We decide to choose the above grid-based scheme, rather than the popular optimization-based scheme in active BS selection. For comparison purpose, we consider an unrealistic, exhaustive search based optimization scheme. It selects the active BS set that consumes minimal energy, while satisfying capacity and coverage constraints.

We first use a simple example to gain insights on why the optimization-based scheme may outperform ours in some scenarios. Figure 16(a) plots the deployment map of nine Region-1 BSes, which are divided into four virtual grids: $\{1,6,8\}$, $\{2,5\}$, $\{3,4,7\}$ and $\{9\}$, following our "northeast" rule. Figure 16(b) marks the active BSes for each hour ( the blue '+' for our grid algorithm and the red circle for the optimal one). We make two observations. First, our algorithm requires at least one active BS for each grid, which may not be necessary in the optimal scheme if the grid can be covered by BSes in neighboring grids. For example, the *grid* scheme has to turn on BS 5 (at midnight) because BSes 2 and 5 belong to one grid and at least one should be on, whereas the *optimal* one can leverage BSes 6 and 8 to cover BS 5, and BSes 3 and 4 to cover BS 2. Therefore, there is no need to turn on BSes 2 and 5 under light traffic. Second, capacity may not be fully utilized due to lack of coordinations among neighboring grids. For example, due to heavy traffic at noon, grid $\{3,4,7\}$ has to turn on two BSes (4 and 7), and grid $\{1,6,8\}$ turns on two BSes (1 and 8). In contrast, the *optimal* scheme can leverage the extra capacity from neighboring grid $\{6,8\}$, thus powering off BS 4, as shown in Figure 16(b). In this case, the performance gap is mainly caused by the unused capacity of BSes, which is not coordinated among grids.

[5]In fact, all $<$ 2-hour intervals have similar and good performance, shown in Section 4.1.

[6]We treat holidays as weekend days, as confirmed by our trace analysis.

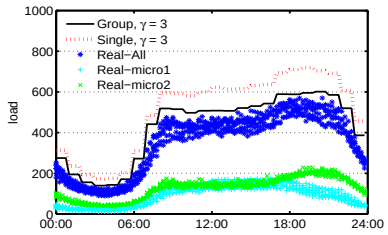



**Figure 13: Profiling Examples.**

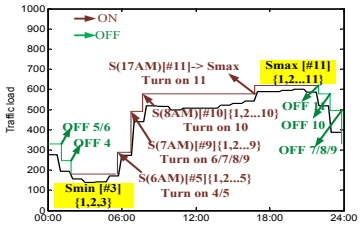



**Figure 14: Illustration of BS graceful selection.**

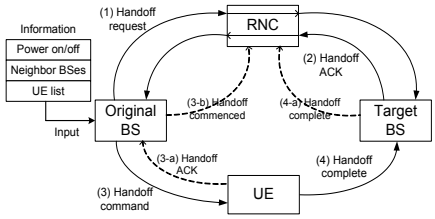
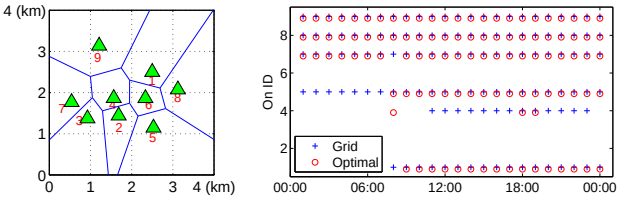



**Figure 15: Handoff procedure for user migration in a 3G network.**




(a) An 9-BSes Example (b) Active BSes Map

**Figure 16: Grid-based vs. optimization-based schemes**

Given the traffic load, and deployment and capacity of each BS, we can formulate the problem as selecting an optimal set of active BSes to minimize energy consumption, similar to the user-cell association problem [11, 16]. We can readily show that this optimization problem is NP hard. Hence, no practical algorithm can reach the optimality. Moreover, in the case where each BS has the same capacity, we can show that the performance gap of these two schemes is upper bounded, as stated by the following theorem, which proof is given in the technical report [25]. We further compare their performance via simulations in Section 6.2.

THEOREM 1. *In the homogeneous capacity setting, our solution has at most $q$ more active BSes than the optimum, where $q$ is the number of grids in the network.*

**Selection of active BSes for idle hour** When deciding the set of active BSes for the idle hour that has the smallest amount of hourly traffic in the 24-hour profile, we devise a different scheme. Rather than select active BSes from all candidates in the grid, we select the active set only from the superset $S_{max}$, which we have calculated for the peak hour. We can use similar selection policy to find the BS set for the idle hour, denoted by $S_{min}$. It is guaranteed to be a subset of $S_{max}$ for the peak hour.

We use the above scheme to minimize the on/off switching of BSes by sticking to the same set of BSes as much as possible. A possible downside is that the computed set may not be optimal since it does not select from all candidates but only those in $S_{max}$. The alternative is to independently derive the set of active BSes for the idle hour. However, the computed set may not be a subset of $S_{max}$, thus incurring more on/off switches during idle-peak migration.

**Smooth transition between idle and peak hours** To minimize on/off switching and reduce energy inefficiency, we devise continuous selection for the rest of the day. It turns out that we need to switch on and switch off each BS at most once during each 24-hour duration. Figure 14 illustrates how our algorithm works for the grid example of Figure 13, where $S_{max}$ has 11 BSes and $S_{min}$ contains 3 BSes.

During the ramp-up transition from the idle hour with smallest traffic volume to the peak hour with heaviest traffic, we use an active node set $S_t$ at hour $t$, which is always a subset of $S_{max}$ but a superset of the previous hour $S_{t-1}$. That is, we find a series of active sets $S\{t\}$ that satisfying $S_{min} = S(t_i) \subseteq S(t_1) \subseteq S(t_2) \cdots \subseteq S(t_p) = S_{max}$, where $t_i < t_1 < t_2 < \cdots < t_p$ denotes the hourly sequence from the idle hour to the peak hour. The algorithmic procedure is similar to that used for the idle hour. When migrating from hour $t-1$ to $t$, we only need to power on those BSes not in $S_{t-1}$, but retain all active BSes in $S_{t-1}$. If $S_{t-1}$ is sufficient, we do not need to power on new BSes. Once a BS appears in $S_{t-1}$, it remains power-on at $t$ and continues to appear in $S_t$. In our example, BSes 4-10 will switch on sequentially based on the prediction of next hourly traffic.

Our solution may incur suboptimal operations for energy savings when the traffic volume experiences a sudden surge (e.g., 11AM - 2PM in the example) at time $t$ or after, before reaching its peak $t_p$. It keeps all current BSes on, though it may be unnecessary. We allow for this sub-optimality to minimize on/off switching. Moreover, our traces show that traffic almost follows diurnal patterns, monotonically increasing during day-time and monotonically decreasing at night. Therefore, the smooth selection works in reality by minimizing on/off switching. Our evaluation in Section 6.2 shows that the power-saving impact is no more than 1% when enabling and disabling smooth selections.

## 5.3 Working within 3G Standards

The above profile-based approach shuts down under-utilized BSes during light-traffic period to save energy. To make it work, we have to be standard compliant and address practical issues: (1) How to let active BSes cover the communication area of those sleep ones; (2) How to effectively migrate existing user clients from the about-to-sleep BSes to other active BSes; (3) How to leverage the 3G infrastructure to share traffic information among local BSes in a grid; (4) How to coordinate the operations of cooling subsystems and Node B communication subsystems during the energy-saving process; (5) How to handle unexpected traffic surge. We now elaborate on these details.

**Adjusting the BS coverage via cell breathing** In our scheme, some BSs need to extend their coverage to serve clients originally covered by neighboring BSs that will power off. To this end, we leverage the well-known "cell breathing" technique that adjusts cell boundaries in today's 3G networks [2, 4]. Cell breathing is traditionally used to adjust the cell size based on the number of client requests to achieve load balancing or capacity increase through micro-cell splitting. We use it for the alternative purpose of power savings. Specifically, the effective service area expands and contracts according to the energy-saving requirement. By increasing the cell radius, an active BS can effectively extend the coverage area to neighboring BSs. Note that most Node B vendors offer products operating over a wide communication range (say, 200m to a few kilometers).

**Alternative solution to BS coverage via duplicate components** An alternative solution to cell breathing is to use dual BBU/RRU subsystems at a BS and switch between these two systems when adjusting the coverage area at peak and idle hours. For example, for a BS in a city area, besides the current subsystem, we install another transmission subsystem that works for rural areas and supports large communication coverage. We then adjust coverage by switching between these two during peak/idle periods that require different transmission ranges. Another alternative is to use lower frequency bands at a given BS and extend its communication range.

**User migration by leveraging the handoff process** When migrating users from the original BS to the equivalent BS for power savings, we leverage the network-controlled handoff (NCHO) mechanism in 3G standards currently used for mobility support.

Figure 15 shows the migration procedure of mobile users to the other active BS when the serving BS decides to power off. For each active UE in the original BS (OBS), the following procedure is performed: (1) The OBS sends a handoff request to the neighboring active BS (ABS) via RNC; (2) The ABS acknowledges the handoff request and reserves resources for the migrating UE; (3) Upon receiving the handoff ACK from the ABS, the OBS sends the UE a handoff command; (4) The UE executes the handoff command via new association with the ABS. Then this handoff process is done in NCHO [1]. In case of handoff failures, the OBS may repeat the above procedure with other active BSes until all UE handoffs succeed or time out. The OBS will defer its power-off if some UEs are still associated with it. Note that our handoff triggering event (i.e., BS power on/off) does not require additional modification to the current 3G NCHO operation except adding one more event type. Thus, the migration process in our power-saving mechanism can be readily made 3G standard compliant.

**Information sharing via RNC** In our group profiling scheme, BSes in the grid should exchange traffic information to compute the envelope for the aggregate traffic. A natural place to exchange such information is via the RNCs. In typical cases, BSes belonging to the same grid also own the same RNC, which is the natural hub for such information exchange and aggregation. In the extreme case that BSes in a given grid do not belong to the same RNC, we can modify the grid construction procedure by imposing the condition that only equivalent BSes belonging to the same RNC can form a grid. The downside of this change is that it may increase the number of grid cells, but with the benefit of reducing inter-RNC message exchange.

**Coordinated operation of cooling and Node B subsystems** Most Node B subsystems require proper operating temperature to function well. When powering off the entire BS for an extended period of time, the ambient temperature may change beyond the desired operating value. Therefore, before powering on the Node B subsystem, we need to power on the cooling/heating subsystem in advance. Our measurements done at three real-life BS machine rooms show that 30 minutes are generally enough for the current cooling/heating system to bring the room temperature to the desired value.

**Emergency BS power-on** While our profile-based approach typically gives a reliable estimate on the traffic envelope, rare-case traffic surge can also occur. To prepare for such transient surges, each active BS monitors its traffic volume. Whenever it sees sudden surge well beyond the envelope specified by the profile, it will notify its RNC. The RNC will subsequently trigger emergency power-on for the neighboring power-off BSes. The power-on number of BSes depends on the traffic surge volume the RNC is notified.

# 6. EVALUATION

We evaluate our power-saving solution using two-month traffic traces collected from four regional 3G networks. We use the first five-week data to construct traffic profiles, and use the remaining three-week traces to assess our solution.

**Evaluation setting** We first evaluate our solution in default parameter settings: (i) profiling parameter $\gamma = 3$; (ii) heterogeneous BS capacity being $110\%$ of the maximum traffic load at a given BS; (iii) power model $P_{tx} = 6L + 600$w and $P_{misc} = 1500$w at normal transmission range; $P_{tx} = 12L + 600$w when expanding to the maximum transmission range. This model states that consumed power still grows linearly with the load but with a larger coefficient, say, $P_a$ doubles at maximum coverage; (iv) the maximal transmission range of 1–2 km, consistent with many available products. We also gauge the effect of various parameters and other power model alternatives, and compare our solution with the optimization-based approach later in this section.

| | **Region 1** | **Region 2** | **Region 3** | **Region 4** |
|---|---|---|---|---|
| $E_{old}$ (Mwh) | 9.81 | 2.63 | 8.58 | 9.18 |
| $E_{bbu}$(Mwh) | 8.3 | 2.3 | 7.5 | 7.9 |
| E Gain | 15.7% | 13.8% | 12.6% | 13.9% |
| $E_{our}$(Mwh) | 4.64 | 1.40 | 5.94 | 7.03 |
| E Gain(%) | 52.7% | 46.6% | 30.8% | 23.4% |
| (min–max) | (34.2–75.9) | (20.6–76.1) | (16.5–46.6) | (9.9–35.4) |
| #miss/BS | 2.83 | 5.23 | 4.37 | 0.12 |
| missRatio(%) | 6.7e-4 | 7.9e-4 | 8.16e-4 | 1.86e-5 |
| #BS(weekday) | 34–97 | 8–32 | 79–122 | 104–142 |
| #BS(weekend) | 34–85 | 8–19 | 79–107 | 103–122 |

**Table 4: Power saving in four regions.**

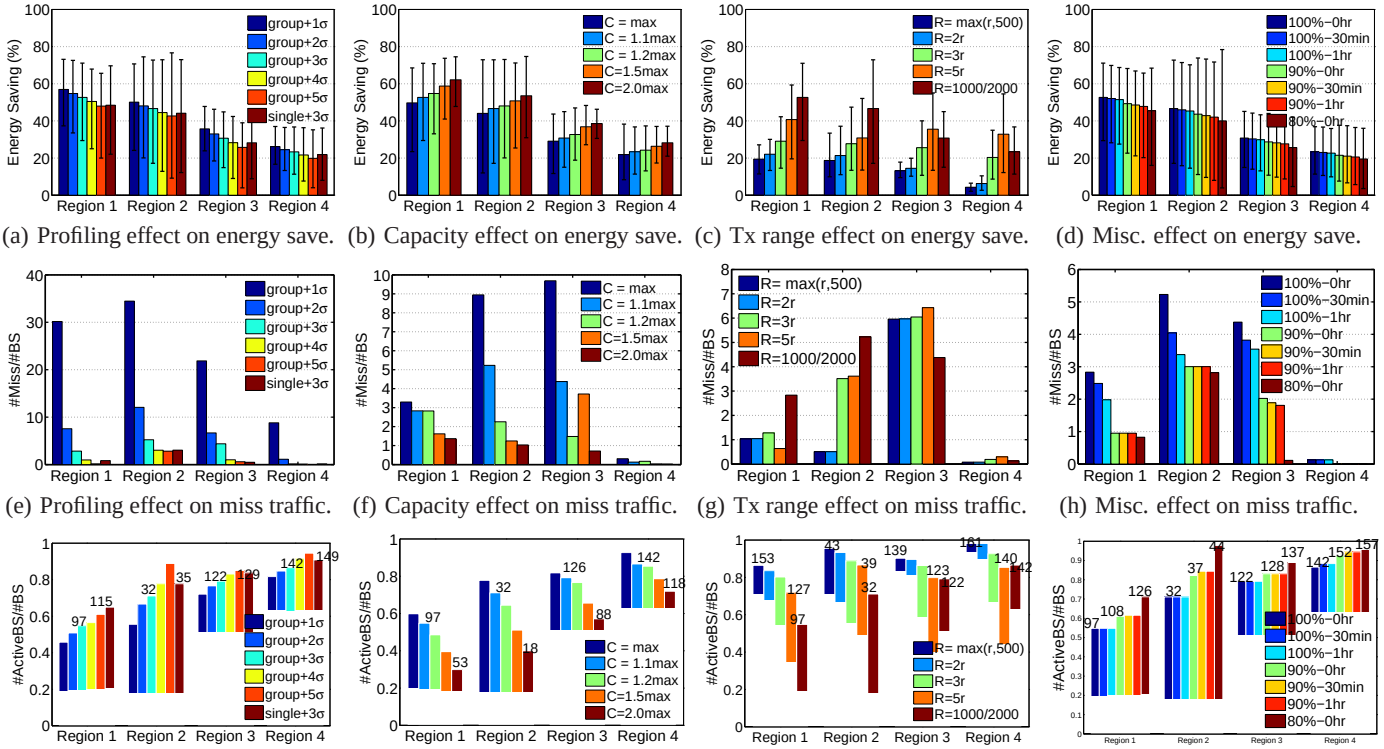
| | Daytime | Midnight | A(sparse) | B (dense) |
|---|---|---|---|---|
| Region 1 | 40.7% | 73.7% | 28.1% | 61.6% |
| Region 2 | 31.2% | 71.6% | 27.7% | 55.3% |
| Region 3 | 20.9% | 45.6% | 8.8% | 51.3% |
| Region 4 | 15.6% | 34.7% | 7.9% | 30.8% |

**Table 5: Power saving in peak/idle hours and subregions.**

**Evaluation results** Table 4 summarizes the results on the above default setting. The table presents the total daily energy consumption of the current 3G network $E_{old}$, BBU-standby solution $E_{bbu}$, and our power saving scheme $E_{our}$, the average energy-saving percentage, and the daily miss traffic (due to profiling inaccuracy or capacity limit) and the active BS count using our scheme. The BBU solution, proposed by some BS vendors [14], aims to turn off some sub-carriers and place BBU into the standby mode when the traffic load is low. We also separate weekday and weekend performances, but they are similar. We only show daily results and active BS sets on weekdays due to space limit.

We make four observations. *First, significant power saving is feasible*. Our profile-based scheme achieves average daily energy savings about 50% in Regions 1 and 2 (dense areas) and 20–30% in Regions 3 and 4 (sparse areas). Compared with the BBU-standby solution, our scheme yields more power savings because the BBU scheme saves $P_\beta$ but cannot eliminate $P_{misc}$. In all cases, 15%–40% BSes are powered on/off in the regional network, i.e., 20–60 BSes each day. Those BSes switch on/off *only once* during each 24-hour period, confirming the operational simplicity of our scheme.

*Second, the power-saving gain is mainly attributed to traffic diversity and deployment density*. Since the wasted energy is unproportionally large under light traffic, our scheme achieves the largest energy savings during idle period. Table 5 shows that, the power-saving gain reaches as high as 70% during night time in Regions 1




(a) Profiling effect on energy save. (b) Capacity effect on energy save. (c) Tx range effect on energy save. (d) Misc. effect on energy save.

(e) Profiling effect on miss traffic. (f) Capacity effect on miss traffic. (g) Tx range effect on miss traffic. (h) Misc. effect on miss traffic.

(i) Profiling effect on active BSes. (j) Capacity effect on active BSes. (k) Tx range effect on active BSes. (l) Misc. effect on active BSes.

**Figure 17: Evaluation results of various effects on energy-saving, miss traffic, and active BS count.**

and 2, and the gain at night time is about 2x the value at daytime in all regions. Deployment density is another crucial factor to power savings. It determines the degree of redundancy to turn off BSes. The gain in dense areas can reach 30–60%, almost 2-6x the value in sparse areas. It also explains why the gain is lower (23.4%) in Region 4, while reaching 52.7% in Region 1. We also assess the impact of grid formation. We find that, the gain is similar no matter what grids are used with different BS sets. The location-dependent density is the key factor of power savings, and grid formation does not affect the inherent density.

*Third, we can save energy by powering off some BSes even during daytime, particularly in dense areas (e.g., Region 1*). Our analysis shows that traffic multiplexing over time and in space is the main contributing factor to such a gain. The current BS deployment does not take the broad system view, but seeks to meet the peak traffic requirement at each location myopically. Consequently, it is inevitable for the operator to over-provision the capacity too aggressively, as observed in Figures 8 and 9.

*Last, our results also reveal the tradeoff between power savings and performance degradation*. Due to occasional unavailability of spare capacity, some user requests will be denied. In principle, our profiling scheme may not always capture extreme-case traffic surge in its profile envelope, thus leading to transient overload beyond the provisioned capacity. However, our study shows that such cases are rare. The average miss ratio is kept as low as $< 0.1\%$, i.e., up to 6 requests per BS each day. If needed, we can use more conservative policies (e.g., use larger profiling margin $\gamma$) or leverage the emergency BS power-on mechanism, to further reduce the miss ratio.

## 6.1 Impact of Various Components

We now study the effect of various components and parameter settings on our energy-saving performance.

**BS power models** The transmission power $P_{tx}$ and the cooling power $P_{misc}$ vary with the sector count and with seasonal changes, respectively. Different vendor products also introduce diversity into power models. Table 6 presents six power models to be assessed. The first five models are homogeneous and the last one assesses the tradeoff between high capacity and high energy efficiency, where BSes with larger capacity consume more power. Because results are similar in other regions, we only show a case in Region 1.

| | | | Region 1 | | |
|---|---|---|---|---|---|
| # | Model | Setting | $E_{old}$ | $E_{our}$ | E gain |
| 1 | Winter | P=1000+6L+600 | 7.7K | 3.6K | 53.2% |
| 2 | Summer | P=2000+6L+600 | 11.9K | 5.3K | 55.4% |
| 3 | Sp/Fall | P=1500+6L+600 | 9.8K | 4.6K | 53.1% |
| 4 | Sp/F-low | P = 1500+4L+600 | 9.5K | 4.2K | 55.8% |
| 5 | Sp/F-high | P = 1500+8L+600 | 10.3K | 5.0K | 51.4% |
| 6 | Hybrid | H, P = 2000+8L+800<br>M, P = 1500+6L+600<br>L, P = 1000+4L+400 | 8.8K | 5.00K | 43.2% |

**Table 6: Energy saving with different power models.**

Our results show that the power model diversity does not much affect the saving gain. It only leads to visible changes in the absolute energy consumption. The power-saving percentage is almost invariant in the five homogeneous cases. It drops about 4-8% in all regions in Model 6, caused by the tradeoff between energy efficiency and capacity. Power models do not affect the miss rate and active sets. The stable power-saving gain in different models indicates that our scheme can work well around the whole year.

**Profiling scheme** We assess the profiling parameter $\gamma$ by varying $\gamma = 1, 2, \ldots, 5$. We also compare the grid-group profiling and the individual BS profiling scheme. Figures 17(a), 17(e), and 17(i)

plot power-saving gain (with min/max bounds), miss requests per BS, and the active BS percentage, respectively. The results show that, when $\gamma$ grows up from 1 to 5, daily energy-saving gain only decreases about 5–10%, which offers large freedom to set $\gamma$. On the other hand, the number of miss requests per BS decreases significantly. The larger $\gamma$ tends to over-estimate the traffic load. We also see that group profiling outperforms the individual one on energy-saving gain and miss rate. Such a gain can be attributed to the effect of traffic aggregation in a local proximity and multiplexing over time, as shown in Figure 8.

**BS capacity** We vary the BS capacity by multiplying variable $\alpha$ and the peak traffic load; we set $\alpha$ from 1 to 2 in our study. Figures 17(b), 17(f), and 17(j) plot power-saving gain, miss requests per BS, and the active BS percentage, respectively. With a larger BS capacity, the network reduces the active BS count while serving the traffic demand, thus reducing power waste. When the BS capacity doubles, energy-saving gain increases by about 10%. However, as we vary BS capacity, the maximum energy-saving gain (when traffic is lightest) is almost invariant. It implies that energy saving under light traffic is constrained by coverage rather than capacity. We also find that increasing BS capacity can offset the profiling inaccuracy, while larger capacity may trigger more BSes to power off, leading to higher miss rate. We note, however, in all cases, the absolute number of miss requests per BS remains small ($<$10, with miss ratio smaller than 0.2%).

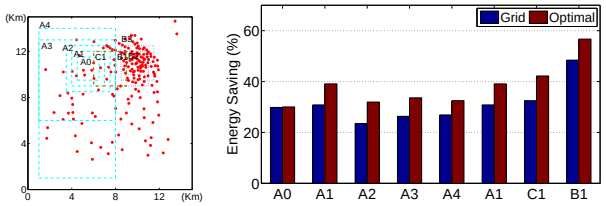
**BS maximum transmission range** We vary the maximum transmission range of a BS by multiplying variable $\alpha$ and the normal transmission range; we vary $\alpha = 1, 2, 3, 5$ in our study. When $\alpha = 1$, we set $R_i = max(500, r_i)$. We also compare them with homogeneous settings of $R$ = 1km or 2km based on BS deployment analysis of [22]. Figures 17(c), 17(g), and 17(k) plot power-saving gain, miss requests per BS, and the active BS percentage, respectively. In general, the larger the transmission range, the more active BSes the network can reduce to cover the entire area. Our results show that we achieve significant power savings when the maximum transmission range is three times larger than the operational range. When it is very small, coverage becomes the limiting factor.

**Other design variants** We also evaluate two more design variants in our scheme. The first is to power on the sleeping BS ahead of the expected working time. It is to give enough time for the cooling system to adjust the ambient temperature inside the BS. The second option is to always reserve a fraction (say, 10%) of the capacity in a BS to be prepared for the worst-case scenario (e.g., unexpected or transient traffic surge). Figures 17(d), 17(h), and 17(l) plot power-saving gain, miss requests per BS, and the active BS percentage, respectively. The results show that, both design variants have little impact on energy saving. The first option (ahead of time) decreases only 1–2% in energy-saving gain, while the 80% resource reservation only reduces 5-10% energy saving gain. Neither visibly affects the miss rate.

## 6.2 Comparing with the Optimization-based Scheme

We now compare our solution with the optimization-based scheme. Specifically, we compare each of the three main design components, i.e., *virtual grid*, *profiling*, and *graceful selection*, with the corresponding idealized or optimization-based solution.

**Virtual grid** Our grid-based scheme can decouple the location-dependent coverage and capacity constraints. The BSes are divided into virtual grids so that BSes in a grid can cover each other, and each grid then makes decision independently based on capacity requirement. We now compare our grid scheme with the optimization-based approach, given the same traffic load. The optimization-based scheme used in our evaluation, uses brute-force search to select the BSes that consume minimum energy while satisfying both capacity and coverage constraints. The exhaustive search offers an upper bound for energy savings, even compared with other optimization-based solutions in the literature [11, 16].



(a) Cases (b) Energy Consumption

**Figure 18: Comparison with the optimization-based scheme in different cases: $A_0 - A_4$ with sparse deployment, $B_1 - B_3$ with dense deployment, $A_1, C_1, B_1$ are 3km$\times$3km areas with varying density.**

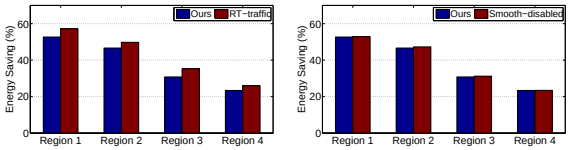
| | A0 | A1 | A2 | A3 | A4 | C1 | B1 | B2 | B3 |
|---|---|---|---|---|---|---|---|---|---|
| #BS | 5 | 9 | 14 | 28 | 54 | 18 | 67 | 54 | 64 |
| Grid(%) | 29.8 | 30.8 | 23.5 | 26.4 | 26.9 | 32.5 | 48.4 | 49.1 | 47.0 |
| Opt(%) | 30.0 | 39.1 | 32.0 | 33.6 | 32.5 | 42.2 | 56.7 | 57.6 | 54.0 |
| $\Delta$(%) | 0.2 | 8.3 | 8.5 | 6.2 | 5.6 | 9.7 | 8.3 | 8.5 | 7 |

**Table 7: Energy consumption using virtual grid and optimization scheme in different cases.**

We compare both schemes in sub-regions of different area size and different deployment density in Region 1: $A0 - A4$ for sparse deployment, $B1 - B3$ for dense deployment, and $A1, C1, B1$ have the same area size but with different deployment density, as shown in Figure 18(a). Figure 18(b) plots the energy-saving percentage using our grid scheme and the optimization-based solution, compared with the All-On option (i.e., all BSes are on). Table 7 also lists the energy-saving gains by both schemes in each sub-region. These results show that the energy-saving gap between our scheme and the optimization scheme is not big, less than 10% in all the cases. We also make interesting observations. When the area size is small, the performance gap also tends to be small. This is because that the optimization-based scheme does not have much room to improve via joint (cross-grid) coordination. As the area size gradually increases, the optimization scheme has larger space to optimize, thus exhibiting larger performance gap. However, when the area size further grows, the gap saturates since the deployment density now becomes the limiting factor. The optimization scheme mainly exploits the local deployment redundancy to improve its power-saving gain. It may turn off more BSes only if each doze-off BS can be covered by several active ones. But this redundancy is ultimately decided by the local deployment density, which is always bounded in reality. Therefore, the optimization scheme cannot yield larger gain as the area increases further. The fundamental reason is that, energy savings are ultimately decided by node deployment density, capacity and coverage. No matter what scheme we use, the selected BSes only work with their inherent deployment and coverage proximity. As long as the deployment density is bounded, the gap between different schemes is also bounded.

**Profile** We use traffic profiles rather than runtime traffic to guide our BS activation and deactivation. The traffic profile approximates the traffic envelope. It is inevitable to overestimate the runtime traffic, and tends to turn on more BSes occasionally. We

compare the energy-saving percentage using our scheme (based on traffic profiles) and the one based on runtime traffic. The results are plotted in Figure 19(a). Our scheme yields 52.7%, 46.6%, 30.8%, and 23.4% of energy savings in Regions 1-4, respectively, whereas the runtime traffic yields 57.1%, 49.8%, 35.3%, and 26.0%, respectively. Their saving-gain difference is between 2.6–4.5%. The reduced energy saving due to the profiling scheme is small (<4.5%) due to two factors. First, our traffic profile estimation offers an accurate traffic upper bound estimate, thus leaving little room for capacity waste. Second, not all the overestimate lead to more active BSes. Each BS capacity is typically a discrete value and may have spare room to accommodate the overestimate incurred by the profiling scheme, thus avoiding more BS activations.




(a) Profiling (b) Graceful selection

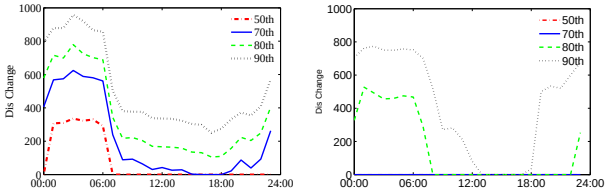
**Figure 19: Comparison of our scheme and other designs on energy saving: (a) using real-time traffic; (b) disabling smooth selection.**

**Graceful selection** We also compare the graceful selection with another solution alternative on energy savings. To reduce the frequency of ON/OFF switches, our scheme selects the active BSes from those already active ones. However, our scheme may lead to retaining unnecessary BSes active or not selecting the most energy-efficient BSes. We compare the energy savings using our scheme (via smooth selection) and the option that disables smooth selection. Figure 19(b) shows the energy-saving percentage of both schemes in four regions. The scheme without smooth selection yields energy-saving gains of 52.8%, 47.2%, 31.2%, and 23.4%, in Regions 1-4, respectively, leading to 0.1–0.7% differences in saving gains compared with our scheme. The energy-saving reduction due to smooth selection is thus negligible (<1%). The reason is that, in rare cases, the traffic envelope is not monotonically increasing. Therefore, the chance is slim when switching off some BSes that are already on for a short time and again turning them on later. In summary, our smooth selection has little negative impact on energy savings, while ensuring the simplicity of smooth BS ON/OFF operations.

## 6.3 Impact on Clients

Our scheme does pay a cost to achieve energy savings on the infrastructure side. It will increase transmit power at client devices when sending uplink data traffic during idle hours (say, late evenings or weekends). In our scheme, when the closest BS powers off, a mobile client will migrate to an active but distant BS, thus incurring additional energy for *uplink transmissions*. However, its impact on the client device is not as severe as it appears. First, the uplink traffic volume is far less than the downlink, which is dominant. The uplink-to-downlink traffic ratio is about 1:8 in the Internet, and 3G networks have similar ratios observed from our traces. Second, the transmission range-extended BSes are conceptually equivalent to the BSes deployed in rural areas, which have larger coverage. Current client devices do not seem to experience severe energy penalty in rural areas. Finally, mobile users are more likely to be uniformly distributed around their serving BSes. Therefore, only a fraction of users will increase their power for uplink transmissions, as shown next.

We quantify the change in transmission range due to our power-saving scheme. Assume uniform distributions for users in their original BSes. Users also associate with their closest active BS. Figure 20 plots the BS-to-client distance change over time in two regions. We see that during daytime, more BSes are active and the distance change is negligible. At night, the distance may increase, e.g., up to 1 Km for 10% clients in Region 1. Table 8 shows the distance change for $n$-th user at a specific time in four regions; it shows that the affected number of clients is still under control.




(a) Region 1 (b) Region 4

**Figure 20: Transmission range change in Regions 1 and 4.**

| | Region 1 | | Region 2 | | Region 3 | | Region 4 | |
|---|---|---|---|---|---|---|---|---|
| | 70th | 90th | 70th | 90th | 70th | 90th | 70th | 90th |
| 4 AM | 588 | 920 | 728 | 991 | 310 | 823 | 0 | 749 |
| 10AM | 64 | 329 | 46 | 159 | 0 | 329 | 0 | 281 |
| 4 PM | 0 | 297 | 0 | 0 | 0 | 17 | 0 | 0 |
| 10PM | 92 | 401 | 176 | 486 | 0 | 341 | 0 | 600 |

**Table 8: Impact on mobile users. The values indicate the BS-to-client distance (m).**

## 6.4 Evaluation Summary

The real trace-based evaluation validates our power-saving solution and shows that significant energy-saving is feasible. It yields up to 52.7% savings in a dense area, and 23.4% in a sparse area. Savings are more significant during night time, e.g., up to 70%, and 1 or 2x larger than those during daytime in all regions. Even during daytime, 20–40% savings are possible by exploiting temporal-spatial multiplexing gains. The traffic miss ratio is also kept lower than 0.1% in the worst case with having the appropriate number of BSes (15–40%) switch on/off at most once during each 24-hour period. Evaluations on various parameters also confirm that our solution is readily applicable to various practical scenarios. For the tradeoff between power-saving gains and the miss rate, our scheme can achieve high power savings as well as low miss rate, e.g., less than 0.1% in our solution. Compared with the optimization-based exhaustive search, our solution achieves effective power savings in a simple and practical manner, while keeping the gap less than 10% in all tested scenarios. On the downside, our scheme may incur increased energy consumption on the client side, but only for uplink traffic and mostly during light-traffic night time.

# 7. RELATED WORK

Energy efficiency in cellular networks has been an active research area in recent years. Many existing studies focus on the client side [6, 7, 27], thus complementing our effort. We focus on the cellular infrastructure side. The overall solution approaches can be classified into two categories: improved component technology and dynamic cell management.

The first approach is to improve energy efficiency of various 3G components, including more efficient power amplifier [18], BBUs with standby mode [14], and optimized cooling [12]. These solutions focus on individual component technology and can work with our scheme. C-RAN [19] proposes to use a cloud-based architecture for energy savings. It deploys RRUs in the field but aggregates all the BBUs into a data center, which uses centralized cooling and

traffic management to be more energy efficient. C-RAN saves system energy but requires an overhaul of current 3G infrastructure. Our design is a near-term solution to the deployed 3G network.

The second approach is to adjust the cell size while turning off idle BSes. Our proposal also conceptually belongs to this category. Current work mostly focuses on the theoretical side by seeking to solve various forms of optimizations [9, 11, 16, 21, 28]. Specifically, [21] studies the optimal time to power off BSes by assuming that all BSes power off simultaneously. [11] formulates the optimal user-BS association as a binary integer programming problem. [9] studies the cell-size optimization given the traffic load. [28] addresses a cost minimization problem that trades-off between energy efficiency and flow performance, and [16] formulates it as a power assignment problem for a specific time interval. Our work differs from all these studies in several aspects. First, we use real traces and measurements, taken from operational networks, in the design and evaluation, without making idealized or simplistic assumptions. Second, we take a novel, grid-based profiling approach, which is distributed rather than centralized. Third, we exploit multiplexing gains to improve energy savings while early studies do not. Finally, we identify and assess various practical factors ignored by early studies in the power-saving operations.

Various energy-efficient techniques have been proposed for other networks, e.g., data center networks [8, 29], the Internet [17] and WLANs [20]. We share the same general guideline to power off idle nodes for maximal power savings. Compared with data centers and the Internet [8, 17, 29], cellular networks exhibit location-dependent non-energy-proportionality. Compared with WLANs [20], cellular networks have different deployment density, node capacity, traffic patterns and power models. They do not leveraging multiplexing for power savings. Our grid-based idea is conceptually similar to GAF [30], which is proposed for energy-efficient ad-hoc routing. Another recent work [24] also studies cellular traffic dynamics, whereas our primary goal is to design a traffic-driven scheme for power savings. The traffic findings are similar, while our traces are also more recent and last longer.

# 8. CONCLUSION

Energy-efficient design has long been an active research area in mobile networks. However, this problem is relatively unaddressed on the 3G infrastructure side, which consumes 99% of overall network energy. In this paper, we propose a location-dependent, profile-based solution, which yields up to $52.7\%$ savings in dense city network, and $23.4\%$ in a mid-sized city with sparse deployment. The key insight gained is to leverage traffic diversity and near-term stability both in time and over space, thus exploiting temporal-spatial multiplexing to save more energy. In the design process, we trade performance increments for simplicity, in that we always retain simple operations rather than squeeze every bit of possible gains. Instead of taking the popular optimization-based approach, we seek to design practical schemes that will work in reality. In a broader scope, our solution explores to build energy-proportional 3G networks using legacy non-energy-proportional base stations.

# Acknowledgments

We appreciate the constructive comments by the anonymous reviewers and UCLA WiNG group members.

# 9. REFERENCES

[1] 3GPP. *TS 29.010: Information element mapping between mobile station - Base station system and base station system mobile-services switching centre (BSS - MSC)*, 2010.

[2] 3GPP. *TS36.902: Self-configuring and self-optimizing network use cases and solutions (V1.0.1)*, 2008.

[3] R. ABI. Mobile networks go green–minimizing power consumption and leveraging renewable energy, 2008.

[4] G. Americas. *The benefits of SON in LTE: self-optimizing and self-organizing networks. White paper*, Dec 2009.

[5] O. Arnold, F. Richter, G. Fettweis, and O. Blume. Power consumption modeling of different base station types in heterogeneous cellular networks. In *Future Network and Mobile Summit'10*, 2010.

[6] N. Balasubramanian, A. Balasubramanian, and A. Venkataramani. Energy consumption in mobile phones: a measurement study and implications for network applications. In *IMC '09*, 2009.

[7] N. Banerjee, A. Rahmati, M. Corner, S. Rollins, and L. Zhong. Users and batteries: interactions and adaptive energy management in mobile systems. In *Ubicomp'09*, Sep 2007.

[8] L. A. Barroso and U. Holzle. The case for energy-proportional computing. In *IEEE Computer*, pages 33–37, 2007.

[9] S. Bhaumik, G. Narlikar, S. Chattopadhyay, and S. Kanugovi. Breathe to stay cool: Adjusting cell sizes to reduce energy consumption. In *Green Networking*, 2010.

[10] Cisco. Cisco visual networking index: Global mobile data traffic forecast update, 2010–2015, Feb 2011.

[11] K. Dufkova, M. Bjelica, B. Moon, L. Kencl, and J.-Y. Le Boudec. Energy savings for cellular network with evaluation of impact on data traffic performance. In *European Wireless 2010*, 2010.

[12] T. Edler and S. Lundber. Energy efficiency enhancements in radio access networks. In *Ericisson Review*, 2004.

[13] Ericsson. RBS 3418 product description.

[14] Ericsson. Energy-saving solutions helping mobile operators meet commercial and sustainability goals worldwide, June 2008.

[15] G. P. Fettweis and E. Zimmermann. Ict energy consumption–trends and challenges. In *11th International Symposium on Wireless Personal Multimedia Communications*, Lapland, Finland, Sep 2008.

[16] G. Fusco, M. Buddhikot, H. Gupta, and S. Venkatesan. Finding green spots and turning the spectrum dial: Novel techniques for green mobile wireless networks. In *DySPAN'11*, Aachen, Germany, 2011.

[17] M. Gupta and S. Singh. Greening of the internet. In *SIGCOMM'03*.

[18] Huawei. Efficient power amplifier: the trend for the development of Node B.

[19] C. M. R. Institute. C-RAN: Road towards green radio access network. In *White Paper, V1.0.0*, April 2010.

[20] A. P. Jardosh, K. Papagiannaki, E. M. Belding, K. C. Almeroth, G. Iannaccone, and B. Vinnakota. Green WLANs: On-demand WLAN infrastructures. *MONET*, 14(6), 2009.

[21] M. A. Marsan, L. Chiaraviglio, D. Ciullo, and M. Meo. Optimal energy savings in cellular access networks. In *GreenComm'09*, 2009.

[22] A. R. Mishra. *Fundamentals of cellular network planning and optimisation:2G/2.5G/3G...evolution to 4G*. Wiley, 2004.

[23] Motorola. Horizon 3G-n macro outdoor data sheet.

[24] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das. Understanding traffic dynamics in cellular data networks. In *INFOCOM'11*, April 2011.

[25] C. Peng, S.-B. Lee, and S. Lu. Green 3G cellular network infrastructure: a traffic-driven approach. Technical report, UCLA CS, 2011.

[26] F. Richter, A. J. Fehske, and G. P. Fettweis. Energy efficiency aspects of base station deployment strategies in cellular networks. In *VTC'09 Fall*, 2009.

[27] A. Schulman, V. Navda, R. Ramjee, N. Spring, P. Deshpande, C. Grunewald, K. Jain, and V. N. Padmanabhan. Bartendr: a practical approach to energy-aware cellular data scheduling. In *MOBICOM'10*, 2010.

[28] K. Son, H. Kim, Y. Yi, and B. Krishnamachari. Base station operation and user association mechanisms for energy-delay tradeoffs in green cellular networks. Technical report, USC, Dec 2010.

[29] N. Tolia, Z. Wang, M. Marwah, C. Bash, P. Ranganathan, and X. Zhu. Delivering energy proportionality with non energy-proportional systems: Optimizing the ensemble. In *HotPower'08*, 2008.

[30] Y. Xu, J. Heidemann, and D. Estrin. Geography-informed energy conservation for ad hoc routing. In *MOBICOM'01*, 2001.