SpeedBalance: Speed-Scaling-Aware Optimal Load Balancing for Green Cellular Networks

Kyuho Son and Bhaskar Krishnamachari
Department of Electrical Engineering, Viterbi School of Engineering
University of Southern California, Los Angeles, CA 90089
Email: {kyuhoson and bkrishna}@usc.edu

***Abstract*—This paper considers a component-level deceleration technique in BS operation, called speed-scaling, that is more conservative than entirely shutting down BSs, yet can conserve dynamic power effectively during periods of low load while ensuring full coverage at all times. By formulating a total cost minimization that allows for a flexible tradeoff between delay and energy, we first study how to adaptively vary the processing speed based on incoming load. We then investigate how this speed-scaling affects the design of network protocol, specifically, with respect to user association. Based on our investigation, we propose and analyze a distributed algorithm, called *SpeedBalance*, that can yield significant energy savings.**

I. Introduction

Recently, potential harmful effects to the environment caused by CO_2 emissions and the depletion of non-renewable resources bring renewed focus on the need to develop more energy-efficient underlying network infrastructures [1]. In particular, the focus of this paper is on reducing the power consumption at base stations (BSs) as they are the key source of heavy energy usage in cellular networks, reported to amount to about 60-80% [2]. From the perspective of mobile network operators, reducing energy consumption is not only a matter of social responsibility towards being green and sustainable but also tightly related to their business survivability in coming years. They are spending huge operational expenditures (OPEX) to pay electricity bills. Moreover, it is expected to grow due to explosive growth in data demand and the possible increase of energy price [3]. According to a study from ABI Research [4], the collective cellular network OPEX will reach $22 billion in 2013. Thus, reining back the spiraling OPEX is crucial to the continuing success of operators.

There have been many studies on dynamic BS switching techniques for energy conservation [2], [5]–[8], which allow the system to entirely shut down some underutilized BSs and transfer the corresponding load to neighboring BSs during low traffic periods such as nighttime. It has substantial potential to obtain energy savings by even reducing static (or standby) power. Nevertheless, the operators are reluctant to turn off their BSs not only due to the technical challenges of implementing it in practice, but also due to concerns about possible serious degradation in user experience: *(i)* users originally in the switched-off cell need to communicate with farther BSs (e.g., more mobile power consumption for file uploading), and *(ii)* there is always a danger of creating coverage holes.

Thus, in this paper, we consider to incorporate a component-level deceleration technique in BS operation that is more conservative than turning off BSs, yet can conserve dynamic power effectively. This technique, called dynamic voltage frequency scaling (DVFS, or simply *speed-scaling*) [9], [10], allows a central processing unit (CPU) to adapt its speed for energy conservation based on incoming processing demand. Note that examples of in-BS processing are increasingly abundant from OFDM modulation, coding, to even security and multimedia conversion. It is also worthwhile mentioning that DVFS lowers heat dissipation as well. As a consequence, it can reduce the power consumption in cooling equipment contributing to a considerable amount of total energy consumption, where this exerting influence is often linear [11].

In the meantime, measurements of real BSs over several days indicate that the power consumption varies only about 2% for a GSM BS and 3% for a UMTS BS over time regardless of its load level [11]. This implies that typical macro BSs deployed today do not adopt dynamic power saving features. More recently, however, Alcatel-Lucent has demonstrated the feasibility of exceptional dynamic power savings on BSs by software upgrades [12] and it can be expected that such BSs will become even more widespread in the near future. Nevertheless, the applications of these features to BSs and their impacts on the design of network protocols in cellular networks have not been fully understood yet.

Our objective and contributions: The goal of this paper is to *(i)* characterize an equilibrium resulting from the interaction between speed-scaling and load balancing for green cellular networks and to *(ii)* propose a distributed iterative optimal speed control and user association policy. The main contributions of the paper are summarized as follows:

1) We develop a theoretical framework for BS energy saving that jointly encompasses speed-scaling and user association. To the best of our knowledge, this work is the first to consider speed-scaling as a tool addressing a flexible tradeoff between delay performance and energy consumption in both networking and processing components of BSs.
2) We first derive an optimal processing speed for two different processors having different capabilities: static speed-scaling and gated-static speed-scaling, and then present the optimal structure of speed-scaling-aware load balancing.

This research was supported in part by NSF NetSE grant 1017881.

Motivated by the above, we propose SpeedBalance, an algorithm that can be implemented in a totally distributed manner. We further evaluate the performance of SpeedBalance through extensive simulations under an acquired 3G cellular topology and traffic trace.

II. System Model and Problem Formulation

A. Model and Notation

1) Network and traffic model: We consider a downlink cellular wireless network with a set of base stations (BSs) $\mathcal{B}$, which serve a region $\mathcal{L} \subset \mathbb{R}^2$. Let $x \in \mathcal{L}$ denote a location and $i \in \mathcal{B}$ be the index of a typical i-th BS. File transfer requests are assumed to arrive following a *spatially inhomogeneous* Poisson point process with arrival rate per unit area $\lambda(x)$ and file sizes which are independently distributed with mean $1/\mu(x)$ at location $x \in \mathcal{L}$, so the traffic load demand is defined as $\gamma(x) \doteq \frac{\lambda(x)}{\mu(x)} < \infty$ [bits/sec]. Note that this captures spatial traffic variability such as a hot spot.

2) Channel model: The average transmission rate of a user located at x and served by BS i is denoted by $c_i(x)$ [bits/sec]. Note that $c_i(x)$ is *location-dependent* but not necessarily determined by the distance from the BS i. Hence, it can capture shadowing effect, e.g., $c_i(x)$ can be very small in a shadowed area where the channel gain is very low.

3) Processing model: Each BS i is assumed to have a processing component such as CPU with a scalable speed s_i [cycles/sec] in $(0, s_{i,max}]$. Flows may have different processing demands. We represent this notion by *processing density* $w(x)$, which is defined as the average number of CPU cycles required per bit for the flow at location x. The processing demand of the traffic load at location x is then $w(x)\gamma(x)$.

4) System utilization and feasible region: Fig. 1 illustrates our system model, where a BS is decomposed into two parts: one part with processing components and the other part with RF functionalities. A routing function $p_i(x)$ specifies the *probability* that a flow at location x is associated with BS i. We will see later that, however, the optimal $p_i(x)$ will turn out to be either 1 or 0, i.e., *deterministic* routing is optimal. As there are processing and transmission resources, we can define two types of *system utilization* (i.e., the fractions of time the processor or network is busy) for BS i as follows:

$$\text{Processing utilization: } \rho_i^{(p)} \doteq \int_{\mathcal{L}} \frac{w(x)\gamma(x)}{s_i} p_i(x)dx, \quad (1)$$

$$\text{Network utilization: } \rho_i^{(n)} \doteq \int_{\mathcal{L}} \frac{\gamma(x)}{c_i(x)} p_i(x)dx. \quad (2)$$

We further denote the vectors containing processing utilizations and network utilizations of all BSs by $\rho^{(p)} = (\rho_1^{(p)}, \cdots, \rho_{|\mathcal{B}|}^{(p)})$ and $\rho^{(n)} = (\rho_1^{(n)}, \cdots, \rho_{|\mathcal{B}|}^{(n)})$, respectively.

Definition 2.1 (Feasibility): The set $\mathcal{F}$ of *feasible* system utilization $\rho = (\rho^{(p)}, \rho^{(n)})$ is given by

$$\mathcal{F} = \Big\{ \rho \mid 0 \le \rho_i^{(p)}, \rho_i^{(n)} \le 1-\epsilon, \quad 0 \le p_i(x) \le 1, \ \textstyle\sum_{i\in\mathcal{B}} p_i(x) = 1, \quad 0 < s_i \le s_{i,max}, \quad \forall i \in \mathcal{B}, \forall x \in \mathcal{L} \Big\}, \quad (3)$$

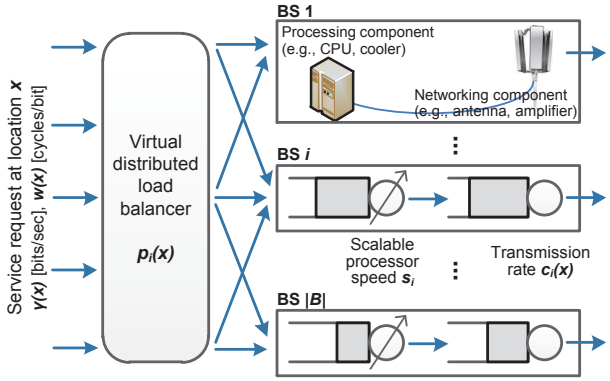


Fig. 1. Flow-level queueing model in the dual-resource environment.

where ϵ is an arbitrarily small positive constant. Hence, the feasible system utilization ρ has the associated processing speed vector $s = (s_1, \cdots, s_{|\mathcal{B}|})$ and routing probability vector $p(x) = (p_1(x), \cdots, p_{|\mathcal{B}|}(x))$ for all $x \in \mathcal{L}$.

B. Problem Formulation

The objective function we consider is

$$\min_{\rho\in\mathcal{F}} \ \mathbb{E}[\mathbf{N}] + \eta\mathbb{E}[\mathbf{P}], \quad (4)$$

where $\mathbf{N}$ is the expected number of flows in the system and $\mathbf{P}$ is the system power consumption.[1] Note that the parameter $\eta \ge 0$, controls the *tradeoff between delay and energy*. When η is zero, we only focus on delay performance, however, as η grows, more emphasis is given to energy conservation.

(i) The cost function of delay performance: We consider the M/GI/1 multi-class processor sharing (PS) system [13]. We focus on this model not only because PS is a tractable model of current scheduling policies, but also because multi-class can reflect the fact that users see different service rates and file sizes based on their locations. Using standard queueing theory, $\mathbb{E}[\mathbf{N}]$, the summation of the expected number of flows in two serial queues for all BSs, is then given by

$$\mathbb{E}[\mathbf{N}] = \sum_{i\in\mathcal{B}} \left[\phi_i^{(p)}(\rho_i^{(p)}) + \phi_i^{(n)}(\rho_i^{(n)}) \right], \quad (5)$$

where $\phi_i^{(p)}(\rho_i^{(p)}) = \frac{\rho_i^{(p)}}{1-\rho_i^{(p)}}$ and $\phi_i^{(n)}(\rho_i^{(p)}) = \frac{\rho_i^{(n)}}{1-\rho_i^{(n)}}$ are the expected number of flows in each queue, respectively.

(ii) The cost function of energy consumption: We consider a general cost function of energy consumption, which consists of two types of powers expended in the processing and networking components, respectively.

$$\mathbb{E}[\mathbf{P}] = \sum_{i\in\mathcal{B}} \left[\psi_i^{(p)}(\rho_i^{(p)}) + \psi_i^{(n)}(\rho_i^{(n)}) \right], \quad (6)$$

The networking components are assumed to gradually consume more power as the activity level increases. Thus, the energy cost for networking is given by

$$\psi_i^{(n)}(\rho_i^{(n)}) = b_i\rho_i^{(n)}, \quad (7)$$

[1]From Little's law and energy-power relationship, the general problem (4) is equivalent to minimizing $\mathbb{E}[\mathbf{D}] + \eta\mathbb{E}[\mathbf{E}]$, where $\mathbf{N}$ is the expected number of flows in the system and $\mathbf{P}$ is the system power consumption.

where $b_i > 0$ is the maximum networking power of BS i, when fully utilized, i.e., $\rho_i^{(n)} = 1$, which includes the power consumptions of Tx antenna, power amplifier and so on.

The remaining is to define the form of the energy cost for processing $\psi_i^{(p)}(\cdot)$, which also depends on the capability of the processor for provisioning its speed. We deal with two different types of processors introduced in [10]: *static speed-scaling (SS)* and *gated-static speed-scaling (GS)*.

We do not know at this moment the explicit form of $\psi_i^{(p)}(\cdot)$ although we will derive it in Section III-A later, which is one of our contributions. For now, we try to express the energy cost with a processing speed s. Let $g(s)$ denote the power consumption when the processor is running at speed s. In the domain of processor design, it has been typically assumed to be polynomial, i.e., $g(s) = as^\beta$. Thus, $\psi_i^{(p)}(\cdot)$ is given by

$$\psi_i^{(p)}(\rho_i^{(p)}) = \begin{cases} a_i s_i^\beta, & \text{when SS,} \quad (8a) \\ a_i \rho_i^{(p)} s_i^\beta, & \text{when GS,} \quad (8b) \end{cases}$$

where $a_i > 0$ and $\beta > 1$ are some constants. Note that, for the case of GS, the energy cost is only incurred during the fraction of time the processor is busy, i.e., $\rho_i^{(p)}$.

III. Speed-Scaling-Aware Optimal Load Balancing

In this paper, we consider not only delay and energy consumption in BS's networking components but also consider delay and energy consumption in BS's processing components. We rewrite our original problem in (4) as follows.

***Speed-scaling-Aware Load Balancing* [SA-LB]:**

$$\min_{\rho \in \mathcal{F}} \quad \Omega(\rho) \quad = \quad \sum_{i \in \mathcal{B}} \Big[\underbrace{\phi_i^{(p)}(\rho_i^{(p)}) + \phi_i^{(n)}(\rho_i^{(n)})}_{\text{delay performance}} + \eta \big(\underbrace{\psi_i^{(p)}(\rho_i^{(p)}) + \psi_i^{(n)}(\rho_i^{(n)})}_{\text{energy consumption}} \big) \Big]$$

A. Speed-scaling Given Processing Demand

We shall start by considering a given processing demand. In this case, we prove that the delay performance and energy consumption of the networking component can be ignored in the original problem in (4) and the problem can be further decomposed into intra-cell speed-scaling subproblems.

Theorem 3.1: For any fixed routing probability $p(x)$, the problem in (4) is reduced to $|\mathcal{B}|$ independent subproblems that find an optimal speed s_i for each BS i.

$$\min_{s_i} \quad \frac{\Gamma_i}{s_i - \Gamma_i} + \begin{cases} \eta a_i s_i^\beta, & \text{when SS,} \quad (9a) \\ \eta a_i \Gamma_i s_i^{\beta-1}, & \text{when GS,} \quad (9b) \end{cases}$$

where $\Gamma_i \doteq \int_{\mathcal{L}} w(x)\gamma(x)p_i(x)dx$. We call this problem *intra-cell optimal speed-scaling*.

Proof: Due to the space limitations, the proof is provided in our technical report [14]. ∎

When the problem in (9) is feasible, differentiating and solving gives the following optimal conditions:

$$\text{for SS,} \quad z_{ss}(s_i) \doteq s_i^{\beta-1}(s_i - \Gamma_i)^2 = \frac{\Gamma_i}{\eta a_i \beta}, \quad (10)$$

$$\text{for GS,} \quad z_{gs}(s_i) \doteq s_i^{\beta-2}(s_i - \Gamma_i)^2 = \frac{1}{\eta a_i (\beta - 1)}. \quad (11)$$

Since the function $z_{ss}(s_i)$ (resp. $z_{gs}(s_i)$) is equal to zero at $s_i = \Gamma_i$ and monotonically increases for $s_i > \Gamma_i$, it will eventually cross the positive constant value $\frac{\Gamma_i}{\eta a_i \beta}$ (resp. $\frac{1}{\eta a_i(\beta-1)}$) just once. Let $s_{i,ss}$ and $s_{i,gs}$ denote the *unique* point that satisfies (10) and (11) for $s_i > \Gamma_i$, respectively. This can be explicitly solved for some β, e.g., $s_{i,gs} = \Gamma_i + \sqrt{\frac{1}{\eta a_i}}$ when $\beta = 2$ and $s_{i,ss} = \frac{1}{2}\left(\Gamma_i + \sqrt{\Gamma_i^2 + 4\sqrt{\frac{\Gamma_i}{3\eta a_i}}}\right)$ when $\beta = 3$.

Substituting $\Gamma_i = \rho_i^{(p)} s_i$ into (10) and (11) and after some simplification, we first obtain the following closed form expression for the optimal speed s_i as a function of $\rho_i^{(p)}$:

$$s_i(\rho_i^{(p)}) = \begin{cases} \sqrt[\beta]{\frac{\rho_i^{(p)}}{\eta a_i \beta (1-\rho_i^{(p)})^2}}, & \text{for SS,} \\ \sqrt[\beta]{\frac{1}{\eta a_i (\beta-1)(1-\rho_i^{(p)})^2}}, & \text{for GS.} \end{cases} \quad (12)$$

We have expressed the energy cost for processing with a processing speed s_i in (8). Now we can write it in a more explicit form as a function of the processing utilization $\rho_i^{(p)}$.

$$\psi_i^{(p)}(\rho_i^{(p)}) = \begin{cases} \frac{\rho_i^{(p)}}{\eta\beta(1-\rho_i^{(p)})^2}, & \text{for SS,} \\ \frac{\rho_i^{(p)}}{\eta(\beta-1)(1-\rho_i^{(p)})^2}, & \text{for GS.} \end{cases} \quad (13)$$

B. Optimal Structure of Speed-scaling-aware Load Balancing

Based on the speed-scaling derived in the previous section, we now investigate the optimal structure of speed-scaling-aware load balancing.

Theorem 3.2: Suppose that the problem **[SA-LB]** is feasible. Let us denote the optimal system utilization $\rho^* = (\rho^{*(p)}, \rho^{*(n)})$, i.e., solution to **[SA-LB]**. Then, the following user association rule[2] for the MT at location x is optimal:

$$i^*(x) = \operatorname*{argmin}_{j \in \mathcal{B}} \left[\frac{\mathcal{M}_j^{(p)}}{w(x)} + \frac{\mathcal{M}_j^{(n)}}{c_j(x)} \right], \quad \forall x \in \mathcal{L}, \quad (14)$$

where $\mathcal{M}_j^{(p)} = [(1 - \rho_j^{*(p)})^{-2} + \eta\psi_j'^{(p)}(\rho_j^{*(p)})]/s_i(\rho_j^{*(p)})$ and $\mathcal{M}_j^{(n)} = (1 - \rho_j^{*(n)})^{-2} + \eta b_j$ are metrics that can be computed at the j-th BS side.

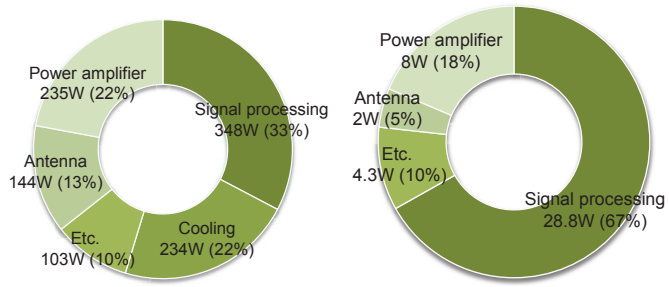
Proof: The proof is a generalization of that of [15], with the additional energy cost. The problem **[SA-LB]** is a convex optimization because its feasible set $\mathcal{F}$ has been proved to be convex and the objective function is the sum of convex functions. Hence, it is sufficient to show that, for all $\rho \in \mathcal{F}$,

$$\left\langle \nabla\Omega(\rho^*), \Delta\rho^* \right\rangle \geq 0, \quad \text{where } \Delta\rho^* = \rho - \rho^*. \quad (15)$$

Let $p(x)$ and $p^*(x)$ be the associated routing probability vectors for ρ and ρ^*, respectively. Then, (14) generates the *deterministic* cell coverage, i.e.,

$$p_i^*(x) = \mathbf{1}\Big\{ i = \operatorname*{argmin}_{j \in \mathcal{B}} \mathcal{M}_j(x) \Big\}, \quad (16)$$

[2]This association rule can be interpreted as saying that each MT selfishly tries to minimize the sum of two types of cost: *(i) processing cost per unit CPU speed* and *(ii) networking cost per unit Tx capacity*.



(a) Macro BS: 1064W in total (b) Micro BS : 43.1W in total

Fig. 2. Maximum power consumption breakdown of LTE BSs based on data obtained from [11].

where $\mathcal{M}_j(x) = \frac{\mathcal{M}_j^{(p)}}{w(x)} + \frac{\mathcal{M}_j^{(n)}}{c_j(x)}$. Then, the inner product $\langle \nabla\Omega(\rho^*), \Delta\rho^* \rangle$ can be calculated such as

$$\begin{aligned} &= \sum_{i\in\mathcal{B}} \Big[\{\phi_i'^{(p)}(\rho_i^{*(p)}) + \psi_i'^{(p)}(\rho_i^{*(p)})\} \cdot (\rho_i^{(p)} - \rho_i^{*(p)}) \\ &\quad + \{\phi_i'^{(n)}(\rho_i^{*(n)}) + \psi_i'^{(n)}(\rho_i^{*(n)})\} \cdot (\rho_i^{(n)} - \rho_i^{*(n)}) \Big] \qquad (17) \\ &= \int_{\mathcal{L}} \gamma(x) \sum_{i\in\mathcal{B}} \mathcal{M}_i(x) \cdot \big(p_i(x) - p_i^*(x)\big) dx. \end{aligned}$$

From (16), as $p_i^*(x)$ is an indicator for the minimizer of $\mathcal{M}_i(x)$, we have the following inequality:

$$\sum_{i\in\mathcal{B}} \mathcal{M}_i(x) \cdot p_i(x) \;\geq\; \sum_{i\in\mathcal{B}} \mathcal{M}_i(x) \cdot p_i^*(x) \qquad (18)$$

Substituting (18) into (17) yields the condition in (15), which completes the optimality proof. ■

C. Distributed Iterative Algorithm

To determine the association in (14), MTs need to know ρ^* *a priori*. However, this will be relaxed in our proposed distributed algorithm, called *SpeedBalance*, which can achieve the global optimum in an iterative manner. The distributed algorithm involves two parts. At the k-th iteration period,

Mobile terminal: ① MTs estimate the transmission rate $c_i(x)$ and receive the system utilization $\rho^{[k]}$, e.g., through broadcast control messages from BSs. ② Then, a new flow request for a MT simply selects the BS $i^{[k]}(x)$ based on the deterministic rule in (14), but using the current system utilization $\rho^{[k]}$ instead of the optimal one ρ^*.

Base station: ① Each BS i adapts its processing speed s_i according to (12). ② It measures the system utilization $\rho_i^{[k+1]}$, calculates the metrics $\big(\mathcal{M}_i^{(p)}, \mathcal{M}_i^{(n)}\big)$, and then broadcasts them to MTs for the next iteration.

IV. Numerical Results

We first investigate the component-level power consumption breakdown of LTE BSs in Fig. 2. This reveals that (i) signal processing contributes to a considerable portion of total power consumption, and (ii) cooling and power amplifier are also major components than Tx antenna does. Note that micro BSs typically do not have cooling components. Etc. (e.g., power supply and battery backup) amounts to about 10%.

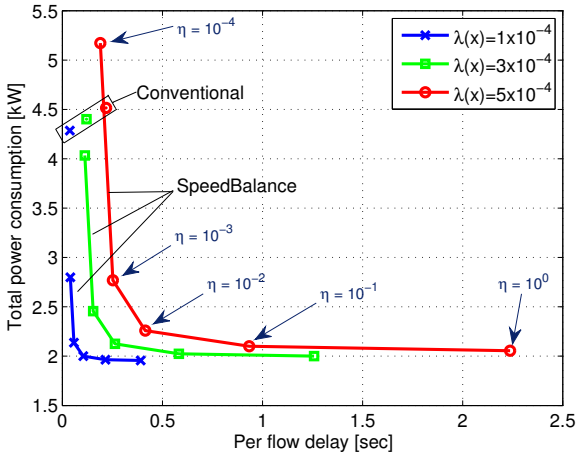


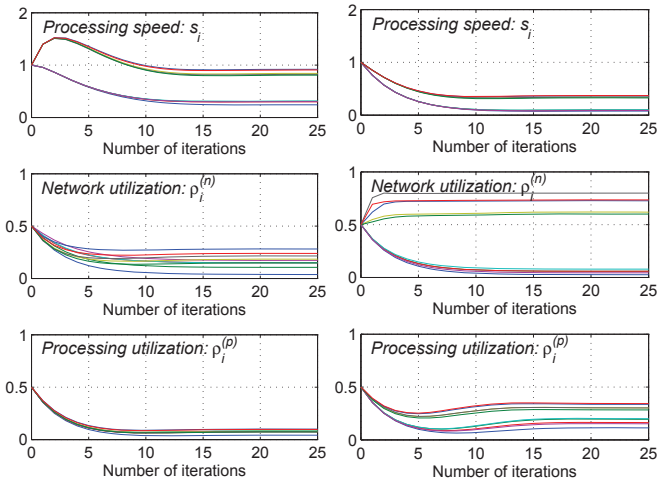
Fig. 3. Energy-delay tradeoff by SpeedBalance w/ SS. As η increases, energy savings can be obtained at the cost of delay increase.

Based on these data in Fig. 2, we choose parameters for our energy cost function. We consider the power amplifier and antenna as networking components, and consider the signal processing and cooling as processing components. To capture the static power of these components, 25% of their total power consumptions is considered as static power. Thus, we set a_i for macro and micro BSs to be 106.6 and 5.3 so that their maximum dynamic power consumptions at the maximum speed $s = 1.6$ [Gcps] are equal to 75% of (348+234)W and 28.8W. We set likewise b_i for macro and micro BSs to be 284.3 and 7.5. The macro and micro BSs have the transmission powers of 43.8dBm and 33dBm, respectively. Each MT's request has exactly one file that is log-normally distributed with mean $1/\mu(x) = 100$ Kbyte and the processing density $w(x)$ over space is considered to be uniform. Other simulation parameters are given in [14].

A. Performance Under A Mixed Macro/Micro BS Topology

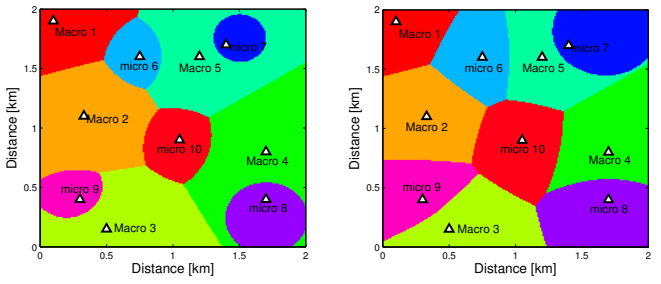
We first verify the energy-delay tradeoff of SpeedBalance and also compare its performance with a conventional scheme using the signal strength-based user association and do not adopt the speed-scaling. Fig. 3 shows tradeoff curves by varying the energy-delay tradeoff parameter η from 10^{-4} to 10^0 for the different values of arrival rate $\lambda(x)$. The results are consistent with our expectations: *the higher η is, the more possible energy savings are possible at the cost of delay.*

In order to examine where and how the energy savings come from, we first plot Figs. 4 (a) and (b) that show the convergence of processing speeds, networking and processing utilizations for the cases of low $\eta = 10^{-3}$ and high $\eta = 10^{-1}$. As can be seen, BSs slow down their processing speeds when η is high (i.e., giving more emphasis on energy conservation) compared to the case of low η. This is one of the main reasons for reduction in power consumption. There is another reason beyond the speed-scaling. Fig. 5 illustrates the snapshots of cell coverage by SpeedBalance for both cases. By comparing two figures, we can clearly see that micro BSs have large



(a) Low $\eta = 10^{-3}$ (b) High $\eta = 10^{-1}$

Fig. 4. Convergence of SpeedBalance w/ SS. ($\lambda(x) = 5 \times 10^{-4}$). The different curve corresponds to each of 10 BSs.



(a) Low $\eta = 10^{-3}$ (b) High $\eta = 10^{-1}$

Fig. 5. Snapshots of cell coverage by SpeedBalance w/ SS. ($\lambda(x) = 5 \times 10^{-4}$). As η increases, the micro BSs indexed by 6 to 10 have larger coverage.

coverages for high η. In other words, more MTs are associated with and served by the energy-efficient micro BSs.

On the other hand, per-flow delay will grow as η increases. This is because reducing the processing speed and concentrating the traffic load in the micro BSs will result in the increase of the processing and networking utilizations as shown in Fig. 4. However, in Fig. 3, it is noteworthy that *the most of energy savings can be obtained at* $\eta = 10^{-3}$ *while not penalizing the delay performance, compared to the conventional scheme.* Thus, we will choose $\eta = 10^{-3}$ throughout the rest of our simulation study.

B. Performance Under A Real 3G BS Deployment Topology

In order to obtain more realistic amount of energy savings, we further consider the real map of BS layout consisting of heterogeneous environments (urban, suburban and rural areas) and normalized traffic trace for our simulation.[3] TABLE I summarizes the average energy use during one day. As expected, *compared to the conventional scheme, significant amounts of energy savings can be achieved by SpeedBalance, e.g., 31.8% and 36.4% for SS and GS in weekdays, 41.7% and 44.8% for SS and GS in weekends.* More energy savings are expected during weekends than weekdays. This is because the traffic load during weekends is relatively lower than that during weekdays. Also note that GS can provide 3.2-4.6% more savings than SS due to its superior characteristic.

[3]Due to the space limitations, the BS layout from [16], traffic trace, and other interesting results are provided in our technical report [14].

TABLE I
AVERAGE ENERGY USE DURING ONE DAY

	Conventional scheme	SpeedBalance w/ SS	SpeedBalance w/ GS
Weekday	617.3kWh	421.3kWh	392.5kWh
Weekend	578.3kWh	336.9kWh	319.0kWh

V. Concluding Remarks

This paper considered speed-scaling to address the tradeoff between delay and energy in both networking and processing components of BSs. By investigating the optimal speed for processors with SS and GS and the optimal structure of speed-scaling-aware load balancing, we proposed a distributed iterative algorithm, SpeedBalance. Extensive simulations showed that compared to the conventional scheme, SpeedBalance can yield significant energy savings about 30-45%.

References

[1] E. Oh, B. Krishnamachari, X. Liu, and Z. Niu, "Towards dynamic energy-efficient operation of cellular network infrastructure," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 56–61, Jun. 2011.

[2] M. A. Marsan, L. Chiaraviglio, D. Ciullo, and M. Meo, "Optimal energy savings in cellular access networks," in *Proc. GreenComm.*, Jun. 2009.

[3] G. Fettweis and E. Zimmermann, "ICT energy consumption–trends and challenges," in *Proc. IEEE WPMC*, Lapland, Finland, Sep. 2008.

[4] "Mobile network energy OPEX to rise dramatically to $22 billion in 2013," ABI Research, Jul. 2008.

[5] K. Son, E. Oh, and B. Krishnamachari, "Energy-aware hierarchical cell configuration: from deployment to operation," in *Proc. IEEE INFOCOM - GCN Workshop*, Shanghai, China, Apr. 2011, pp. 289–294.

[6] Z. Niu, Y. Wu, J. Gong, and Z. Yang, "Cell zooming for cost-efficient green cellular networks," *IEEE Commun. Mag.*, vol. 48, no. 11, pp. 74–79, Nov. 2010.

[7] C. Peng, S.-B. Lee, S. Lu, and H. Luo, "Traffic-driven power saving in operational 3g networks," in *Proc. ACM MobiCom*, Las Vegas, NV, Sep. 2011.

[8] K. Son, H. Kim, Y. Yi, and B. Krishnamachari, "Base station operation and user association mechanisms for energy-delay tradeoffs in green cellular networks," *IEEE JSAC*, vol. 29, no. 8, Sep. 2011.

[9] J. R. Lorch and A. J. Smith, "Improving dynamic voltage scaling algorithms with PACE," in *Proc. ACM SIGMETRICS 2001*, Annapolis, MD, June 2001, pp. 50–61.

[10] A. Wierman, L. L. H. Andrew, and A. Tang, "Power-aware speed scaling in processor sharing systems," in *Proc. IEEE INFOCOM*, Rio de Janeiro, Brazil, Apr. 2009, pp. 2007–2015.

[11] O. Arnold, F. Richter, G. Fettweis, and O. Blume, "Power consumption modeling of different base station types in heterogeneous cellular networks," in *Proc. ICT MobileSummit*, Florence, Italy, Jun. 2010.

[12] "Alcatel-Lucent demonstrates up to 27 percent power consumption reduction on base stations deployed by china mobile," Mobile World Congress, Barcelona, Feb. 2009.

[13] J. Walrand, *An introduction to queueing networks*. Prentice Hall, 1998.

[14] K. Son and B. Krishnamachari, "Speedbalance: Speed-scaling-aware optimal load balancing for green cellular networks," [Online] Available at http://anrg.usc.edu/~kyuho/TR_SpeedBalance.pdf, *Tech. Report*, 2011.

[15] H. Kim, G. de Veciana, X. Yang, and M. Venkatachalam, "Distributed α-optimal user association and cell load balancing in wireless networks," *accepted to IEEE/ACM Trans. Netw.*, 2011 (to appear).

[16] K. Son, S. Lee, Y. Yi, and S. Chong, "REFIM: A practical interference management in heterogeneous wireless access networks," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 6, pp. 1260–1272, Jun. 2011.