On Energy-Delay Tradeoff in Base Station Sleep Mode Operation

Zhisheng Niu, *Fellow, IEEE*, Jianan Zhang, Xueying Guo, and Sheng Zhou
Tsinghua National Laboratory for Information Science and Technology (TNList)
Tsinghua University, Beijing, China

***Abstract*—Base station (BS) sleep mode operation is one of the effective ways to save energy, but it may lead to longer delay to the customers. In order to evaluate the tradeoffs between energy consumption and customer delay, we model the BS sleep mode operation as an N-policy $M/G/1$ vacation queue with setup and close-down times, where the BS enters sleep mode if no customers arrive during the close-down time after the queue becomes empty and it starts to setup when it sees N customer arrivals during its sleep period. Several closed-form formulas are derived to demonstrate the tradeoffs between energy consumption and mean delay by changing the close-down time and N. It is shown that the relationship between the energy consumption and the mean delay is linear by changing the close-down time. Besides, larger N reduces the energy consumption, but there may exist $N > 1$ that minimizes the mean delay. We also investigate the bound on given percentile of overall delay. We observe that the delay bound is nearly linear in mean delay in the cases tested. Therefore, similar tradeoffs exist between energy consumption and the delay bound.**

I. Introduction

Recently, it has been reported that information and communication technology (ICT) industry is becoming a significant part of the world energy consumption. Cellular networks are among the main energy consumers in the ICT field. Specifically, base stations (BSs) account for over 80% of the cellular network energy consumption [1]. Therefore, in order to support increasing data transmission rate, energy efficiency is key in future BS operations [2].

Sleep mode operation is an effective way to save energy while maintaining acceptable quality of service (QoS) [2], [3], [4], [5], [6], [7], [8]. To save energy, a BS can be turned off when the traffic load is light, but the quality of service will deteriorate accordingly. In [3], the authors investigate energy saving sleep mode operations while maintaining acceptable throughput received by users. Based on a Markovian model, they solve a set of balance equations and obtain the probability that users achieve the target throughput. [4], [5] consider sleep mode operations with blocking probability constraint in a cellular network.

In this paper, we consider the energy-delay tradeoff because delay performance is a key metric in mobile multimedia communications, i.e., *how much energy can be traded off by a certain amount of delay*? In [9], it has been shown that the tradeoffs do exist between average transmission power and average buffer delay by changing the transmission rate, but the BS sleep operation was not considered. We consider energy-delay tradeoffs in BS sleep mode operation and assume fixed transmission rate when the BS is turned on. When a BS is turned off, customers have to wait until the BS wakes up and therefore experience longer delay. We aim to investigate how much energy can be traded off by the queueing delay in terms of the BS sleep time, setup time, and close-down time. We focus on the cases where the sleep mode operations do not affect customer arrival processes, and do not consider the benefit of coordinated multi-point [10] or cell zooming [11], where a customer can be served by a neighboring BS.

We derive closed-form formulas to demonstrate the tradeoffs between energy consumption and mean queueing delay, which depend on control policies. For example, by changing the close-down time, BS energy consumption is a linear decreasing function of mean delay. However, by changing the number of customers that intrigue BS setup, the energy consumption is not necessarily decreasing in delay. Moreover, we demonstrate by numerical studies that the nearly linear relationship between the mean delay and the delay bound on given percentile of delay exists. Therefore, the tradeoffs between energy consumption and mean delay well indicate the tradeoffs between energy consumption and the delay bound.

II. BS Sleep Mode Modeling

A. $M/G/1$ *Vacation Queue with Setup and Close-down Times*

We model a BS as a N-policy $M/G/1$ queue with close-down and setup times. Customers arrive in a Poisson process with parameter λ. Service times follow a general distribution with mean $1/\mu$. When the queue becomes empty, the server keeps waiting during the close-down time, which follows a general distribution. If new customers arrive during the close-down time, the server immediately starts its service without setup. But when the close-down time expires and no customers arrive, the server will be turned to sleep mode. During the sleep period, if N customers have arrived, the server starts to setup and then to serve new customers. This queuing model was introduced in [12]. The switch-over cost of the BS is considered as the setup time and the power consumption during the setup time. To find the tradeoffs between energy and delay, and to design effective sleep policies, we consider the effects of the close-down time and N on energy and delay.

In general, longer close-down time will lead to shorter waiting time of customers, because customers which arrive during the close-down time will be served immediately without setup time. However, since the BS is idle in close-down time, it consumes more power compared with the case where the BS enters sleep mode immediately after serving all customers.

The switch-over cost is an impediment of frequently turning on and off the BS. Both longer close-down time and larger N avoid frequent switch-over and the energy cost.

In an N-policy $M/G/1$ queue with close-down and setup times, let B, S, and D denote the service time for a customer, the setup time and close-down time for the server, respectively. Let $\widetilde{B}(s)$, $\widetilde{S}(s)$, and $\widetilde{D}(s)$ denote the Laplace-Stieltjes transforms of their probability density functions. The probability that no customers arrive during the close-down time is $\widetilde{D}(\lambda)$. The mean length of a cycle $E[C]$ (time between two successive epochs at which the queue becomes empty) is given by [13]:

$$\begin{aligned} E[C] &= [1-\widetilde{D}(\lambda)](\frac{1}{\lambda}+\frac{E[B]}{1-\rho}) \\ &+ \widetilde{D}(\lambda)(\frac{N}{\lambda}+\frac{E[S]+NE[B]}{1-\rho}) \\ &= \frac{1-\widetilde{D}(\lambda)+\widetilde{D}(\lambda)(N+\lambda E[S])}{\lambda(1-\rho)}, \end{aligned}$$

where $\rho=\lambda E[B]$.

We then obtain the Laplace-Stieltjes transform of the density function for customer's sojourn time, which includes both the waiting time in the queue and the service time,

$$\begin{aligned} \widetilde{T}(s) &= \frac{\widetilde{B}(s)}{E[C]}\{\frac{1-\widetilde{D}(\lambda)}{\lambda} \\ &+ \frac{(1-\widetilde{D}(\lambda))E[B]}{1-\rho}\frac{(1-\rho)(1-\widetilde{B}(s))}{E[B][s-\lambda+\lambda\widetilde{B}(s)]} \\ &+ \widetilde{D}(\lambda)\frac{N}{\lambda}\frac{\widetilde{S}(s)}{N}\frac{[\lambda/(s+\lambda)]^N-[\widetilde{B}(s)]^N}{\lambda/(s+\lambda)-\widetilde{B}(s)} \\ &+ \widetilde{D}(\lambda)\frac{1-\widetilde{S}(s)(\widetilde{B}(s))^N}{s-\lambda+\lambda\widetilde{B}(s)}\}. \end{aligned} \qquad (1)$$

Here we correct an error in the Laplace-Stieltjes transform of the density function for waiting time in the queue ($\widetilde{W}(s)=\widetilde{T}(s)/\widetilde{B}(s)$) on page 136 of [13].

Differentiate Eq. (1) at $s=0$, we obtain the mean sojourn time, which is in consistent with the result obtained by [12]:

$$\begin{aligned} E[T] &= E[B]+\frac{\lambda E[B^2]}{2(1-\rho)} \\ &+ \frac{\widetilde{D}(\lambda)[N(N-1)+2N\lambda E[S]+\lambda^2E[S^2]]}{2\lambda[\widetilde{D}(\lambda)(N+\lambda E[S])+1-\widetilde{D}(\lambda)]}. \end{aligned} \qquad (2)$$

Let P_{ON}, P_{SL}, P_{ST} and P_{CD} denote the power consumption of the BS during its on, sleep, setup, and close-down phases, respectively. The long time average operation power is:

$$\begin{aligned} E[P] &= \frac{1}{E[C]}[\frac{1-\widetilde{D}(\lambda)}{\lambda}(P_{\mathrm{CD}}-P_{\mathrm{SL}}) \\ &+ \widetilde{D}(\lambda)E[S](P_{\mathrm{ST}}-P_{\mathrm{SL}})] \\ &+ (1-\rho)P_{\mathrm{SL}}+\rho P_{\mathrm{ON}}. \end{aligned} \qquad (3)$$

B. Impact of Close-down Time

We investigate the closed form relationship between mean power and mean sojourn time by changing close-down time.

From Eq. (2) and Eq. (3), we can see the close down time D affects $E[T]$ and $E[P]$ through the probability that no customers arrive during the close-down time ($\widetilde{D}(\lambda)$). In Eq. (2), $E[T]$ is monotonically increasing in $\widetilde{D}(\lambda)$. This can be explained as: smaller $\widetilde{D}(\lambda)$ means more customers arrive during the close-down time and be served immediately without setup time.

In addition, we obtain $\widetilde{D}(\lambda)$ as a function of $E[T]$ from Eq. (2), then substituting $\widetilde{D}(\lambda)$ in Eq. (3), we obtain the linear relationship between $E[P]$ and $E[T]$, given by

$$\begin{aligned} E[P] &= \rho P_{\mathrm{ON}}+(1-\rho)P_{\mathrm{CD}} \\ &+ (1-\rho)\frac{2\lambda[E[T]-E[B]-E[B^2]\lambda/2(1-\rho)]}{E[S^2]\lambda^2+2NE[S]\lambda+N(N-1)} \\ &\times [E[S]\lambda(P_{\mathrm{ST}}-P_{\mathrm{SL}}) \\ &- (N+E[S]\lambda)(P_{\mathrm{CD}}-P_{\mathrm{SL}})]. \end{aligned}$$

C. Impact of N

The server is turned on when there are N customers present in the queue. Larger N leads to longer cycle. Therefore, the server goes to setup less frequently. For simplicity, consider the special case where the close-down time equals to zero, i.e. $D=0$ ($\widetilde{D}(\lambda)=1$). In this case, mean power of N-policy is equivalent to a 1-policy system which has $E[S]/N$ mean setup time, given by

$$E[P]=\rho P_{\mathrm{ON}}+(1-\rho)P_{\mathrm{SL}}+\frac{\lambda(1-\rho)E[S]}{N+\lambda E[S]}(P_{\mathrm{ST}}-P_{\mathrm{SL}}). \qquad (4)$$

However, the mean sojourn time is not necessarily monotonically increasing in N, given by

$$E[T]=E[B]+\frac{\lambda E[B^2]}{2(1-\rho)}+\frac{N(N-1)+2N\lambda E[S]+\lambda^2E[S^2]}{2\lambda(N+\lambda E[S])}. \qquad (5)$$

If we generalize N to real numbers, in condition that $\sqrt{C_S^2+1/(\lambda E[S])}>1$, there exists an $N_{\text{delay optimal}}$ that minimizes $E[T]$, i.e.,

$$N_{\text{delay optimal}}=\lambda E[S](\sqrt{C_S^2+1/(\lambda E[S])}-1),$$

where C_S^2 is the squared coefficient of variation of the setup time. If $\sqrt{C_S^2+1/(\lambda E[S])}\leq 1$, $E[T]$ is monotonically increasing in N for $N>0$.

By changing N, the relationship between $E[P]$ and $E[T]$ is:

$$E[T]=\frac{C_S^2\lambda E[S]+1}{2\lambda}A+\frac{E[S]}{2A}+E[B]+\frac{\lambda E[B^2]}{2(1-\rho)}-\frac{1}{2\lambda},$$

where

$$A=\frac{E[P]-(\rho P_{\mathrm{ON}}+(1-\rho)P_{\mathrm{SL}})}{(P_{\mathrm{ST}}-P_{\mathrm{SL}})(1-\rho)}.$$

D. Delay Bound

To meet the required QoS guarantee, customers need to be served within their tolerable delay with high probability. Consider the delay bound denoted by $T^{\varepsilon}_{\max}$, which means that the probability that the overall delay of a customer exceeds $T_{\max}$ is ε. Let ε be arbitrarily small to satisfy the QoS requirement. We obtain the probability density function of the overall delay by inverse Laplace transform of Eq. (1). Then we calculate the tail probability and obtain $T^{\varepsilon}_{\max}$.

III. Numerical Results

In this section we provide numerical results to demonstrate the impacts of close-down time and N on BS performance and energy saving, and the tradeoffs between mean power and mean sojourn time. We also investigate the relationship between mean sojourn time and $T^{0.01}_{\max}$. In all these cases, customers arrive as a Poisson process with rate λ. When the BS is transmitting data at rate μ, the power is P_{ON}. When the BS is in setup or close-down phases, $P_{\mathrm{ST}} = P_{\mathrm{CD}} = 0.9P_{\mathrm{ON}}$, which is justified in [2] as the power consumption in idle state. We assume that BS power consumption during sleep is $0.2P_{\mathrm{ON}}$.

A. Impact of Close-down Time

Figure 1 depicts the effects of the close-down time on the mean sojourn time and mean power. The setup time is deterministic and equals to $1/\mu$. The close-down and service times follow exponential distribution. We consider different load conditions, and assume $N = 1$. Since the distribution of close-down time is given, as the mean close down time ($E(D)$) increases, the probability that no customers arrive during the close-down time ($\widetilde{D}(\lambda)$) decreases. Hence, we changes the mean close down time to see the relationship. We observe that as the close-down time increases, the mean power increases and mean sojourn time decreases. In light load conditions, sleep mode brings more benefits on energy saving.

Figure 2 depicts the linear relationship between the mean power and mean sojourn time for different N. Other parameters, such as λ, μ, $E[S]$, C_S^2, C_B^2, have effects on the slope of the linear function, but the linear relationship always exists.

B. Impact of N

Figure 3 depicts the effects of N on the mean sojourn time and mean power. Mean setup time equals to $1/\mu$. We consider both light load and heavy load conditions, i.e, $\rho = \lambda/\mu = \lambda E[S] = 0.1$, and 0.8. We also consider the effects of the deviation of setup time, and let $C_S^2 = 0$, and 25. The deviation of setup time does not affect the mean power, which is given by Eq. (4). Let T_{wait} denote the waiting time of a customer that arrives during the sleep time until the server starts to setup, i.e., the time interval between one customer arrival and the epoch when the N^{th} customer arrives during the sleep phase. In light load conditions, the inter-arrival time between customers is long, and T_{wait} dominates the mean sojourn time for large N. In such cases, mean sojourn time is increasing in N. In heavy load conditions, T_{wait} is comparable with the setup time. By increasing N, the server goes to setup less often, and the benefit may outweigh the cost of longer T_{wait}, especially when the deviation of setup time is large. Therefore, there may exist $N > 1$ that minimizes the mean sojourn time given by Eq. (5).

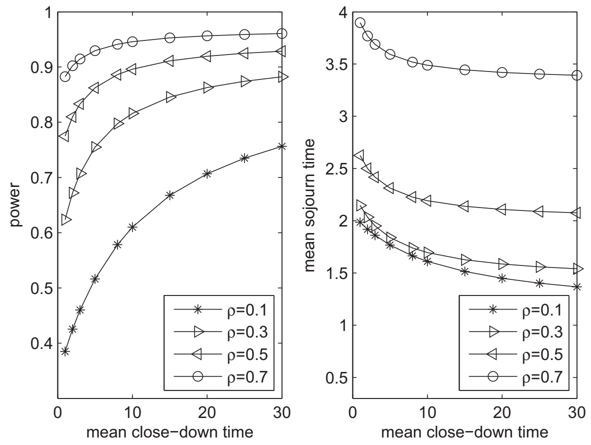

Fig. 1. Mean sojourn time (normalized by $1/\mu$) and mean power (normalized by P_{ON}) vs. mean close-down time (normalized by $1/\mu$) in an 1-policy $M/M/1$ queue with exponentially distributed close-down time and deterministic setup time. $E[S] = 1/\mu$.

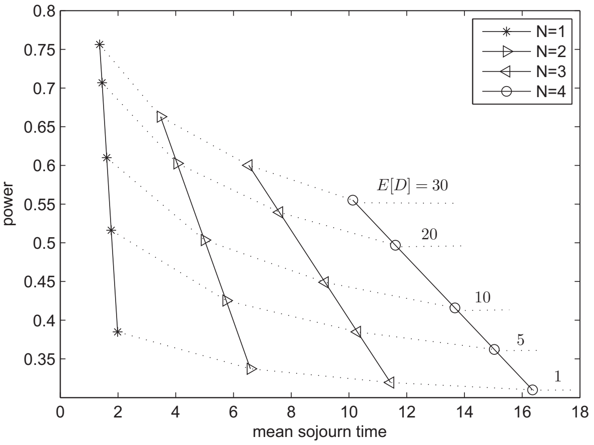

Fig. 2. Mean sojourn time (normalized by $1/\mu$) vs. mean power (normalized by P_{ON}) in an N-policy $M/M/1$ queue with exponentially distributed close-down time and deterministic setup time (changing the close-down time). $\rho = 0.1$, $E[S] = 1/\mu$.

Figures 4 and 5 depict the relationships between the mean sojourn time and mean power. Since larger N always reduces mean power, but not necessarily increases the mean sojourn time, mean power may not be a monotonically decreasing function in the mean sojourn time as depicted in Fig. 5.

C. Mean Delay vs. Delay Bound

We consider the relationship between the mean sojourn time and $T^{0.01}_{\max}$. From cases we studied, the relationship between $E[T]$ and $T^{0.01}_{\max}$ is almost linear and depicted in Fig. 6. We obtain these cases by changing the close-down time. For simplicity, we assume that the setup times follow

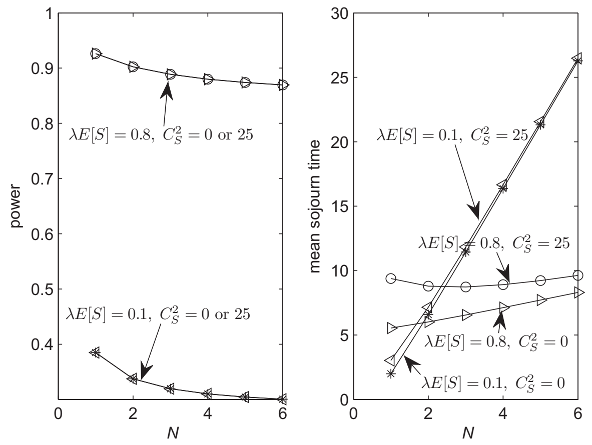

Fig. 3. Mean sojourn time (normalized by $1/\mu$) and mean power (normalized by P_{ON}) vs. N in an N-policy $M/M/1$ queue with close-down and setup times. $E[S]=1/\mu$.

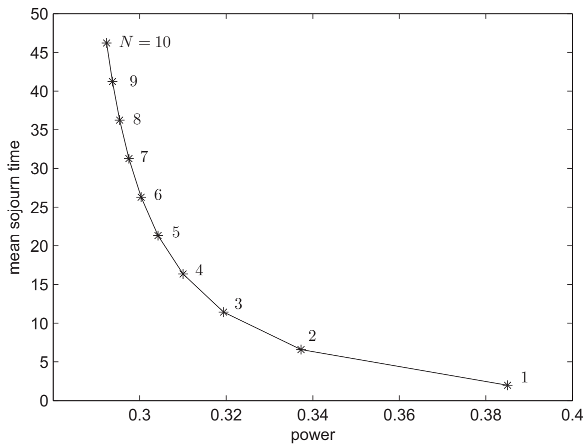

Fig. 4. Mean sojourn time (normalized by $1/\mu$) vs. mean power (normalized by P_{ON}) in an N-policy $M/M/1$ queue with close-down and setup times (changing N). $E[S]=1/\mu$, $\lambda E[S]=0.1$, $C_S^2=0$.

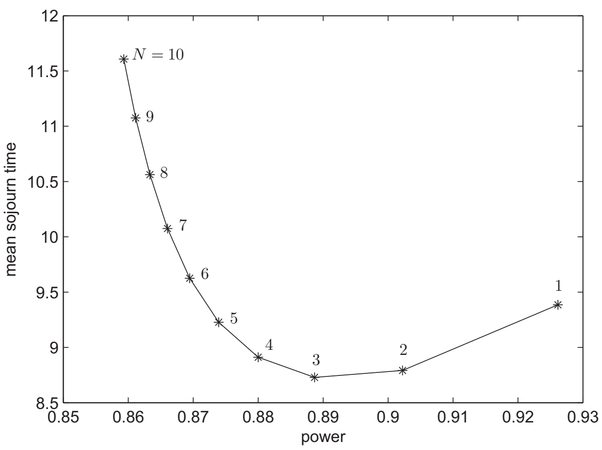

Fig. 5. Mean sojourn time (normalized by $1/\mu$) vs. mean power (normalized by P_{ON}) in an N-policy $M/M/1$ queue with close-down and setup times (changing N). $E[S]=1/\mu$, $\lambda E[S]=0.8$, $C_S^2=25$.

exponential distribution, and thus simplify the calculations of inverse Laplace transforms and tail probabilities. We consider cases where the service times follow exponential or hyper-exponential distribution, where $p_1=0.8,\ p_2=0.2,\ \mu_1=8,\ \mu_2=2/9,\ C_B^2=7.125$. We observe that although larger deviation of setup and service times leads to significantly larger $T_{\max}^{0.01}$, the nearly linear relationship still exists.

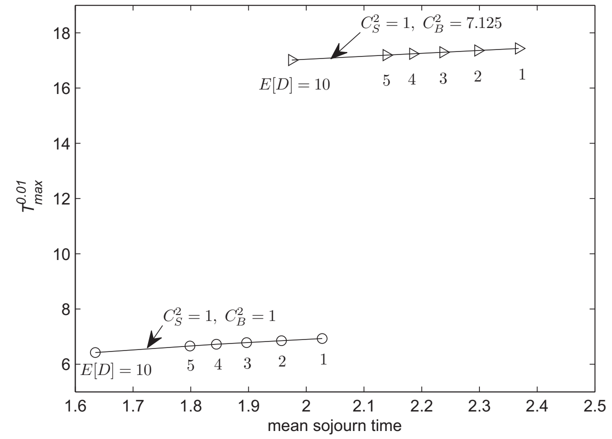

Fig. 6. $T_{\max}^{0.01}$ vs. mean sojourn time (both normalized by $1/\mu$) in a 1-policy $M/M/1$ queue with close-down and setup times (changing the close-down time).

Cases in Fig. 7 have the same distributions of service, setup, close-down times as in Fig. 6. The only difference is that we aim to investigate the effect of N, and therefore we obtain different sojourn times by changing N rather than the close-down time. We observe that by changing N from 1 to 5, the mean sojourn time is also nearly linear with $T_{\max}^{0.01}$. Moreover, $T_{\max}^{0.01}$ is not very sensitive to the deviation of service time. One reason is that the effects of setup time on service delay diminish as N increases. Another reason is that T_{wait} dominates the delay as N increases. Note that we consider light load conditions for the BS sleep mode operation.

IV. Conclusion

Energy can be traded by delay in BS sleep mode operation. The tradeoffs between energy consumption and delay depend on BS control policies. We derive closed form relationships between mean power and mean overall delay based on an N-policy $M/G/1$ queue with setup and close-down times. By changing the close-down time, mean power is a monotonically decreasing linear function of the mean delay. By increasing N, mean power decreases, but there may exist $N>1$ that minimizes the mean delay, in which case energy may not be monotonically decreasing in delay.

We observe nearly linear relationship between the mean delay and the bound on given percentile customer delay from the cases we tested. The nearly linear relationship is not very sensitive to the distributions of service time. Therefore, similar tradeoffs exist between mean power and the delay bound. For the control policies discussed in this paper, by limiting the

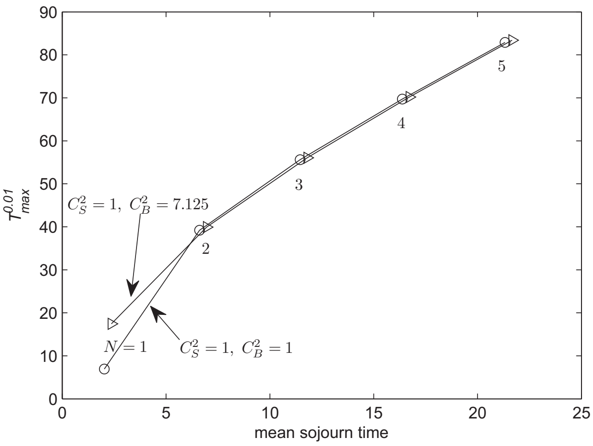

Fig. 7. $T_{\max}^{0.01}$ vs. mean sojourn time (both normalized by $1/\mu$) in an N-policy $M/M/1$ queue with close-down and setup times (changing N).

mean delay to a corresponding level, they guarantee with given high probability that customers be served within their tolerable delay.

Acknowledgment

The authors would like to thank Dr. Yiqun Wu for insightful discussions. This work was supported in part by the National Basic Research Program of China (973 Program: No. 2012CB316001), the Nature Science Foundation of China (No. 61021001, No.60925002) and Hitachi Ltd.

References

[1] F. Richter, A. J. Fehske, and G. Fettweis, "Energy efficiency aspects of base station deployment strategies for cellular networks," in *IEEE Vehicular Technology Conference*, 2009.

[2] M. A. Imran and *et, al*, "Energy efficiency analysis of the reference systems, areas of improvements and target breakdown," EARTH, Tech. Rep., 2011. [Online]. Available: https://bscw.ict-earth.eu/pub/bscw.cgi/d71252/EARTH_WP2_D2.3_v2.pdf

[3] S.-E. Elayoubi, L. Saker, and T. Chahed, "Optimal control for base station sleep mode in energy efficient radio access networks," in *IEEE INFOCOM*, 2011, pp. 106–110.

[4] J. Gong, S. Zhou, and Z. Niu, "A dynamic programming approach for base station sleeping in cellular networks," *IEICE Trans. Commun.*, vol. E95.B, pp. 551–562, Feb. 2012.

[5] E. Oh and B. Krishnamachari, "Energy savings through dynamic base station switching in cellular wireless access networks," in *IEEE GLOBECOM*, Dec. 2010, pp. 1 –5.

[6] F. B. I. Ashraf and L. Ho, "Power savings in small cell deployments via sleep mode techniques," in *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications Workshops*, 2010, pp. 307–311.

[7] S. W. Fuhrmann and R. B. Cooper, "Sleep mode techniques for small cell deployments," *IEEE Commun. Mag.*, vol. 49, no. 8, pp. 72–79, Aug. 2011.

[8] S. Mclaughlin, P. Grant, J. Thompson, H. Haas, D. Laurenson, C. Khirallah, Y. Hou, and R. Wang, "Techniques for improving cellular radio base station energy efficiency," *IEEE Wireless Commun.*, vol. 18, no. 5, pp. 10 –17, Oct. 2011.

[9] R. A. Berry and R. G. Gallager, "Communication over fading channels with delay constraints," *IEEE Trans. Inform. Theory*, vol. 48, no. 5, pp. 1135–1149, 2002.

[10] R. Irmer, H. Droste, P. Marsch, M. Grieger, G. Fettweis, S. Brueck, H.-P. Mayer, L. Thiele, and V. Jungnickel, "Coordinated multipoint: Concepts, performance, and field trial results." *IEEE Commun. Mag.*, vol. 49, no. 2, pp. 102–111, 2011.

[11] 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.

[12] M. Yadin and P. Naor, "Queuing systems with a removable service station," *Operations Research Quarterly*, vol. 14, no. 4, pp. 393–405, Dec. 1963.

[13] H. Takagi, *Queueing analysis: a foundation of performance evaluation. Volume 1: Vacation and Priority Systems*. Elsevier Science, 1991.