Maximizing Energy Efficiency for Loss Tolerant Applications: The Packet Buffering Case

M. Majid Butt
Department of Computer Science and Engineering
Qatar University, Qatar
Email: majid.butt@qu.edu.qa

Eduard A. Jorswieck
Department of Electrical Engineering and Information Technology
Dresden University of Technology, Germany
Email: eduard.jorswieck@tu-dresden.de

Björn Ottersten
Interdisciplinary Center For Security, Reliability and Trust (SnT)
University of Luxembourg, Luxembourg
Email: bjorn.ottersten@uni.lu

***Abstract*—Energy efficient communication has emerged as one of the key areas of research due to its impact on industry and environment. Any potential degree of freedom (DoF) available in the system should be exploited smartly to design energy efficient systems. This paper proposes a framework for achieving energy efficiency for the data loss tolerant applications by exploiting the multiuser diversity and DoFs available through the packet loss pattern. For a real time application, there is a constraint on the maximum number of packets to be dropped successively that must be obeyed. We propose a channel-aware energy efficient scheduling scheme which schedules the packets such that the constraint on the average packet drop rate and the maximum number of successively dropped packets is fulfilled for the case when a finite number of unscheduled packets can be buffered. We analyze the scheme in the large user limit and show the energy gain due to buffering on the proposed scheme.**

I. Introduction

The quality of service (QoS) requirements characterize the behavior of the network traffic. If real time services require strict QoS guarantees, there are few degrees of freedom (DoF) available to exploit to improve the energy efficiency or throughput of the system. For example, multimedia traffic is constrained by a fixed set of parameters like average and/or maximum tolerable packet delay. Also, there is a certain bound on the average packet drop rate which must be obeyed to maintain the quality of experience (QoE) of the user.

On the other side, reducing costs of transmission is one of the most important factors for the network providers due to increasing cost of energy. Energy efficient transmission techniques must exploit any DoF available in the supported services to design the systems at the physical layer.

In the literature, energy efficient scheduling has been discussed in different settings for delay limited systems [1]–[3]. Similarly, the work in [4] addresses the importance of packet dropping mechanisms from energy point of view. The average packet drop rate is considered to be one of the most important parameters for system design. Only little attention is paid to the pattern of the dropped packets. For example, consider a scenario where the average packet drop rate θ_{tar} is quite small but a large number of packets are dropped successively due to deeply faded wireless channel. In spite of fulfilling an average packet drop rate guarantee, the users will experience a jitter in the perceived QoE (for a multimedia application). Thus, QoS service must also be defined in terms of maximum number of packets allowed to be dropped successively in addition to the average packet drop probability. This additional parameter characterizing the pattern of the dropped packets is termed the continuity constraint parameter N [5]. The authors in [6] discuss an optimal dropping scheme with the objective to minimize/maximize the packet drop gap. Traditionally, such problems are handled at upper layers of communications through link adaptation or automatic repeat request (ARQ) mechanisms. However, bringing this information to the physical layer design shows significant merits as the information can effectively be used for opportunistic scheduling purpose.

The work in [5] proposes an opportunistic scheduling scheme which exploits the DoF available through continuity constraint and average packet drop parameters and aims at minimizing the average system energy. The work characterizes the effects of the parameters θ_{tar} and N on the system energy. However, the proposed scheme does not allow buffering of data packets which is an integral part of the resource allocation mechanisms. This work generalizes the framework in [5] for the case when buffering of packets is allowed for a finite number of time slots. This additional DoF poses new challenges in terms of modeling and analysis of the problem. In addition to the QoS parameters θ_{tar} and N, the maximum size of the buffer provides another trade-off for energy efficiency.

The main contribution of this work is to generalize the framework proposed in [5] and analyze it. The generalized framework is more complex due to involvement of another DoF but provides better results in terms of energy efficiency as demonstrated through the numerical examples in Sec. V.

The rest of this paper is organized as follows. Section II introduces the system model and fundamental assumptions. We discuss and analyze the proposed scheme in Section III. The optimization problem is formulated in Section IV. We provide the numerical evidence of the gain of our scheme in Section V and conclude with the main contributions of the work in Section VI.

II. System Model

This work follows the system model in [5], [7]. We consider a multiple-access system with K users randomly placed within a circular area. Every user is provided an average rate $R_k = \frac{C}{K}$ where C denotes the spectral efficiency of the system.

We consider an uplink scenario while the system is time-slotted. Each user k experiences a channel gain $h_k(t)$ in slot t. The channel gain $h_k(t)$ is the product of path loss s_k and small-scale fading $f_k(t)$. The path loss is a function of the distance between the transmitter and the receiver and remains constant within the time scale considered in this work. Small-scale fading depends on the scattering environment. It changes from slot to slot for every user and is independent and identically distributed across both users and slots, but remains constant during the time span of a single time slot.

$E_k^R(t)$ and $E_k(t)$ denote the received and the transmitted energy for each user k such that

$$E_k^R(t) = h_k(t)E_k(t). \tag{1}$$

Note that the distribution of $h_k(t)$ differs from user to user. The channel state information is assumed to be known at the both transmitter and receiver sides.

The continuity constraint requires us to allow scheduling of multiple users simultaneously in the same time slot. If only a single user is scheduled per time slot, the continuity constraint cannot be satisfied without allowing outage when multiple users have already dropped N packets. The scheme follows the results for the asymptotic user case analysis which implies that there is no limit on the number of users scheduled simultaneously. Those scheduled users are separated by superposition coding and successive interference cancelation (SIC). Let $\mathcal{K}$ denote the set of users to be scheduled and Φ_k be the permutation of the scheduled user indices that sorts the channel gains in increasing order, i.e. $h_{\Phi_1} \leq \cdots \leq h_{\Phi_k} \leq \cdots \leq h_{\Phi_{|\mathcal{K}|}}$. Then, the energy of the scheduled user Φ_k with rate R_{Φ_k}, is given by [7], [8]

$$E_{\Phi_k} = \frac{N_0}{h_{\Phi_k}} \left(2^{\sum_{i \leq k} R_{\Phi_i}} - 2^{\sum_{i < k} R_{\Phi_i}} \right). \tag{2}$$

where N_0 denotes the noise power spectral density.

III. Modeling of the Scheme for the Finite Buffered Data Packets

A constant arrival of a single packet with normalized size $\frac{C}{K}$ is assumed for simplicity. However, the scheme is not restricted to this assumption as a random packet arrival process can be modeled as a constant arrival process where multiple arrived packets in the same time slot are merged as a single packet with random packet size following the framework in [2], [9]. The packet arrival occurs at the start of time slot and the scheduling is performed afterwards taking into account the newly queued packet. All the arriving packets are queued sequentially, i.e., the oldest arrived packet is the head of line (HOL). If a single packet has to be scheduled or dropped, it has to be HOL packet.

The continuity constraint and buffer size parameters for a user k are denoted by N and B, respectively; and assumed to be identical for all the users. The variables $i \leq N$ and $j \leq B$ denote the number of successively dropped and buffered packets for a user k at time t, respectively. A packet arriving at time t is not dropped immediately if not scheduled but buffered for B time slots and dropped then (if still not scheduled).

We use Markov decision process (MDP) to model the scheme. The summation of the number of already *successively* dropped and (already) buffered packets at time t defines the state of a Markov process. It is denoted by p. Thus,

$$p = i + j. \tag{3}$$

At the start of the Markov process ($p = 0$), the packet is not dropped if not scheduled as packets can be buffered for B time slots resulting $i = 0$ and $p = j$ for $p \leq B$. When the buffer is completely filled with packets, the unscheduled HOL packet is dropped onwards. Note that dropping operation is limited to a single packet as this is enough to make room for the newly arrived[1] packet at time t. Thus, the variable i increases and j is fixed to B for $p > B$. Thus, the maximum number of states in a Markov chain is $B + N + 1$ where $M = B + N$ denotes the termination state.

Let α_{pq} denote the transition probability from a state p to q. Furthermore, we denote transition probabilities associated with the scheduling, buffering and dropping decisions by the notation $\hat{\alpha}_{pq}$, $\tilde{\alpha}_{pq}$ and $\dot{\alpha}_{pq}$, respectively. We define α_{pq} as

$$\alpha_{pq} = \Pr(S_{t+1} = q | S_t = p) = \begin{cases} \hat{\alpha}_{pq} & \forall p, q \leq \min(p, B) \\ \tilde{\alpha}_{pq} & p < B, q = p + 1 \\ \dot{\alpha}_{pq} & p \geq B, q = p + 1 \\ 0 & \text{else} \end{cases} \tag{4}$$

We define a scheduling threshold.

Definition 1 (Scheduling Threshold κ_{pq}): It is defined as the minimum small scale fading value f required to make a state transition from state p to q such that

$$\hat{\alpha}_{pq} = \Pr\big(\kappa_{pq} < f \leq \kappa_{p(q-1)}\big) \quad 0 \leq q \leq \min(p, B). \tag{5}$$

where κ_{p0-} is defined to be infinity with S_{0-} denoting a dummy state before S_0. The threshold definition uses fading instead of channel gain to avoid near-far effect which gives unfair advantage to the users near the base station.

A. The Proposed Scheme

We assume infinitely large number of users in the system. In the large user limit, the scheduling decisions of the users decouple and the multiuser system can be modeled as a single user system following the work in [2], [10]. Every user makes his own scheduling decision independent of the other users.

The purpose of the scheduling scheme is to maximize the use of available fading conditions by scheduling as many packets as possible. Thus, the fading is quantized in such a way that the discrete set of state-dependent scheduling thresholds determines the intervals for the optimal scheduling decisions. In a state $p \geq q$, the scheduler makes a state transition to state q such that

$$q = \arg\min_{\acute{q}} \kappa_{p\acute{q}} < f \leq \kappa_{p(\acute{q}-1)} \quad 0 \leq \acute{q} \leq \min(p, B). \tag{6}$$

[1] The newly arrived packet remains in a separate *temporary* buffer momentarily before scheduling decision as it arrived at the start of the time slot.

For a state transition $\hat{\alpha}_{pq}, q \leq \min(p, B)$, the number of the scheduled packets is given by

$$L(p, f) = \min(p, B) - q + 1, \tag{7}$$

where q is determined uniquely by (6). Obviously, a user can only schedule as many packets as buffered. Thus, the maximum scheduled packets for a state $p < B$ are limited to $p - q + 1$ (due to constant arrival model) while they are fixed to $B - q + 1$ for $p \geq B$. Note that scheduling of packets starts with the HOL packet and ends with the most recently arrived packet. Equation (6) chooses q which maximizes the scheduling of packets for a state p and fading f. To meet the continuity constraint with probability one, κ_{MB} is set to zero to allow transmission of the HOL packet in state M.

We deduce the following properties of scheduling from (6).

***Property** 1:* The next state (in case of scheduling) is limited by the minimum of p and B. If $p \leq B$, q cannot exceed p, otherwise it is limited to B.

Thus, up to $\min(p, B) + 1$ buffered packets can be scheduled depending on small scale fading in a state p.

***Property** 2:* Scheduling thresholds follow the monotonic decrease property that

$$\kappa_{p(q+1)} \leq \kappa_{pq} \quad \forall p, 0 \leq q < \min(p, B). \tag{8}$$

If $f \leq \kappa_{p\min(p,B)}$, no scheduling occurs. In this case, the next state q equals $p + 1$ but a packet can be dropped or buffered depending on the conditions in (9) and (11) below. We have

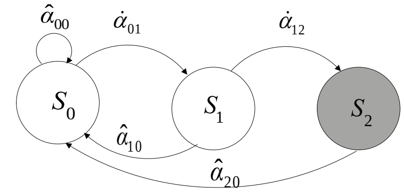
$$\begin{aligned} \tilde{\alpha}_{pq} &= \Pr(f \leq \kappa_{pp}), \qquad p < B, q = p + 1 && (9) \\ &= 1 - \sum_{q=0}^{p} \hat{\alpha}_{pq} && (10) \\ \dot{\alpha}_{pq} &= \Pr(f \leq \kappa_{pB}), \qquad p \geq B, q = p + 1 && (11) \\ &= 1 - \sum_{q=0}^{B} \hat{\alpha}_{pq} && (12) \end{aligned}$$

where κ_{pp} and κ_{pB} denote the minimum thresholds to schedule at least one packet. We explain (9) and (11) in detail.

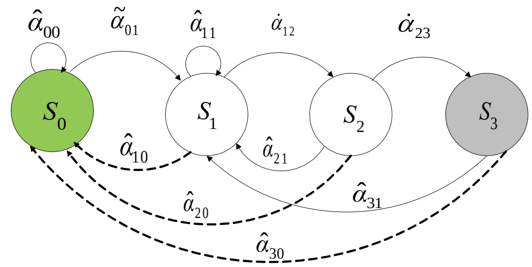
- $\tilde{\alpha}_{pq} = \Pr(f \leq \kappa_{pp}), \qquad p < B, q = p + 1$
  If $p < B$, the HOL packet is buffered with the option that it can be scheduled in one of the $B - p$ time slots in future.
- $\dot{\alpha}_{pq} = \Pr(f \leq \kappa_{pB}), \qquad p \geq B, q = p + 1$
  If $p > B$, the HOL packet has to be dropped as the buffer is already full. The best option by continuity constraint point of view is to drop HOL packet to make room for the newly arrived packet.

Consider the state diagram in Fig. 1 for the case when $N = 1, 2$ and $B = 0, 1, 2$. The corresponding state transition matrix for a system with $N = 1$ and $B = 2$ is given by

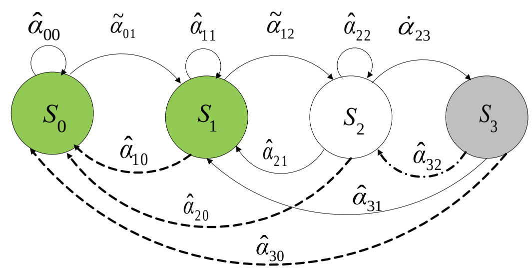
$$\mathbf{Q} = \begin{pmatrix} \hat{\alpha}_{00} & \tilde{\alpha}_{01} & 0 & 0 \\ \hat{\alpha}_{10} & \hat{\alpha}_{11} & \tilde{\alpha}_{12} & 0 \\ \hat{\alpha}_{20} & \hat{\alpha}_{21} & \hat{\alpha}_{22} & \dot{\alpha}_{23} \\ \hat{\alpha}_{30} & \hat{\alpha}_{31} & \hat{\alpha}_{32} & 0 \end{pmatrix}. \tag{13}$$



(a) N=2, B=0 as in [5].



(b) N=2, B=1



(c) N=1, B=2

Fig. 1. State transition diagrams for different buffer size and continuity constraint parameters. The green colored states represent the buffering states while grey state is the M^{th} state. The superscript notation with transition probabilities depict the actions associated with the transitions.

It should be noted that the number of states in an MDP are the same for the parameter sets $N = 2, B = 1$ and $N = 1, B = 2$ but the transition probability matrix $\mathbf{Q}$ differs and captures the effect of each parameter on the system energy. The parameter set $N = 2, B = 1$ requires optimization of 2 thresholds per state for $p \geq 1$ while the parameter set $N = 1, B = 2$ requires 3 thresholds per state for $p \geq 2$. We evaluate the energy efficiency of both the cases numerically in Sec. V.

IV. The Optimization Problem

The number of scheduled packets are considered as virtual users (VU) for the analysis purpose. It is known that the average energy consumption of the system per transmitted information bit at the large system limit $K \to \infty$ is then given by [5], [7]

$$\left(\frac{E_{\mathrm{b}}}{N_0}\right)_{\mathrm{sys}} = \log(2) \int_0^\infty \frac{2^{C\,\mathrm{P}_{h,\mathrm{VU}}(x)}}{x}\, \mathrm{d}\mathrm{P}_{h,\mathrm{VU}}(x) \tag{14}$$

where $\mathrm{P}_{h,\mathrm{VU}}(\cdot)$ denotes the cumulative distribution function (cdf) of the fading of the scheduled VUs. It comprises of the small scale fading and the path loss components of the VUs. However, in the large system limit, the state transitions depend only on the small scale fading distribution as the path loss for VUs follows the same distribution as the path loss of the users.

Now, we formulate the optimization problem according to the model described in Section III.

$$\min_{\mathbf{Q}\in\Omega} \left(\frac{E_{\mathrm{b}}}{N_0}\right)_{\mathrm{sys}} \tag{15}$$

$$\text{s.t.}: \begin{cases} \mathcal{C}_1 : 0 \le \sum_{m=0}^{\min(p,B)} \alpha_{pm} \le 1 & 0 \le \alpha_{pm} \le 1, \\ & 0 \le p \le M \\ \mathcal{C}_2 : \theta_r \le \theta_{\mathrm{tar}} & \mathbf{Q} \in \Omega \\ \mathcal{C}_3 : \sum_{q=0}^{M} \alpha_{pq} = 1 & 0 \le p \le M \\ \mathcal{C}_4 : B + N = M & B < \infty, N < \infty \end{cases} \tag{16}$$

where Ω denotes the set of permissible matrices for $\mathbf{Q}$ and θ_r is the average packet drop rate for a fixed $\mathbf{Q}$ and given by

$$\theta_r = \sum_{p=B}^{M-1} \alpha_{p(p+1)}\pi_p = \sum_{p=B}^{M-1} \big(1 - \sum_{m=0}^{\min(p,B)} \alpha_{pm}\big)\pi_p \ . \tag{17}$$

Equation (17) results by combining $\mathcal{C}_1$ and $\mathcal{C}_3$ in (16) while π_p denotes the steady state probability of the state p and follows the property

$$\sum_{p=0}^{M} \pi_p = 1 \ . \tag{18}$$

The forward transition for the state $p \ge B$ represents the events of dropping the packet and the summation over the corresponding transition probabilities $\alpha_{p(p+1)}$ gives the average dropping probability. The summation in (17) starts from state B as the unscheduled packets are buffered for $p < B$. For a fixed p, the corresponding channel-dependent optimal scheduling thresholds can be computed from the optimized $\vec{\alpha_p^*} = [\alpha_{p0}^*, \dots \alpha_{p\min(p,B)}^*]$ using (5).

The probability density function (pdf) of the small scale fading of the scheduled VUs is given by

$$\mathrm{p}_{f,\mathrm{VU}}(y) = \sum_{p=0}^{M} c_p \pi_p L(p,y)\, \mathrm{p}_f(y) \tag{19}$$

where $\mathrm{p}_f(y)$ and c_p denote the small scale fading distribution and a normalization constant while $L(p,y)$ is given by (7). The cdf of the VUs can be written as a sum of integrals

$$\begin{aligned} \mathrm{P}_{f,\mathrm{VU}}(y) &= \sum_{p=0}^{M} c_p \pi_p \Big(L(p,y) \int_{\kappa_{pq}}^{y} \mathrm{p}_f(\xi)\mathrm{d}\xi \\ &+ \sum_{b=1}^{\min(p,B)-q} b \int_{\kappa_{p(\min(p,B)-b+1)}}^{\kappa_{p(\min(p,B)-b)}} \mathrm{p}_f(\xi)\mathrm{d}\xi \Big) \end{aligned} \tag{20}$$

Using linear algebra (20) yields

$$\begin{aligned} \mathrm{P}_{f,\mathrm{VU}}(y) &= \sum_{p=0}^{M} c_p \pi_p \Big(L(p,y) \mathrm{P}_f(y) \\ &- \sum_{b=0}^{\min(p,B)-q} \mathrm{P}_f\big(\kappa_{p(\min(p,B)-b)}\big)\Big) \end{aligned} \tag{21}$$

since no users are scheduled for $y < \kappa_{p\min(p,B)}$. The channel distribution for the scheduled VUs can be computed using (21) and the path loss distribution.

A. Heuristic Optimization

The optimization problem is to compute a set of transition probabilities that result in minimum system energy in (14). For every state p, an optimal $\vec{\alpha^*} = [\alpha_{p0}^*, \dots \alpha_{p\min(p,B)}^*]$ needs to be computed. The computation of optimal $\vec{\alpha^*}$ under constraints in (16) is a stochastic optimization problem and requires heuristic optimization techniques like genetic algorithms, neural networks, etc., which provide acceptable solutions.

We choose Simulated Annealing (SA) to compute the solution for the optimization problem. SA is believed to help avoiding local minima by probabilistically allowing a candidate configuration to be the best known solution temporarily even if the configuration is not the best available solution at that time. $M+1$ state MDP models the problem for finite B, N while matrix $\mathbf{Q}$ represents a candidate configuration for SA. In SA, one transition probability (that fulfills the conditions in (16)) in $\mathbf{Q}$ is varied randomly in a single iteration and the objective function is computed *only* if the constraint $\mathcal{C}_2$ is satisfied. If the solution improves the previous best solution, the new configuration, i.e., the proposed $\mathbf{Q}$, is selected as the current best solution, discarded otherwise. However, depending on certain probability, the new configuration can be selected as the best solution even if it does not improve the previous best solution. This helps to avoid local minima. After a fixed number of iterations, a solution is computed which is considered optimal. We omit the details of SA scheme here due to space limitations but the reader is referred to [11] for the details of SA algorithm.

V. Numerical Results

In this section, we provide some numerical examples to demonstrate the potential gain of our scheme. K users are uniformly distributed in a circular cell except a forbidden region of radius δ around the access point and the path loss follows the distribution in [7]. We assume Rayleigh fading with mean one and the path loss is exponential with an exponent 2. The value of C is fixed to 0.5 bits/s/Hz. We use fast annealing cooling schedule for SA simulations.

We focus on the effects of buffering on a system constrained by parameters θ_{tar}, N. Fig. 2 shows system energy for various combinations of parameter sets $N, B, \theta_{\mathrm{tar}}$. For a fixed N, the system becomes more energy efficient as B increases. Thus, flexibility in latency requirements helps to combat the transmission challenges emerging from the finite packet dropping parameters. Note that energy gain by increasing parameter B is nearly constant and independent of θ_{tar} while the gain due to increasing N depends on θ_{tar}. For a fixed $M = 3$, the energy efficiency for the parameters $B = 2, N = 1$ is substantially larger than the case with parameters $B = 1, N = 2$ at small θ_{tar}; but the energy efficiency for the case $B = 1, N = 2$ outperforms the energy efficiency for the case $B = 2, N = 1$

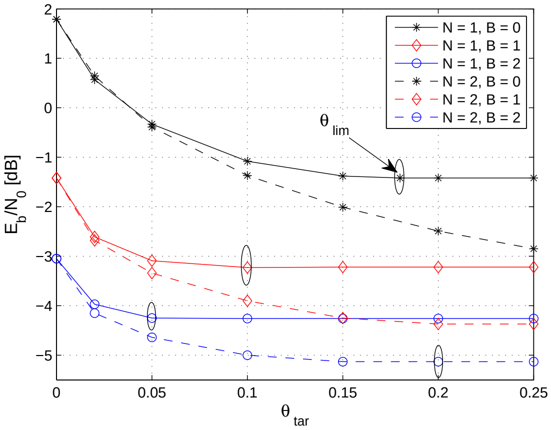


Fig. 2. The system energy as a function of average dropping probability, continuity constraint and buffer size parameters. The values for $\theta_{\rm lim}$ have been pointed with elliptical shapes for every curve.

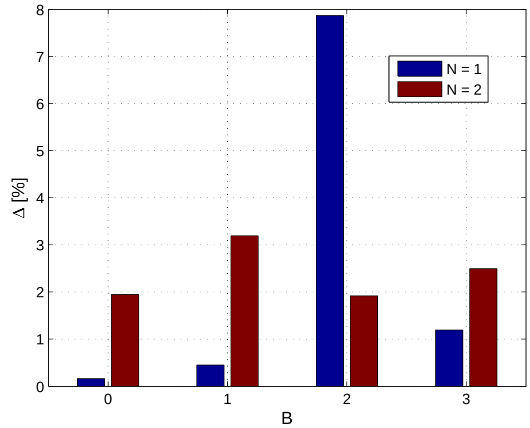


Fig. 3. The accuracy measure $\Delta[\%]$ as a function of buffer size and continuity constraint parameters while $\theta_{\rm tar} = 0.05$ for all simulations. The number of temperature iterations are 100 while $50 * (M+1)$ random configurations of $\mathbf{Q}$ are simulated at one temperature.

at large $\theta_{\rm tar}$. Thus, it is important to realize the operating region for the system to maximize the advantage from DoFs.

For a fixed N, the system energy saturates at some $\theta_{\rm lim} = \theta_{\rm tar}$ and $\theta_{\rm tar} > \theta_{\rm lim}$ does not improve the energy efficiency (c.f. Lemma 1 in [5])[2] where $\theta_{\rm lim}$ is the solution of (15) without applying $\mathcal{C}_2$ in (16). As evident from Fig. 2, an increase in value of B results in decrease in the value of $\theta_{\rm lim}$.

The parameter $\theta_{\rm tar} = 0$ is a special case where N becomes irrelevant as zero average packet drop rate implies that the system is lossless and thus, $N > 0$ does not help. However, if $\theta_{\rm tar} = 0$, an increase in value of B does help to make the system energy efficient as shown in Fig. 2. A system with parameters $B = 1, N = 1, \theta_{\rm tar} = 0$ is almost as much energy efficient as a system with $B = 0, N = 1$ and $\theta_{\rm tar} \simeq 0.20$, i.e. additional freedom in average dropping rate.

To measure the relative accuracy of the computed solution for the SA algorithm, we define a parameter Δ by

$$\Delta = 1 - \frac{\theta_r^*}{\min(\theta_{\rm tar}, \theta_{\rm lim})} \tag{22}$$

where θ_r^* is computed for a given $\theta_{\rm tar}$ by using (17) for the optimal solution $\mathbf{Q}^*$. The small Δ implies that the computed solution is sufficiently exploiting the DoF inherited by the system through parameter $\theta_{\rm tar}$. We observe in Fig. 3 that Δ is quite small for the computed solutions. As SA is a heuristic algorithm, there is no consistent pattern in the values of Δ. In general, for a fixed number of temperature iterations, the computation of the solution is expected to be hard as the number of parameters involved increases, .i.e., sparse $\mathbf{Q}$ and large number of thresholds.

VI. Conclusions

This work considers a framework for achieving energy efficiency by exploiting multiuser diversity. The proposed scheme is modeled and analyzed for the loss tolerant applications which are parameterized by average packet loss and maximum number of successively dropped packets. The optimization problem is modeled and formulated using Markov decision process and large system results. Specifically, this work focuses on analyzing the effect of buffer size (or allowed delay) on the problem as compared to the work in [5] where no buffering is allowed. The results characterize the additional gain resulting from the DoF available through buffering of data on the overall system energy efficiency.

[2] Please refer to [5] for the details of Lemma 1 and 2.

Acknowledgements

This work was made possible by NPRP 5 - 782 - 2 - 322 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

References

[1] A. Fu, E. Modiano, and J. N. Tsitsiklis, “Optimal transmission scheduling over a fading channel with energy and deadline constraints,” *IEEE Transactions on Wireless Communications*, pp. 630–641, March 2006.

[2] M. M. Butt, “Energy-performance trade-offs in multiuser scheduling: Large system analysis,” *IEEE Wireless Communications Letters*, vol. 1, no. 3, pp. 217–220, June 2012.

[3] M. J. Neely, “Optimal energy and delay tradeoffs for multiuser wireless downlinks,” *IEEE Trans. Inform. Theory*, vol. 53, no. 9, pp. 3095–3113, September 2007.

[4] ——, “Intelligent packet dropping for optimal energy-delay tradeoffs in wireless downlinks,” *IEEE Trans. on Automatic Control*, vol. 54, no. 3, pp. 565–579, March 2009.

[5] M. M. Butt and E. A. Jorswieck, “Energy efficient multiuser scheduling: Exploiting the loss tolerance of the application,” in *IEEE Global Communications Conference (Globecom)*, Anaheim, CA, USA, Dec 2012.

[6] C. W. L. Coralville and M. S. Andersland, “Control of consecutive packet loss in a packet buffer,” Patent US 5 629 936, May 13, 1997.

[7] G. Caire, R. Müller, and R. Knopp, “Hard fairness versus proportional fairness in wireless communications: The single-cell case,” *IEEE Trans. Inform. Theory*, vol. 53, no. 4, pp. 1366–1385, April 2007.

[8] D. Tse and S. Hanly, “Multi-access fading channels-part I: Polymatroid structure, optimal resource allocation and throughput capacities,” *IEEE Trans. Inform. Theory*, vol. 44, no. 7, pp. 2796–2815, November 1998.

[9] M. M. Butt, K. Kansanen, and R. R. Müller, “Hard deadline constrained multiuser scheduling for random arrivals,” in *WCNC*, Cancun, Mexico, March 2011.

[10] P. Viswanath, D. N. Tse, and V. Anantharam, “Asymptotically optimal water-filling in vector multiple-access channels,” *IEEE Trans. Inform. Theory*, vol. 47, no. 1, pp. 241–267, January 2001.

[11] S. Kirkpatrick, C. Gelatt, and M. Vecchi, “Optimization by simulated annealing,” *Science*, vol. 220, no. 4598, pp. 671–680, May 1983.