EXTENDED ABSTRACT: ANTICIPATORY NETWORKING FOR ENERGY SAVINGS IN 5G SYSTEMS

E. Pollakis, and S. Stańczak

Fraunhofer Heinrich Hertz Institute, Einsteinufer 37, 10587 Berlin, Germany
email: {emmanuel.pollakis,slawomir.stanczak}@hhi.fraunhofer.de

ABSTRACT

In this paper, we devise novel techniques for saving energy in 5G wireless systems. By means of anticipated transmission rates we find user-cell assignments and scheduling policies that help to identify energy-efficient network topologies. In particular, the objective of this paper is to find a user-cell association and rate allocation over time that provides the requested Quality of Service (QoS) to all users while attempting to reduce the energy consumption by identifying the set of active cells consuming the least amount of energy. We formalize this problem as a non-convex optimization problem that accounts for the requirements of two emerging application types, i.e., delay-tolerant and buffered delay-sensitive applications. We use an energy consumption model that specifically includes the static energy consumption and the dynamic, load dependent energy consumption of cells. We apply relaxation techniques to find feasible anticipated schedules for rate allocation and user-cell assignments. Our approach is characterized by its broad applicability, good performance and low complexity making it amenable to online implementation. We characterize achievable energy saving gains by means of simulations in a realistic network scenario under realistic traffic patterns.

I. INTRODUCTION

With the advent of the Internet of Things (IoT) billions of new devices will be connected wirelessly to the mobile communication system of the fifth generation (5G). A wide spectrum of use cases mandates 5G to support extreme objectives for delay, capacity and energy necessitating a high network adaptability to varying user requirements. Todays networks are operated in a static manner with fixed network settings providing the maximum quality of service (QoS) at all times. Eventhough such a mode of operation might satisfy delay and capacity requirements, it will lead to unacceptable high energy consumption in 5G networks. Therefore, it is of utmost importance to develop mechanisms supporting energy savings in both peak hours as well as in off-peak hours.

Most existing energy saving techniques, such as cell deactivation or sleep mode, are designed for stationary users and static user demands. These techniques usually have bad performance in bursty traffic situations. The achievable energy savings are nullified by short time increases of traffic demand in certain areas. Proactive resource allocation and user assignment (PRAUA) is a promising approach to enable 5G to stand up to its high promises by improving service quality and reducing energy consumption at the same time. In particular, PRAUA helps 'smear' the traffic requirements in time and space allowing for the energy savings techniques to be valid over a longer period of time. Thereby, PRAUA exploits the knowledge about users' mobility which can be obtained from side information or estimated with sufficient accuracy due to the high regularity in human mobility [1]. This information along with learned path loss maps [2] is used to proactively build user-cell assignment and resource allocation schedules that greatly support energy savings in cellular communication systems during off-peak hours. In particular, we develop algorithms that schedule data transmissions for new service applications when it is favorable for energy savings. The developed mechanisms target two new service types enabled by the storage capabilities at user devices: delay-tolerant and buffered delay-sensitive applications. File transfer is a famous representative of the former one where it is required to transfer a certain amount of data before a deadline time with no limitations on the instantaneous transmission rate. Buffered delay-sensitive applications include services like stored content streaming (music/video). Such applications require a constant instantaneous data rate where data can be pulled either directly from the access network or from a pre-filled buffer (depicted in Fig. 1). These two service types allow to delay or bring forward the transfer of data to users which is the fundamental concept we exploit for energy savings. To increase the degrees of freedom for the PRAUA by multi-connectivity to multiple cells and to exploit the mutual information received from them we propose the use of fountain coding [3]. This helps to find better solutions for the problem at hand and increases the robustness of our solution.

Proactive scheduling has been considered in [4] to improve the QoS of users traveling through the service area of several cells. The presented framework plans the resource allocation over a certain time horizon for fixed user-cell assignments to maximize the throughput to users. The authors of [5] propose a predictive framework for video streaming applications to increase the energy-efficiency in wireless networks. The problem is composed as a mixed integer linear program (MILP) where decisions on multiuser rate allocation, video segment quality, and base station transmit power are jointly optimized. A heuristic multi

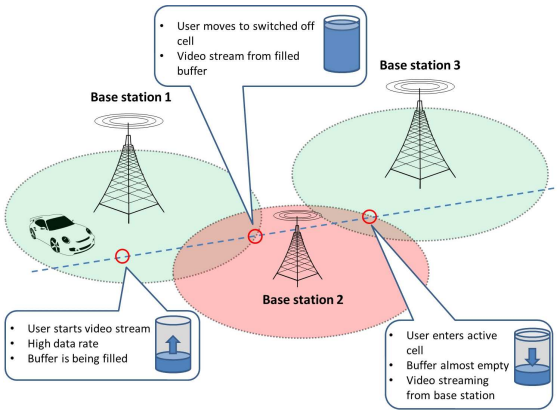

Fig. 1. Toy-example of a buffered delay-sensitive application schedule.

stage algorithm is used to derive solutions for the MILP problem by first allocating rates to users and then determining the segment quality and active base station set. The reasoning is based on the observation that efficient rate-allocation schemes provide power savings. Other analytical justifications for the perfomance are not given. The work in [6] is most closely related to ours. By proactive resource allocation and video quality decisions the authors reduce the energy consumption of the whole network by solving a mixed integer non-linear problem (MINLP). An algorithm is proposed that decomposes the association and resource allocation problem in a master problem and several sub-problems to make the problem tractable. Thereby, the authors leverage energy costs and video quality taking into account backhaul costs. The resulting integer programs are solved directly by mathematical solvers and the authors argue to achieve decent scalability. However, this is achieved by assuming the allocation of an equal number of resource blocks to all users in the master problem.

In contrast to the above work, we explicitly incorporate the user-cell assignment in the optimization framework and target energy savings with a guaranteed QoS level instead of maximizing the QoS. Furthermore, we use mathematically justified relaxation techniques instead of heuristics to derive solutions ensuring good scalability. In the following we summarize the main contribution of our work:

- We propose an optimization framework that exploits the knowledge of user-cell trajectories and learned path-loss maps. It finds user-cell association and rate schedules that provide the requested QoS of users and reduces the energy consumption of future cellular communication networks.
- Our model for energy consumption is general enough to capture static energy consumption (cooling, basic power conversion etc.) as well as dynamic load dependent energy consumption.
- We exploit the end user devices' storage capabilities to implement delay-tolerant and buffered delay-sensitive applications with PRAUA.
- With the introduction of PRAUA we stretch the applicability of cell sleep and switching on/off techniques in the time horizon leading to improved energy savings.
- The use of fountain coding is proposed to improve the QoS of users while being able to deactivate more cells.
- We present relaxation techniques that are able to give good solutions to this problem in reasonable time making it amenable for online implementation. Thereby, it has theoretical justification for its good performance.

II. SUMMARY

In the following we sketch the basic idea of our approach. We exploit the possibility to preload and store data on user devices which will serve as an enabling concept to save energy in the communication system by disengaging certain cells[1]. By delaying the service provision of some users we may avoid to activate cells that are only needed when the traffic demand is of bursty nature. The result is a service use at user level that receives and buffers data in high capacity cells whereas it avoids access to cells that are overloaded or switched of for reasons of energy savings. We use the predicted routes of users and the learned path loss coverage maps to find such a user-cell association and rate allocation policy under the exploitation of the users' buffers. When a user is predicted to pass an area without coverage it will be allocated more resources right before, so that the data can be loaded in the buffer for a delay-sensitive non-realtime application (bridging the coverage hole). For a delay-tolerant applicaton the user can be denied service in certain coverage regions when either before or after the provided service is high enough. To enable the multi-connectivity of users with low coordination overhead we propose to use fountain codes [3] for mutual information combining. In the concept of fountain coding a potential infinitely long stream of encoded symbols is generated from a finite set of data symbols. Decoding is possible as soon as a particular amount of code symbols is received error free with no requirement for a consecutive order.

II-A. Scenario and System Model

We consider a cellular heterogeneous communication system employing an OFDM-based resource allocation. We are interested in switching off capacity units of the network, e.g. sectors, cells or the entire base station[2]; the corresponding decisions are performed at a central network controller. The set of all cells is denoted by $\mathcal{M} = \{1, 2, ..., M\}$. Each cell i has total number of resource blocks B_i to allocate to its users. There are N users in the system to be served and we denote the set of all users as $\mathcal{N} = \{1, 2, ..., N\}$. The time is divided into K time slots of equal duration Δ_k. For each time slot, the objective is to find a resource allocation and user-cell association. Each user is equipped with a buffer and

[1]Notice, that we are considering only capacity cells for deactivation. Basic coverage for other users and services has to be secured at all times. We consider therefore a basic coverage by some legacy network.

[2]In the text that follows we will use cells as a placeholder for any type of network element.

we denote the buffer level (in bits) of user j in slot k by $d_j^{(k)}$ with $d_j^{(0)} = 0$ (empty buffer at start). In this work we assume a sufficiently large buffer and refer to the technology specific spectral efficiency per resource block of the link from cell i to user j in slot k as $\omega_{i,j}^{(k)}$.

Assumption 1: A reliable estimate of the users' routes and the supported spectral efficiency per resource unit along those routes is available at the central controller.

The task of our optimization framework is to provide a schedule of resource allocations satisfying the QoS requirements of users while trying to reduce the energy consumption. If a user j is served by cell i in slot k we denote the effective transmit data rate as $r_{i,j}^{(k)} := b_{i,j}^{(k)}\omega_{i,j}^{(k)}$ where $b_{i,j}^{(k)}$ is the number of resource units allocated to user j by cell i in slot k. We collect the rates allocated by cell i to all users at time k in vector $\boldsymbol{r}_i^{(k)} = \left[r_{i,1}^{(k)}, r_{i,2}^{(k)}, ..., r_{i,N}^{(k)}\right]^T$. We further use $\mathbf{R}^{(k)} = \left[\boldsymbol{r}_1^{(k)}, \boldsymbol{r}_2^{(k)}, ..., \boldsymbol{r}_M^{(k)}\right]$ to refer to all rates allocated over all cells to all users in slot k.

Definition 1 (Instantaneous Cell Load): Given the rate assignment matrix $\mathbf{R}^{(k)}$ for slot k, the load of cell i, denoted by $\rho_i^{(k)}\left(\mathbf{R}^{(k)}\right) \in [0,1]$ or simply $\rho_i^{(k)}$ for notational simplicity, is defined to be the ratio of the number of resource blocks allocated to users served by cell $i \in \mathcal{M}$ in slot k to the total number of resource blocks B_i available at this cell, i.e., $\rho_i^{(k)} = \frac{\sum_{j\in\mathcal{N}} b_{i,j}^{(k)}}{B_i}$.

We use $\boldsymbol{\rho}_i := [\rho_i^{(1)}, \ldots, \rho_i^{(K)}]^T \in [0,1]^K$ to denote the vector of cell loads at cell i for all time slots and denote the collection of all cell loads over time by $\mathbf{P} := [\boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_M]^T \in [0,1]^{M\times K}$. A consequence of Definition 1 is the following fact:

Fact 1: The load at cell i satisfies $\rho_i^{(k)} > 0$ if and only if (iff) cell i serves at least one user in slot k.

In other words, $|\boldsymbol{\rho}_i\ \mathbf{1}|_0 = 0$ iff cell i serves no user in all time slots K, where $\mathbf{1} \in \mathbb{R}^K$ is a vector of ones and $|\cdot|_0$ is the l_0-norm[3]. If $|\boldsymbol{\rho}_i\ \mathbf{1}|_0 = 0$ cell i can be deactivated for energy saving reasons.

Buffered delay-sensitive applications: In the case of a buffered delay-sensitive application each user has a strict per time slot data rate requirement $r_j^{\min}$. Whenever the scheduling algorithm allocates a higher data rate to a user in time slot k, i.e., $r_{i,j}^{(k)} > r_j^{\min}$, then the additional transferred data is saved in the users buffer $d_j^{(k)} = d_j^{(k-1)} + \Delta_k(r_{i,j}^{(k)} - r_j^{\min})$. If the user is not allocated a sufficiently high rate in slot k it loads the missing data from its buffer. In this case the buffer level decreases as $d_j^{(k)} = d_j^{(k-1)} - \Delta_k(r_j^{\min} - r_{i,j}^{(k)})$. In every time slot k users require the minimum data rate either streamed from a cell or loaded from its buffer which is stated as[4]

$$\sum_{i\in\mathcal{M}} r_{i,j}^{(k)} + \frac{d_j^{(k-1)}}{\Delta_k} \geq r_j^{\min}. \tag{1}$$

The buffer level of user j at the end of time slot k is therefore described by

$$0 \leq d_j^{(k)} = d_j^{(k-1)} + \sum_{i\in\mathcal{M}} \Delta_k r_{i,j}^{(k)} - \Delta_k r_j^{\min}. \tag{2}$$

Since each base station has only B_i resource units to allocate to users we have the condition

$$\sum_{j\in\mathcal{N}} \frac{b_{i,j}^{(k)}}{B_i} = \sum_{j\in\mathcal{N}} \frac{r_{i,j}^{(k)}}{B_i\omega_{i,j}^{(k)}} \leq \rho_i^{(k)}. \tag{3}$$

Delay-tolerant applications: For delay-tolerant applications users are typically interested in maximizing throughput at the expense of delay. Hence, the QoS requirement of user j is fulfilled if the requested amount of data D_j can be transferred within a predefined number of time slots K. To satisfy the users QoS constraint it suffices to guarantee

$$\sum_{i=1}^{M}\sum_{k=1}^{K} \Delta_k r_{i,j}^{(k)} \geq D_j, \tag{4}$$

which can be interpreted as an average per slot data rate requirement $\bar{r}_j = \frac{D_j}{K} = \frac{1}{K}\sum_{k=0}^{K} \Delta_k r_{i,j}^{(k)}$. The buffer level of user j is implicitly included in (4).

II-B. Problem statement

We are now in the position to state the optimization problem that aims at finding the optimal set of active cells, user-cell assignments and rate allocations while consuming the least amount of energy. The objective function $E : [0,1]^{M\times K} \to \mathbb{R}_+$ is a combination of static and dynamic sources of energy consumption. In more detail, each active cell has static energy consumption of e_i per time slot and a load dependent part which is captured by a concave or convex function $f_i : [0,1]^K \to \mathbb{R}_+$. The total network energy consumption is thus given by

$$E(P) = \sum_{i\in\mathcal{M}} K\ e_i\,|\boldsymbol{\rho}_i\ \mathbf{1}|_0 + f_i\left(\boldsymbol{\rho}_i\right). \tag{5}$$

The above model assumes that cells are deactivated before the first time slot and stay inactive for all K time slots. The model can easily be adapted to modes of operation where so called micro-sleeps of cells are allowed. Such a mode of operation and the comparison with the former mode will be presented elsewhere.

The complete optimization problem for *buffered delay-*

[3]For a scalar $x \in \mathbb{R}$, the l_0-norm is defined as $|x|_0 := 1$ if $x \neq 0$ and $|x|_0 :=$ 0otherwise.

[4]Note, that this definition allows users to be served by multiple cells as well as the buffer in a time slot. In such cases fountain coding is used to implement mutual information combing.

sensitive applications can be written as

$$\text{min.} \sum_{i\in\mathcal{M}} K\ e_i\,|\boldsymbol{\rho}_i\ \mathbf{1}|_0 + f_i\,(\boldsymbol{\rho}_i) \tag{6a}$$

$$\text{s. t.:} \sum_{j\in\mathcal{N}} \frac{r_{i,j}^{(k)}}{B_i\omega_{i,j}^{(k)}} \le \rho_i^{(k)} \qquad \forall i,k \tag{6b}$$

$$\sum_{i\in\mathcal{M}} r_{i,j}^{(k)} + \frac{d_j^{(k-1)}}{\Delta_k} \ge r_j^{\min} \qquad \forall j,k \tag{6c}$$

$$d_j^{(k-1)} + \sum_{i\in\mathcal{M}} \Delta_k r_{i,j}^{(k)} - \Delta_k r_j^{\min} = d_j^{(k)} \qquad \forall j,k \tag{6d}$$

$$0 \le d_j^{(k)} \qquad \forall j,k, \tag{6e}$$

where the optimization variables are $r_{i,j}^{(k)} \in \mathbb{R}_+$ and $\rho_i^{(k)} \in [0,1]$. Thereby, (6b) assures that cells are not overloaded and (6c) guarantees that users receive the required instantaneous data rate. Constraint (6d) represents the flow of data in and out of the users' buffer.

Problem (6) can be written in a more compact form as

$$\text{min.} \sum_{i\in\mathcal{M}} K\ e_i\,|\boldsymbol{\rho}_i\ \mathbf{1}|_0 \tag{7a}$$

$$\text{s. t.:} \sum_{j\in\mathcal{N}} \frac{r_{i,j}^{(k)}}{B_i\omega_{i,j}^{(k)}} \le \rho_i^{(k)} \qquad \forall i,k \tag{7b}$$

$$\sum_{l=1}^{k}\left(\sum_{i\in\mathcal{M}} r_{i,j}^{(l)} - r_j^{\min}\right) \ge 0 \qquad \forall j,k, \tag{7c}$$

since the buffer level at the end of time slot k can be stated as the data surplus of the aggregated data transmitted up to time slot k.

The problem formulation for *delay-tolerant applications* uses (4) and we deduce

$$\text{min.} \sum_{i\in\mathcal{M}} K\ e_i\,|\boldsymbol{\rho}_i\ \mathbf{1}|_0 \tag{8a}$$

$$\text{s. t.:} \sum_{j\in\mathcal{N}} \frac{r_{i,j}^{(k)}}{B_i\omega_{i,j}^{(k)}} \le \rho_i^{(k)} \qquad \forall i,k \tag{8b}$$

$$\sum_{k=0}^{K}\sum_{i\in\mathcal{M}} r_{i,j}^{(k)} = \frac{D_j}{\Delta_k} \qquad \forall j \tag{8c}$$

with the optimization variables being $r_{i,j}^{(k)} \in \mathbb{R}_+$ and $\rho_i^{(k)} \in [0,1]$. The main difference to problem 7 is in constraint (8c) where an average data rate per user is required.

Problem 7 and problem 8 exhibit similar structure as the problem considered in [7]. Thus, we can apply similar reformulation techniques in combination with the Majorization-Minimization method to derive algorithms that find good solutions in reasonable time.

III. OUTLOOK

In the full version of this paper we will detail on the algorithms to find solutions to Problem 7 and Problem 8 in a computationally efficient way. Furthermore, we use a realistic simulation setup for dense urban communication scenarios from the METIS project [8], [9] to evaluate the energy saving gains and compare them with a base line approach. Preliminary simulation results indicate large energy saving potential.

Acknowledgements

This work has been partly supported by the framework of the research project ComGreen under the grant-number 01ME11010, which is funded by the German Federal Ministry of Economics and Technology (BMWi).

Part of this work has been performed in the framework of the FP7 project ICT-317669 METIS, which is partly funded by the European Union. The authors would like to acknowledge the contributions of their colleagues in METIS, although the views expressed are those of the authors and do not necessarily represent the project.

IV. REFERENCES

[1] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.

[2] M. Kasparick, R. Cavalcante, S. Valentin, S. Stanczak, and M. Yukawa, "Kernel-based adaptive online reconstruction of coverage maps with side information," *Vehicular Technology, IEEE Transactions on*, vol. PP, no. 99, pp. 1–1, 2015.

[3] D. MacKay, "Fountain codes," *Communications, IEE Proceedings-*, vol. 152, no. 6, pp. 1062–1068, Dec 2005.

[4] H. Abou-zeid, H. Hassanein, and S. Valentin, "Optimal predictive resource allocation: Exploiting mobility patterns and radio maps," in *Global Communications Conference (GLOBECOM), 2013 IEEE*, Dec 2013, pp. 4877–4882.

[5] ——, "Energy-efficient adaptive video transmission: Exploiting rate predictions in wireless networks," in *IEEE Transactions on Vehicular Technology*, vol. 63, no. 5, Jun 2014, pp. 2013–2026.

[6] A. Galanopoulos, G. Iosifidis, A. Argyriou, and L. Tassiulas, "Green video delivery in LTE-based heterogeneous cellular networks," in *World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2015 IEEE 16th International Symposium on a*, June 2015, pp. 1–9.

[7] E. Pollakis, R. Cavalcante, and S. Stanczak, "Base station selection for energy efficient network operation with the majorization-minimization algorithm," in *Signal Processing Advances in Wireless Communications (SPAWC), 2012 IEEE 13th International Workshop on*, June 2012, pp. 219–223.

[8] METIS Project, "Deliverable D1.1: Scenarios, requirements and KPIs for 5G mobile and wireless systems," METIS Project (Mobile and wireless communications Enablers for the Twenty-twenty Information Society), Tech. Rep. ICT-317669-METIS/D1.1, 2013.

[9] ——, "Deliverable D6.1: Simulation guidelines," METIS Project (Mobile and wireless communications Enablers for the Twenty-twenty Information Society), Tech. Rep. ICT-317669-METIS/D6.1, 2013.