Daniel Nowak, Tobias Mahn, Hussein Al-Shatri, Alexandra Schwartz and Anja Klein, ”A Generalized Nash Game for Mobile Edge Computation Offloading,” in *Proc. 6th IEEE International Conference on Mobile Cloud Computing, Services, and Engineering (Mobile Cloud))*, March 2018.

A Generalized Nash Game for Mobile Edge Computation Offloading

Daniel Nowak*, Tobias Mahn†, Hussein Al-Shatri†, Alexandra Schwartz*, Anja Klein†
∗ Graduate School of Computational Engineering, TU Darmstadt, Germany, {nowak, schwartz}@gsc.tu-darmstadt.de
† Communications Engineering Lab, TU Darmstadt, Germany, {t.mahn, h.shatri, a.klein}@nt.tu-darmstadt.de

***Abstract*—The number of tasks performed on wireless mobile devices instead of stationary computers is steadily increasing. Due to the limited resources of battery and computation capability on these devices, computation offloading became a relevant concept. We consider several mobile users with a splittable computation task each, which try to minimize their own computation time. All of them are connected to a central access point, where a cloudlet with limited computation power can be utilized for offloading fractions of their tasks. To account for the selfishness of the users and their individual goals, we propose a game theoretic framework resulting in a nonconvex generalized Nash game. The decision of each mobile user, which fraction of his task to offload, depends on the offloading decisions of the others, since they share the communication resources and computation capabilities of the cloudlet. We prove the existence and uniqueness of the generalized Nash equilibrium and propose an algorithm for computing it efficiently. In addition, we show that the price of anarchy of our model is one and investigate the advantage of computation offloading numerically. Furthermore, we extend our model to a scenario, where mobile users are able to offload parts of their computation in repeated sessions during a given time period. This allows every mobile user to offload multiple successive tasks.**

I. Introduction

Computation offloading has become an active field of research over the last few years [1]. Although wireless mobile devices are steadily increasing their computation capabilities, the limited battery capacity prohibits the local execution of highly demanding tasks on the devices. Even before the smartphone era, studies showed that an increased battery life is a crucial feature for customers [2]. A later study, that evaluated the typical charging behavior and battery levels of over 4000 smartphones users, revealed a cautious attitude of most users towards keeping as much charge as possible [3]. Thus, offloading of demanding applications like video processing or augmented reality to a more powerful server instead of processing the data locally is a desirable approach to save battery power and computation time. The remote server can be a distant cloud service, but also mobile edge computing has become a desirable concept for 5G cellular networks [4], [5].

While many publications consider central optimization algorithms to determine a globally minimal solution for completion time or energy consumption in the network [6], [7], these formulations neglect the individual behavior of users. Since users can be considered selfish and pursue their own goals, like battery level and individual completion time, the central solution might not be optimal for them. Therefore, they may be unsatisfied with the forced offloading decision by the central entity.

For this a game-theoretic approach can be considered, see for example [8]–[11]. In a game users can be modeled as individual players competing against each other. Although this might not lead to the globally optimal solution (price of anarchy), it can be an intuitive option to describe the behavior of the users as multiple independent decision makers. Therefore, it is an interesting question to investigate whether a globally optimal result can be achieved when applying a game theoretic model to computation offloading.

The aforementioned computation offloading publications can be divided into two groups, publications [8]–[10] consider non-splittable tasks, whereas [11] considers a splittable task model. Usually, multiple offloading tasks are positioned in a queue and it is only possible to compute the task locally or to offload it completely. For many applications this is a valid choice when the whole set of information is required to complete the computation, e.g. for sorting data. On the other hand, splittable tasks can occur in data processing applications [6]. One possible application mentioned in [11] is a future concept for public WiFi access points, where the provided connection not only grants internet access, but also holds the ability to offload computation tasks partially to an attached server with computation capability (cloudlet). The authors of [11] model tasks in their offloading scenario as a constant stream, which allows them to use the tools of queuing theory. In contrast, we assume only a single splittable task per user, which leads to non-smooth constraints in the game theoretic formulation and prevents the use of queuing theory. Nonetheless, we show the existence of a generalized Nash equilibrium. To introduce the possibility of multiple tasks per user in a given time period, we further extend our model with a dynamic offloading scheme. We consider an access point with attached cloudlet that has a finite computation capability which results in a competition about resources among the users.

The paper is structured as follows: In Section II we introduce the scenario, a communication model and a computation model. Section III contains the game formulation, existence result and an algorithm to alternatively solve the offloading problem in a closed-form. Additionally we investigate the price of anarchy here. Finally we apply our theoretical results numerically in Section IV, where we examine the impact of computation offloading for the users.

II. SYSTEM MODEL

A. Overview

For our mobile-edge computation offloading system model we consider K mobile users (MUs) which are connected to one wireless access point (AP) with an attached cloudlet. Each MU has a uniformly splittable task he wants to compute in the shortest possible time. For this, he can offload fractions of the computation task to a cloudlet service, which is offered by the AP. This way, he can simultaneously compute on his local mobile device and use the computation power of the cloudlet service to shorten his overall computation time. Since the cloudlet has limited computation power, offloading might or might not be beneficial. To model the individual behavior of each MU, we use a game theoretic framework for the offloading model. In order to achieve this, we first introduce the communication model for transmitting the task parts to the cloud. Afterwards, the computation model is explained, which is separated into local computation at the K MUs and offloaded computation.

B. Communication Model

We assume a communication model with K MUs with index k, $k = 1, \dots, K$. The MUs transmit their offloaded data utilizing a TDMA scheme with bandwidth B. Hence, there is no interference while offloading. Let p_k denote the transmit power of mobile device k, h_k the channel coefficient from the k-th user to the AP and σ^2 the power of white Gaussian noise. Then, the transmission rate of MU k is given by the Shannon channel capacity

$$R_k = B \log_2 \left(1 + \frac{p_k |h_k|^2}{\sigma^2}\right).$$

The AP is assumed to have one antenna, i.e. being able to receive one signal at a time. To avoid scheduling problems, we assume all transmissions have to be completed before the computation starts, which can be seen as an upper bound for transmission time. After the computation the AP transmits the result back to each MU. The result is expected to have a negligible size compared to the original task size W_k and is not considered in the offloading scenario, similar to [12]. Let x_k be the percentage of the task that user k offloads. Further, let W_k be the size of user k's computation task in bits. Then the accumulated offloading time for all MUs is given by

$$\sum_{k=1}^{K} x_k \frac{W_k}{R_k}. \tag{1}$$

After all transmissions are finished, the computation on the cloudlet begins.

C. Computation Model

Each MU is able to split his computation task into two parts: the local part and the offloaded part. The local part $(1 - x_k)$ is computed on the mobile device of user k and the computation starts at the same moment the transmission for offloading starts. Since different types of tasks may have different computation complexities, let L_k denote the task size of MU k in CPU cycles. Further let f_k denote the computation power of his mobile device in CPU cycles per second. Then, the time MU k needs for the local part of his computation, is given by

$$T_k^{\text{local}} = (1 - x_k)\frac{L_k}{f_k}.$$

The time of the offloaded tasks to finish depends on the offloading decisions of all MUs and the availability of the cloudlet. The total time needed is the sum of the transmission time (1) to the AP, the computation time at the cloudlet and possibly the time $C \geq 0$ necessary to finish some already running tasks on the cloudlet. This enables us to look at a dynamic scenario, where multiple MUs can offload several tasks in a given time frame, which will be explained in detail in Section IV. Let f_c denote the computation power of the cloudlet in CPU cycles per second. Then the total time for all offloaded tasks to be finished is

$$T^{\text{offload}} = \sum_{k=1}^{K} x_k \underbrace{\left(\frac{W_k}{R_k} + \frac{L_k}{f_c}\right)}_{=:O_k} + C,$$

where $x_k O_k$ represents the amount of time that MU k contributes to the total time for offloading. Throughout the paper we assume $O_k \in (0, \infty)$ for all MUs k, i.e all MUs can reach the AP, the cloudlet has nonzero computation power and communication and computation are not instantaneous. Let T_k denote the time for user k until his computation is finished. It depends on his decision, whether to offload a part of the computation $(x_k > 0)$ or to compute only on his local device $(x_k = 0)$:

$$T_k = \begin{cases} \max\{T_k^{\text{local}}, T^{\text{offload}}\}, & \text{if } x_k > 0, \\ T_k^{\text{local}}, & \text{if } x_k = 0. \end{cases} \tag{2}$$

Since each MU tries to minimize his overall computation time T_k, we use a non-cooperative game theoretic framework, which is explained in the following.

III. GAME FORMULATION

In this section, the game theoretic approach for the computation offloading scenario is explained in detail. Each MU solves his own optimization problem with his respective design variables being his offloading decision x_k and his completion time T_k, which he seeks to minimize. The completion time has to fulfill two constraints:

1) $T_k \geq T_k^{\text{local}}$, i.e. the completion time is as least as long as the local computation time.
2) Only if MU k chooses to offload and thus $x_k > 0$: $T_k \geq T^{\text{offload}}$, i.e. MU k has to wait for the offloaded computation tasks to finish.

The second constraint results in a *vanishing constraint* (see for example [13], [14]), which is modeled by

$$x_k T^{\text{offload}} \leq x_k T_k.$$

Note that T^{offload} also depends on x_k. For $x_k > 0$, this constraint is equivalent to the desired constraint stated in 2) and for $x_k = 0$ it vanishes, i.e. is automatically fulfilled. The complete optimization problem of MU k is then given by:

$$\begin{aligned} \min_{x_k, T_k} \quad & T_k, \\ \text{s.t.} \quad & (1 - x_k)\frac{L_k}{f_k} \leq T_k, \\ & x_k \left[\sum_{l=1}^{K} x_l O_l + C\right] \leq x_k T_k, \\ & x_k \in [0,1], \quad T_k \in [0, \bar{T}]. \end{aligned} \tag{3}$$

Here, $\bar{T}$ is an upper bound which can be chosen for example as $\bar{T} = \max_k \left\{\frac{L_k}{f_k}\right\}$. The *feasible set* of (3) are all (x_k, T_k) that satisfy the given constraints. Furthermore, let $z_{-k} = (x_l, T_l)_{l \neq k}$ denote the strategy vector of all MUs besides k. Since the feasible set for each MU depends on the decisions of the other MUs, the K coupled optimization problems (3) are called a *generalized Nash game*, which we denote by Γ. To solve his optimization problem MU k needs information about the other players (offloading decisions x_{-k} and the offloading factors O_{-k}). This assumption can be weakened and is explained at the end of this section. Next, we introduce the concept of generalized Nash equilibria:

Definition 1. *A strategy vector $\bar{z} = \left(\bar{x}_k, \bar{T}_k\right)_{k=1}^{K}$ is called a* generalized Nash equilibrium *(GNE) of the game Γ, if $(\bar{x}_k, \bar{T}_k)$ is the optimal solution of the respective optimization problem (3) with x_{-k} fixed to $\bar{x}_{-k}$ for each $k = 1, \ldots, K$.*

In a generalized Nash equilibrium, no MU will deviate from his offloading strategy, if all other MUs keep theirs. One typical way to ensure the existence of generalized Nash equilibria are convexity assumptions, see for example [15]. Unfortunately the vanishing constraint for the computation offloading is nonconvex. We show in the next section that we can prove the existence of a generalized Nash equilibrium nonetheless and even give a closed form solution.

A. Game Properties

As mentioned before, the game Γ is nonconvex and thus we need a different approach from the standard existence theory. For this, we first need a few statements which enable us to explicitly derive the best response map for each player and therefore, the best response map of the game. All proofs are given in the respective section of the Appendix. Lemma 1 ensures that for all optimal solutions of (3) the inequality constraints corresponding to (2) hold with equality.

Lemma 1. *For all optimal solutions $(\bar{x}_k, \bar{T}_k)$ of (3) we have*

$$(1 - \bar{x}_k)\frac{L_k}{f_k} = \bar{T}_k.$$

If $\bar{x}_k > 0$, then it additionally holds that

$$\sum_{l \neq k} x_l O_k + \bar{x}_k O_k + C = \bar{T}_k.$$

With this knowledge, we can easily derive the best response map for MU k for a given z_{-k}. Although the best response function is defined partially, it is continuous. The condition for purely local computation basically says that MU k can compute his whole task locally at least as fast as the time needed by all other users for transmissions to the AP and the following computations at the cloudlet.

Lemma 2. *The best response map for MU k, i.e. the global solution of problem (3) for a given z_{-k}, is given by*

$$S_k\left(z_{-k}\right) = \begin{cases} \left(0, \frac{L_k}{f_k}\right), & \text{if } \sum_{l \neq k} x_l O_l + C \geq \frac{L_k}{f_k}, \\ \left(\underbrace{\frac{\frac{L_k}{f_k} - \sum_{l \neq k} x_l O_l + C}{\frac{L_k}{f_k} + O_k}}_{=:\bar{x}_k}, (1 - \bar{x}_k)\frac{L_k}{f_k} \right), & \text{else}, \end{cases} \tag{4}$$

where $S_k\left(z_{-k}\right) = S_k\left(x_{-k}\right)$ is single valued and continuous.

Since we have shown that there exists a continuous best response map for each user k, the existence of a generalized Nash equilibrium is not difficult to prove.

Theorem 1. *The nonconvex generalized Nash game Γ has a generalized Nash equilibrium.*

Standard algorithms for finding a generalized Nash equilibrium, like Gauss-Seidel (see [15]), would require an enormous amount of signaling between all users, which is not desired. In the following we present a closed form for the generalized Nash equilibrium. Note that this form still depends on the optimal set $\mathcal{A}$ of users who offload a fraction of their task, whose existence and uniqueness is discussed later.

Theorem 2. *Every generalized Nash equilibrium $(\bar{x}, \bar{T})$ of the game Γ is of the form*

$$\bar{x}_k = \max\left\{1 - \frac{\frac{f_k}{L_k}\left[C + \sum_{k \in \mathcal{A}} O_l\right]}{1 + \sum_{k \in \mathcal{A}} O_l \frac{f_l}{L_l}}\ ,\ 0\right\}$$

and $\bar{T}_k = (1 - \bar{x}_k)\frac{L_k}{f_k}$ for all $k = 1, \ldots, K$, where

$$\mathcal{A} = \left\{k \mid \frac{L_k}{f_k} > \frac{C + \sum_{l \in \mathcal{A}} O_l}{1 + \sum_{l \in \mathcal{A}} O_l \frac{f_l}{L_l}}\right\}.$$

The set $\mathcal{A}$ of active MUs in Theorem 2 is defined in an implicit way, which is not desirable from an algorithmic point of view. A solution for this can be achieved by reordering the MUs according to their computation time without offloading, which can be interpreted as the "need to offload":

$$\frac{L_1}{f_1} \geq \frac{L_1}{f_2} \geq \ldots \geq \frac{L_K}{f_K}. \tag{5}$$

In addition to providing an explicit formulation, we can thus show that the set $\mathcal{A}$ and the generalized Nash equilibrium is unique.

Theorem 3. *There is exactly one set $\mathcal{A}$, such that*

$$\mathcal{A} = \left\{ k \mid \frac{L_k}{f_k} > \frac{C + \sum_{l\in\mathcal{A}} O_l}{1 + \sum_{l\in\mathcal{A}} O_l \frac{f_l}{L_l}} \right\}.$$

In the case of $\frac{L_1}{f_1} \geq \ldots \geq \frac{L_K}{f_K}$ it is of the form

$$\bar{\mathcal{A}} = \left\{ k \mid \frac{L_k}{f_k} > \frac{C + \sum_{l=1}^{k} O_l}{1 + \sum_{l=1}^{k} O_l \frac{f_l}{L_l}} \right\}.$$

Consequently the game Γ has exactly one generalized Nash equilibrium.

As a consequence of Theorem 3 we propose Algorithm 1, which computes the set $\bar{\mathcal{A}}$ of active players and thus also the generalized Nash equilibrium of the game Γ.

Algorithm 1 Computation of a GNE for the game Γ

Require: Ordering of MUs, such that $\frac{L_1}{f_1} \geq \ldots \geq \frac{L_K}{f_K}$
  $\mathcal{A} = \emptyset$
  $k = 1$
  while $k \leq K$ **and** $\left(\frac{L_k}{f_k} > \frac{C+\sum_{i=1}^{k} O_i}{1+\sum_{i=1}^{k} O_i \frac{L_i}{f_i}} \right)$ **do**
    $\mathcal{A} \leftarrow \mathcal{A} \cup \{k\}$
    $k \leftarrow k + 1$
  end while
  Set $(\bar{x}, \bar{T})$ according to Theorem 2
  return Generalized Nash equilibrium $(\bar{x}, \bar{T})$

As stated earlier, conventional algorithms to derive GNEs require nearly all information for every MU and a lot of signaling. This can be circumvented by Algorithm 1 due to the closed form solution. It can easily be computed at the cloudlet (or AP), returning for each user the amount of offloading that he should perform, i.e. the information only needs to be present at the central entity. The question that arises is whether a completely centralized solution without individual decisions (see for example [6], [7]) differs from a game theoretic approach. This will be answered in the next subsection.

B. Comparison to a Centralized Solution

We compare our game theoretic approach to a centralized one with the respective model. In a centralized offloading scenario the cloudlet strives to minimize the overall completion time which includes the completion times of all MUs and of the cloudlet itself. This can be formulated as the following optimization problem:

$$\begin{aligned} \min_{\{x_k\},T} \quad & T, \\ \text{s.t.} \quad & (1 - x_k)\frac{L_k}{f_k} \leq T, \qquad \forall k \\ & \sum_{k=1}^{K} x_k O_k + C \leq T, \\ & x_k \in [0, 1], \qquad \forall k \\ & T \geq 0, \end{aligned} \tag{6}$$

where T is the overall completion time. For (6) we can also prove the corresponding counterpart to Lemma 1:

Lemma 3. *For each optimal solution $(\hat{x}, \hat{T})$ of (6) ,where $\hat{x} = (\hat{x}_k)_{k=1}^{K}$, holds*

$$\sum_{k=1}^{K} \hat{x}_k O_k + C = \hat{T}$$

and for all k with $\hat{x}_k > 0$ holds

$$(1 - \hat{x}_k)\frac{L_k}{f_k} = \hat{T}.$$

Following from this one can prove the following (compare Theorem 2 for the game):

Theorem 4. *Every optimal solution $(\hat{x}, \hat{T})$ of (6) is of the form*

$$\hat{x}_k = 1 - \frac{\frac{f_k}{L_k}\left[C + \sum_{k\in\mathcal{A}} O_l\right]}{1 + \sum_{k\in\mathcal{A}} O_l \frac{f_l}{L_l}}$$

for all $k \in \mathcal{A}$ and $\hat{x}_k = 0$ else, where

$$\mathcal{A} = \left\{ k \mid \frac{L_k}{f_k} > \frac{C + \sum_{l\in\mathcal{A}} O_l}{1 + \sum_{l\in\mathcal{A}} O_l \frac{f_l}{L_l}} \right\}.$$

With this we can connect the unique GNE of the game Γ with the solution of (6):

Theorem 5. *The optimal solution $(\hat{x}, \hat{T})$ of* (6) *and the unique GNE $(\bar{x}_k, \bar{T}_k)_{k=1}^{K}$ of the game Γ coincide in the sense that*

$$\hat{x}_k = \bar{x}_k \quad \forall k, \quad \text{and} \quad \hat{T} = \max\{C, \bar{T}_1, \ldots, \bar{T}_K\}.$$

For $\mathcal{A} = \emptyset$ due to a large $C > 0$ the computation time $\hat{T}$ of the central optimization problem can be larger than the computation time of each MU. Then the computation time for each MU k for the centralized optimization is equal to $\frac{L_k}{f_k}$, since $\hat{x}_k = 0$ for all k. Thus the computation time for the MUs in the centralized optimization can be described with

$$\hat{T}_k^{\text{central}} := (1 - \hat{x}_k)\frac{L_k}{f_k}$$

due to Lemma 3. With this it follows immediately that the price of anarchy (see [16]), which displays the ratio of the worst possible Nash equilibrium and the social optimum to measure the effectiveness of the system, is given by

$$\text{PoA} = \frac{\max_k \bar{T}_k}{\max_k \hat{T}_k^{\text{central}}} = 1,$$

i.e. there is no loss for the system. The proofs for Lemma 3 and Theorem 4 are similar to the proofs of Lemma 1 and Theorem 2, respectively and are omitted. The proof of Theorem 5 is straight forward.

IV. Numerical Results

For the simulation scenario we place K mobile users randomly distributed in a circle around the access point with the distance uniformly chosen in $[10, 100]$ m. Every MU k is assumed to have a single-core processor handling a single task at once and its CPU frequency f_k to be uniformly chosen from the set of possible frequencies $\{0.8, 0.9, 1.0, 1.1, 1.2\}$ GHz. Each MU has a single task of size $W_k = 10$ MB that has to be computed locally or is (partially) offloaded. Computation of one bit of the task requires 1000 CPU cycles at the MU's or cloudlet's processor. The server connected to the AP is more powerful than the MU devices and is therefore assumed to have a single core computation capability $f_c = 5$ GHz. At first we assume that the queue of the cloudlet is empty which results in $C = 0$.

The channel coefficient h_k from the AP to each MU k is modeled by $1/d_k^\alpha$ with $\alpha = 3$, where d_k is the normalized distance in the range of $[0, 1]$. The SNR at the transmitter is chosen to be 20 dB and the overall available bandwidth is $B = 10$ MHz. Every data point in Figures 1 and 2 consists of 1000 Monte Carlo runs, where the MUs are randomly placed every run.

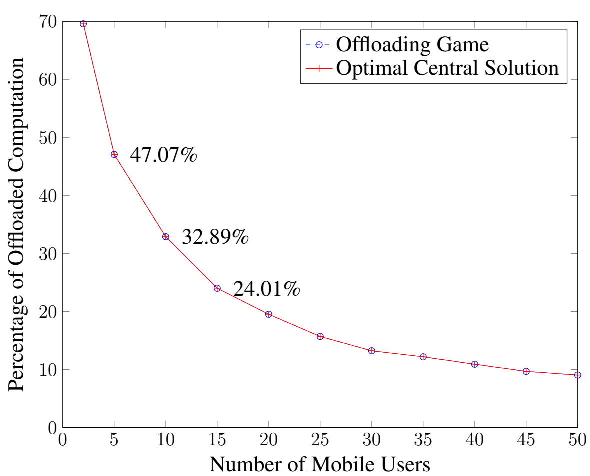

Fig. 1. Average percentage of offloaded data per user $\left(\frac{\sum x_k}{K}\right)$ for different numbers of mobile users K

Figure 1 illustrates the dependency of the scenario on the number of MUs connected to the AP. Due to the random placement of MUs around the AP, the average percentage of offloaded data per MU is shown. As the number of connected users increases, the desire to offload fractions of the task falls rapidly. While in our simulated system with 5 MUs each user offloads on average 47% of his task, in a system with 15 MUs it is only half of the amount at an average of 24%. In Figure 1 it is also visible that the game theoretic formulation solved using Algorithm 1 and the central optimization problem (6) lead to the exact same results and thus a price of anarchy of one as proven in Subsection III-B.

In Figure 2 the computation time at the cloudlet (= maximum computation time for all MUs since $C = 0$) is shown versus the number of CPU cycles required to successfully compute one bit of the task $(\frac{L}{W})$ for $K = 10$ users. While the task size is still fixed at $W_k = 10$ MB, the computational complexity of the tasks is varying. Algorithm 1 is again used to compute the generalized Nash equilibrium of the offloading game. The simulation shows that when the task becomes more complex and more CPU cycles are required, the gap between a system employing computation offloading and a system with full local computation will strongly increase.

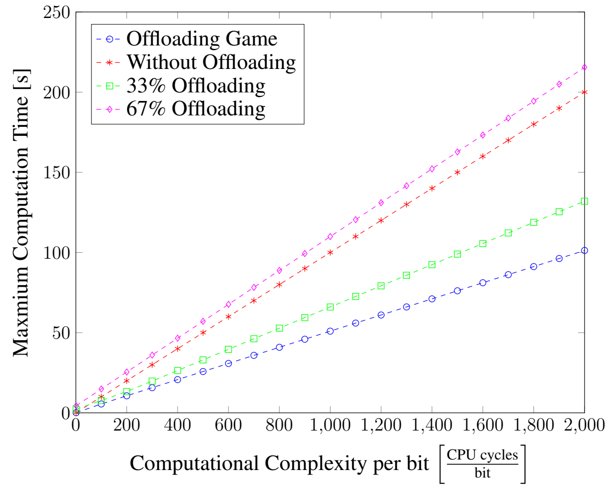

Fig. 2. Maximum computation time for different computational complexities

The additional two curves in Figure 2 represent reference scenarios, where each MU connected to the access point offloads a fixed percentage of his task. The value of 33% was chosen according to the result shown in Figure 1 that in a system with 10 MUs each user offloads on average nearly 33% of his task. While the 33% curve shows a significant decrease of the maximum computation time compared to a system without offloading, the offloading game leads to an even better solution. For the highest simulated value of 2000 CPU cycles per bit, the optimal game solution reduces the computation time of the system to 102 seconds compared to 132 seconds with the fixed amount of 33% offloading. The curve for 67% of offloaded data is included to demonstrate that offloading is not always beneficial due to the limited computation resources at the cloudlet. In this case the cloudlet is overburdened which results in an increase of the maximum computation time compared to the scenario where nothing is offloaded. In a scenario with more MUs, bigger task sizes or higher complexity even smaller amounts of fixed offloading could lead to an overload of the cloudlet and to increased computation times, as visualized (for increased number of MUs) in Figure 1.

A dynamic scheme for computation offloading is exemplary demonstrated in Figure 3. Every 20 seconds a session starts, where MUs are able to offload parts of their computation to the nearby cloudlet. A MU can only offload parts of his task, if he did not participate in the previous session or if he finished all his previous computation, i.e. he cannot offload several tasks at once. For each session a random

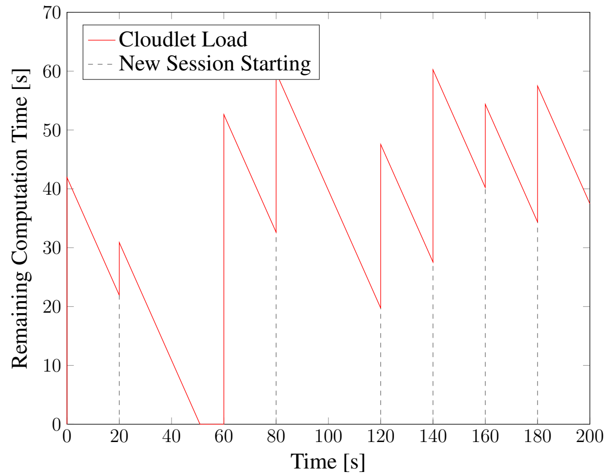

Fig. 3. Exemplary timeline for multiple offloading sessions at the cloudlet

number of MUs uniformly chosen in $\{0, \ldots, 10\}$ decides to compute a task. The remaining simulation parameters are equal to the described model at the beginning of Section IV. If computations from the previous session are still executed while MUs want to offload in a new session, the remaining time C (height of the gray dotted lines in Figure 3) is be added to the computation time of the new session at the cloudlet. Due to this the cloudlet is able to handle simple queues, whereas the offloading decisions of the MUs are influenced by the remaining computation time C on the cloudlet.

V. Conclusion

We proposed a game theoretic formulation for the presented computation offloading model for a single access point with attached cloudlet that serves multiple mobile users. In the following, we proved the existence of the unique generalized Nash equilibrium, which can also be derived by a simple algorithm. As this algorithm can be computed at the central entity, the amount of needed signaling can be limited as well as the information sharing between the mobile users, here we also proved that the price of anarchy of the game theoretic formulation is one.

We showed that our formulation could be used for a dynamic offloading scheme and that offloading is an efficient method to reduce the total computation time of the system. Additionally, we showed that offloading performance will decrease if the number of MUs exceeds a critical number.

In future work we plan to extend the model and investigate, whether an increase in the price of anarchy will occur. Additionally, a reasonable extension could be a local algorithm running on the mobile devices with limited available information about the other players connected to the same access point.

Acknowledgment

This work is supported by the 'Excellence Initiative' of the German Federal and State Governments and the Graduate School of Computational Engineering at Technische Universität Darmstadt.

The work of Tobias Mahn, Hussein Al Shatri, and Anja Klein has been performed in the context of the DFG Collaborative Research Center (CRC) 1053 MAKI - subproject B03.

References

[1] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "Mobile edge computing: Survey and research outlook," *CoRR*, vol. abs/1701.01090, 2017.

[2] K. Kumar and Y.-H. Lu, "Cloud computing for mobile users: Can offloading computation save energy?" *Computer*, vol. 43, no. 4, pp. 51–56, 2010.

[3] D. Ferreira, A. K. Dey, and V. Kostakos, "Understanding human-smartphone concerns: a study of battery life," in *International Conference on Pervasive Computing*. Springer, 2011, pp. 19–33.

[4] S. Barbarossa, S. Sardellitti, and P. Di Lorenzo, "Communicating while computing: Distributed mobile cloud computing over 5G heterogeneous networks," *IEEE Signal Processing Magazine*, vol. 31, no. 6, pp. 45–55, 2014.

[5] Y. C. Hu, M. Patel, D. Sabella, N. Sprecher, and V. Young, "Mobile edge computing - a key technology towards 5g," *ETSI White Paper*, vol. 11, 2015.

[6] Q. H. Le, H. Al-Shatri, and A. Klein, "Optimal joint power allocation and task splitting in wireless distributed computing," in *International ITG Conference on Systems, Communication and Coding*, 2016.

[7] M.-H. Chen, B. Liang, and M. Dong, "Joint offloading decision and resource allocation for multi-user multi-task mobile cloud," in *IEEE International Conference on Communications (ICC)*. IEEE, 2016, pp. 1–6.

[8] Y. Ge, Y. Zhang, Q. Qiu, and Y.-H. Lu, "A game theoretic resource allocation for overall energy minimization in mobile cloud computing system," in *ACM/IEEE International Symposium on Low Power Electronics and Design*, ser. ISLPED '12, 2012, pp. 279–284.

[9] X. Chen, "Decentralized computation offloading game for mobile cloud computing," *IEEE Transactions on Parallel and Distributed Systems*, vol. 26, no. 4, pp. 974–983, 2015.

[10] M.-H. Chen, M. Dong, and B. Liang, "Multi-user mobile cloud offloading game with computing access point," in *5th IEEE International Conference on Cloud Networking (Cloudnet)*. IEEE, 2016, pp. 64–69.

[11] V. Cardellini, V. D. N. Personé, V. Di Valerio, F. Facchinei, V. Grassi, F. L. Presti, and V. Piccialli, "A game-theoretic approach to computation offloading in mobile cloud computing," *Mathematical Programming*, vol. 157, no. 2, pp. 421–449, 2016.

[12] H. Al-Shatri, S. Müller, and A. Klein, "Distributed algorithm for energy efficient multi-hop computation offloading," in *IEEE International Conference on Communications (ICC)*. IEEE, 2016, pp. 1–6.

[13] W. Achtziger and C. Kanzow, "Mathematical programs with vanishing constraints: optimality conditions and constraint qualifications," *Mathematical Programming*, vol. 114, no. 1, pp. 69–99, 2008.

[14] T. Hoheisel, "Mathematical programs with vanishing constraints," Ph.D. dissertation, Julius-Maximilians-Universität Würzburg, 2009.

[15] F. Facchinei and C. Kanzow, "Generalized nash equilibrium problems," *Annals of Operations Research*, vol. 175, no. 1, pp. 177–211, 2010.

[16] E. Koutsoupias and C. Papadimitriou, "Worst-case equilibria," *Comput. Sci. Rev.*, vol. 3, no. 2, pp. 65–69, 2009. [Online]. Available: http://dx.doi.org/10.1016/j.cosrev.2009.04.003

[17] J. Sherman and W. J. Morrison, "Adjustment of an inverse matrix corresponding to a change in one element of a given matrix," *The Annals of Mathematical Statistics*, vol. 21, no. 1, pp. 124–127, 03 1950.

[18] M. Bartlett, "An inverse matrix adjustment arising in discriminant analysis." *The Annals of Mathematical Statistics*, vol. 22, pp. 107–111, 1951.

Appendix

Proof of Lemma 1

Case 1: $\bar{x}_k = 0$:
Then every $T_k \geq \frac{L_k}{f_k}$ is feasible and thus the minimal value is attained at

$$\bar{T}_k = \frac{L_k}{f_k} = (1 - \bar{x}_k)\frac{L_k}{f_k}.$$

Case 2: $\bar{x}_k > 0$:
First note that
$$\sum_{l\neq k} x_l O_l + C \leq \frac{L_k}{f_k} \tag{7}$$
holds. Otherwise the solution $(0, \frac{L_k}{f_k})$ strictly dominates $(\bar{x}_k, \bar{T}_k)$, which would be a contradiction to the optimality. Now choose $\hat{x}_k$ as the solution of
$$(1-\hat{x}_k)\frac{L_k}{f_k} = \sum_{l\neq k} x_l O_l + \hat{x}_k O_k + C.$$
It follows that
$$\hat{x}_k = \frac{\frac{L_k}{f_k} - \sum_{l\neq k} x_l O_l - C}{\frac{L_k}{f_k} + O_k}.$$
Note that $0 \leq \hat{x}_k \leq 1$ due to (7). Then $(\hat{x}_k, \hat{T}_k)$ with $\hat{T}_k = (1-\hat{x}_k)\frac{L_k}{f_k}$ is feasible.

Assume for contradiction $(1-\bar{x}_k)\frac{L_k}{f_k} < \bar{T}_k$. Then we obtain
$$\begin{aligned}\hat{T}_k &= (1-\hat{x}_k)\frac{L_k}{f_k} = \left(1 - \frac{\frac{L_k}{f_k} - \sum_{l\neq k} x_l O_l - C}{\frac{L_k}{f_k} + O_k}\right)\frac{L_k}{f_k} \\ &= \left(\frac{O_k + \sum_{l\neq k} x_l O_l + C}{\frac{L_k}{f_k} + O_k}\right)\frac{L_k}{f_k} \\ &\leq \left(\frac{O_k + \bar{T}_k - \bar{x}_k O_k}{\frac{L_k}{f_k} + O_k}\right)\frac{L_k}{f_k} \\ &= \frac{\bar{T}_k \frac{L_k}{f_k} + (1-\bar{x}_k)\frac{L_k}{f_k} O_k}{\frac{L_k}{f_k} + O_k} < \frac{\bar{T}_k \frac{L_k}{f_k} + \bar{T}_k O_k}{\frac{L_k}{f_k} + O_k} = \bar{T}_k,\end{aligned}$$
a contradiction to the optimality of $(\bar{x}_k, \bar{T}_k)$. It follows
$$(1-\bar{x}_k)\frac{L_k}{f_k} = \bar{T}_k$$
for all optimal solutions.

Now assume $\sum_{l\neq k} x_l O_k + \bar{x}_k O_k + C < \bar{T}_k$. Then we have
$$\begin{aligned}\hat{x}_k &= \frac{\frac{L_k}{f_k} - \sum_{l\neq k} x_l O_l - C}{\frac{L_k}{f_k} + O_k} > \frac{\frac{L_k}{f_k} - \bar{T}_k + \bar{x}_k O_k}{\frac{L_k}{f_k} + O_k} \\ &= \frac{\frac{L_k}{f_k} - (1-\bar{x}_k)\frac{L_k}{f_k} + \bar{x}_k O_k}{\frac{L_k}{f_k} + O_k} = \bar{x}_k.\end{aligned}$$
Thus it follows that
$$\hat{T}_k = (1-\hat{x}_k)\frac{L_k}{f_k} < (1-\bar{x}_k)\frac{L_k}{f_k} = \bar{T}_k,$$
which is a contradiction to the optimality of $(\bar{x}_k, \bar{T}_k)$. ■

Proof of Lemma 2

For $\sum_{l\neq k} x_l O_l + C \geq \frac{L_k}{f_k}$ all feasible strategies with $x_k > 0$ are strictly dominated by $(\bar{x}_k, \bar{T}_k) = (0, \frac{L_k}{f_k})$.

For $\sum_{l\neq k} x_l O_l + C < \frac{L_k}{f_k}$ we know from Lemma 1 that every optimal solution $(\bar{x}_k, \bar{T}_k)$ with $\bar{x}_k > 0$ satisfies
$$(1-\bar{x}_k)\frac{L_k}{f_k} = \bar{T}_k = \sum_{l\neq k} x_l O_l + \bar{x}_k O_k + C,$$
hence the unique solution is
$$\begin{aligned}\bar{x}_k &= \frac{\frac{L_k}{f_k} - \sum_{l\neq k} x_l O_l - C}{\frac{L_k}{f_k} + O_k} \quad \text{and} \quad (8)\\ \bar{T}_k &= (1-\bar{x}_k)\frac{L_k}{f_k} = \frac{O_k + \sum_{l\neq k} x_l O_l + C}{\frac{L_k}{f_k} + O_k}\frac{L_k}{f_k}. \quad (9)\end{aligned}$$
Since $S_k(z_{-k})$ is piecewise continuous and continuous at $\sum_{l\neq k} x_l O_l + C = \frac{L_k}{f_k}$, it is continuous in general. ■

Proof of Theorem 1

We define the best response map $S(z)$ of the game by
$$S(z) = S_1(z_{-1}) \times \ldots \times S_K(z_{-K}),$$
where $z = (x_k, T_k)_{k=1}^K$. From Lemma 2 we know that $S(z)$ is continuous and that $S: Z \to Z$, where $Z = [0,1]^K \times [0, \bar{T}]^K$ is convex and compact. With Brouwer's fixed point theorem it follows that $S(z)$ has at least one fixed point on Z, which is a generalized Nash equilibrium of the game. ■

Proof of Theorem 2

Let $(\bar{x}, \bar{T})$ be a generalized Nash equilibrium of the game Γ and define
$$\mathcal{A} := \{k \mid \bar{x}_k > 0\}.$$
In case $\mathcal{A} = \emptyset$ we have $\bar{x}_k = 0$ for all MUs k and thus the formulae are correct due to Lemma 2. Now consider $\mathcal{A} \neq \emptyset$. Without loss of generality reorder the MUs, such that $\mathcal{A} = \{1, \ldots, n\}$ with $n = |\mathcal{A}|$. We know that each MU $k \in \mathcal{A}$ has $\bar{x}_k > 0$ and therefore
$$\begin{aligned}&(1-\bar{x}_k)\frac{L_k}{f_k} = \bar{T}_k = \sum_{l\neq k} \bar{x}_l O_k + \bar{x}_k O_k + C,\\ \Leftrightarrow \quad &\sum_{l\in\mathcal{A}\setminus\{k\}} \bar{x}_l O_l + \bar{x}_k\left(O_k + \frac{L_k}{f_k}\right) = \frac{L_k}{f_k} - C,\end{aligned}$$
for all $k \in \mathcal{A}$. Thus, all $\bar{x}_k$ for $k \in \mathcal{A}$ are given by the solution of the $n \times n$ linear system of equalities
$$\begin{pmatrix} O_1 + \frac{L_1}{f_1} & O_2 & \cdots & O_n \\ O_1 & O_2 + \frac{L_2}{f_2} & & \vdots \\ \vdots & & \ddots & \\ O_n & \cdots & & O_n + \frac{L_n}{f_n}\end{pmatrix} \underbrace{\begin{pmatrix}\bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_n\end{pmatrix}}_{=:\bar{x}} = \underbrace{\begin{pmatrix}\frac{L_1}{f_1} - C \\ \frac{L_2}{f_2} - C \\ \vdots \\ \frac{L_n}{f_n} - C\end{pmatrix}}_{=:b},$$
which is equivalent to
$$\left(\underbrace{\operatorname{diag}\left(\frac{L_k}{f_k}\right)}_{=:A} + e \cdot \underbrace{\begin{pmatrix}O_1 \\ \vdots \\ O_n\end{pmatrix}^T}_{=:v^T}\right)\bar{x} = b,$$

where e is a n-dimensional column vector of all ones. Since A is a nonsingular matrix and $v^T A^{-1} e \neq -1$ we can compute $\bar{x}$ using the Sherman-Morrison formula (see [17], [18]):

$$\begin{aligned}\bar{x} &= \left(A^{-1} - \frac{A^{-1} e v^T A^{-1}}{1 + v^T A^{-1} e}\right) b \\ &= \left(\operatorname{diag}\left(\frac{f_k}{L_k}\right) - \frac{\operatorname{diag}\left(\frac{f_k}{L_k}\right) e v^T \operatorname{diag}\left(\frac{f_k}{L_k}\right)}{1 + v^T \operatorname{diag}\left(\frac{f_k}{L_k}\right) e}\right) b \\ &= e - \left(\frac{f_l}{L_l}\right)_{l \in \mathcal{A}} \left[C + \frac{\sum_{l \in \mathcal{A}} O_l - \sum_{l \in \mathcal{A}} O_l \frac{f_l}{L_l} C}{1 + \sum_{l \in \mathcal{A}} O_l \frac{f_l}{L_l}}\right].\end{aligned}$$

This means, for each $k \in \mathcal{A}$ it holds that

$$\begin{aligned}\bar{x}_k &= 1 - \frac{f_k}{L_k} \left[C + \frac{\sum_{l \in \mathcal{A}} O_l - \sum_{l \in \mathcal{A}} O_l \frac{f_l}{L_l} C}{1 + \sum_{l \in \mathcal{A}} O_l \frac{f_l}{L_l}}\right] \\ &= 1 - \frac{\frac{f_k}{L_k} \left[C + \sum_{l \in \mathcal{A}} O_l\right]}{1 + \sum_{l \in \mathcal{A}} O_l \frac{f_l}{L_l}}.\end{aligned}$$

Lemma 1 yields the representation of $\bar{T}_k$. Additionally this formula implies for all $k \in \mathcal{A}$

$$0 < \bar{x}_k = 1 - \frac{\frac{f_k}{L_k} \left[C + \sum_{l \in \mathcal{A}} O_l\right]}{1 + \sum_{l \in \mathcal{A}} O_l \frac{f_l}{L_l}} \Leftrightarrow \frac{L_k}{f_k} > \frac{C + \sum_{l \in \mathcal{A}} O_l}{1 + \sum_{l \in \mathcal{A}} O_l \frac{f_l}{L_l}}.$$

With Lemma 1 it follows that for all $k \notin \mathcal{A}$

$$\frac{L_k}{f_k} \leq \sum_{l=1}^{K} \bar{x}_l O_l + C = \sum_{l \in \mathcal{A}} \bar{x}_l O_l + C$$

from which we can conclude

$$\begin{aligned}\frac{L_k}{f_k} &\leq \sum_{l \in \mathcal{A}} O_l \left(1 - \frac{\frac{f_l}{L_l} \left[C + \sum_{i \in \mathcal{A}} O_i\right]}{1 + \sum_{i \in \mathcal{A}} O_i \frac{f_i}{L_i}}\right) + C \\ &= \sum_{l \in \mathcal{A}} O_l - \frac{\sum_{l \in \mathcal{A}} O_l \frac{f_l}{L_l} \left[C + \sum_{i \in \mathcal{A}} O_i\right]}{1 + \sum_{i \in \mathcal{A}} O_i \frac{f_i}{L_i}} \\ &\quad + \frac{C \left[1 + \sum_{i \in \mathcal{A}} O_i \frac{f_i}{L_i}\right]}{1 + \sum_{i \in \mathcal{A}} O_i \frac{f_i}{L_i}} \\ &= \frac{\sum_{l \in \mathcal{A}} O_l \left[1 + \sum_{i \in \mathcal{A}} O_i \frac{f_i}{L_i}\right]}{1 + \sum_{i \in \mathcal{A}} O_i \frac{f_i}{L_i}} \\ &\quad - \frac{\sum_{l \in \mathcal{A}} O_l \frac{f_l}{L_l} \sum_{i \in \mathcal{A}} O_i + C}{1 + \sum_{i \in \mathcal{A}} O_i \frac{f_i}{L_i}} \\ &= \frac{\sum_{l \in \mathcal{A}} O_l + C}{1 + \sum_{l \in \mathcal{A}} O_l \frac{f_l}{L_l}}.\end{aligned}$$

Together this yields

$$\mathcal{A} = \left\{k \mid \frac{L_k}{f_k} > \frac{C + \sum_{l \in \mathcal{A}} O_l}{1 + \sum_{l \in \mathcal{A}} O_l \frac{f_l}{L_l}}\right\}$$

as well as

$$\bar{x}_k = \max\left\{1 - \frac{\frac{f_k}{L_k} \left[C + \sum_{k \in \mathcal{A}} O_l\right]}{1 + \sum_{k \in \mathcal{A}} O_l \frac{f_l}{L_l}}\ ,\ 0\right\}$$

for all $k = 1, \ldots, K$. ■

Proof of Theorem 3

W.l.o.g. reorder the MUs such that (5) holds.

First note, that at least one such set $\mathcal{A}$ always exists, since we know of the existence of a GNE and its explicit formulation from Theorem 2.

Due to the ordering of the MUs we know $k \in \mathcal{A} \Rightarrow k-1 \in \mathcal{A}$ and thus $\mathcal{A}$ is of the form $\{1, \ldots, n\}$ for some $n \leq K$. Note that the case $n = 0$ is possible, in which $\mathcal{A} = \emptyset$. Let $\mathcal{B} = \{1, \ldots, m\}$ be another such set where we assume without loss of generality $m > n$. Then $n+1, \ldots, m \notin \mathcal{A}$ and thus

$$\begin{aligned}&\frac{L_{n+1}}{f_{n+1}} \leq \frac{C + \sum_{l=1}^{n} O_l}{1 + \sum_{l=1}^{n} O_l \frac{f_l}{L_l}} \\ \Rightarrow\ &\frac{L_{n+1}}{f_{n+1}} \left(1 + \sum_{l=1}^{n} O_l \frac{f_l}{L_l} + O_{n+1} \frac{f_{n+1}}{L_{n+1}}\right) \leq C + \sum_{l=1}^{n+1} O_l \\ \Rightarrow\ &\frac{L_{n+1}}{f_{n+1}} \leq \frac{C + \sum_{l=1}^{n+1} O_l}{1 + \sum_{l=1}^{n+1} O_l \frac{f_l}{L_l}}.\end{aligned}$$

Since $\frac{L_{n+2}}{f_{n+2}} \leq \frac{L_{n+1}}{f_{n+1}}$ we can repeat this process and derive

$$\frac{L_m}{f_m} \leq \frac{C + \sum_{l=1}^{m} O_l}{1 + \sum_{l=1}^{m} O_l \frac{f_l}{L_l}} = \frac{C + \sum_{l \in \mathcal{B}} O_l}{1 + \sum_{l \in \mathcal{B}} O_l \frac{f_l}{L_l}},$$

which is a contradiction to $m \in \mathcal{B}$. Consequently there is only one such set $\mathcal{A}$ and since all Nash equilibria are given by Theorem 2, the Nash equilibrium of game Γ is unique.

Next we show that the unique set $\mathcal{A} = \{1, \ldots, n\}$ is given by $\bar{\mathcal{A}}$. To show this we first prove $k \in \bar{\mathcal{A}}$ implies $k-1 \in \bar{\mathcal{A}}$ for $k > 1$.

$$\begin{aligned}k \in \bar{\mathcal{A}} \Rightarrow\ &\frac{L_k}{f_k} > \frac{C + \sum_{l=1}^{k} O_l}{1 + \sum_{l=1}^{k} O_l \frac{f_l}{L_l}} \\ \Rightarrow\ &\frac{L_k}{f_k} \left(1 + \sum_{l=1}^{k-1} O_l \frac{f_l}{L_l}\right) + O_k > C + \sum_{l=1}^{k-1} O_l + O_k \\ \Rightarrow\ &\frac{L_{k-1}}{f_{k-1}} \geq \frac{L_k}{f_k} > \frac{C + \sum_{l=1}^{k-1} O_l}{1 + \sum_{l=1}^{k-1} O_l \frac{f_l}{L_l}} \Rightarrow k-1 \in \bar{\mathcal{A}}.\end{aligned}$$

Now we know that $\bar{\mathcal{A}}$ is of the form $\bar{\mathcal{A}} = \{1, \ldots, m\}$ for some m. It remains to show $n = m$. This is done by verifying $n \in \bar{\mathcal{A}}$, which follows directly from the formula for $\mathcal{A}, \bar{\mathcal{A}}$, and $n+1 \notin \bar{\mathcal{A}}$, which follows from

$$\begin{aligned}n+1 \notin \mathcal{A} \Leftrightarrow\ &\frac{L_{n+1}}{f_{n+1}} \leq \frac{C + \sum_{l=1}^{n} O_l}{1 + \sum_{l=1}^{n} O_l \frac{f_l}{L_l}} \\ \Leftrightarrow\ &\frac{L_{n+1}}{f_{n+1}} \left(1 + \sum_{l=1}^{n} O_l \frac{f_l}{L_l}\right) + O_{n+1} \leq C + \sum_{l=1}^{n+1} O_l \\ \Leftrightarrow\ &\frac{L_{n+1}}{f_{n+1}} \left(1 + \sum_{l=1}^{n+1} O_l \frac{f_l}{L_l}\right) \leq C + \sum_{l=1}^{n+1} O_l \\ \Leftrightarrow\ &\frac{L_{n+1}}{f_{n+1}} \leq \frac{C + \sum_{l=1}^{n+1} O_l}{1 + \sum_{l=1}^{n+1} O_l \frac{f_l}{L_l}} \quad \Leftrightarrow n+1 \notin \bar{\mathcal{A}}.\end{aligned}$$

With this we see that $\bar{\mathcal{A}}$ is the desired set. ■