A Distributed Offloading Market for 5G Heterogeneous Networks

Endre H. Hjort Kure[1,2*], Sabita Maharjan[1,2], Stein Gjessing[1,2], and Yan Zhang[1,2]

[1]Simula Research Laboratory, Norway
[2]Department of Informatics, University of Oslo, Norway
[*]Corresponding author: endre.hjort.kure@simula.no

***Abstract*—Concerns have been raised regarding the economical viability for each operator to have a full regional 5G coverage. A possible solution is to have traffic offloaded to competitors. In this work we present a new scheme for optimal offloading in a stochastic environment. This is more in line with the conditions 5G base stations will face with changing link and traffic conditions. The problem is formulated as a Stackelberg game, and the players' utility functions are derived though queuing models. Numerical results illustrate that our scheme provides a global optimal resource allocation up to a threshold. The threshold is a function of the traffic load and the number of offloading candidates. Beyond the threshold players still have incentives to participate, but the market equilibrium is not globally optimal.**

I. Introduction

Mobile traffic is expected to have a compounded annual growth of 47% until 2021 [1] with the majority being downlink traffic. This forces the industry to identify ways to increase both energy efficiency and end users' quality of experience (QoE).

Traffic offloading addresses this issue by using radio access networks (RANs) that are closer to the end user or that are less congested, to carry the traffic. Therefore, it increases energy efficiency, as a RAN closer to the user uses less transmission power for the same information. Network delay is an important element of both QoE and service level agreement (SLA), and therefore a parameter the operators need to control in order to satisfy the customers' expectations. Offloading may therefore reduce delay as RANs closer to the UE[1] can have better link conditions and be less congested than a larger access point, serving a larger number of UEs. The two major RAN technologies are WiFi (IEEE 801.11) and LTE, with 5G expected to become a third major technology.

Researchers have already raised concerns that a fully developed 5G RAN could be too costly for each operator to develop and maintain on their own and suggest infrastructure sharing as a viable option [2]. Moreover, sharing of small cells across operators' domains is technically feasible [3]. A typical scenario is that a subset of operators deploy the 5G network in a region, while the remaining operators rent capacity from it. Similar solutions are emerging from companies such as Cloudberry[2] which is renting out small cells in indoor environment that operators can use. We investigate a scenario with 5G small cells base stations named Access Points (AP) forming a RAN. Although, we focus on offloading from LTE to 5G as an example, our results also apply to offloading to the other types of RAN technologies such as LTE and WiFi.

Marshal and Meo [4] were one of the first to identify the benefits of offloading, focusing on the total energy saving potential. However, RANs are often owned by different entities, and incentive structures are needed for cooperation. Aram et al. [5] investigated a similar problem, focusing on the benefits of pooling resources. However, they used a cooperative game, that may result in organization structures, such as cartels that have limited applicability. Many researches have investigated traffic offloading using non-cooperative game models, which respect the need for all involved RANs to be incentivized without the dangers of creating a cartel. Poularakis et al. [6] proposed an architecture for a single operator, offering both mobile broadband and WiFi, to lease back WiFi capacity from its users in the form of a Stackelberg game. The aim was to optimize pre-cached access of files for the UEs. Shah-Mansouri et al. [7] investigated interactions between a macro base station and two third party networks offering non-overlapping access points, where one was price setting and the other price taking. A similar setup was investigated by Li et al. [8] but with the third party having overlapping networks. Assuming fixed timeslots, both Wang et al. [9] and Gao et al. [10] extended the problem to a multi-leader and multi-follower Stackelberg game.

Common assumptions in the above work are deterministic radio link conditions or traffic flows. Resource use is therefore optimized for short time horizons. It mandates that the offloading schemes are re-optimized every time the number of packets or radio link conditions change. However, both aspects are stochastic, and an operator has limited control on the downlink traffic generated per UE and changes in radio link conditions. The majority of the work assumes that all information is public, which means that every player has to broadcast changes in link conditions or traffic needs. This may not be valid from a commercial viewpoint, as it weakens a player's bargaining position and raises privacy concerns. In addition, measurements have limited time validity and the broadcasting will increase traffic overhead, which may constrain the UEs' throughput.

[1]We will use UE as a collective term on end user and user equipment
[2]http://cloudberrymobile.com/en/about-us/the-company, visited 3.03.18

It is therefore necessary to develop a scheme based on the stochastic nature of traffic and radio conditions, and optimize resource use based on statistical properties. Modelled by these concerns a scheme for a 5G environment is presented in this work. We investigate the market potential and interactions between rational entities whom all require incentives to participate in the offloading market.

The paper is structured as follows; In Section II we provide the scenario along with the power and traffic models. In Section III we describe the market model and formulate a distributed algorithm to obtain the market equilibrium. Section IV we describe the globally optimum, where the overall gain is highest, but all involved entities do not necessarily have incentives for participating in the solution. The global optimum is used as a benchmark to assess the market equilibriums for numerical results given in Section V. Sensitivity analyses on the equilibrium are also presented in Section V. Finally in Section VI, the numerical results are illustrated and discussed, and the paper's conclusion is presented.

II. System model

With stochastic traffic and link conditions, neither the seller nor the buyer knows precisely the actual cost/benefit of offloading. With varying link conditions a seller may sell more capacity than what it should, resulting in congestion and delay for the UEs. Likewise, if a fixed amount of traffic flow for a UE is bought, the amount can cause delay if it is not sufficient.

Our scenario consists of a single traffic offloading seller, that controls several 5G APs that constitutes an AP RAN. Multiple LTE operators, each covering the same geographical area with a macro base station (MBS), can use the AP RAN to decrease their UEs' average delay in the network and their own power cost by offloading traffic. The AP RAN operator aims to maximize its earnings from offloading trade and charges a fixed price for providing the LTE operators access. Since the pay-off of the offloading LTE operators and the AP RAN are tightly coupled, we approach the scenario using a game theoretic framework. The LTE operators' interactions fit into the structure of a non-cooperative game. If all the operators allocate traffic to the same AP, the resource use and delay will increase at that AP, affecting all involved LTE operators. The AP RAN operator sets its price as in [7]. Thus, the hierarchical routine of the AP RAN operator's price setting fits into a Stackelberg game framework where the AP RAN operator is the leader and the followers' response is the collective optimal response of the LTE operators. A flow can only be delivered through one MBS or AP at a time. In the following sections, we will first describe the traffic model used to capture the stochastic nature of traffic and then the power model that links traffic to power usage. The market models used to model the players' interactions are presented in Section III.

A. The traffic model

Each flow is a product of the number of UEs and the stochastic downlink flow generation process per UE. To simplify the model, all UEs are assumed to generate flow requests independently, such that the number of flows to be delivered to a UE follows a Poisson process with intensity λ [11]. The intensity is constant in this work. The total flow requests per operator will also follow a Poission process with intensity $N\lambda$, where N is the number of UEs. As the file size is not known at the beginning of the downlink session and the link conditions are stochastic, the service time is approximated with a general distribution. With IP based traffic, the downlink data arrives the MBS or the AP in a first-in first-out (FIFO) manner, making each MBS and AP a shared queue as TCP/UDP packets are processed upon arrival. They can therefore be modelled as an M/G/1-process shared queue (M/G/1-PS), which is often used to model RAN technologies [12, 13, 14].

B. Power model

We consider the power usage as a function of base and transmission usage as in [15]. A similar relationship is also observed for WiFi [16]. With an M/G/1-PS model, a transmitter can only have two states, P^A when active in serving a downlink requests and $P^{IN} \ll P^A$ when inactive, given by the probability of no flows in the queue. The probability of being in an inactive state is $(1-\frac{\lambda}{\mu})$ [17], where λ is the traffic intensity and μ is the serving rate. The expected power usage will be:

$$P^{total} = \frac{\lambda}{\mu}(P^A - P^{IN}) + P^{IN} \tag{1}$$

III. The market model

The game is modelled in two stages where an AP RAN operator determines a price, c^{AP}, for letting $\mathcal{I} := [1..I]$ LTE operators get access to its network. The AP RAN consists of $\mathcal{B} := [1..B]$ APs. The LTE operators' goal is to minimize both average delay for their UEs and resources used to facilitate the UEs' traffic. Each operator i has a MBS and there are $\mathcal{D} = \mathcal{I} \cup \mathcal{B}$ transmitters in the area. At the beginning of the game each operator has N_i users where N_{bi} are within the coverage of AP_b, both being common knowledge for all involved players. After the price is announced, the LTE operators decide on an allocation vector $\vec{x}_i = [x_d^i \; \forall d \in \mathcal{D}]$, describing the probability of letting a downlink flow go through an MBS or an AP. Figure 1 shows that the game can be visualized as a balancing act of queues where both resource use and delay are minimized. The AP RAN operator and the LTE operators all operate in different spectrums, avoiding interference with each other. None of the APs have overlapping coverage. As the game consist of two stages and is solved with backwards induction, the last stage is presented first.

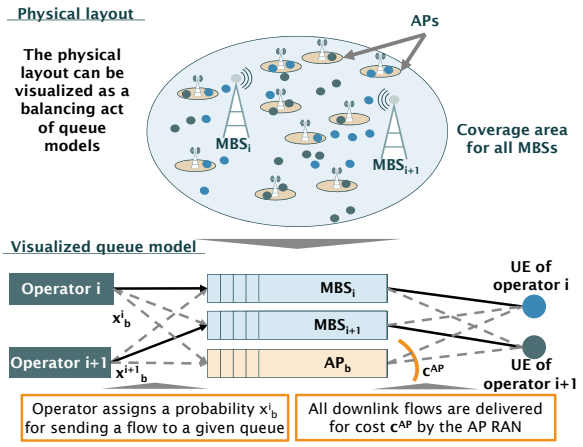

Fig. 1. The setup illustrated for I=2 and B=10. Each UE has at least one MBS or AP it can be served by. The LTE operators decide on optimal allocation vector $\vec{x}_i$, while the AP RAN operator sets a fixed price c^{AP} for every flow delivered to the UEs.

A. The LTE operators' utility

Once the price c^{AP} per flow is announced by the AP RAN operator, each operator i assigns its allocation vector $\vec{x}_i$ dictating the probability of assigning a flow to an MBS or an AP. Let $\vec{x}_{-i}$ denote the strategy profiles of every operator except operator i. The cost function, given by (2) consists of three parts, the waiting time, the energy cost related to carrying traffic and the net offloading earnings related the other MBS.

$$\begin{aligned} K_i(\vec{x}_i,\vec{x}_{-i}) &= \beta_i \sum_{d\in\mathcal{D}} \frac{x_d^i}{(\mu_d - \sum_{j\in I} N_j\lambda_j x_d^j)} \\ &+ (\frac{\sum_{j\in I} N_j\lambda_j x_i^j}{\mu_i})(P_i^A - P_i^{IN})C_i^P + P_i^{IN}C_i^P \\ &+ \sum_{j\in I|j\neq i} (C_j N_i\lambda_i x_j^i - C_i N_j\lambda_j x_i^j) + \sum_{b\in\mathcal{B}} c^{AP} N_i\lambda_i x_b^i \end{aligned} \tag{2}$$

The delay (total time in queue) per M/G/1-PS queue is $\frac{1}{\mu-\lambda}$ [17]. Each operator has a trade-off rate β_i, in units of $\frac{cost}{\text{wait time reduction per time slot}}$, which is the cost equivalence between resource use and reduced waiting time. The exact trade-off would depend on many factors that may vary with time such as the expected traffic load, current traffic load, type of customers the LTE operator is serving and its market position. As an upper bound, β_i is given by the alternative cost of adding a new MBS to the system, and hence for low traffic cases β_i would be higher than for high traffic cases since the relative gain in delay differs significantly. However, as the offloading is mainly intended for MBSs with resource scarcity and overall mobile traffic has a high exponential growth rate, β_i can be approximated as a constant calculated on a high utilized MBS. The power cost is a function of the power model and the power price, C_i^P. Each LTE operator accepts traffic from other operators in its MBS for the pre-negotiated price C_i. Therefore, an LTE operator also has a net offloading earning with regards to other LTE operators. The problem is formulated as (3) along with constraints (3a)-(3c).

$$\arg\min_{\vec{x}_i} \; K_i(\vec{x}_i,\vec{x}_{-i}) \tag{3}$$

$$0 \le x_d^i \le \frac{N_{di}}{N_i} \;\forall d\in\mathcal{D} \tag{3a}$$

$$\sum_{d\in\mathcal{D}} x_d^i = 1 \tag{3b}$$

$$\sum_{j\in\mathcal{I}|j\neq i} N_j\lambda_j x_d^j + N_i\lambda_i x_d^i \le (1-\epsilon_d)\mu_d \;\forall d\in\mathcal{D} \tag{3c}$$

The probability is non-negative and cannot be higher than the expected traffic from that area as constrained by (3a). Constraint (3b) ensures that all traffic is assigned to an MBS or an AP. For queues to be stable, $\mu_d > \sum_{i\in I} N_i\lambda_i x_d^i$, and a safety margin ϵ_d is used to denote the share of capacity that is never allocated. Constraint (3c) ensures that the queues are stable.

Lemma 1. *Problem (3) is a convex problem, as the utility function $K_i(\vec{x}_i,\vec{x}_{-i})$ given by (2) is strictly convex within the problem's constraints given by (3a) and (3c).*

Proof. The second order partial derivatives of $K_i(\vec{x}_i,\vec{x}_{-i})$ are all non-negative for stable queues as given by (4) and (5) where $\mu_d^{-i} = \mu_d - \sum_{j\in\mathcal{I}|j\neq i} N_j\lambda_j x_d^j$ is always positive due to (3c).

$$\frac{\partial^2 K_i(\vec{x}_i,\vec{x}_{-i})}{\partial (x_d^i)^2} = \beta_i \frac{2N_i\lambda_i\mu_d^{-i}}{(\mu_d^{-i} - N_i\lambda_i x_d^i)^3} \;\forall d\in\mathcal{D} \tag{4}$$

$$\frac{\partial^2 K_i(x)}{\partial x_d^i \partial x_{p\neq d}^i} = 0 \;\forall d,p\in\mathcal{D} \tag{5}$$

The associated Hessian matrix, H_i will always have a positive diagonal given by (4). The only condition that would yield $\vec{x}^T H_i \vec{x} = 0$ is when $\vec{x} = \vec{0}$ and therefore $H_i \succ 0$. □

Lemma 1 ensures that the local minimum is global within the closed and bounded set of $\vec{x}$ and problem (3) is solved as a convex optimization problem. The solution will be the LTE operator's best response to the other operators.

Lemma 2. *For a given price c^{AP}, the iterative best response obtained by solving (3) will converge to a unique Nash Equilibrium (NE).*

Proof. The problem in (3) can be formulated as maximizing $-K_i(\vec{x}_i,\vec{x}_{-i})$, and according to Lemma 1 it would be a strict concave function. Hence, there exists at least one NE [18]. Let $G(x)$ be a matrix containing all the second order partial derivatives for all $-K_i(\vec{x}_i,\vec{x}_{-i})$ as structured in [18]. The matrix will have only non-positive elements, as both (4) and (6) are always positive, while (5) is always zero. The diagonal of $G(x)$ will only contain negative values.

$$\frac{\partial^2 K_i(\vec{x}_i,\vec{x}_{-i})}{\partial x_d^{j\neq i} \partial x_d^i} = \beta_i \frac{N_j\lambda_j(\mu_d^{-i} + N_i\lambda_i x_d^i)}{(\mu_d^{-i} - N_i\lambda_i x_d^i)^3} \tag{6}$$

It is a sufficient condition that the matrix $B := [G(x) + G(x)^T] \prec 0$ for the NE to be unique [18]. Matrix B is a symmetric matrix and since all diagonal entries in G(x) are negative, only $\vec{x} = \vec{0}$ could produce $\vec{x}^T B\vec{x} = 0$ and hence $B \prec 0$. Therefore there exists only one NE per price c^{AP}. □

The NE is the collective response to the given c^{AP} from the LTE operators to the AP RAN operator.

B. AP RAN operator's utility

The AP network's utility function, $U(c^{AP}, \vec{X}(c^{AP}))$, consists of two parts; revenue gained by selling capacity and the power cost of serving the traffic. $\vec{X}^*(c^{AP})$ denotes the NE for the LTE operators for a given c^{AP}. Let us introduce $y_d^{*i} = N_i\lambda_i x_d^{*i}$, such that $U(c^{AP}, \vec{X}^*(c^{AP}))$ can be written as:

$$\begin{aligned} U(c^{AP}, \vec{X}^*(c^{AP})) = &\sum_{\substack{b\in\mathcal{B},\\ i\in\mathcal{I}}} c^{AP} y_b^{i*} - \sum_{b\in\mathcal{B}} P_b^{IN} C_b^P \\ &- \sum_{b\in\mathcal{B}} \Big(\frac{\sum_{i\in\mathcal{I}} y_b^{i*}}{\mu_b}\Big)(P_b^A - P_b^{IN})C_b^P \end{aligned} \tag{7}$$

The AP RAN operator will only accept non-negative values of c^{AP} as specified by (8a). The AP RAN operator's problem can be stated as (8) constrained by (8a).

$$\arg\max_{c^{AP}} \; U(c^{AP}, \vec{X}^*(c^{AP})) \tag{8}$$

$$c^{AP} \geq 0 \tag{8a}$$

Eq. (7) is linear and the price range of c^{AP} is bounded between lower value given by (8a) and upper value $c^{AP,max}$. If the price exceeds $c^{AP,max}$ the LTE operators will prefer to not use the AP RAN. External conditions, such as traffic, power price or difference in β_i may lead to different price ranges where the operators are price sensitive. Constrain (3b) links the $\vec{x}$, as a decrease in the share of traffic offloaded will increase the traffic on an operator's MBS. $x_i^i = 1 - \sum_{d\in\mathcal{D}|d\neq i} x_d^i$ is combined with (2) to create (9) that is later used to find $c^{AP,max}$. For convenience we use $\underline{\vec{x}}_i = [x_d^i \; \forall d \in \mathcal{D} | d \neq i]$ to denote the modified strategy space, μ_d^{-i} the modified service rate where the other operators' strategies are accounted for, $P_i^{\text{base}} = P_i^{IN} C_i^P$ and $P_i^{\text{trans}} = (P_i^A - P_i^{IN})C_i^P$. Eq. (9) is by all comparison equivalent to eq. (2) constraint by (3b) as they will have the same solution space.

$$\begin{aligned} Q_i(\underline{\vec{x}}_i, \underline{\vec{x}}_{-i}) = &\beta_i \frac{1 - \underline{\vec{x}}_i}{(\mu_i^{-i} - N_i\lambda_i(1 - \underline{\vec{x}}_i))} \\ &+ \beta_i \sum_{d\in\mathcal{D}|d\neq i} \frac{x_d^i}{(\mu_d^{-i} - y_d^i)} \\ &+ \Big(\frac{\sum_{j\in I|j\neq i} y_i^j + N_i\lambda_i(1 - \underline{\vec{x}}_i)}{\mu_i}\Big) P_i^{\text{trans}} + P_i^{\text{base}} \\ &+ \sum_{j\in I|j\neq i} (C_j y_j^i - C_i y_i^j) + \sum_{b\in\mathcal{B}} c^{AP} y_b^i \end{aligned} \tag{9}$$

The relation between an LTE operator's reaction and the price c^{AP} can be derived by setting $\frac{\partial Q_i(\underline{\vec{x}}_i, \underline{\vec{x}}_{-i})}{\partial x_d^i} = 0$ which gives:

$$\begin{aligned} c^{AP} = &\frac{\beta_i \mu_i^{-i}}{N_i\lambda_i(\mu_i^{-i} - N_i\lambda_i(1 - \underline{\vec{x}}^i))^2} + \frac{1}{\mu_i} P_i^{\text{trans}} \\ &- \sum_{d\in\mathcal{D}|d\neq i} \frac{\beta_i \mu_d^{-i}}{N_i\lambda_i(\mu_d^{-i} - N_i\lambda_i x_d^i)^2} \end{aligned} \tag{10}$$

The upper limit of $c_i^{AP,max}$ corresponds to the price where operator i prefers to route all its traffic through its own MBS given that it is the only operator in the network, that is $\{x_d^j = 0 \; \forall j \in \mathcal{I}, d \in \mathcal{D}\}$.

$$c_{bi}^{AP,max} = \frac{\beta_i \mu_i}{N_i\lambda_i(\mu_i - N_i\lambda_i)^2} - \sum_{d\in\mathcal{D}|d\neq i} \frac{\beta_i}{N_i\lambda_i\mu_d} + \frac{1}{\mu_i} P_i^{\text{trans}} \tag{11}$$

The upper bound price $c^{AP,max}$ is the highest one for all APs over all the LTE operators, i.e:

$$c_i^{AP,max} = \sup_{b\in\mathcal{B}} \; c_{bi}^{AP,max} \tag{12}$$

$$c^{AP,max} = \sup_{i\in\mathcal{I}} \; c_i^{AP,max} \tag{13}$$

If c^{AP} is larger than $c_i^{AP,max}$, the LTE operator i will not use the AP RAN. As the operators mutually influence each others' strategies, several local maximums in $U(c^{AP}, \vec{X}^*(c^{AP}))$ may exist, each with an ϵ-Stackelberg NE (ϵ-SNE). Let $\mathcal{S} := [0, c_i^{AP,max} \; \forall i \in \mathcal{I}]$ be the order set containing only unique values. For each interval given by the elements in $\mathcal{S}$, $[\tilde{c}_n^{AP}, \tilde{c}_{n+1}^{AP}]$ will denote a price interval in which a distinct number of LTE operators do not use the AP RAN. All ϵ-SNE are found through binary search within the limits of each pair $[\tilde{c}_n^{AP}, \tilde{c}_{n+1}^{AP}]$, where $n \in \mathcal{N}$ is the number of elements. ϵ indicates the accuracy of the solution. If $c^{AP,max} = c_i^{AP,max} \; \forall i \in \mathcal{I}$, then ϵ-SNE is unique as all operators are price sensitive for the price interval.

C. The distributed algorithm

We propose a distributed algorithm to solve the ϵ-SNE problem, which is summarized in Algorithm 1. An underlying assumption is that all players know the processing capacity μ_d for all the MBSs and all the APs. Algorithm 1 is used by the AP RAN operator and consists of two stages. First, the AP RAN asks the LTE operators for their $c_i^{AP,max}$. Then based on the replies it performs binary search between each of the price pairs, $[\tilde{c}_n^{AP}, \tilde{c}_{n+1}^{AP}] \in \mathcal{S}$. For convenience, the binary search is done in a price array $\vec{p}$, with $\vec{p}[r+1] - \vec{p}[r] = \epsilon$ where r denotes the index and $[\underline{R}, \overline{R}]$ gives the index' value range. Also, for readability we have shortened the notation of the utility function $U(c^{AP}, \vec{X}^*(c^{AP}))$ to $U(c^{AP})$. Algorithm 2 is used by the LTE operators to find the NE for $\vec{x}_i$ for price c^{AP}. The algorithm stops when all the LTE operators are satisfied that future changes in $\vec{x}_i$ are less than γ, that is $\sup([\Delta y_d^i \; \forall i \in \mathcal{I}, d \in \mathcal{D}]) < \gamma$ where $\Delta y_d^i = N_i\lambda_i |y_d^{i,k} - y_d^{i,k+1}|$. The NE

is achieved when a stop signal is sent from all involved LTE operators.

Algorithm 1 finds the local optimum per interval and then selects the global optimum among those with the time complexity of $\mathcal{O}\big(I \cdot log(\frac{c^{AP,max}}{I \cdot \epsilon})\big)$. For each comparison, Algorithm 2 is called two times, in which the LTE operators' problems are solved. Each LTE operator's problem can be solved with interior-point methods, which are solvable in polynomial time [19], making the overall time complexity of Algorithm 1 polynomial.

Algorithm 1: The AP RAN's algorithm

Result: Find the optimal price c^{*AP}
$\mathcal{S} \leftarrow$ ordered set of $[0, c_i^{AP,max}\ \forall i \in \mathcal{I}]$, $c^{*AP} = 0$;
for $[\tilde{c}_n^{AP}, \tilde{c}_{n+1}^{AP}] \in \mathcal{S}$ **do**
  // Binary search;
  $c^{AP} = 0$;
  $\vec{p}[\underline{R}] = \tilde{c}_n^{AP}$,$\vec{p}[\overline{R}] = \tilde{c}_{n+1}^{AP}$;
  while $\underline{R} \neq \overline{R}$ **do**
    $r_1 = \lfloor \frac{\underline{R}+\overline{R}}{2} \rfloor$, $r_2 = r_1 + 1$;
    if $U(\vec{p}[r_1]) \geq U(\vec{p}[r_2])|$ *Algorithm 2* **then**
      $c^{AP} = \vec{p}[r_1], \overline{R} = r_1$
    end
    else
      $c^{AP} = \vec{p}[r_2], \underline{R} = r_2$
    end
  end
  if $U(c^{*AP}) \leq U(c^{AP})$ **then**
    $c^{*AP} = c^{AP}$
  end
end

Algorithm 2: The LTE operators algorithm

Result: Find the best response to price c^{AP}
while $\sup([\Delta y_d^i\ \forall i \in \mathcal{I}, d \in \mathcal{D}]) \geq \gamma$ **do**
  for $i \in I$ **do**
    AP RAN broadcast μ_b^{-i} and c^{AP};
    LTE operator i reports its y_d^i for all μ_b^{-i} using Eq. (2) ;
  end
end

IV. Globally optimal solution

Minimizing the average waiting time in queue and energy usage to the entire system can be expressed as:

$$U^{global} = \Big(\sum_{d \in \mathcal{D}} \frac{\beta^{avg}}{|\mathcal{D}|} \frac{1}{\mu_d - \sum_{i \in \mathcal{I}} N_i \lambda_i x_d^i} + \sum_{d \in \mathcal{D}} P_d^{IN} C_d^P + \sum_{d \in \mathcal{D}} \big(\frac{\sum_{i \in \mathcal{I}} N_i \lambda_i x_d^i}{\mu_d}\big)(P_d^A - P_d^{IN}) C_d^P \Big) \tag{14}$$

Eq. (14) equivalent to minimizing all the LTE operators' utility functions and subtracting the AP RAN operator's utility function, $\sum_{i \in \mathcal{I}} K_i(\vec{X}^*(c^{AP})) - U(c^{AP}, \vec{X}^*(c^{AP}))$. β^{avg} is the average of β_i. The global optimal solution will later be used as a benchmark on the market's performance. Therefore the function is a subject to constraints (3a) - (3c) for all the LTE operators and the problem is given by (15).

$$\min_{x_d^i} \ U^{global} \quad \text{s.t constraints (3a) - (3c)} \ \forall i \in \mathcal{I} \tag{15}$$

As (14) is equivalent to the collection of players utility functions, it can be shown to be convex following the same logic as described in Lemma 1. All the second order partial derivatives, (16a)-(16c), of (14) are non-negative. The corresponding Hessian matrix, H^{global}, has only positive values in the diagonal and hence $H^{global} \succ 0$.

$$\frac{\partial^2 U^{global}}{\partial (x_d^i)^2} = \frac{\beta^{avg}}{|\mathcal{D}|} \frac{2\lambda_i^2}{(\mu_d - \sum_{j \in \mathcal{I}} N_i \lambda_j x_d^j)^3} \tag{16a}$$

$$\frac{\partial^2 U^{global}}{\partial x_d^i \partial x_d^k | k \neq i} = \frac{\beta^{avg}}{|\mathcal{D}|} \frac{2\lambda_i \lambda_k}{(\mu_d - \sum_{j \in \mathcal{I}} N_i \lambda_j x_d^j)^3} \tag{16b}$$

$$\frac{\partial^2 U^{global}}{\partial x_d^i \partial x_p^j | p \neq d} = 0 \tag{16c}$$

Problem (15) is solved as a convex optimization problem.

V. Numerical results

We conduct sensitivity analyses to gain insight into how the game evolves, focusing on the market equilibrium's behaviour. We start by defining a base case with I=2 operators, each with an MBS with coverage radius of 100 meters that overlaps each other perfectly. The APs are parametrized as 5G cell small BSs, assumed to have a coverage radius of 10 meters. The UEs of each operator are distributed according to a homogeneous poisson point process with λ_{UE}. All UEs have a downlink request process of $\lambda_i = 1$. The processing speed of each MBS $\mu_i = 1000$. Since offloading traffic has highest potential when resources are scarce, we focus the analyses on a high traffic density case, where 80% of resources are used on average per MBS, making $\lambda_{UE} = 0.03 \frac{UE}{m^2}$. We analyse the average case of users density, where $N_i = \lambda_{UE} 100^2 \pi$. For a given N_i, the number of UEs per AP follows a binomial distribution, with the average being $N_{bi} = \frac{N_i}{100}$. The processing rate of the APs are similar to that of the LTE MBSs with $\mu_b = \mu_i$. The MBSs are parametrized with $P_i^{IN} = 780W$ and $P_i^A = 1344W$, with power data from Auer et al. [15], and the APs are parametrized with data from Fisusi et al. [20] where $P_i^{IN} = 1W$ and $P_i^A = 8.2W$. The average system power price of Nordpool for 2013-2016[3] along with a time slot duration of 5 minutes gives $C_d^P = 2.4 \cdot 10^{-6} \frac{Euro}{\text{W per time slot}}$. Roaming cost between MBSs are symmetric with $C_i = 1$ and all precision parameters (ϵ and γ) and safety margins (ϵ_d) are set to 10^{-9}. The operator's β_i is calculated based on the alternative cost. An

[3]https://www.nordpoolgroup.com/historical-market-data/ visited 12.16.16

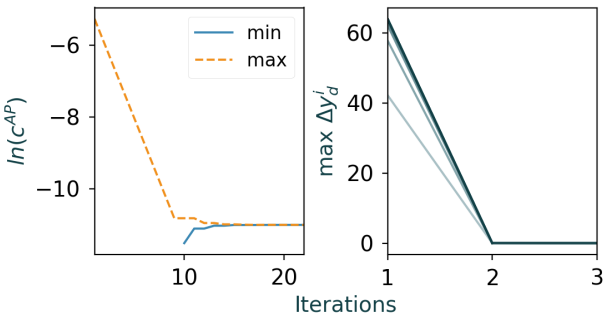

Fig. 2. The left plot shows the natural logarithm of the price range of $\mathcal{S}$ per iteration in the binary search as Algorithm 1 progresses. In the base case there exists only one ϵ-SNE. As the natural logarithm is not defined when $c^{AP} = 0$, the lower bound (blue line) is not plotted before $c^{AP} > 0$. Each green line in the right plot shows convergence of the maximum change of Δy_d^i for each iteration of Algorithm 2.

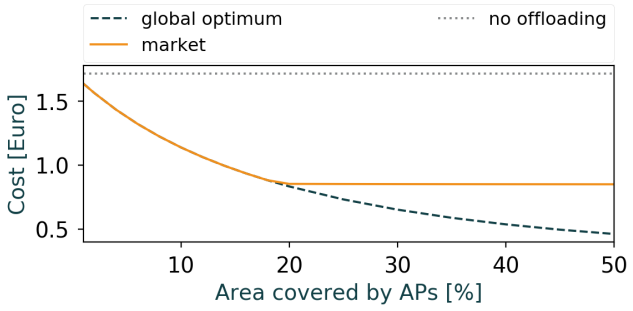

Fig. 3. Comparison of the system's utility for the global, the market and the no offloading solutions as a function of the AP RAN's area coverage.

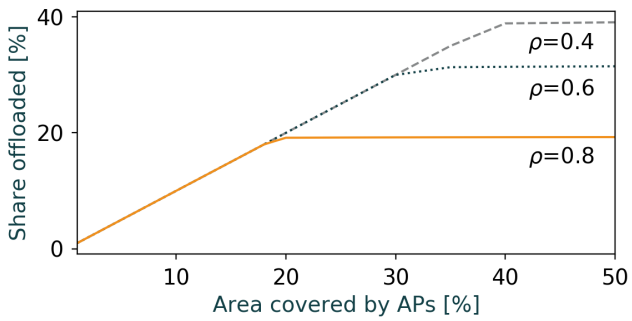

Fig. 4. The share of traffic offloaded as a function of the AP RAN's area coverage. The calculations are done for different MBS utilizations, $\rho_i = \frac{N_i \lambda_i}{\mu_i}$, given no offloading for the LTE operators. The base case scenario is plotted with an orange line.

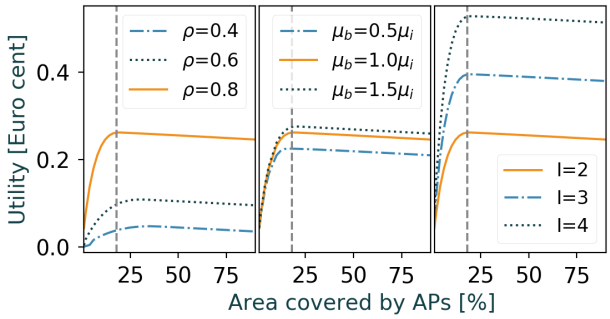

Fig. 5. The AP network operator's utility as a function of ρ, relative service capacity $\frac{\mu_b}{\mu_i}$ and number of operators I. The grey vertical dotted lines denote the optimal AP RAN size for the base case as described above. The base case scenario is plotted with an orange line in all plots.

additional MBS with an expected lifetime of 10 years is approximated to cost 10,000 Euros. The additional MBS would decrease the expected waiting time with 67% giving $\beta_i = 171 \frac{Euro}{\text{waiting time per time slot}}$. The value is driven by the energy use and the LTE operator's MBS utilization $\rho_i = \frac{\lambda_i N_i}{\mu_i}$. The calculations were implemented in Python with associated libraries [21, 22].

Fig. 2 shows the convergence of the distributed algorithm to the ϵ-SNE for Algorithm 1 in the left plot and to the NE for Algorithm 2 given c^{AP} in the right plot.

Fig. 3 shows the system's utility at global optimum, the suggested market and the no offloading solution plotted as a function of the AP RAN's area coverage. The suggested market and the global optimum are the same until a threshold is reached, where the market solution diverges from the global optimum as the AP densification increases.

Fig. 4 shows the share of LTE traffic offloaded, $\frac{\vec{x}_i}{x_i^0}$, plotted for different utilizations of the MBS (given no offloading), $\rho = \frac{N_i \lambda_i}{\mu_i}$. In the global optimal solution, it is best to offload as much traffic as possible to the AP RAN, and hence the system's utility decreases with increased AP densification at the account of the AP RAN operator's utility. That is why in the market model, the share of offloaded traffic reaches a saturation point. The saturation point decreases in ρ, reducing the share of traffic offloaded when the LTE operators have higher resource scarcity.

The offloading share reaches a threshold as the relative cost of P^{IN} increases with every new AP added to the AP RAN. Hence, there exists an optimal network size for the AP RAN in terms of number of APs. The marginal utility of adding an extra AP decreases with the increase of the AP RAN's coverage, creating a concave curve for the AP RAN operator's utility as a function of the network size. Fig. 5 shows the AP RAN operator's utility for various MBS utilizations, processing speeds at the APs and number of I operators. The optimal network size for the base case scenario is shown with a dotted vertical line in each plot.

VI. Discussion

The proposed distributed algorithm achieves the global optimal solution as seen in Fig. 3, where the market and the global optimum solution follow each other. However, the market equilibrium reaches a threshold where increased energy cost is balanced with the marginal delay reduction of adding a new AP. The threshold depends on market conditions such as traffic load. As observed in Fig. 4, the lower the traffic load is in the MBS (given no offloading), the larger the potential for the share of traffic that can be offloaded. This is seemingly

counter intuitive, as higher MBS utilization suggests resource scarcity. This observation can be explained with Fig. 5, where the utility function of the AP RAN increases in LTE resource scarcity $\rho_i = \frac{N_i \lambda_i}{\mu_i}$. Hence, when resources are scarce, the AP RAN can demand a higher price, which results in a lower share of traffic being offloaded. As the utility function is concave and skewed towards the lower end of the AP RAN area coverage, there is an optimal network size. The physical interpretation of the observation is that the AP RAN has best returns on the first AP it installs, with diminishing returns as more AP's are added. The AP network's utility as a function of the RAN size maintains its shape for different values of μ_b and I. When the radio resource scarcity is low in the MBSs, that is ρ is low, the peak of the AP RAN's utility function shifts, and the optimal network size increases resulting in the counter intuitive observation seen in Fig. 4. Based on these results, a AP RAN operator could maximize its return by adapting the number of AP online depending on the resource scarcity at the LTE operators side. With such a configuration the AP RAN should have less APs online when the LTE operators have resource scarcity, thus pressing the price and maximizing its own return.

In this study we have investigated offloading markets in a stochastic environment which we believe would provide a better abstraction. We have proposed a distributed algorithm and identified the conditions under which a market would yield the global optimum result. We focused on evenly dispersed UEs. For cases with hot-spots, the AP RAN's pricing power will increase, and it is expected that the AP RAN operator's utility will increase, but not necessarily the share of the offloaded traffic.

Acknowledgment

This research is supported by the project 240079/F20, funded by the Research Council of Norway.

References

[1] Cisco VNI Mobile, "Cisco Visual Networking Index (VNI) Update Global Mobile Data Traffic Forecast 2016 - 2021," Cisco Public Information, Tech. Rep., 2017.

[2] E. J. Oughton and Z. Frias, "The cost, coverage and rollout implications of 5G infrastructure in Britain," *Telecomm. Policy*, 2017.

[3] F. Boccardi, H. Shokri-Ghadikolaei, G. Fodor, E. Erkip, C. Fischione, M. Kountouris, P. Popovski, and M. Zorzi, "Spectrum Pooling in MmWave Networks: Opportunities, Challenges, and Enablers," *IEEE Commun. Mag.*, vol. 54, no. 11, pp. 33–39, 2016.

[4] M. A. Marsan and M. Meo, "Energy efficient management of two cellular access networks," *ACM SIGMETRICS Perform. Eval. Rev.*, vol. 37, no. 4, pp. 69–73, 2010.

[5] A. Aram, C. Singh, S. Sarkar, and A. Kumar, "Cooperative Profit Sharing in Coalition Based Resource Allocation in Wireless Networks," in *Ieee Infocom*, 2009.

[6] K. Poularakis, G. Iosifidis, and L. Tassiulas, "A framework for mobile data offloading to leased cache-endowed small cell networks," in *Int. Conf. Mob. Ad Hoc Sens. Syst.*, 2014.

[7] H. Shah-mansouri and V. W. S. Wong, "An Incentive Framework for Mobile Data Offloading Market under Price Competition," *IEEE Trans. Mob. Comput.*, vol. 16, no. 11, pp. 2983 – 2999, 2017.

[8] M. Li, T. Q. Quek, and C. Courcoubetis, "Economics in mobile data offloading with uniform pricing," in *IEEE Int. Conf. Commun.*, 2017.

[9] K. Wang, F. C. M. Lau, L. Chen, and R. Schober, "Pricing Mobile Data Offloading: A Distributed Market Framework," in *IEEE Trans. Wirel. Commun.*, vol. 15, no. 2, 2016, pp. 913 – 927.

[10] L. Gao, G. Iosifidis, J. Huang, and L. Tassiulas, "Economics of mobile data offloading," in *INFOCOM*, 2013.

[11] R. R. Tyagi, F. Aurzada, K. D. Lee, and M. Reisslein, "Connection Establishment in LTE-A Networks: Justification of Poisson Process Modeling," *IEEE Syst. J.*, vol. 11, no. 4, pp. 1–12, 2017.

[12] F. Mehmeti and T. Spyropoulos, "Performance modeling, analysis and optimization of delayed mobile data offloading under different service disciplines," *Ieee/Acm Trans. Netw.*, vol. 25, no. 1, pp. 550–564, 2017.

[13] J. Yang, X. Zhang, and W. Wang, "Two-stage base station sleeping scheme for green cellular networks," *J. Commun. Networks*, vol. 18, no. 4, pp. 600–609, 2016.

[14] J. Yang, W. Wang, and X. Zhang, "Hysteretic Base Station Sleeping Control for Energy Saving in 5G Cellular Network," in *Veh. Technol. Conf.*, 2017.

[15] G. Auer, V. Giannini, C. Desset, I. Gódor, P. Skillermark, M. Olsson, M. A. Imran, D. Sabella, M. J. Gonzalez, O. Blume, and A. Fehske, "How much energy is needed to run a wireless network?" *IEEE Wirel. Commun.*, vol. 18, no. 5, pp. 40–49, 2011.

[16] K. Gomez, R. Riggio, T. Rasheed, and F. Granelli, "Analysing the energy consumption behaviour of WiFi networks," in *2011 IEEE Online Conf. Green Commun.*, 2011, pp. 98–104.

[17] J. Sztrik, *Basic Queueing Theory: Foundations of System Performance Modeling*. Saarbrücken: GlobeEdit, 2016.

[18] J. B. Rosen, "Existence and Uniqueness of Equilibrium Points for Concave N-Person Games," *Econometrica*, vol. 33, no. 3, pp. 520–534, 1965.

[19] S. Boyd and L. Vandenberghe, *Convex Optimization*, 7th ed. New York: Cambridge University Press, 2010, vol. 25, no. 3.

[20] A. Fisusi, D. Grace, and P. Mitchell, "Energy saving in a 5G separation architecture under different power model assumptions," *Comput. Commun.*, vol. 105, no. 1, pp. 89–104, 2017.

[21] M. S. Andersen, J. Dahl, and L. Vandenbergheg, "CVXOPT: A Python package for convex optimization," 2018.

[22] E. Jones, T. Oliphant, and P. Peterson, "SciPy: Open source scientific tools for Python," 2018.