Green Resource Management for Over-The-Top Services in 5G Networks using Matching Theory

Eftychia Datsika*, Angelos Antonopoulos†, Nikos Passas‡, Georgios Kormentzas§, Christos Verikoukis†
*IQUADRAT Informatica S. L., Barcelona, Spain
†Telecommunications Technological Center of Catalonia (CTTC/CERCA), Castelldefels, Spain
‡National and Kapodistrian University of Athens, Department of Informatics & Telecommunications, Athens, Greece
§University of the Aegean, Department of Information & Communication Systems Engineering, Karlovassi, Greece
Email: edatsika@iquadrat.com, {aantonopoulos, cveri}@cttc.es, passas@di.uoa.gr, gkorm@aegean.gr

***Abstract*—Nowadays, over-the-top (OTT) applications can be accessed via Internet connections over cellular networks. Mobile network operators (MNOs) strive to accommodate mobile traffic through energy efficient shared networks. The OTT service providers (OSPs) need to interact with MNOs and require resources for offering desired quality-of-service (QoS) levels to users of different categories, with different QoS requirements, which imply prioritization of certain OTT flows. Resource scheduling should respect network neutrality, which forbids OSP prioritization. Furthermore, OSPs request resources periodically, according to their performance goals, causing delay in flows' accommodation due to i) the time required for information exchange between OSPs and MNOs, affected by network congestion, and ii) the time required for flows to receive resources, affected by the number of active flows. The intervention of OSPs may induce additional energy cost for the MNOs and affect the mobile network energy efficiency. Acknowledging the lack of OSP-oriented resource management approaches, we introduce a matching theoretic flow prioritization (MTFP) algorithm and investigate delay and energy efficiency through extensive simulations. Our study shows that MTFP improves both metrics comparing to the best effort approach, whereas its performance is affected by the OTT traffic level and the frequency of resource allocation process.**

Keywords*—Over-the-top services, Energy efficiency, Wireless network virtualization, Resource management, Matching theory.

I. Introduction

The global mobile data traffic is expected to increase sevenfold until 2021 [1], stressing the need for high network capacity in the fifth generation (5G) wireless networks and motivating the design of energy efficient mobile network deployments [2]. Aiming to serve the users without excessive expenditures, the mobile network operators (MNOs) share their infrastructure and spectrum. Different network sharing architectures are specified in the long term evolution advanced (LTE-A) mobile communication standard [3]. Shared LTE-A networks can be virtualized, i.e., resources can be abstracted into virtual slices (VSs) managed by MNOs in isolation.

Modern applications based on mobile Internet connectivity, i.e., over-the-top (OTT) applications (e.g., YouTube, Skype, etc.) have introduced OTT service providers (OSPs) that offer their communication services over the MNOs' networks. A large portion of mobile data is related to user equipment terminals (UEs) that generate OTT application flows. The OSPs benefit from the popularity of their applications, thus they are motivated to improve the quality-of-service (QoS) of flows, which may have different QoS demands, e.g., low latency for gaming or high data rate for video streaming. Each application may also involve different user categories, e.g., free or premium users, which generate flows of dissimilar importance. However, when resources are allocated to the UEs, the flows' priorities are not considered and VSs are allocated in a best effort manner [4]. The OSPs are not involved in the VS allocation, do not control their performance indicators, e.g., grade of service (GoS), and cannot not apply flow prioritization, as MNOs fully control the UEs' connections.

The intervention of OSPs in resource management of flows can be profitable for both OSPs and MNOs, as delivering high quality services is a goal of both parties. Notably, cooperation of OSPs and MNOs for the joint deployment of network infrastructure has demonstrated their common interests [5]. However, it is not clear how the resources can be shared among OSPs, when flows of different priorities coexist. The network resources should be shared impartially among OTT applications, thus, prioritization should be applied at OTT application flow level, while fairness should be guaranteed at OSP level, as dictated by the network neutrality rules [6].

The VS allocation for OSPs can be challenging for the MNOs with respect to the energy efficiency of the shared network infrastructure. The use of OTT applications may induce additional energy cost when resources are allocated to the UEs, e.g., OTT voice calls result in high signaling overhead [7]. Aiming to restrain the mobile network's energy consumption, various energy saving strategies have been proposed. The MNOs may opt to switch off underutilized eNBs in order to reduce the energy consumption by means of cooperative [8] or non-cooperative game theory [9], [10]. However, despite that the impact of OSP-oriented resource allocation on the energy efficiency can be a motive or impediment for MNO-OSP cooperation, it has not been extensively studied yet.

The VSs encompass resources of the core network (CN), i.e., bandwidth in CN links, and the radio access network (RAN), i.e., spectrum. Hence, end-to-end resources are allocated to flows [11]. RAN resource scheduling periodically allocates spectrum to UEs, in VS allocation rounds, according to network-related parameters (e.g., congestion of links, cellular links' channel conditions, etc.), and MNOs' performance goals (e.g., maximization of spectral efficiency, etc.). Given the periodicity of resource allocation process and the dynamic number of flows concurrently requesting resources, flows may not receive resources in each round, experiencing time delay during their service time. Moreover, when OSPs' policies are considered, the shared network coordinator (e.g., a centralized controller) should periodically interact with the OSPs. As

978-1-5386-3180-5/18/$31.00 ©2018 IEEE

information about the flows needs to be conveyed from the RAN to the OSPs and vice versa, the CN links may also experience congestion, increasing the overall delay, i.e., the time needed for the reception of resources by the flows, as arranged by the RAN scheduling scheme, and the time required for the transmission of flows' information through the CN.

Most resource allocation approaches refer to RAN resources of a single evolved NodeB base station (eNB) [12], or a RAN where eNBs and/or spectrum resources are shared among MNOs [13]. Other resource management schemes refer to virtual MNOs (MVNOs) that do not own spectrum or infrastructure. MVNOs can adjust their resource allocation demands to the MNOs' pricing strategies reaching a Stackelberg equilibrium [14], increase their benefit using market equilibrium theory [15] or declare preferences over VSs using matching theory [3]. Although the schemes for MVNOs could potentially apply to OSPs, novel resource management schemes are required, as these schemes do not consider the OSPs' policies and the network neutrality issue that arises when flows with different priorities access the network. Also, the OSPs may have dynamic preferences over resources that are not expressed by strictly defined utility functions required by the existing approaches, and may entail intractable combinatorial problems. The matching theory can facilitate the interaction among MNOs and OSPs without yielding optimization problems of high complexity [16]. Even though the approaches for MNOs or MVNOs can be applied periodically as scheduling techniques, no insights for the additional delay they may induce are provided. The fact that end-to-end resources are allocated to the flows has not been studied as a factor that could affect the performance of the proposed schemes.

In this paper, motivated by the deficiencies of the existing approaches and the importance of satisfying the QoS of the OTT applications, we introduce a novel method that enables the intervention of OSPs in the VS allocation process with the aid of matching theory, which allows the OSPs to express interest for resources. The contribution of this work is twofold:

(i) *Design of an efficient matching theoretic flow prioritization (MTFP) algorithm:* We introduce a novel resource allocation algorithm that allows the OSPs to i) declare their preferences over the network resources in each VS allocation round, exploiting the distributed nature of matching theory, and ii) manage their user prioritization policies, without violating the network neutrality rules.

(ii) *Extensive assessment of the performance of MTFP algorithm in terms of induced delay and energy efficiency of the mobile network:* We investigate the average delay experienced by the flows and the energy efficiency of the shared mobile network when the MTFP algorithm is employed, considering different OTT application traffic levels and VS allocation frequencies.

The remainder of the paper is organized as follows. The system model is described in Section II and, in Section III, the MTFP algorithm is presented. The simulation results are discussed in Section IV. In Section V conclusions are drawn.

II. Network Architecture and System Model

We next describe the architecture of a shared LTE–A network and the considered system model.

A. Shared LTE-A network

In the shared LTE-A network of Fig. 1, different MNOs manage cooperatively the RAN elements e.g., collocated eNBs, spectrum resource blocks (RBs) and the CN elements connected to RAN, e.g., routers. The resource management is performed using the software defined networking (SDN) framework that offers a virtualization controller (VC) [17]. For the interaction with the VC, suitable network application programming interfaces (APIs) are provided (MNOs' and OSPs' APIs). In the RAN, the spectrum of each eNB is sliced and the VSs offered to OSPs include sets of RBs. The resource scheduling process is performed periodically, in VS allocation rounds, with a frequency determined by the network's capabilities and congestion levels, allowing the transmission of the UEs' information from RAN to VC and the exchange of the required information between OSPs and VC. This process may last longer than the regular resource scheduling performed per transmission time interval (TTI). In the CN, the aggregation of the flows' information is performed via the available routers. Thus, when VSs are assigned to OSPs, specific bandwidth is reserved in each CN link.

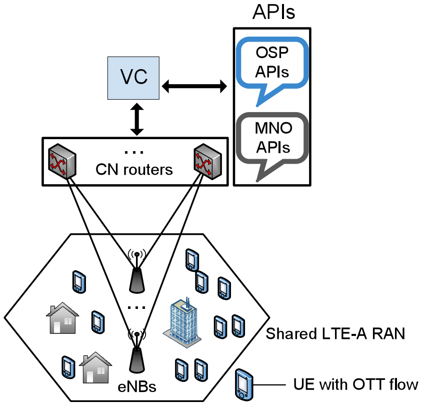


Fig. 1: Considered shared LTE-A network

In order to decide about the required VSs, the OSPs should be aware of the status of the UEs related to the flows, e.g., the downlink channel conditions. Each UE can connect to an eNB and report the channel quality indicator (CQI), which determines the modulation and coding scheme (MCS) used for the downlink transmissions related to the UEs' flows. The VC can provide the information about flows to the OSPs' APIs, allowing the OSPs to estimate the QoS levels using the metrics they prefer and adjust their requirements regarding the VSs.

B. System model

We consider the cell of a shared RAN jointly operated by N MNOs with co-located eNBs (Fig. 2). Each MNO owns an eNB $n \in \mathcal{N}$ and spectrum, shared with the other MNOs. A number of W RBs is available, whereas U UEs are subscribers of either of the MNOs. A set of $\mathcal{M}$ OSPs may co-exist in the network. Each UE generates flows related to different OTT applications, thus, each flow corresponds to a specific UE and OSP. Assuming a set of $\mathcal{J}$ flows and m a specific OSP, we denote $\mathcal{J}^{(m)}$ the set of flows related to the OSP m. A set $\mathcal{J}_{(n)}$ of flows is associated with eNB n.

The OSPs' policies deem each flow to be of higher or lower priority p_j. A number of $\mathcal{K}$ priority classes exists. Flows belonging to different OTT applications may have different priorities, even when the flows are related to the same UE. The flows are generated by U UEs following a Poisson distribution

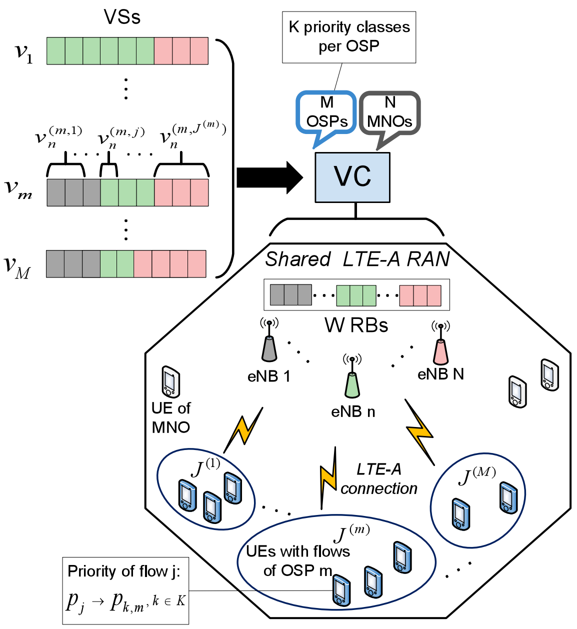


Fig. 2: VS allocation in the considered network

with rate λ (flows/UE). The mean duration of each flow is exponentially distributed with mean equal to $1/\mu$. Each OSP needs to acquire a set of RBs in order to serve the flows that are associated with UEs in either of the available eNBs. The VC allocates v_m RBs to the VS that corresponds to OSP $m \in \mathcal{M}$. Each flow $j \in \mathcal{J}^{(m)} \subset \mathcal{J}$ needs a number of $v_n^{(m,j)} \leq v_m$ RBs that provides it with a downlink data rate $r_{\text{srv}}^{(m,j)}$.

A UE that generates flows can report CQIs to each eNB n in every TTI [12]. Given an $\text{MCS}_n^{(m,j)}$ and a number of allocated RBs $v_n^{(m,j)}$ to the UE related to flow j, the achievable downlink data rate is given by:

$$r_n^{(m,j)} = \frac{L\left(\text{MCS}_n^{(m,j)}, v_n^{(m,j)}\right)}{\text{TTI}}, \quad (1)$$

where the transport block size $L(\text{MCS}_n^{(m,j)}, v_n^{(m,j)})$ can be found as in [3]. The value $\text{MCS}_n^{(m,j)}$ may be different in each round for a specific UE. Assuming downlink channels with Rayleigh fading, the Signal–to–Noise Ratio (SNR) is a random variable with average value γ and probability density function:

$$f(x) = \frac{1}{\gamma} e^{-\frac{x}{\gamma}} u(x), \quad (2)$$

where $u(x)$ is the unit step function and $[\gamma_{thr}^{(i)}, \gamma_{thr}^{(i+1)}]$ is the SNR range that corresponds to MCS i.

The VS allocation is performed periodically in successive VS allocation rounds. The OSPs request RBs on behalf of their flows in a VS allocation step exponentially distributed with mean value $\mathbb{E}[t]$. While a UE's connection to an OTT application is active, the corresponding flow experiences several VS allocation rounds. However, in each round, RBs may or may not be allocated to a flow, thus, it experiences an average delay $\mathbb{E}[D]$, due to the time spent in fruitless rounds.

The network energy efficiency is affected by the total data rate demand in each eNB, i.e., the number of served flows and their data rate requirements, and the channel conditions of the UEs, i.e., the total number of RBs used by the corresponding eNBs. Using Eq. (1), we define the energy efficiency η_n per eNB n in a VS allocation round as:

$$\eta_n = \frac{\sum_{m \in \mathcal{M}} \sum_{j \in \mathcal{J}^{(m)} \cap \mathcal{J}_{(n)}} r_n^{(m,j)}}{P_n}, \quad (3)$$

where the power consumption P_n of eNB n is equal to [18]:

$$P_n = P_C^{(n)} + \delta P_{RB}^{(n)}, \quad (4)$$

considering, for each eNB n, the constant power consumption $P_C^{(n)}$ related to signal processing, cooling and battery backup, the power consumption δ that scales with the average radiated power due to amplifier and feeder losses and the power consumption $P_{RB}^{(n)}$ for the transmission of one RB. Given W available RBs, the transmission power $P_{Tx}^{(n)}$ and a_n the number of antennas of eNB n, the value $P_{RB}^{(n)}$ is calculated as:

$$P_{RB}^{(n)} = \frac{P_{Tx}^{(n)}}{a_n W}. \quad (5)$$

Using Eq. (3), we derive the overall network efficiency as:

$$\mathbb{E}[\eta] = \frac{\sum_{n \in \mathcal{N}} \eta_n}{|\mathcal{N}|}, \quad (6)$$

assuming a number of $\mathcal{N}$ eNBs in the shared network.

III. Matching theoretic flow prioritization

We thereupon describe the VS allocation problem for OSPs and propose a matching theoretic flow prioritization algorithm.

A. VS allocation and involved parties' preferences

The VS allocation involves the assignment of RBs to each flow according to the CQI and MCS values of the UE related to each flow, and the required QoS levels, i.e., acceptable data rate and flow priority, as defined by the corresponding OSP's policy. At each VS allocation round, each OSP m requests RBs in the eNBs that offer the requested downlink data rates $\sum_{j \in \mathcal{J}^{(m)}} r_{\text{srv}}^{(m,j)}$ and aims to minimize the GoS_m, defined as the ratio of the number of flows that do not achieve the required data rate with the allocated resources over the total number of flows $\mathcal{J}^{(m)}$ and estimated as:

$$GoS_m = 1 - \frac{1}{|\mathcal{J}^{(m)}|} \sum_{j \in \mathcal{J}^{(m)}} \sum_{n \in \mathcal{N}} [r_n^{(m,j)}(v_{n,m}^{(j)}) \geq r_{\text{srv}}^{(m,j)}], \quad (7)$$

The allocation of RBs may not be possible for all flows at each VS allocation round. Each OSP prefers that flows with higher priority, i.e., lower p_j value, receive the required RBs first in each round, ensuring that they experience lower delay than flows of lower priority. Among flows with the same priority, those that have lower demands of RBs, e.g., experience better channel conditions or have lower data rate demands, should be served first, increasing the number of flows that can be served.

The MNOs are interested in minimizing the expected number of flows of all OSPs that do not achieve the required data rates, defined as expected GoS $\mathbb{E}[GoS]$, respecting the priorities defined by the OSPs, without violating the network neutrality rules. The value $\mathbb{E}[GoS]$ is equal to:

$$\mathbb{E}[GoS] = \frac{\sum_{m \in \mathcal{M}} GoS_m}{|\mathcal{M}|}. \quad (8)$$

Flows with lower priorities may be lead to starvation, as the RBs may not be sufficient. Hence, the eNBs update the priorities depending on whether each flow has previously received resources or not, respecting the OSPs' policies and ensuring that all flows receive resources at some point. The higher the priority of a flow, the more likely it is that it receives resources in a round, experiencing lower delay.

B. Formulation of matching process using contracts

In VS allocation, the flows offer contracts, whereas eNBs rank the offered contracts. Each contract is a combination of parameters that associate a flow with an eNB, i.e., the flow's priority and the RBs required for achieving the desired QoS in a specific eNB. A flow must be associated with exactly one eNB and an eNB serves multiple flows (many-to-one matching). A contract c related to flow j and eNB n is represented by a vector (j, n, q), where q is the cost of contract $q = (p_j, v_n^{(m,j)})$ that is defined as a real number with the integer part equal to the flow's priority p_j and a decimal part equal to the RBs $v_n^{(m,j)}$ required by the UE related to flow j in order to achieve $r_{\text{srv}}^{(m,j)}$, when the UE is connected to eNB n, as given by Eq. (1). A flow ranks the contracts by ascending q value, as this will provide it with higher priority in resource allocation, increasing its chances of receiving RBs and reducing the experienced delay. Moreover, each eNB aims at establishing contracts with flows at the minimum possible cost q. Among flows of the same priority, those with lower resource demands (fewer RBs) are preferred. The accepted contracts confirm the agreement between flows and eNBs and form the chosen set, whereas the rest of the contracts form the rejected set. We denote as C the set of all possible contracts.

Definition 1. *Given the set of all possible contracts C and $C' \subset C$ a subset of C, the chosen set $S_j(C')$ of a flow j either contains only one element (the flow's preferred contract out of C') or is empty, if there is no acceptable contract c in C' for flow j. Similarly, the chosen set $S_n(C')$ of an eNB n either contains the eNB's preferred contracts out of C' or is empty, if there is no acceptable contract c in C' for eNB n.*

The remaining options from the set of contracts that are not accepted from anyone form the set of rejected contracts.

Definition 2. *Given the set of all possible contracts C, a subset C' of C, and $S_J(C') = \cup_{j \in \mathcal{J}} S_j(C')$ and $S_N(C') = \cup_{n \in \mathcal{N}} S_n(C')$ the chosen sets of all flows and eNBs, respectively, the sets of contracts that are rejected by all flows and all eNBs are defined as $R_F(C') = C' \backslash S_J(C')$ and $R_N(C') = C' \backslash S_N(C')$. The rejected sets of a flow j and an eNB n are defined as $R_j(C')$ and $R_n(C')$, respectively.*

At the end of the matching procedure, a stable association between eNBs and flows is achieved, if there can be no allocation strictly preferred by any eNB and there exists no flow that would prefer to reject the contract it has received.

Definition 3. *A set of contracts $C' \subset C$ results in a stable VS allocation if and only if*

(i) $S_N(C') = S_J(C') = C'$ *(*individual rationality*)*

*(ii) there exists no eNB $n \in \mathcal{N}$ and set of contracts $C'' \neq S_n(C')$ such that $C'' = S_n(C' \cup C'') \subset S_J(C' \cup C'')$ (*nonexistence of blocking contracts*).*

The first condition dictates that if only the contracts in C' are available, then they are all chosen. When the condition does not hold, it means that there exist a flow or eNB that prefers to reject a contract. The second condition imposes that there exist no set of contracts C'' that could be added and would be selected by both eNB n and the flows related to n.

The property of substitutability for the eNBs' preferences is a sufficient condition for achieving a stable allocation [16]:

Definition 4. *The contracts in C are substitutes for eNB $n \in \mathcal{N}$, if for all subsets $C' \subset C'' \subset C$, it holds that $R_n(C') \subset R_n(C'')$, where R_n is the set of contracts rejected by n.*

According to the property of substitutability of eNBs' preferences over contracts, every contract rejected from C' is also rejected from C'', and if a contract is chosen by an eNB from some available contracts, then that contract will still be selected from any smaller set that includes it.

C. Proposed matching algorithm

We present a matching theoretic flow prioritization (MTFP) algorithm that matches the flows accessing a shared network with the eNBs' resources, considering the flows' priorities and can be applied in each VS allocation round.

Algorithm 1 consists of two phases (i.e., initialization and negotiation) that take place in each VS allocation round. During the initialization phase, all UEs report their CQIs and each eNB transmits this information to the VC. The OSPs update the information about the flows' priorities and required QoS. In the negotiation phase, at each matching iteration, the flows start ranking their preferences over the available set of contracts, according to the priorities set by their OSPs, and submit their requests for assignment to the most preferred contracts with the corresponding eNBs using the VC. The eNBs update the flows' priorities and sort the available contracts. Therefore, two sets of contracts are created. The first set is the chosen set S_N of contracts, which contains the most preferred contracts based on the OSPs' preferences. The second set is the rejected set R_N of contracts, which is the complement of the chosen set. The negotiation phase is repeated while the rejected flows submit requests for assignment to their next preferred set of contracts, until no more requests are added to the rejected set R_N. Once contracts are finalized, the VSs for all OSPs are created and the requested RBs are allocated to each eNB.

Proposition 1. The MTFP algorithm converges to a stable eNB-flow matching after a finite number of iterations.

Proof: The number of possible contracts that can be allocated is finite. The algorithm converges when no more flows are added to R_N, thus every flow is associated with an eNB through a contract and the property of substitutability (Definition 4) characterizes the eNBs' preferences. ■

IV. Performance Evaluation

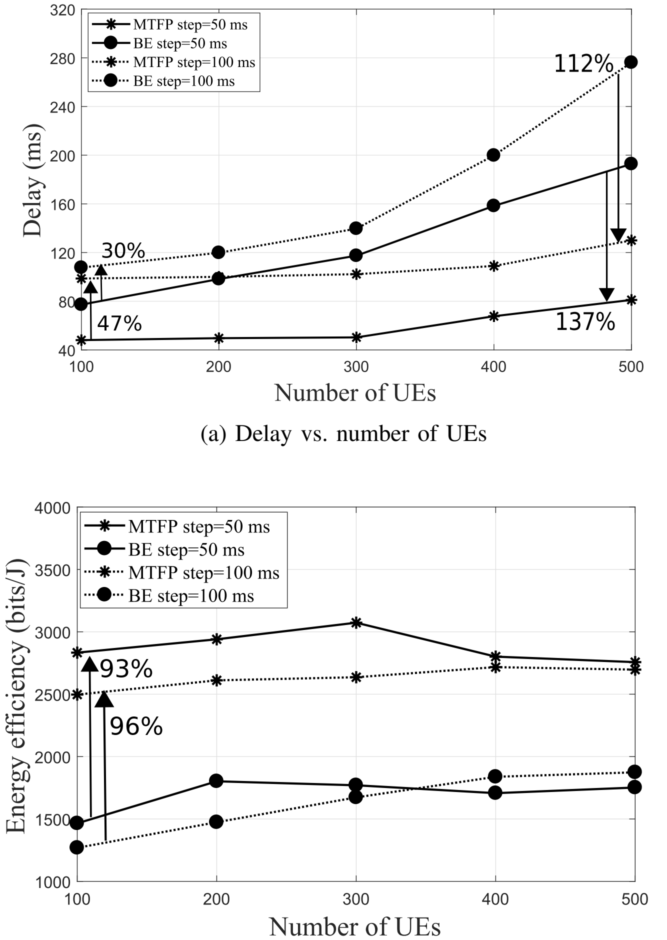
We study the performance of the MTFP algorithm in terms of delay and energy efficiency considering different scenarios and the settings described in Section IV-A. The simulation results are discussed in Section IV-B.

A. Simulation Setup

We consider the network in Fig. 2 with $N = 2$ MNOs and $|\mathcal{M}| = 2$ OSPs, e.g, YouTube or Skype. Each OSP has $|\mathcal{K}| = 2$ priority classes, i.e., a high priority class of premium users requesting 1 Mb/s downlink data rate and a low priority class, i.e., free users, requesting 0.5 Mb/s. High priority characterizes

Algorithm 1 Matching theoretic flow prioritization (MTFP) algorithm

```
Input: CQIs of UEs, rate constraints and priorities of flows
Output: Stable allocation per VS allocation round
  Initialization phase:
  -The UEs with active flows submit their CQIs to eNBs.
  -The eNBs submit the flows' information to VC.
  -Each OSP $m$ checks each flow's $j$ status and assigns the priorities $p_j$ and requested data rate $r_{\text{srv}}^{(m,j)}$.
  Negotiation phase: // Start matching iterations
  Repeat:
  -The flows estimate the RBs required at each eNB $n$ and sort the available contracts $c \in C$ according to cost $q \in Q$.
  -Each flow $j \in \mathcal{J}$ creates the chosen set $S_j(C')$ and the rejected set $R_j(C') = C' \backslash S_j(C'), C' \subset C$.
  -Each eNB $n \in \mathcal{N}$ updates the priorities of flows served in the previous round ($p_j =$ initial $p_j + 1$).
  -Each flow with $R_j(C') \neq \emptyset$ submits the next preferred contract from $S_j(C')$ to the VC.
  -The eNBs check if the flows that submit contracts have been previously served:
  $\forall$ flow $j \in \mathcal{J}$:
  if flow $j$ rejected in the previous round then
      Set $p_j =$ initial $p_j$.
  end if
  -Each eNB $n$ accepts most preferred contracts and rejects the others, creating the chosen set $S_n(C')$ and the rejected set $R_n(C') = C' \backslash S_n(C'), C' \subset C$.
  Until convergence to a stable allocation.
  -The VC allocates RBs considering the number of available RBs $W$ and transmits the required information to eNBs.
```



(a) Delay vs. number of UEs

(b) Energy efficiency vs. number of UEs

Fig. 3: Effect of number of connected UEs

50% of the flows. Three modulation schemes (QPSK, 16-QAM and 64-QAM) and $W = 100$ RBs are available, whereas U UEs ($U/2$ per MNO), generate flows following a Poisson distribution with rate λ (flows/hour/UE). Each flow has an exponentially distributed duration with $1/\mu = 180$ s. The mean step value $\mathbb{E}[t]$ is set to 50 ms and 100 ms, simulating different CN congestion levels. We set $P_C^{(n)} = 354.44$ W, $P_{Tx}^{(n)} = 46$ dBm, $\delta = 21.45$ and $a_n = 2$ $\forall n \in \mathcal{N}$ [18]. Given the lack of approaches that perform resource allocation for OSPs in shared network, we compare the MTFP algorithm with a best effort (BE) approach that allocates randomly the RBs to the flows. In a simulation period of two hours, we study the effect of different numbers of UEs and OTT flow generation rates on delay and energy efficiency, comparing the MTFP algorithm with the BE approach and using different step values.

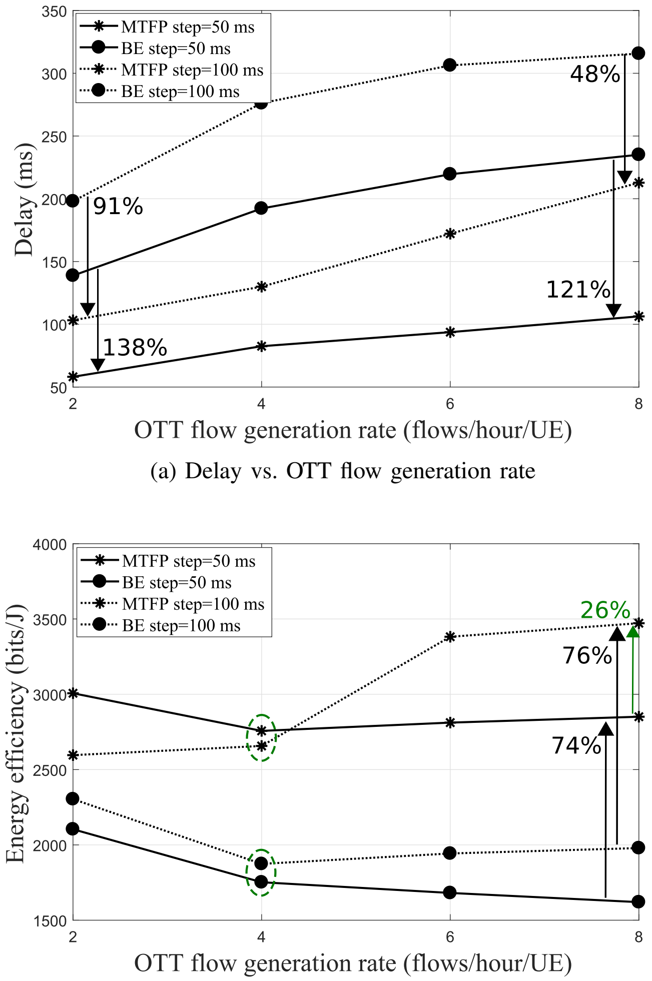
B. Simulation Results

We study the effect of number of connected UEs on the delay experienced by the flows, using the MTFP and BE approaches, considering $U = \{100, 200, \ldots, 500\}$ UEs.

As shown in Fig. 3 (a), the increase of the number of UEs leads to higher experienced delay, as more flows compete for resources. Still, MTFP achieves lower delay than BE, reaching a reduction of 137% and 112% for step values of 50 and 100 ms ($U = 500$), respectively, as RBs are allocated in a way that the highest possible number of flows are accommodated in each VS allocation round. In contrast, BE does not consider the OSPs' policies and allocates randomly the RBs to the flows. We also observe that, for both schemes, the delay is higher when the step value increases, reaching values up to 47% and 30% higher for MTFP and BE ($U = 100$), respectively. As the information exchange takes longer to be completed, each round lasts longer and the impact of lost rounds on the delay is higher, increasing the average experienced delay.

In Fig. 3(b), we see that MTFP outperforms the BE approach in terms of energy efficiency, reaching 93% and 96% increase, for step equal to 50 and 100 ms, respectively ($U = 100$). With MTFP, RBs are allocated in accordance with the flows' downlink channel conditions and QoS demands and the total data rate increases, improving the energy efficiency. As eNBs are always active and no switching off scheme is applied, i.e., P_n (Eq. (3)) is always considered, it is more efficient that more flows are served by each eNB n.

We next focus on the effect of different flow generation rates, assuming $U = 500$ UEs and $\lambda = \{2, 4, 6, 8\}$ flows/hour/UE. In Fig. 4(a), we observe that for both approaches, higher number of flows induces higher delay, as more flows concurrently request RBs in each round. As expected, the increase of step value affects the delay negatively. However, MTFP still performs better, resulting in delay values



(a) Delay vs. OTT flow generation rate

(b) Energy efficiency vs. OTT flow generation rate

Fig. 4: Effect of OTT flow generation rate

$121\% - 138\%$ and $48\% - 91\%$ lower than those of BE, for the step values of 50 ms and 100 ms, respectively.

Figure 4(b) shows that MTFP increases the energy efficiency by 74% and 76% ($\lambda = 8$) for step=50 and 100 ms, respectively, comparing to BE. Also, as λ increases, the energy efficiency improvement attenuates, as more RBs become occupied, providing the highest total data rate that is feasible per VS allocation round. With MTFP and $\lambda > 4$, although the higher step value (100 ms) produces higher delay (Fig. 4(a)), it improves the energy efficiency up to 26% ($\lambda = 8$), as it leads to fewer rounds with low RB utilization.

V. Conclusions

In this paper, a matching theoretic flow prioritization (MTFP) algorithm has been presented. Considering the network characteristics, i.e., different numbers of UEs generating flows, flow generation rates and VS allocation steps, we have extensively studied the performance of MTFP. The MTFP algorithm achieves lower delay and higher energy efficiency, compared to a best-effort scheme. In high data traffic cases, the longer duration of VS allocation rounds increases the delay but improves the energy efficiency. We believe that our work provides useful insights for network resource management that respects network neutrality and QoS demands without inflating the energy consumption of the mobile network.

Acknowledgement

This work has been supported by the research projects SPOT5G (TEC2017-87456-P), SEMIoTICS (780315) and CASPER (645393).

References

[1] "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016–2021," 2017.

[2] R. Bassoli, M. Di Renzo, and F. Granelli, "Analytical Energy-efficient Planning of 5G Cloud Radio Access Network," in *IEEE International Conf. on Commun.*, 2017, pp. 1–4.

[3] E. Datsika, A. Antonopoulos, N. Zorba, and C. Verikoukis, "Matching Game Based Virtualization in Shared LTE-A Networks," in *IEEE Global Commun. Conf.*, Dec. 2016, pp. 1–6.

[4] A. Ahmad, A. Floris, and L. Atzori, "QoE-centric Service Delivery: A Collaborative Approach among OTTs and ISPs," *Computer Networks*, vol. 110, pp. 168–179, 2016.

[5] P. Di Francesco, J. Kibiłda, F. Malandrino, N. J. Kaminski, and L. A. DaSilva, "Sensitivity Analysis on Service-Driven Network Planning," *IEEE Trans. on Networking*, vol. 25, no. 3, pp. 1417–1430, June 2017.

[6] A. Antonopoulos, E. Kartsakli, C. Perillo, and C. Verikoukis, "Shedding Light on the Internet: Stakeholders and Network Neutrality," *IEEE Commun. Mag.*, vol. 55, no. 7, pp. 216–223, May 2017.

[7] M. Yan, C. A. Chan, W. Li, C. L. I, S. Bian, A. F. Gygax, C. Leckie, K. Hinton, E. Wong, and A. Nirmalathas, "Network Energy Consumption Assessment of Conventional Mobile Services and Over-the-Top Instant Messaging Applications," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3168–3180, Dec. 2016.

[8] M. A. Marsan and M. Meo, "Energy Efficient Management of Two Cellular Access Networks," *ACM SIGMETRICS Performance Evaluation Review*, vol. 37, no. 4, pp. 69–73, 2010.

[9] B. Leng, P. Mansourifard, and B. Krishnamachari, "Microeconomic Analysis of Base-Station Sharing in Green Cellular Networks," in *IEEE Conf. on Comp. Commun.*, Apr. 2014, pp. 1132–1140.

[10] A. Bousia, E. Kartsakli, A. Antonopoulos, L. Alonso, and C. Verikoukis, "Game-theoretic Infrastructure Sharing in Multioperator Cellular Networks," *IEEE Trans. on Vehicular Technology*, vol. 65, no. 5, pp. 3326–3341, May 2016.

[11] A. Gudipati, D. Perry, Li E. Li, and S. Katti, "SoftRAN: Software Defined Radio Access Network," in *ACM SIGCOMM Workshop on Hot Topics in Software Defined Networking*, Aug. 2013, pp. 25–30.

[12] F. Capozzi, G. Piro, L. A. Grieco, G. Boggia, and P. Camarda, "Downlink Packet Scheduling in LTE Cellular Networks: Key Design Issues and a Survey," *IEEE Commun. Surveys & Tutorials*, vol. 15, no. 2, pp. 678–700, Second 2013.

[13] M. Srinivasan, V. J. Kotagi, and C. S. R. Murthy, "A Q-Learning Framework for User QoE Enhanced Self-Organizing Spectrally Efficient Network Using a Novel Inter-Operator Proximal Spectrum Sharing," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 11, pp. 2887–2901, Nov. 2016.

[14] T. D. Tran and L. B. Le, "Stackelberg Game Approach for Wireless Virtualization Design in Wireless Networks," in *IEEE International Conf. on Commun.*, 2017, pp. 1–6.

[15] G. Zhang, K. Yang, J. Wei, K. Xu, and P. Liu, "Virtual Resource Allocation for Wireless Virtualization Networks using Market Equilibrium Theory," in *IEEE INFOCOM WKSHPS*, Apr. 2015, pp. 366–371.

[16] J. W. Hatfield and P. R. Milgrom, "Matching with Contracts," *The American Economic Review*, vol. 95, no. 4, pp. 913–935, 2005.

[17] A. Nakao, P. Du, Y. Kiriha, F. Granelli, A. A. Gebremariam, T. Taleb, and M. Bagaa, "End-to-end Network Slicing for 5G Mobile Networks," *Journal of Inf. Processing*, vol. 25, pp. 153–163, 2017.

[18] F. Richter, A. J. Fehske, and G. P. Fettweis, "Energy Efficiency Aspects of Base Station Deployment Strategies for Cellular Networks," in *IEEE Vehicular Technology Conf. Fall*, Sept. 2009, pp. 1–5.