LTE Offloading: When 3GPP Policies Are Just Enough

LTE Offloading: When 3GPP Policies Are Just Enough

Francesco Malandrino
Politecnico di Torino, Torino, Italy

Claudio Casetti
Politecnico di Torino, Torino, Italy

Carla-Fabiana Chiasserini
Politecnico di Torino, Torino, Italy

***Abstract*—We investigate the effectiveness of the 3GPP offloading policy framework, called Access Network Discovery Selection Function (ANDSF). We consider geographical areas where both LTE and WiFi are available and present a model describing multi-RAT networks as visible by the operator, as well as the offloading policy rules that apply to them. Our model captures user behavior and allows us to express any 3GPP policy in a compact and convenient way. We use the model to develop a dynamic offloading scheme, which is fully compatible with 3GPP specifications and dynamically adapts to changing traffic patterns. We analyse it in a typical two-tier 3GPP scenario, comparing its performance to those of three alternate offloading strategies. We also investigate the effectiveness for data offloading of the current and proposed features of 3GPP ANDSF.**

I. Introduction

Increase in the traffic demand by mobile users is one of the most serious challenges faced by today's cellular networks. For example, the authors of [1] have raised the concern that traffic demand could increase so much as to endanger the profitability of cellular networks.

One of the most popular approaches to face this issue is *data offloading*, i.e., diverting traffic from the cellular infrastructure onto other networks in a multi-Radio Access Technology (RAT) system. The RATs to offload to include WiFi networks [2]–[4] and device-to-device communication [5], [6]. In particular, some operators are aggressively embracing WiFi offloading as a cost-effective solution to increase capacity and data rates of their networks [7].

3GPP Releases 11 and 12 [8] do include an offloading policy framework, supporting policies for both inter-system mobility and inter-system routing. Inter-system mobility policies describe how User Equipments (UEs) should select the network (LTE, WiMax, WiFi) to access when they can connect only to one. Inter-system routing policies, instead, describe how the UEs that can connect to multiple networks should route their traffic through the different radio interfaces. These policies are network-based, i.e., they are imparted by the operator's network for UEs to follow. Also, policies cannot depend on the UE profile, but they target specific geographical areas, time of day, and type of content.

In this work, we deal with 3GPP inter-system routing policies and focus on how the 3GPP framework can be used for dynamic data offloading. Our contribution is threefold. Firstly, we provide a compact model of 3GPP policies which, although remarkably simpler, is completely compatible with LTE specifications. Secondly, we use such a model to define a practical, dynamic offloading strategy. This strategy capitalizes on the limited knowledge that the operator can have about other networks. We show that it outperforms trivial rules like "use WiFi if possible and LTE otherwise", as well as more sophisticated schemes. Finally, and perhaps most interestingly, we rank the features of 3GPP policies according to their capability to improve the user experience.

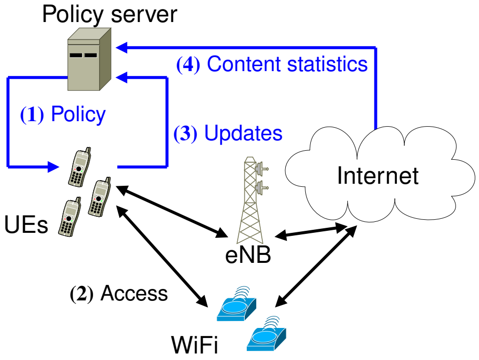


Fig. 1. System scenario. The operator-owned policy server pushes a policy to users (1). Users access the Internet (2) following such a policy, and report to their operator about the performance they are experiencing (3). The operator also collects performance statistics from the cellular infrastructure, and use them to refine the current policy if needed.

The latter contribution is especially relevant from a practical viewpoint. Current LTE UEs are not designed to implement policies, and upcoming ones will require to know which features specified by 3GPP yield the most significant benefits.

The rest of the paper is organized as follows. We outline the multi-RAT scenario and briefly describe how 3GPP policies work, in Sec. II. Both network system and 3GPP policies are modeled in Sec. III. Our dynamic offloading strategy, which is fully compliant with 3GPP policy specifications, is proposed in Sec. IV. The performance of our solution is compared to that of three other strategies in Sec. VI, in the scenario described in Sec. V. Finally, we draw our conclusions in Sec. VII.

II. System scenario and 3GPP policies

We consider a multi-RAT network scenario, as depicted in Fig. 1. UEs attempting to access the Internet may connect to different radio interfaces, namely, LTE eNBs and WiFi hotspots. We refer to them as Points of Access (PoAs). In this work, we focus on the more interesting case where WiFi hotspots are privately-owned access points (either domestic or commercial), with which the operator has an agreement [9]. The case where they are owned by the operator itself could be easily considered as well.

According to the standard [8], UEs rely on operator's policies in order to select the PoA to use when routing their traffic flows. We therefore envision the presence of operator-owned policy servers (see Fig. 1), which are in charge of collecting and processing useful information, and of issuing the policy to be applied. In particular, such servers (i) are

978-1-4799-4937-3/14/$31.00 ©2014 IEEE

aware of the PoAs available in the multi-RAT network area, and (ii) collect performance reports from UEs and cellular infrastructure. Based on such information, the policy servers determine the policy to be used for each type of content, in different areas of the network and in well-defined time intervals. Such areas and intervals correspond to homogeneous traffic conditions, e.g., rush hour traffic in a busy freeway, or Saturday afternoon at a shopping mall. The policies designed by the operator are then pushed to the UEs, along with the list of PoAs that are available in their proximity. UEs periodically return feedback to the server, including their own position, the type of traffic they have routed through a PoA and the throughput they have experienced. We remark that servers can dynamically adapt the policies so as to reflect network conditions and account for user feedback.

Below, we provide an outline of 3GPP policies, which is useful to understand the model we present in the next section.

A. Policies and ANDSF

A policy mandates which access technology (e.g., WiFi or LTE) and, possibly, the specific PoA a user should connect to for a given data transfer. This information is conveyed to the users through one or more *rules*, called Access Network Discovery and Selection Functions (ANDSFs). Each rule applies subject to some *conditions* concerning the UE. Such conditions may concern: its geographic location; the network coverage, e.g., which WiFi hotspots are available; current date and time; the host and ports to/from which it transfers data; the type of service being provided. Furthermore, rules may have different priorities: the highest-priority rule, among those whose conditions are satisfied, is enacted. As an example, the semantics of a policy could be translated as: "if you are in a shopping mall at peak time, and you are not under the coverage of one of the operator's picocells, then use WiFi to download web content if possible, falling back to LTE macrocells if needed". In this case, the policy includes four conditions (location, time, network coverage and content type) and two rules (use WiFi and fall back to LTE). Both rules are valid if all conditions hold. Also, the rule selecting WiFi has higher priority over the other.

In more detail, a policy can be described as a set of ANDSF rules, whose structure is depicted in Fig. 2. The figure reflects the policy description given in [8] through XML nodes; a thick solid line highlights the rule features that are most relevant to our study. Such features are detailed below.

RulePriority: number stating rule priority within the policy.

PrioritizedAccess: it expresses the priority of available networks. The *AccessTechnology* sub-node indicates the priority given to a network technology, e.g., LTE or WiFi, while the *AccessId* sub-node refers to a specific PoA.

ValidityArea: it states the conditions related to the user position that must hold for the rule to be enacted. Its sub-nodes indicate whether such conditions refer to the availability of some network technology or specific PoAs in the area where the user is (e.g., *3GPP_Location*, *WiFi_Location* and their sub-nodes: *LAC* (location area code) and *SSID* (service set identification)). The rule may also require the UE to be at a specific location, e.g., corresponding to a stadium or a mall (*Geo_Location* sub-node).

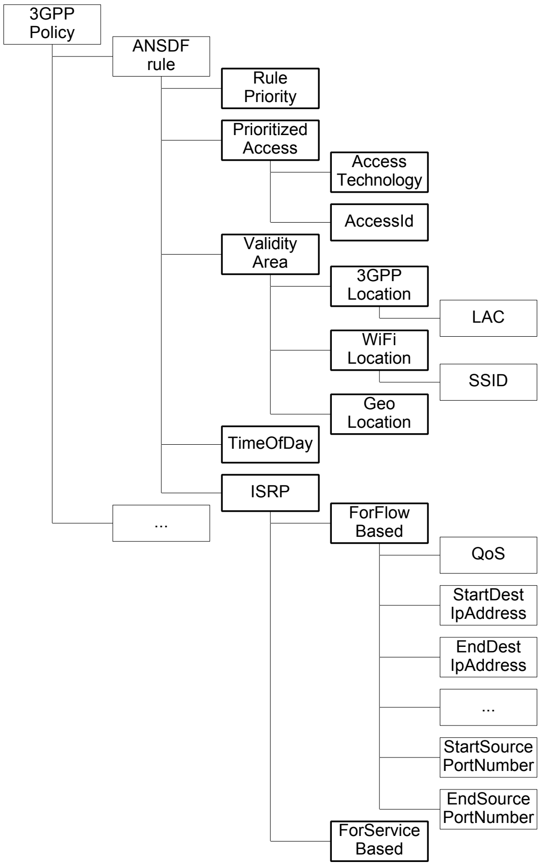


Fig. 2. Policy and ANDSF rule structure, as specified in 3GPP R12. Refer to [8] for the full description.

TimeOfDay: it indicates the time interval (e.g., in a day or week) during which the rule applies.

ISRP (Inter-System Routing Policy): it expresses the conditions related to the data to be transferred. It refers either to the type of traffic flow (sub-node *ForServiceBased*), or to the QoS traffic class and the source/destination hosts and ports (sub-node *ForFlowBased*).

The application of the rules included in the policy issued by the policy server follows the steps outlined in Fig. 3. UEs first sort rules by priority. Then, for each rule, they check the conditions on time, location, network availability, and traffic flow. If all of them are met, the rule is applied.

III. Model and policy definition

In this section, we describe how we model the multi-RAT network (Sec. III-A) and the 3GPP policies (Sec. III-B) described above, as well as the user behavior (Sec. III-C).

A. Multi-RAT networks

We build a fairly simple model, deliberately ignoring those aspects that cannot be known to operator policy servers when they define their offloading policies.

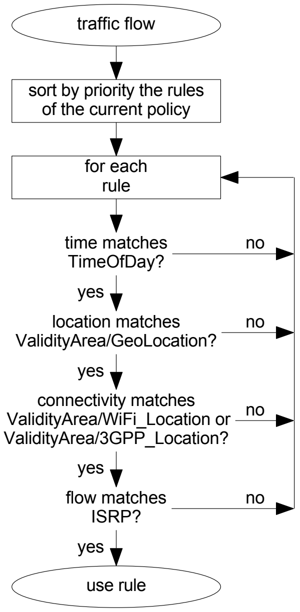

traffic flow
sort by priority the rules of the current policy
for each rule
time matches TimeOfDay?
no
yes
location matches ValidityArea/GeoLocation?
no
yes
connectivity matches ValidityArea/WiFi_Location or ValidityArea/3GPP_Location?
no
yes
flow matches ISRP?
no
yes
use rule

Fig. 3. Flow diagram representing how policy rules apply.

We denote by $\mathcal{U}$ the set of users, or, equivalently, UEs, that are present in the area covered by the multi-RAT network, and we divide the network area into a set $\mathcal{T}$ of non-overlapping tiles. Time is divided into a set $\mathcal{K}$ of *time slots*. These are periods during which external conditions (e.g., user mobility and density, type of requested content, WiFi operational hours) are expected to remain homogeneous. The set of content items is denoted by $\mathcal{C}$.

We also define $\mathcal{A}$ as the set of available PoAs. It includes WiFi hotspots, as well as cellular base stations (e.g., LTE eNBs). Each PoA $a \in \mathcal{A}$ has coverage area $T_a \subseteq \mathcal{T}$. We indicate by $n_c^k(a)$ the average number of operator's subscribers interested in content c that are simultaneously connected to PoA a, during slot k. The policy server can easily compute such number based on the statistics collected from the operator-controlled networks (e.g., LTE) and on the feedback received from the users whose traffic has been offloaded to privately-owned hotspots. However, note that, if a is an LTE eNB, $n_c^k(a)$ coincides with the actual number of users connected to a and interested in content c. On the contrary, if a is a privately-owned hotspot, there may be other users connected to a (in addition to the $n_c^k(a)$ operator's subscribers), of which the policy server is unaware.

Two important quantities depend on the number of users simultaneously connected to a PoA. The first is the connection establishment time, i.e., the time required to join a PoA and start operating within that network. If the connection establishment takes longer than a given (technology-specific) timeout, the PoA is considered to be unavailable. As already mentioned, in this work we do not address issues related to vertical handovers or fast network connection techniques, as our focus is uniquely on data offloading policies. However, our model could easily account for the above aspects.

The second quantity is the throughput enjoyed by users after they have successfully joined a PoA. This quantity is PoA-specific, e.g., domestic WiFi hotspots may serve different numbers of users (hence experience different traffic loads) or have cable subscriptions with different transfer speeds. Unless otherwise specified, we consider that the policy server (i.e., the operator) is unaware of the real-time traffic load in privately-owned WiFi hotspots. Also, users do not sense in advance the throughput they could receive from a hotspot.

TABLE I
MAPPING BETWEEN OUR MODEL AND ANDSF RULES

Feature	ANDSF	Model
Dependency on time	*TimeOfDay* node	x-values depend on time slot $k \in \mathcal{K}$
Dependency on location	*ValidityArea/ Geo_Location* node	x-values depend on tile $t \in \mathcal{T}$
Dependency on service	*ISRP/PerServiceBased* node	x-values depend on content item $c \in \mathcal{C}$
Dependency on network technology or PoA availability	*ValidityArea/ WiFi_Location* and similar nodes, and their subnodes, e.g., LAC and SSID	x-values depend on tile $t \in \mathcal{T}$, time slot $k \in \mathcal{K}$ and PoA $a \in \mathcal{A}$
Priority of network technology or PoA	*Prioritized_Access/ AccessTechnology* or *AccessId* subnode (optional)	x-values depend on PoA a; they can also be the same for all PoAs using the same network technology

B. Policies

Given the system model defined above, we introduce a set of tile-, time-, content- and PoA-specific values $x_c^k(a,t)$. These quantities express the fraction of data of content c, originating from users in tile t, that should be transmitted through PoA a in time slot k. Thus, $0 \leq x_c^k(a,t) \leq 1$, $\forall\, t \in \mathcal{T}, k \in \mathcal{K}, c \in \mathcal{C}, a \in \mathcal{A}$. In the following, we first show that the way we define the x quantities reflects the dependency of ANDSF rules on system parameters. Then, we prove by construction that any set of x-values corresponds to a valid 3GPP policy. Thus, finding an efficient policy actually means to determine the x-values that lead to user satisfaction.

1) Mapping dependencies: Table I shows how different features (first column) are expressed through the ANDSF rule syntax of 3GPP policies (second column) and through the control variables of our model (third column). The first three rows, representing the fact that a policy changes according to time, location and service (i.e., content), are quite straightforward. The last two call for a more detailed discussion.

Firstly, the dependency of 3GPP policy on the presence of a network technology, or a PoA, makes a rule hold only if such technology, or PoA, is available at the user location. This feature allows a policy server to issue rules such as "if a WiFi hotspot is available, then try WiFi first". The fact that our x-values depend on the PoA lets us easily account for that. Note that, when rules do not refer to specific PoAs but to a network technology, we just assign the same x-value to all PoAs that use the same network technology.

Secondly, rules may assign different priorities to network technologies or PoAs (see the element *PrioritizedAccess* and its subnodes). Our x-values are not priorities, but fractions. In order to reflect such ANDSF feature, we first consider the highest-priority technology, or PoA. If available, the x's corresponding to it will be set to positive values, while the x quantities referring to other technologies/PoAs will be set to zero. As a consequence, transmissions will be attempted only through the highest-priority network. If instead such network is unavailable, we repeat the same procedure considering the second technology/PoA in the priority list, and so on. In this way, priorities are always honored.

Finally, and most importantly, we remark that 3GPP policy rules have no direct way to express how the traffic should be *split* across several available network technologies or PoAs. We are able to do so by exploiting client port ranges (nodes *ISRP/ForFlowBased/StartSourcePortNumber* and *ISRP/ForFlowBased/EndSourcePortNumber*). The client port number assigned to each traffic flow is picked randomly between a minimum and a maximum value, namely, 49152 and 65535 Therefore, if we want 25% of the traffic to be transmitted through WiFi and 75% through LTE, we can route through the former those flows with client port in the range $49152-53247$, and through the latter all the others. In this way, we can compile policies that specify how much traffic should be offloaded toward a given network or PoA. This additional feature is fully compliant with the 3GPP syntax and allows the definition of policies that can split traffic through different PoAs with, virtually, any desired granularity.

2) From x-values to a 3GPP policy: Algorithm 1 shows how to translate a set of x-values, i.e., our representation of a policy, into a 3GPP policy as defined in [8]. It also proves, by construction, that any policy that can be expressed in terms of x-values is a valid 3GPP policy.

The algorithm takes the x-values as input (line 0). Then, for each content, time slot and tile, we create a rule stub (line 3) and set the corresponding validity area, time of day, and service attributes (lines 4-6). These attributes directly map onto the elements of our system model, as shown in Tab. I.

Next, in line 8, we craft the actual PoA-specific rules. We start by making a copy of the stub rule we created earlier (line 9). In line 10, we set the *AccessId* field, i.e., the PoA to which the rule refers, to a. In lines 11-15, we map each of the $x^k_c(a,t)$ values onto a set of integers in the $49152-65535$ range. The newly-created rule is then added to the policy (line 16). Finally, the policy is ready to be pushed toward the UEs (line 17).

C. User behavior

A UE selects the PoA to connect with, according to the policy rules received from the server. It therefore proceeds as shown in Fig. 3. We stress that verifying the matching between the traffic flow and the *ISRP* node also implies checking whether the client port falls in the value range specified by the rule. In this way, considering the overall user traffic, the desired amount of data routed through the different PoAs will meet the fractions x that originated the policy.

Algorithm 1 Mapping x-values onto a 3GPP policy

```
Require: $x^k_c(a,t), \forall c \in \mathcal{C}, k \in \mathcal{K}, a \in \mathcal{A}, t \in \mathcal{T}$
 1: pol←new Policy()
 2: for all $c \in \mathcal{C}, k \in \mathcal{K}, t \in \mathcal{T}$ do
 3:     stub←new Rule()
 4:     set_node(stub, ValidityArea/Geo_Location, $t$)
 5:     set_node(stub, TimeOfDay, $k$)
 6:     set_node(stub, ISRP/PerServiceBased, $c$)
 7:     port_so_far←MinPortNo
 8:     for all $a \in \mathcal{A}$: $x^k_c(a,t) > 0$ do
 9:         rule←copy(stub)
10:         set_node(rule, PrioritizedAccess/AccessId, $a$)
11:         begin_port←port_so_far
12:         end_port←begin_port+ $\lceil$MaxPortNo $\cdot x^k_c(a,t)\rceil$
13:         set_node(rule, ISRP/PerFlowBased/StartSourcePortNumber, begin_port)
14:         set_node(rule, ISRP/PerFlowBased/EndSourcePortNumber, end_port)
15:         port_so_far←end_port+1
16:         add_rule(pol, rule)
17: return pol
```

Once the UE has selected the PoA, it tries to connect to it. If no connection can be established after a timeout, the attempt is declared as failed and another PoA is selected, again according to the policy rules. The UE keeps trying until the transfer succeeds or all available PoAs have been polled. The extension to the case where the whole procedure is aborted after a given timeout is straightforward. The UE periodically returns a feedback including the PoA(s) it could not connect with (if any), its own location, the type of traffic transferred through each PoA and the throughput experienced there during the connection time.

IV. The dynamic offloading scheme

Expressing a policy through the x-values introduced above is very convenient and allows us to represent any data offloading strategy. Here, we propose a practical, dynamic offloading scheme, which, within a given time slot k, makes the system quickly adapt to changes in the traffic load conditions of the PoAs, as well as in their availability. We remark that the proposed strategy is fully compliant with 3GPP policy specifications.

While the behavior and performance of operator-controlled cellular networks can be predicted, this is not true for other networks used for data offloading. The only established fact about PoAs where data are offloaded is that association time and per-user throughput are monotonic with the number of UEs, i.e., their behavior does not improve by adding more users. It follows that the policy server should increase the number of UEs connected to a PoA that is performing well, while it should remove UEs from a network that is performing poorly. Lacking any additional information, and inspired by the well-known additive-increase-multiplicative-decrease (AIMD) behavior of the TCP congestion control, we propose the scheme in Algorithm 2.

Consider that the policy server has to dynamically identify the policy to be enacted during time slot k (e.g., weekday afternoon). We define as *iteration period* the time interval between two consecutive updates of the policy (or, equivalently, of the x-values). An iteration period is at most as long as the time slot. At every update, the policy server bases its decision on the feedback received from the UEs about their performance during the previous iteration. Let us denote by $\delta_c^k(a,u)$ the throughput that user u has experienced for content c while being connected to PoA a, and that has been included by u in its feedback. Clearly this value refers to the UE performance under the previous policy. Also, let $\mathcal{B} \subseteq \mathcal{A}$ be the set of LTE eNBs that are available in the multi-RAT network area; similarly, $\mathcal{H} \subseteq \mathcal{A}$ denotes the set of available WiFi hotspots.

The AIMD offloading algorithm takes as input the time slot k, the average number of users, $n_c^k(a)$, simultaneously connected to a PoA during slot k and the throughput value, $\delta_c^k(a,u)$, sent by each UE in its feedback. For each tile, we identify the LTE eNB whose signal is the strongest, i.e., $b^\star$ in line 2. Then, for every WiFi hotspot h covering the tagged tile, we compare the average per-user throughput provided by h in slot k to that offered by $b^\star$ (line 4). If h performs better than $b^\star$, its x-value is incremented by the quantity $1/n_c^k(h)$ (line 5); otherwise, it is halved (line 7). Note that $1/n_c^k(h)$ corresponds to the throughput fraction that the average user experiences for content c while being connected to WiFi hotspot h. Thus, intuitively, in line 5 the traffic that should be routed through h is increased by one flow.

Algorithm 2 The AIMD dynamic offloading scheme

```
Require: $k$, $\delta_c^k(a,u)$, $n_c^k(a)$, $\forall a \in \mathcal{A}, c \in \mathcal{C}$
1: for all $t \in \mathcal{T}, c \in \mathcal{C}$ do
2:     $b^\star \leftarrow \arg\max_{b \in \mathcal{B} \wedge t \in T_b} \mathrm{RSS}(b)$
3:     for all $h \in \mathcal{H} \colon t \in T_h$ do
4:         if $\frac{1}{n_c^k(h)} \sum_u \delta_c^k(h,u) \geq \frac{1}{n_c^k(b^\star)} \sum_u \delta_c^k(b^\star,u)$ then
5:             $x_c^k(h,t) \leftarrow x_c^k(h,t) + 1/n_c^k(h)$
6:         else
7:             $x_c^k(h,t) \leftarrow x_c^k(h,t)/2$
```

Once the x-values have been updated for each content, tile and PoA, the policy server uses Alg. 1 to map the x-values onto the new policy to be issued to UEs.

V. Reference scenario and benchmark strategies

In this section, we describe our reference scenario, remarking, however, that our scheme works for any topology and under any assumption on connection establishment time and user throughput.

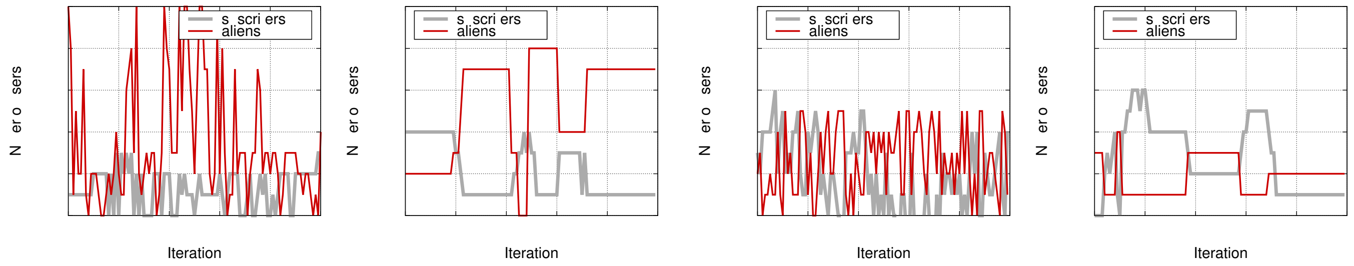
We focus on a two-tier network scenario covering 12.34 km^2, and including 57 LTE macrocells and 4 WiFi hotspots per macrocell. The macrocell deployment is taken from the LTE scenario typically used within 3GPP for performance evaluation [10]. Macrocell eNBs operate over a 10 MHz band at 2.6 GHz. eNBs are located at 19 tri-sectorial sites, at a distance of 500 m from each other. According to ITU specifications [11], we assume that eNB antennas are 25-m high and transmit at a power level of 43 dBm. With regard to the WiFi technology, we consider privately-owned hotspots, both domestic and commercial, using the IEEE 802.11n standard. They operate over a 20-MHz band at 5.2 GHz. The 802.11n antennas are omnidirectional and at a height of 2.5 m off the ground; they irradiate power at 15 dBm. Over the multi-RAT network area there are 3420 uniformly-distributed operator's subscribers, who move according to the cave-man model [12] with average speed of 1 m/s. The antenna of the UE is assumed to be, on average, at 1.5 m off the ground. UEs are involved in the download of one of the following types of content, each of them representative of a different QoS class: a 100-Mbyte, delay-tolerant data item (hereinafter referred to as DT), and a 20-Mbyte data item requiring 500 kb/s as minimum guaranteed rate (hereinafter referred to as GR). Content items are downloaded one at a time, and UEs return feedback once every iteration period. We assume the iteration period to be equal to 1 minute.

We assume that all performance metrics related to the user traffic within the LTE network are perfectly known to the policy server. In particular, we compute the throughput experienced by a UE connected to an eNB as follows. We use the ITU signal propagation model for urban environment [11] to obtain the values of SINR corresponding to different UEs. Then, by exploiting the experimental results in [13], for each UE, we map the SINR onto throughput per radio resource. Finally, for each GR traffic flow, we consider that enough radio resources are reserved so as to provide it with a 500 kb/s service rate (see the LTE guaranteed bit rate (GBR) bearer). Round-robin scheduling is assumed to be implemented at the eNBs, hence the remaining resources are evenly shared among all connected UEs. As for the connection establishment time, it is neglected as cellular interfaces are always active on UEs.

In the case of WiFi hotspots, as mentioned, the policy server cannot have complete, real-time knowledge of their status: it can only leverage the feedback from subscribers whose traffic has been offloaded to such PoAs. In order to derive our numerical results, we compute the data rate employed by a user at a given distance from a WiFi hotspot, by adopting the propagation model in [14, Ch. 5] and assuming the modulation and coding scheme MCS 0. Also, we set the data payload to 1 kB, the transmission opportunity limit for GR flows to 3 ms and the other MAC-layer parameters to the values reported in [14, Ch. 12]. The user throughput is then computed by considering the average behavior of the 802.11n MAC layer. Based on recent experimental results [15], we assume the association time to be Gaussian-distributed with mean and standard deviation equal to 5 s and 3 s, respectively.

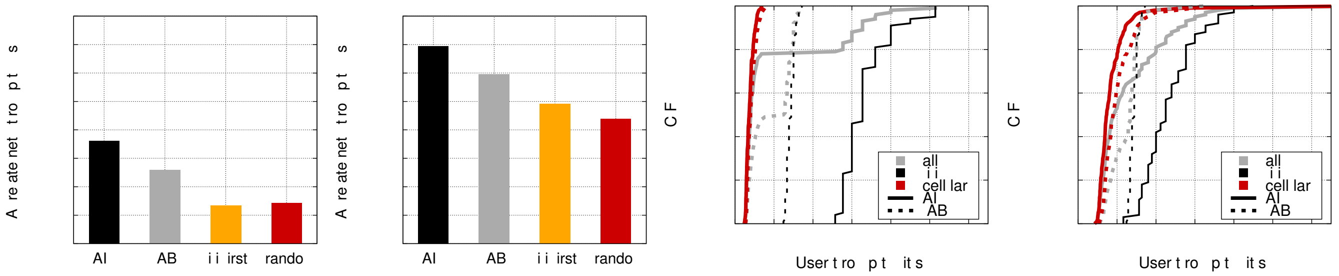
In the next section, we evaluate the performance of the AIMD dynamic offloading scheme and compare it to that of the following alternate strategies.

Multi-Armed Bandit (MAB). We formulate data offloading in multi-RAT networks as a multi-armed bandit problem [16]



(a) Commercial, freq.=1 (b) Commercial, freq.=1/10 (c) Domestic, freq.=1 (d) Domestic, freq.=1/10

Fig. 4. Ability of AIMD to adjust the number of offloaded traffic flows to the number of aliens in a WiFi hotspot, as the latter number varies once every 1 (a,c) and 10 (b,d) iteration periods. Commercial (a,b) and domestic (c,d) scenarios.



(a) Commercial (b) Domestic (c) Commercial (d) Domestic

Fig. 5. Aggregate network throughput and CDF of the user throughput for the different offloading schemes. Commercial (a,c) and domestic (b,d) scenarios.

where "arms" correspond to different policies, i.e., sets of x-values. The MAB scheme is obtained by solving such problem through the ϵ-greedy algorithm, which has been shown [17] to consistently perform close to the optimum.

WiFi-first. UEs always connect to the available WiFi network from which they receive the strongest signal, if any is available. Otherwise, they connect to the LTE network. This is the strategy commonly implemented in current smartphones.

Random. Among the available PoAs, UEs pick one to connect to with uniform probability.

VI. Results

We consider two different case studies: in the former all WiFi hotspots are commercial, in the latter they are domestic. We name the scenarios after the type of hotspots they include. In either case, there is a number of users connected to the hotspots that are not operator's subscribers and of which the policy server is unaware. We refer to such users as *aliens* and assume that they have the same behavior and request the same content items as operator's subscribers do. In the commercial scenario, we take the time evolution of the number of aliens per hotspot from the measurements in [18]. This trace mostly includes open hotspots offered by shops and restaurants, where the average number of connected users (aliens) is 5. In the domestic case, instead, the number of aliens is uniformly distributed between 0 and 5, hence its average is significantly lower than in the commercial scenario.

We start by considering only one type of service, namely, DT content. Under these traffic conditions, we investigate the ability of our AIMD offloading scheme to adapt to variations in the traffic load of WiFi hotspots. Fig. 4 shows the number of operator's subscribers (i.e., number of traffic flows) that are offloaded to one of the 228 WiFi hotspots over time (thick, grey curves), as the number of aliens at the PoA varies (red, thin curves). Similar results are obtained for the other hotspots; they are omitted for brevity. The plots differ by type of scenario (commercial and domestic) and by frequency with which the number of aliens changes (once every iteration period and once every ten periods). Both scenarios highlight that AIMD swiftly increases the fraction of offloaded traffic when aliens leave the PoA, and decreases it as new aliens become active. However, the range of choices is more limited when AIMD is run in a commercial scenario. The large number of aliens overcrowding the hotspot leaves little maneuvering room to AIMD. Thus, even though the number of aliens fluctuates wildly between 0 and 10, AIMD is often forced into selecting the fewest possible users (i.e., 0-1) by its multiplicative decrease behavior.

Next, we fix the frequency with which the number of aliens varies to once every iteration period, and we compare the performance of AIMD to that of the three benchmark strategies (see Sec. V). We note up front that, under the domestic scenario, the total throughput provided by LTE and WiFi to operator's subscribers is always higher than in the commercial case. This is due to the larger spare bandwidth at domestic hotspots: on average, the whole WiFi network has 1140 alien users in the commercial scenario and 648 in the domestic one. Figs. 5(a) and (b) show that our solution consistently ranks above all other offloading schemes. This is hardly surprising for the WiFi-first and Random schemes: disregarding contention and congestion issues when selecting a WiFi network is hardly a sensible choice. In particular,

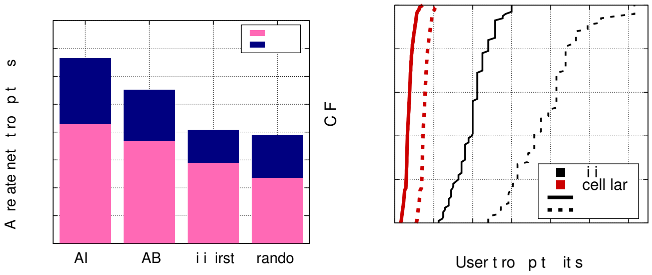


Fig. 6. Different types of content: Aggregate network throughput (left) and CDF of the user throughput with AIMD (right) in the commercial case.

recall that the WiFi-first strategy connects to WiFi whenever under coverage, thus overloading the involved hotspots. This is particularly evident in the more crowded commercial scenario where WiFi-first yields poorer results than the Random scheme. Having said that, we remark that AIMD outperforms the more clever strategy based on the MAB approach. This is due to the fact that the throughput of WiFi hotspots does not exhibit any handy statistical property, such as memoryless behavior, which in [17] is shown to be amenable to a MAB approach. Instead, AIMD leverages the monotonicity of throughput as a function of the offered load, thus crafting a simple, yet effective, policy.

Figs. 5(c) and (d) present the cumulative distribution function (CDF) of subscribers' throughput obtained through AIMD and MAB, in the commercial and domestic scenarios, respectively. From the plots, it is clear that performance cannot improve without affecting fairness. In both commercial and domestic scenarios, UEs using LTE (red curves) experience lower throughput than those whose traffic is offloaded toward WiFi (black curves). This unbalancing is not always a consequence of policies. WiFi is not a viable choice for UEs too far away from a hotspot, so LTE eNBs tend to be much more crowded than WiFi hotspots. As for the gain of AIMD over MAB, this is due to the fact that AIMD offloads to WiFi just the amount of traffic that can be supported by hotspots without experiencing an exceedingly high number of packet collisions. In other words, AIMD always makes sure to back off from overloading WiFi hotspots, so that they can efficiently serve the connected UEs (as well as the aliens). It follows that, under the AIMD scheme, 60% of the UEs connected to WiFi enjoy a throughput between 6 and 10 Mb/s in the commercial case, and between 4 and 8 Mb/s in the domestic case (black, solid curves). On the contrary, under MAB, the user throughput is about 2.5 Mb/s for almost all of the UEs connected to WiFi (black, dotted curves), in both scenarios. Under AIMD, the difference between commercial and domestic is due to the fact that AIMD tends to offload to WiFi hotspots more traffic in the former case, as the available bandwidth is much higher (fewer alien users). Recall however that the aggregate throughput with domestic hotspots is higher than with commercial ones (see Figs. 5(a) and (b)). Due to the heavier offload executed by AIMD in the domestic scenario, LTE throughput is higher than in the commercial case (red, solid curves).

Fig. 6 presents similar performance metrics, but now users ask for DT and GR items with equal probability. Due to lack of room, only results for the commercial case are shown, as this is the more challenging scenario for GR content. The left plot again highlights the high values of aggregate throughput provided by AIMD and the low values obtained by WiFi-first and Random. Interestingly, the comparison with Fig. 5(a) highlights that WiFi-first now outperforms the Random strategy. Indeed, WiFi networks (which are mainly used by WiFi-first) become more efficient thanks to the long transmission opportunity assigned to GR traffic. For the same reason, the aggregate throughput obtained under any of the considered schemes is significantly larger than with DT content only (by about 1 Gb/s for AIMD). Also, due to the priority awarded to GR traffic by eNBs and WiFi hotspots, 2/3 of such throughput are enjoyed by GR flows.

The behavior of the user throughput in Fig. 6 (right) confirms the above findings about AIMD. Additionally, it further underscores that AIMD always offloads toward WiFi a sufficiently low number of flows so as to avoid overloading the hotspots. As a result, 80% of the offloaded subscribers downloading GR traffic (black, dotted curve) get a throughput larger than 6 Mb/s, which is much higher than the one provided by LTE (red, dotted curve). Clearly, the excellent performance of GR flows comes at the expenses of DT traffic, both in LTE and WiFi (red and black, solid curves, resp.).

Performance of partial or enhanced ANDSF features. We now aim at assessing the impact of the different ANDSF features on the offloading performance. We focus on AIMD and MAB, and analyse the aggregate throughput in the multi-RAT network when some features are not enabled. We label as "no-space" the case where decisions (i.e., x-values) cannot depend on tiles, as "no-time" the case where the policy is never updated during the time slot, and as "no-others" the case where decisions about a user traffic flow only depend on the feedback that has been provided by the user itself. Note that, given a strategy, the latter case has UEs autonomously applying it without the support of the policy server.

Fig. 7 (left plots) shows that, with respect to the case where all features are implemented (labeled as "all"), disabling the space-dependency has a negligible impact in the commercial scenario. A more noticeable loss of performance can be observed in the domestic case: our strategy loses around 300 Mb/s in aggregate throughput, while MAB about 200 Mb/s. Indeed, the arrival/departure process of aliens is quite homogeneous across the various commercial hotspots, hence the same offloading policy can be applied to all of them without degrading the system throughput. Domestic hotspots instead are characterized by fewer aliens, and even a small variation in their number represents a non-negligible relative change. Thus, adapting the policy to the traffic load of a specific geographical area becomes more relevant. The effect with respect to the "all" case is similar, but more evident, when time dependency is removed. This is due to the time-varying nature of the user traffic load, which impairs the benefit of a policy that does not adapt to new conditions as they set in. However, the most severe performance degradation is observed when feedback from other users is not exploited, i.e., in the case where each UE can decide based on its own experience only. Indeed, both AIMD and MAB can promptly react to conditions changes only if a significant amount of information (i.e., feedback) is available.

We then consider the following possible additional feature to the standard ANDSF: UEs can "monitor" traffic in nearby WiFi hotspots while being connected to LTE, and, thus, assess the number of concurrent users (aliens and subscribers) connected to each of them.Such information is reported to (and exploited by) the policy server. As we can see from Fig. 7 (right plots), allowing such a capability brings little benefit (beside being a questionable practice in terms of energy consumption and privacy).

In summary, not only does AIMD outperform the alternate strategies when full-fledged ANDSF rules are implemented, but it is also more effective even in case of partial implementation. Furthermore, space-dependency showed to be the least effective feature thus suggesting that it could be neglected in early implementations. Of utmost importance, instead, is the feedback from all UEs, complementing the role of the policy-issuing server with that of collector of information. As for possible additions to the standard, the monitoring feature may not be worth investing in, given the little improvement, the high energy cost for UEs and, finally, the privacy concerns. Furthermore, our AIMD scheme with the standard ANDSF performs better than the MAB strategy with monitoring. We can thus conclude that optimizing the strategy with which policies are determined is a more promising approach than adding brand-new capabilities to UEs.

VII. Conclusions

Much though 3GPP standards already provide for offloading strategies to relieve the growing congestion of cellular networks, research efforts have largely neglected investigating their effectiveness. This paper is a first attempt at bridging such gap by introducing a simple model that captures the nuances of 3GPP offloading policies. We used such model to propose a practical, dynamic offloading scheme and benchmark its performance against three other strategies. As a side-product of our investigation, we could assess how different elements of a policy can provide high overall throughput for users of a cellular operator.

Acknowledgment

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

References

[1] Credit Suisse, "U.S. wireless networks running at 80% of capacity," http://benton.org/node/81874, 2011.

[2] K. Lee, J. Lee, Y. Yi, I. Rhee, S. Chong, "Mobile data offloading: How much can WiFi deliver?," *IEEE/ACM Trans. on Networking,* vol. 21, no. 2, pp. 536–550, 2013.

[3] M. Bennis *et al.*, "When cellular meets WiFi in wireless small cell networks," *IEEE Comm. Mag.*, vol. 51, no. 6, 2013.

[4] F. Malandrino, C. Casetti, C.-F. Chiasserini, M. Fiore, "Content downloading in vehicular networks: What really matters," *IEEE INFOCOM*, 2011.

[5] X. Bao, Y. Lin, U. Lee, I. Rimac, R. Choudhury, "DataSpotting: Exploiting naturally clustered mobile devices to offload cellular traffic," *IEEE Infocom MiniConference,* 2013.

[6] F. Malandrino, C. Casetti, C.-F. Chiasserini, Z. Limani, "Fast resource scheduling in HetNets with D2D support," *IEEE INFOCOM*, 2014.

[7] R. Metz, "How your Facebook ID can get you more Wi-Fi access," *MIT Tech. Rev.*, 2013.

[8] 3GPP TS 24.312 v. 12.0, "Access Network Discovery and Selection Function (ANDSF) Management Object (MO)," 2013.

[9] W. Dong, S. Rallapalli, R. Jana, L. Qiu, K. K. Ramakrishnan, L. Razoumov, Y. Zhang, T. W. Cho, "iDEAL : Incentivized dynamic cellular offloading via auctions," *IEEE Infocom,* 2013.

[10] 3GPP Technical Report 36.814, "Further advancements for E-UTRA physical layer aspects," 2010.

[11] ITU-R, "Guidelines for evaluation of radio interface technologies for IMT-Advanced", *Report ITU-R M.2135-1,* 2009.

[12] D.J. Watts, *Small worlds: The dynamics of networks between order and randomness,* Princeton University Press, 1999.

[13] D. Martín-Sacristán *et al.,* "3GPP long term evolution: Paving the way towards next 4G," *Waves*, 2009.

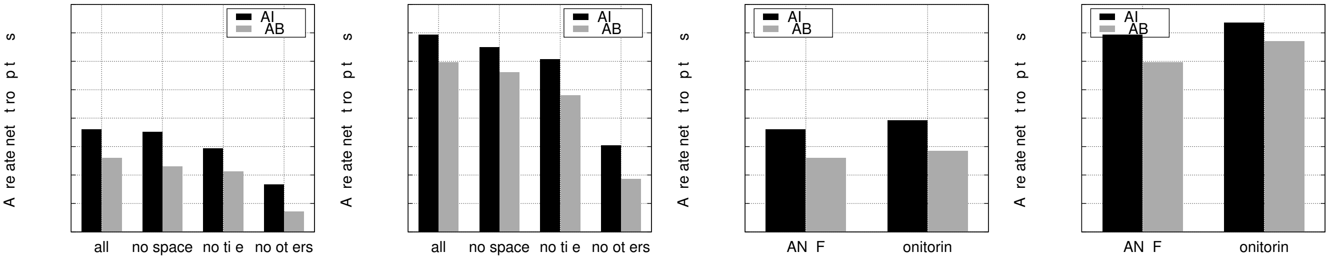
[14] E. Perahia, R. Stacey, *Next generation wireless LANs,* Cambridge University Press, 2nd Ed., 2013.

[15] S. Busanelli, M. Martalò, G. Ferrari, G. Spigoni, N. Iotti "Vertical handover between WiFi and UMTS networks: Experimental performance analysis," *Int. J. of Energy, Inf. and Comm.*, 2011.

[16] K. Liu, Q. Zhao, "Distributed learning in multi-armed bandit with multiple players," *IEEE Trans. on Signal Proc.*, vol. 58, no. 11, pp. 5667–5681, 2010.

[17] J. Vermorel, M. Mohri, "Multi-armed bandit algorithms and empirical evaluation," *Springer LNCS*, 2005.

[18] N. Eagle, A. Pentland, "Reality mining: sensing complex social systems," *Personal and Ubiquitous Computing*, 2005.



(a) Commercial (b) Domestic (c) Commercial (d) Domestic

Fig. 7. Performance of AIMD and MAB in the commercial (a,c) and domestic (b,d) scenarios. The left plots compare the following cases: all ANDSF features are enabled ("all"); time dependency is disabled ("no-time"); space dependency is disabled ("no-space"); other UE feedbacks cannot be exploited ("no-others"). The right plots compare the case where standard ANDSF features are exploited ("ANDSF") to that where UEs can also monitor traffic in WiFi hotspots ("monitoring").