# Energy efficiency and traffic offloading in WLANs with caching and mesh capabilities

Rosario G. Garroppo, Gianfranco Nencioni, Gregorio Procissi, Luca Tavanti
Dipartimento di Ingegneria dell'Informazione
Università di Pisa, Italy
Email: luca.tavanti@iet.unipi.it

Bernard Gendron
CIRRELT and DIRO
Université de Montréal, Canada
Email: bernard.gendron@cirrelt.ca

***Abstract*—In this paper we study a wireless access network based on the WLAN technology and enriched with features such as caching and mesh networking. This system is analysed in terms of energy efficiency and traffic offloading, two objectives that are somewhat in contrast, but both relevant to network and service providers as they directly impact the operational cost. To this aim, we developed a mathematical model of the system and solved it to optimality by means of integer linear programming. We can thus show how much saving can be achieved both in terms of energy and traffic, also considering various tradeoff points among the two contrasting objectives. A basic degree of quality of service is also accounted for.**

## I. Introduction

Medium and large wireless Local Area Networks (WLANs) are now a popular technique for providing Internet access to the increasing number of mobile user devices. Along with them, many vendors and networks operators have been selling and deploying WLANs enriched with mesh capabilities (wireless mesh networks, WMNs). Recently, there has been an increasing interest in extending the capabilities of both WLANs and WMNs to meet the recent advances in various technological and social fields.

In particular, some researchers have proposed to broaden the Content Distribution Network (CDN) model to include also the access gateway. The idea is that the GW can be exploited, either as a replacement or as an addition to the CDN, to store the contents destined to the users, thus reducing the congestion of the Internet and improve the quality of the delivery service [1], [2], [3]. Another topical field is energy efficiency, a.k.a. "green" networking. To this aim, various approaches have been explored for reducing the power consumption of carrier-grade WLANs [4], [5]. Remarkably, there have already been attemps at merging the CDN and mesh paradigms, as exposed for example in [6], [7], as well as studies on power saving solutions in wireless mesh networks [8].

Given these premises, we build our work on the system architecture that arises from joining the WLAN access system with the CDN and the mesh paradigms. Specifically, we address a Content-Delivery Wireless Mesh Network (CDWMN), i.e. an enterprise wireless LAN in which the GWs, besides granting Internet access to the user terminals (UTs), are also able to perform (limited) content caching and to handle wireless multi-hop paths. In delivering the various multimedia contents to the end-users, the CDWMN should strive to achieve the best energy efficiency and traffic offloading performance. Indeed, both goals are important for the network/service provider, because they both translate in cost saving. At the same time, a minimum level of quality of service (QoS) should also be guaranteed in order to satisfy the customers' expectations.

The main contribution of the paper is therefore an investigation on the potential of the CDWMN from both an energy-aware and a traffic-aware perspective. In detail, we devise a resource allocation and routing scheme (named OETAC, i.e. Optimal Energy and Traffic Allocation for CDWMN) that jointly considers all the distinctive CDWMN features mentioned above for delivering the contents in an efficient manner. The objective is either to minimise the overall energy consumption or to maximise the traffic that can be offloaded from the core network. Since the two objectives are partially contrasting, we study them both singularly and in combination. Accordingly, we formulate a mathematical programming model to analyse the performance of this scheme, especially with reference to the basic (and currently most employed) approach of having a set of independent cache-less GWs.

The final goal and result of our work is to quantify the maximum overall power and traffic saving that can be achieved by the OETAC strategy. Yet, in order to make the optimisation practical, both goals are to be achieved under some QoS constraint. In our case, this constraint is the average per-flow delay. We can thus provide an indication on how much a CDWMN can be green and how much traffic it can offload under the optimal configuration. To the best of our knowledge, no previous work has performed a similar study.

## II. OETAC: Optimal Energy and Traffic Allocation for CDWMN

### A. Architecture of the CDWMN

The CDWMN system fits in particular those wireless Internet service providers that do not own or cannot access a wired distribution system to directly connect the GWs among them. Several application scenarios falls in this case, such as municipal wireless, sparse campuses, open areas for sport and artistic events, rural communities, disaster relief.

The reference CDWMN architecture is shown in Fig. 1. The access gateways (GWs) are connected to the Internet by means of a high speed connection (e.g. Gigabit Ethernet, GPON, wireless point-to-point). The GWs can also be part of the CDN service, which decides the contents (CNs) that are cached on each GW. Then, each GW is equipped with two different wireless interfaces. One interface is used to serve

978-0-9836283-9-2 

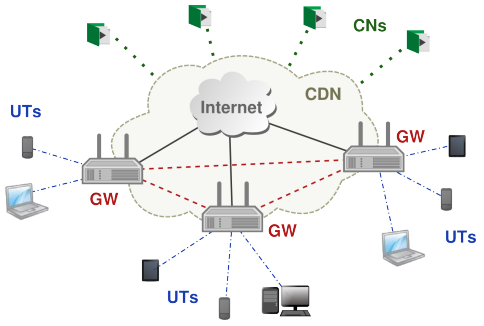


Fig. 1. Architecture of the CDWMN. Solid lines represent wired links, dashed lines wireless links, and dotted lines (towards the contents, CNs) virtual links.

the assigned user terminals (UTs), while the other is used to connect with the neighbouring GWs. The two interfaces work on non-overlapping channels in order to avoid mutual interference. A UT can be assigned (associated) to one GW only. Typically, the UTs of a given user will be associated to the GW from which they receive the strongest signal. However, we do not prevent UTs from associating with a different GWs if a radio link is available and the optimum allocation requires so.

Therefore, according to the illustrated CDWMN architecture, there are three possibilities for a GW to deliver a content to an assigned UT: $(i)$ the CN is cached on the GW itself, $(ii)$ the CN is cached on another GW that can be reached either directly or by means of a multi-hop connection, $(iii)$ the CN must be downloaded from the Internet.

### B. The OETAC approach

OETAC exploits the distinctive features of the CDWMN to achieve energy efficiency and traffic offloading. Internet connectivity sharing, i.e. the possibility to associate any UT to any reachable GW, allows to power off some GWs. Caching on the GWs can be used to save Internet bandwidth (and energy). The mesh service enables the dissemination of the CNs cached in any GW to any UT in the network.

To this purpose, we build a mathematical program that takes as input the placement of the contents on the GWs, the device connectivity, and the user demands, and decides (outputs) the UT-GW associations and the routing of the CNs. Note that our focus is on already deployed WLANs, i.e. we do not solve the problem of choosing in which candidate sites the GWs shall be deployed. We assume this has been done in a previous phase, for example on the basis of the peak traffic demand pattern. Due to the already widespread adoption of enterprise WLANs, this hypothesis matches quite well with reality. Also, it allows to apply our method to existing networks, not just to the future ones. Similarly, our analysis starts from a given content placement pattern, laid out by a suitable CDN approach (see e.g. [2], [9], [10]). Therefore our work does not integrate the placement of the CNs, but is complementary to the CDN deployment.

In the CDWCN abstraction used to formulate the mathematical model there are three kinds of nodes (UTs, GWs, and CNs) and three kinds of links (GW-UT, GW-GW, and CN-GW). The GW-UT and GW-GW are physical (wireless) links, and therefore are characterized by a data rate dependent on the modulation and radio propagation rules. Conversely, the CN-GW links are logical, because we assume that every content on the Internet can be accessed by every GW with no data rate restrictions.

### C. Notation

In the addressed problem we define the following sets and parameters:

- $\mathcal{G}$, the set of deployed gateways (GWs);
- $\mathcal{U}$, the set of user terminals (UTs);
- $\mathcal{C}$, the set of available contents (CNs);
- $\mathcal{D} = \{(c,u) : u \in \mathcal{U}, c \in \mathcal{C}\}$, the set of content demands;
- $\mathcal{E}' = \{(g,u) : u \in \mathcal{U}, g \in \mathcal{G}, r_{gu} > 0\}$, the set of GW-UT edges;
- $\mathcal{E}'' = \{(g,h) : g,h \in \mathcal{G}, r_{gh} > 0\}$, the set of GW-GW edges;
- $r_{ij}$ $[b/s]$, the average data rate between a vertex $i \in \mathcal{G}$ and a vertex $j \in \{\mathcal{U} \bigcup \mathcal{G}\}$ (for example $r_{gu}$ is the data rates between a GW $g$ and a UT $u$);
- $b_c$ $[b/s]$, the rate needed for retrieving CN $c$;
- $L$ $[b]$, the average packet size, including all protocol headers;
- $P^{GW}$ $[W]$, the power consumption of a GW;
- $E^I$ $[J/b]$, the average energy for retrieving a bit from the Internet;
- $E^W$ $[J/b]$, the average energy for transmitting a bit through a wireless link (either GW-GW or GW-UT);
- $t_{cg}$, a binary flag that is set to 0 if CN $c$ is cached in GW $g$, to 1 otherwise.

We then define the following binary variables:

- $q_g \in \{0,1\}$ , which is set to 1 if GW $g$ is powered on;
- $x_{gu} \in \{0,1\}$, which is set to 1 if UT $u$ is assigned to GW $g$;
- $y_{hg}^{cu} \in \{0,1\}$, which is set to 1 if CN $c$ is delivered to UT $u$ through link $(h,g)$;
- $z_g^{cu} \in \{0,1\}$, which is set to 1 if UT $u$ retrieves CN $c$ from GW $g$ (i.e. either $c$ is cached in $g$ or $c$ is downloaded from the Internet by $g$); note that $u$ needs not to be assigned to $g$.

### D. Mathematical programming model

The objective function of our problem is the combination of two elements. The first term is the total consumed power, which is built by three terms: the power consumption of powered-on GWs, the power consumption for retrieving

contents from the Internet, and the power consumption for transferring contents among the GWs:

$$P_{ee} = P^{GW} \sum_{g \in \mathcal{G}} q_g + \sum_{(c,u) \in \mathcal{D}} b_c \Big\{ E^I \sum_{g \in \mathcal{G}} (t_{cg}\, z_g^{cu}) + E^W \sum_{(h,g) \in \mathcal{E}''} y_{hg}^{cu} \Big\}. \quad (1)$$

Note that we do not follow the greedy approach of minimising the energy consumed by the sole CDWMN, but we also account for the energy required by the Internet for bringing the data up to the GWs.

The second term is proportional to the traffic that is not offloaded from the Internet:

$$P_{to} = E^I \sum_{g \in \mathcal{G}} \sum_{(c,u) \in \mathcal{D}} (t_{cg}\, z_g^{cu}\, b_c), \quad (2)$$

where the term $E^I$ is used to convert the traffic measure to an energy value in order to make the objective function homogeneous.

The two parts are merged by means of a "tuning" parameter, $\gamma$, into our objective function:

$$\begin{aligned} \Omega &= (1-\gamma)\, P_{ee} + \gamma\, P_{to} \\ &= (1-\gamma) P^{GW} \sum_{g \in \mathcal{G}} q_g + E^I \sum_{g \in \mathcal{G}} \sum_{(c,u) \in \mathcal{D}} b_c\, t_{cg}\, z_g^{cu} \\ &\quad + (1-\gamma) E^W \sum_{(h,g) \in \mathcal{E}''} \sum_{(c,u) \in \mathcal{D}} b_c\, y_{hg}^{cu} \end{aligned} \quad (3)$$

The tuning parameter $\gamma$ can take real values in the $[0, 1]$ interval, with 0 implying a pure energy efficient optimisation, and 1 a traffic oriented one.

The objective of the mathematical program is to minimise $\Omega$ subject to:

$$x_{gu} - z_g^{cu} - \sum_{(h,g) \in \mathcal{E}''} y_{hg}^{cu} = 0 \quad \forall (c,u) \in \mathcal{D}\ \forall (g,u) \in \mathcal{E}', \quad (4)$$

$$\sum_{(g,h) \in \mathcal{E}''} y_{gh}^{cu} - \sum_{(h,g) \in \mathcal{E}''} y_{hg}^{cu} - z_g^{cu} = 0 \quad \forall (c,u) \in \mathcal{D}\ \forall (g,u) \notin \mathcal{E}', \quad (5)$$

$$\sum_{g \in \mathcal{G}} z_g^{cu} = 1 \qquad \forall (c,u) \in \mathcal{D}, \quad (6)$$

$$\sum_{(g,u) \in \mathcal{E}'} x_{gu} = 1 \qquad \forall u \in \mathcal{U}, \quad (7)$$

$$x_{gu} \le q_g \qquad \forall (g,u) \in \mathcal{E}', \quad (8)$$

$$y_{hg}^{cu} \le q_g\, q_h \qquad \forall (h,g) \in \mathcal{E}''\ \forall (c,u) \in \mathcal{D}, \quad (9)$$

$$z_g^{cu} \le q_g \qquad \forall g \in \mathcal{G}\ \forall (c,u) \in \mathcal{D}, \quad (10)$$

$$\frac{\sum_{(g,u) \in \mathcal{E}'} \frac{\sum_{(c,u) \in \mathcal{D}} b_c\, L}{r_{gu}^2}\, x_{gu}}{2\,(1 - \sum_{(g,u) \in \mathcal{E}'} \frac{\sum_{(c,u) \in \mathcal{D}} b_c\, x_{gu}}{r_{gu}})} \le d_U \qquad \forall g \in \mathcal{G}, \quad (11)$$

$$\frac{\sum_{(g,u) \in \mathcal{E}'} \frac{\sum_{(c,u) \in \mathcal{D}} b_c\, L}{r_{gu}^2}\, (y_{hg}^{cu} + y_{gh}^{cu})}{2\,(1 - \sum_{(g,u) \in \mathcal{E}'} \frac{\sum_{(c,u) \in \mathcal{D}} b_c\, (y_{hg}^{cu} + y_{gh}^{cu})}{r_{gu}})} \le d_G / H \qquad \forall g \in \mathcal{G}. \quad (12)$$

Equations (4) and (5) are the flow conservation constraints, equations (6) impose that each UT $u$ retrieves the content $c$ from exactly one GW (single source of demand $(c,u)$), equations (7) impose that each UT must be assigned to exactly one GW, equations (8) and (9) impose that, if a GW is powered off, no UT or GW can be connected to it, equations (10) impose that no content can be retrieved from a powered-off GW, and equations (11) and (12) impose that the average waiting times in the GW-UT and GW-GW interfaces does not exceed a predefined bound. The next Section illustrates how (11) and (12) have been obtained.

The above formulated model is an Integer Linear Programming (ILP) problem. Thus it can be solved by means of a Mixed-Integer Linear Programming solver, such as the IBM ILOG CPLEX Optimizer.

### *E. Delay bound implementation*

In order to provide a reasonable limit for the flow delays, we took advantage of the queuing theory. In particular, we modelled the WLAN system as a network of M/D/1 queues with heterogeneous traffic classes/flows. Each queue represents a GW interface, in which the packet arrivals are exponential, with parameter $\lambda_i$, and the server implements a FCFS discipline with deterministic serving time $1/\mu_i$; for both parameters, $i$ is the index of the traffic flow. Therefore the average waiting time in the queue is determined by the Pollaczek-Khinchine formula [11]:

$$E[T_W] = \frac{\sum_i \lambda_i\, E[{T_{S,i}}^2]}{2\,(1 - \sum_i \rho_i)}, \quad (13)$$

where $\rho_i = \lambda_i / \mu_i$, and $T_{S,i} = 1/\mu_i$ is the serving time of class $i$. We can now express the general queuing parameters in terms of the parameters of our problem. Specifically, let $L_i = L\, \forall i$, and $b_i$ the data rate of flow $i$ (note that a flow does not necessarily coincide with a CN demand). Then, the arrival rate $\lambda_i$ is equal to $b_i/L$, the serving time $\mu_i$ is equal to $r_i/L$, and consequently $\rho_i = b_i/r_i$. Therefore, (13) becomes:

$$E[T_W] = \frac{\sum_i b_i\, L / r_i^2}{2\,(1 - \sum_i b_i / r_i)}. \quad (14)$$

The last step is to convert and adapt (14) to fit our model. To this aim, we must separate the GW-UT links from the GW-GW links. For the former, we have that the flow index $i$ corresponds to the GW-UT edge $(g,u)$: $i \to (g,u)$. Accordingly, $r_i \to r_{gu}$. Then, the flow data rate $b_i$ is built by the sum of all demands routed over the $(g,u)$ link: $b_i \to \sum_{(c,u) \in \mathcal{D}} b_c$. Thus, operating the defined substitutions and imposing a delay bound of $d_U$, we obtain (11). Note that (11) also impose a capacity constraint, since $E[T_W]$ is bounded if and only if $\sum_i \rho_i < 1$, i.e. if the utilization of the queue is less than one. We obtain (12) in a similar fashion, with the sole difference that we must account for both directions of the $(g,h)$ link and for the average number of hops $H$.

TABLE I. PARAMETER VALUES FOR THE TESTED SCENARIOS.

| Parameter | Minimum | Standard | Maximum |
|---|---|---|---|
| $\|\mathcal{G}\|$ | 10 | 30 | 50 |
| $\|\mathcal{U}\|/\|\mathcal{G}\|$ | 2 | 4 | 6 |
| $\|\mathcal{C}\|/\|\mathcal{U}\|$ | 100 | 1000 | 10000 |
| $S/\|\mathcal{U}\|$ | 0.25 | 2.5 | 25 |
| $\alpha$ | 0.5 | 0.65 | 0.8 |

Note that, though this model is somewhat simplistic, it nevertheless allows us to describe the delay of the various flows in a reasonable way. This is sufficient for the purpose of our work, because our goal is assessing the energy-traffic optimisation tradeoff, whereas providing a thorough modelling of the WLAN system is out of the scope.

## III. PERFORMANCE EVALUATION

The performance of OETAC has been tested over a series of 11 network scenarios. Each scenario is characterised by different values of the input parameters, which are reported in Table I. Besides the sizes of the sets $\mathcal{G}, \mathcal{U}$, and $\mathcal{C}$, $S$ is the cache size on each GW (number of CNs), and $\alpha$ is the popularity exponent of the content demand. The "standard" scenario (all values in the central column) is used as the reference one. Starting from it, we have changed one parameter value per scenario. For each scenario, we generated and solved twenty instances. The results have then been averaged over the whole set of instances and scenarios.

### A. Scenario and parameter values

GWs and UTs have been placed in a fictitious test area of varying size in order to keep the GW density constant. In each instance, the positions of the GWs and UTs have been randomly determined. However, to avoid heavily unbalanced instances, the test field has been divided into a regular grid and the GWs and UTs have been evenly distributed across the grid squares.

The energy consumed by every GW has been set to $P^{GW}$ = 15 W, which is a typical value for enterprise wireless routers. The energy consumption for retrieving a bit of content via Internet has been set to $E^I$ = 39 $\mu$J/b [12], and the energy consumption for transferring a bit of content over a wireless link to $E^W$ = 0.02 $\mu$J/b [13].

To determine the GW-GW and GW-UT data rates (i.e. the various $r_{gu}$ and $r_{gh}$), we employed a simplified version of the COST-231 path loss model, which allows to account for various propagation aspects, such as the presence of walls and other obstacles, and the use of a realistic path loss exponent.

The content demand has been modelled according to a content popularity that follows the widely adopted Zipf distribution [14], with the popularity exponent $\alpha$ varying from $0.5$ to $0.8$ (as shown in Table I) based on the empirical observation reported in [15], [16].

The cached contents in the GWs have been modelled by means of the same Zipf distribution. We reckon that this is a sensible behaviour for a "smart" caching strategy, which would presumably store with the highest probability the most popular contents [14]. For each GW we assumed a number of cached contents equal to $S$ (the system is in the steady state). We have considered a somewhat cooperative caching strategy, in which the contents to be cached are selected so as the same contents are not cached in adjacent GWs.

Finally, we have considered the discrete distribution of the data rate of the contents ($b_c$) based on [17].

As for the delay aspect, we set $d_U$ = 50 ms and $d_G$ = 50 ms. Consequently, the delay that is introduced by the CDWMN should be, at most, in the order of 50/100 ms, which is generally suitable for most multimedia contents. For the $H$ parameter we have obtained an upper bound by performing the same experiments described in Section III using the capacity constraints (16)-(17) (see below) in place of the delay bounds (11)-(12), and then by extracting the 99-percentile of the average hop count. The resulting value is $H = 1.095$.

### B. Reference benchmarks

The performance of our model is assessed in terms of several metrics. For most of them we adopted a common benchmark, which is the currently most widespread connection model in enterprise WLANs. In short, the GWs form separate connectivity islands, do not perform content caching, are not shared nor connected with each other, and are always active. The UTs simply associate to the GW from which they receive the strongest signal. In this case, the power consumption of this network model, named SCI (Single Connectivity Islands) for convenience, can be readily computed as:

$$P_{\text{SCI}} = |\mathcal{G}| \cdot P^{GW} + E^I \sum\nolimits_{(c,u)\in\mathcal{D}} b_c. \qquad (15)$$

Two further benchmarks are built by replacing (11) and (12) with the classical capacity constraints:

$$\sum_{(g,u)\in\mathcal{E}'} \left\{ \frac{\sum_{(c,u)\in\mathcal{D}} b_c}{r_{gu}} \cdot x_{gu} \right\} \leq q_g \qquad \forall g \in \mathcal{G}, \qquad (16)$$

$$\sum_{(h,g)\in\mathcal{E}''} \sum_{(c,u)\in\mathcal{D}} \left\{ \frac{b_c}{r_{hg}} \cdot (y^{cu}_{hg} + y^{cu}_{gh}) \right\} \leq q_g \qquad \forall g \in \mathcal{G}, \ (17)$$

and by setting $\gamma$ to either 1 or 0, in order to perform, respectively, pure energy efficiency (EE) or traffic offloading (TO) optimisation. The resulting problems, which have no QoS guarantees, yield an upper bound on the energy or traffic performance of the system. However, as it will be shown later on, the obtained solutions are hardly of any practical use because of the unbounded delay.

### C. Results

The most prominent measure of energy efficiency is the amount of consumed power. Fig. 2 depicts the power consumption of the OETAC approach for various values of $\gamma$ – indicated as OETAC($\gamma$). EE and TO have also been reported. All values are normalised to the power consumption of SCI.

At first, we can see that in all cases OETAC provides a fair amount of power saving with respect to SCI. This amount, however, is variable, as the traffic oriented version provides quite a small energy efficiency improvement. More in detail, OETAC(0) can save up to 14.9% of power, which gradually reduces as $\gamma$ increases, touching the minimum (6.9%) for $\gamma = 1$. From a comparison with the non-QoS versions, it emerges

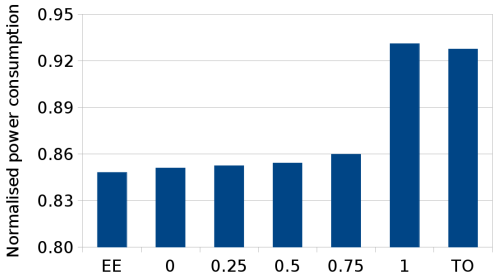


Fig. 2. Normalised power consumption of OETAC as a function of the $\gamma$ value. Lower is better.

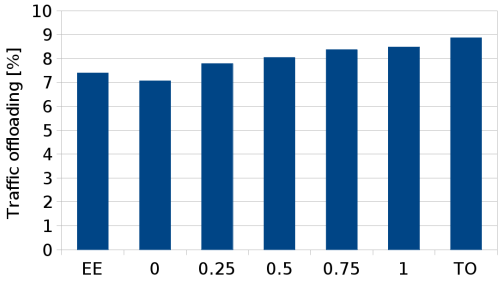


Fig. 3. Percentage of traffic offloading as a function of the $\gamma$ value. Higher is better.

that the efficiency loss of OETAC is very limited. EE yields a 15.2% saving, and TO 7.2%. Thus, the energy loss due to the delay constraints is indeed minimal.

Fig. 3 shows the outcome of the experiments from the traffic offloading point of view. In this case the best result is, obviously, achieved by OETAC(1), which allows to save up to 8.5% of Internet traffic. A gradual decrease goes along with the reduction of $\gamma$, touching a minimum for $\gamma = 0$ (7.1%). Again, EE and TO performs better than OETAC(0) and OETAC(1), respectively, but the gap is almost negligible (0.4%).

The third parameter we examine is the delay $d$ experimented by the demands (in all scenarios and all instances). The delay is computed, for each demand, as the sum of all waiting times in the queues plus all transmission times. Table II reports four statistics about the delay observation set: the mean value ($\bar{d}$), the 99-percentile, and the percentage of observed values greater than the 50 and 100 ms bounds used for (11)-(12). Incidentally, we recall that (11)-(12) impose a limit on the *average queuing times* of the demands, which is a statistical property of an aleatory variable of which we are now observing the realisations. Therefore it may well occur that $d > d_U$ or even $d > d_U + d_G$.

As already anticipated, the delay figures of EE and TO do not allow for a plain transposition of the provided resource allocation into the real world. In both cases, $\bar{d}$ is well beyond the QoS requirement, with a fair number of demands exceedings both QoS bounds. Conversely, OETAC promises much lower delays, with no more than one hundredth of the demands failing to achieve the QoS objective of 100 ms.

By means of the cumulative distribution function (CDF), illustrated in Fig. 4, we can complete the picture about the delay performance. The behaviour is very similar across all $\gamma$ values, with OETAC(0) having a sligthly steeper inclination,

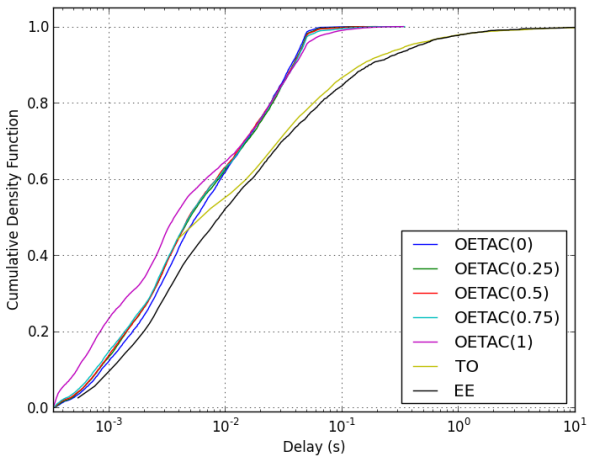


Fig. 4. CDF of OETAC as a function of the $\gamma$ value. Note that the x-axis is in logarithmic scale.

and OETAC(1) a imperceptibly more gradual trend. For all them, the knee is exactly at 50 ms. The performance difference with either EE or TO is apparent.

Further insights can be obtained from the percentage of GWs that are powered off by the various models (last row of Table II). It emerges that OETAC(1) and TO do not deactivate any GW. This is the obvious consequence of having no GW component in the objective function. Since both approaches focus exclusively on maximing the offloaded traffic, turning off some GWs might be inconvenient. The other interesting result is about EE and OETAC(0). Both approaches target energy efficiency only, and they yield the same highest fraction of inactive GWs. However, OETAC(0) is subject to the more stringent delay bounds. Therefore it appears that meeting the QoS parameter can be achieved without loosing much energy efficiency. For the other values of $\gamma$, OETAC gradually increases the number of active GWs in order to allow for more traffic to be carried over the CDWMN. Obviously, this leads to less power saving, as shown in Fig. 2.

The last metric we analyse is the hit rate ($\eta$), which measures the capability to find and retrieve the demanded contents from within the CDWMN. The hit rate is defined as:

$$\eta = \frac{\sum_{(c,u)\in\mathcal{D}} \sum_{g\in\mathcal{G}} (1 - t_{cg}) \cdot \bar{z}_g^{cu}}{|\mathcal{D}|} \cdot 100. \qquad (18)$$

Fig. 5 plots the bars for all approaches. A further distinction is performed on the basis of the location of the content. With the term "local" we refer to contents cached on the GW a UT is currently assigned to ($\bar{x}_{gu} = 1$), whereas "remote" indicates contents cached on other GWs in the CDWMN. This implies that a remote CN must be retrieved by means of a multi-hop path, and might incur a higher delay ($d_G + d_U$) with respect to a local CN (whose delay is solely $d_U$).

As expected, the greater the $\gamma$, the more the objective function is traffic oriented, and the highest is the achieved hit rate. The bigger jump is between OETAC(0) and OETAC(0.25), suggesting that even a moderate lean towards the traffic

TABLE II. DELAY FIGURES AND FRACTION OF POWERED-OFF GWS FOR OETAC, EE, AND TO.

| | EE | OETAC(0) | OETAC(0.25) | OETAC(0.5) | OETAC(0.75) | OETAC(1) | TO |
|---|---|---|---|---|---|---|---|
| Mean delay value, $\overline{d}$ [ms] | 451 | 15.2 | 15.4 | 15.3 | 15.7 | 16.5 | 1519 |
| 99-percentile of $d$ [ms] | 2082 | 53.9 | 58.8 | 60.5 | 72.0 | 100.2 | 2494 |
| $d > 50$ms [%] | 23.9 | 1.55 | 2.28 | 2.55 | 2.90 | 4.77 | 21.8 |
| $d > 100$ms [%] | 15.6 | 0.08 | 0.19 | 0.27 | 0.55 | 1.00 | 13.5 |
| Powered-off GWs [%] | 49.8 | 49.8 | 45.6 | 43.4 | 38.5 | 0 | 0 |

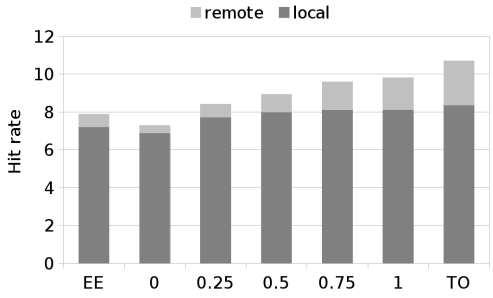


Fig. 5. Hit rate (local plus remote) as a function of the $\gamma$ value.

offloading goal is enough to sensibly improve the performance of the optimisation model. Also worth noting is the fact that the local hit rate ceils quite rapidly around 8%, and the subsequent increases in $\eta$ are driven mostly by the remote part. In other words, increasing $\gamma$ pushes for a higher utilisation of the mesh facility, and therefore to activate more GWs in order to support the delivery along the delay-bounded paths. Lastly, note how EE and TO get sensibly higher $\eta$ values. However, as already anticipated, these are obtained at the expenses of very large delays, since EE makes use of many multi-hop paths with few active GWs, thus congesting them, whereas TO strives for maximising the use of remote caches, thus building very long multi-hop paths.

## IV. CONCLUSION

The paper presented and discussed OETAC, a method for optimising the energy efficiency and traffic offloading performance of a wireless access network with caching and mesh capabilities under delay bounds. From a large set of experiments, based on practical rate and consumption figures, it emerged that OETAC enables a sensible saving in both objectives. We also showed that the combination of the two objectives permits to simultaneously achieve almost the same gains as the two separate approaches, thus merging the benefits of both. Finally, we proved that employing realistic delay bounds has the advantage of making the optimal allocation practically deployable and offering an acceptable level of service to the end users, whilst at the same time it does not produce any meaningful loss in neither energy efficiency nor traffic offloading. As a result, the OETAC approach can indeed be profitable for network and service operators, which would have an efficient and viable tool for quantifying and choosing the energy/traffic tradeoff point for their networks.

## ACKNOWLEDGMENT

This work was partially supported by the Italian Ministry of University and Research under the "GreenNet" FIRB project.

## REFERENCES

[1] F. den Hartog, B. Bastiaans, M. Blom, M. Pluijmaekers, and R. van der Mei, "The use of Residential Gateways in Content Delivery Networking," in *Proc. Australian Telecommunication Networks and Applications Conference (ATNAC)*, Sydney, 2004.

[2] V. Valancius, N. Laoutaris, L. Massoulié, C. Diot, and P. Rodriguez, "Greening the internet with nano data centers," in *CoNEXT*, 2009, pp. 37–48.

[3] S. Spagna, M. Liebsch, R. Baldessari, S. Niccolini, S. Schmid, R. Garroppo, K. Ozawa, and J. Awano, "Design principles of an operator-owned highly distributed content delivery network," *IEEE Communications Magazine*, vol. 51, no. 4, pp. 132–140, 2013.

[4] J. Lorincz, A. Capone, and D. Begusic, "Optimized network management for energy savings of wireless access networks," *Computer Networks*, vol. 55, no. 3, pp. 514–540, Feb. 2011.

[5] A. P. Couto da Silva, M. Meo, and M. A. Marsan, "Energy-performance trade-off in dense WLANs: A queuing study," *Computer Networks*, vol. 56, no. 10, pp. 2522–2537, 2012.

[6] V. Manetti, R. Canonico, W. De Donato, G. Ventre, A. Mauthe, and G. Tyson, "Next Generation CDN Services for Community Networks," in *NGMAST*, 2009, pp. 89–94.

[7] A. Alasaad, S. Gopalakrishnan, H. Nicanfar, and V. Leung, "Green content distribution in Wireless Mesh Networks with infrastructure support," in *ICC*, 2012, pp. 5896–5900.

[8] A. De La Oliva, A. Banchs, and P. Serrano, "Throughput and energy-aware routing for 802.11 based mesh networks," *Computer Communications*, vol. 35, no. 12, pp. 1433–1446, 2012.

[9] S. Borst, V. Gupta, and A. Walid, "Distributed Caching Algorithms for Content Distribution Networks," in *Proc. IEEE INFOCOM*, 2010.

[10] Z. Al-Arnaout, Q. Fu, and M. Frean, "A divide-and-conquer approach for content replication in WMNs," *Computer Networks*, 2013, in press (DOI 10.1016/j.comnet.2013.09.016).

[11] L. Kleinrock, *Queueing Systems*. Wiley Interscience, 1976, vol. II: Computer Applications.

[12] J. Baliga, R. Ayre, K. Hinton, W. V. Sorin, and R. S. Tucker, "Energy Consumption in Optical IP Networks," *Journal of Lightwave Technology*, vol. 27, no. 13, pp. 2391–2403, Jul. 2009.

[13] D. Halperin, B. Greenstein, A. Sheth, and D. Wetherall, "Demystifying 802.11n power consumption," in *Proc. International conference on Power aware computing and systems*. USENIX Association, 2010.

[14] G. Haßlinger and F. Hartleb, "Content delivery and caching from a network providers perspective," *Computer Networks*, vol. 55, no. 18, pp. 3991–4006, 2011.

[15] P. Gill, M. Arlitt, Z. Li, and A. Mahanti, "Youtube traffic characterization: a view from the edge," in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, ser. IMC '07, 2007.

[16] X. Cheng, C. Dale, and J. Liu, "Statistics and Social Network of YouTube Videos," in *IWQoS*, 2008.

[17] P. Ameigeiras, J. J. Ramos-Munoz, J. Navarro-Ortiz, and J. Lopez-Soler, "Analysis and modelling of YouTube traffic," *Transactions on Emerging Telecommunications Technologies*, vol. 23, no. 4, pp. 360–377, 2012.