Performance Analysis of Mobile Data Offloading in Heterogeneous Networks

Fidan Mehmeti, *Student Member, IEEE,* and Thrasyvoulos Spyropoulos, *Member, IEEE*

Abstract—An unprecedented increase in the mobile data traffic volume has been recently reported due to the extensive use of smartphones, tablets and laptops. This is a major concern for mobile network operators, who are forced to often operate very close to (or even beyond) their capacity limits. Recently, different solutions have been proposed to overcome this problem. The deployment of additional infrastructure, the use of more advanced technologies (LTE), or offloading some traffic through Femtocells and WiFi are some of the solutions. Out of these, WiFi presents some key advantages such as its already widespread deployment and low cost. While benefits to operators have already been documented, it is less clear how much and under what conditions the user gains as well. Additionally, the increasingly heterogeneous deployment of cellular networks (partial 4G coverage, small cells, etc.) further complicates the picture regarding both operator- and user-related performance of data offloading. To this end, in this paper we propose a queueing analytic model that can be used to understand the performance improvements achievable by WiFi-based data offloading, as a function of WiFi availability and performance, user mobility and traffic load, and the coverage ratio and respective rates of different cellular technologies available. We validate our theory against simulations for realistic scenarios and parameters, and provide some initial insights as to the offloading gains expected in practice.

Index Terms—Mobile data offloading, Queueing theory, Probability generating functions, HetNets.

1 Introduction

Lately, an enormous growth in the mobile data traffic has been reported. This increase in traffic demand is due to a significant penetration of smartphones and tablets in the market, as well as Web 2.0 and streaming applications which have high-bandwidth requirements. Furthermore, Cisco [1] reports that by 2017 the mobile data traffic will increase by 13 times, and will climb to 13.2 exabytes per month, with approximately 5.2 billion users. Mobile video traffic will comprise 66 % of the total traffic, compared to 51% in 2012 [1].

This increase in traffic demand is overloading the cellular networks (especially in metro areas) forcing them to operate close to (and often beyond) their capacity limits causing a significant gradation of user experience. Possible solutions to this problem could be the complete upgrade to LTE or LTE-advanced, as well as the deployment of additional network infrastructure [2]. However, these solutions may not be cost-effective from the operators' perspective: they imply an increased cost (for power, location rents, deployment and maintenance), without similar revenue increases, due to flat rate plans, and the fact that a small number of users consume a large amount of traffic (3% of users consume 40% of the traffic [3]). As a result, LTE has only been sporadically deployed, and it is unclear whether providers will choose to upgrade their current deployments, and if in fact the additional capacity would suffice [2], [4].

• *F. Mehmeti and T. Spyropoulos are with the Department of Mobile Communications, EURECOM, France.*
E-mail: mehmeti@eurecom.fr, spyropou@eurecom.fr

This research was partially supported by Intel Mobile Communications, Sophia-Antipolis. A subset of initial results have been presented at the IEEE Globecom 2013 conference.

A more cost-effective way of alleviating the problem of highly congested mobile networks is by offloading some of the traffic through Femtocells (SIPTO, LIPA [5], [6]), and the use of WiFi. In 2012, 33% of total mobile data traffic was offloaded [1]. Projections say that this will increase to 46% by 2017 [1]. Out of these, data offloading through WiFi has become a popular solution. Some of the advantages often cited compared to Femtocells are: lower cost, higher data rates, lower ownership cost [2], etc. Also, wireless operators have already deployed or bought a large number of WiFi access points (AP) [2]. As a result, WiFi offloading has attracted a lot of attention recently.

The current approach to offloading is that of *on-the-spot offloading*: when there is WiFi availability, all data is sent over WiFi, otherwise traffic is transmitted over the cellular network. This is easy to implement, as smart phones currently are only able to use one interface at a time[1]. More recently, *delayed offloading* has been proposed [8], [9]: if there is currently no WiFi availability, (some) traffic can be delayed up to a chosen time threshold, instead of being sent immediately over the cellular interface. If up to that point, no AP is detected, the data is transmitted over the cellular network. Nevertheless, delayed offloading is still a matter of debate, as it is not known to what extent users would be willing to delay a packet transmission. It also requires disruptive changes in higher layer protocols (e.g. TCP) [10].

As a result, in this paper, we will focus on on-the-spot offloading. Although on-the-spot offloading is already

1. Using both interfaces in parallel, as well as per flow offloading (IFOM) are currently being considered in 3GPP [5], [7].

used and there is some evidence that it helps reduce the network load [1], [2], it is not clear how key factors such as WiFi availability, average WiFi and cellular performance, and existence of advanced cellular technologies (e.g. LTE) affect this performance. What is more, it is still a matter of debate if offloading offers any benefits to the user as well (in terms of performance, battery consumption, etc.). Studies suggest that such benefits might strongly depend on the availability and performance of WiFi networks, or type of user mobility [8], [9], etc. Finally, the increasingly heterogeneous deployement of current and future cellular environments, with partial coverage by different technologies (GPRS, EDGE, HSPA/HSPDA, LTE) and a growing number of "small cells" (e.g. femto, pico) utilized, further complicates these questions.

To this end, in this paper we propose a queueing analytic model for performance analysis of on-the-spot mobile data offloading. Our contributions can be summarized as follows:

- We consider a simple scenario where the user can choose between WiFi and a single cellular technology, and derive general formulas, as well as simpler approximations, for the expected delay and offloading efficiency (Section 2).
- We generalize our analysis to the case where multiple cellular technologies (and respective rates) are available to a user, e.g. depending on her location, and/or different rates are offered by the same technology (e.g. rate adaptation, indoor/outdoor, etc.) (Section 3).
- We validate our model in scenarios where most parameters of interest are taken from real measured data, and which might diverge from our assumptions. (Section 4).
- We use our model to provide some preliminary answers to the questions of offloading efficiency and delay improvements through WiFi-based offloading (Section 4.5).

We discuss some related work in Section 5, and conclude in Section 6.

Before proceeding to our model and results, we would like to stress that the proposed model and analysis clearly is not meant to capture all the details of the various wireless access technologies, as such a task would be beyond the scope of a single paper, and would be too complex to yield useful insights. Nevertheless, we believe it is an important first step towards analytical tools that can be used to understand the performance of offloading in current and future network deployments, and to allow operators to make informed design decisions.

2 Analyzing Offloading in Homogeneous Cellular Networks

2.1 Problem Setup

We consider a mobile user that enters and leaves zones with WiFi coverage. The average time until a user enters such a zone, and the time she stays there, depend on the user's mobility (e.g. pedestrian, vehicular), the environment (e.g. rural, urban), the access point (AP) density, etc. E.g. increasing AP density would (i) increase the total WiFi coverage time, (ii) shorten the transition times between WiFi covered areas. On the other hand, higher user speeds would also shorten the expected time until one finds WiFi connectivity again, but would also lead to shorter connections (with a given AP).

In this section, we will first assume a "simple" cellular coverage environment, where the user has access to a single cellular data access technology, in addition to WiFi. Without loss of generality, we assume that there is always cellular network coverage (we allow coverage holes in the next section). Whenever there is coverage by some WiFi AP, all traffic will be switched over to WiFi. As soon as the WiFi connectivity is lost, the traffic will be transmitted through the cellular network.

Definition 1. *[WiFi Availability Model] We model the WiFi network availability as an ON-OFF alternating renewal process [11].*

- *Time consists of ON-OFF cycles,* $\left(T_{ON}^{(i)}, T_{OFF}^{(i)}\right), i \in \mathcal{N}^+$*, as shown in Fig. 1. ON periods represent the presence of WiFi connectivity, while during OFF periods only cellular access is available.*
- $T_{ON}^{(i)} \sim Exponential(\eta_w)$*, and independent from other (ON or OFF) periods. The data transmission rate during these periods is denoted with* μ_w*.*
- $T_{OFF}^{(i)} \sim Exponential(\eta_c)$*, and independent from other (ON or OFF) periods, with an offered data rate equal to* μ_c*.*

Transmission rates in practice: Transmission rates in real networks are not stable and are affected by signal quality (e.g. through rate adaptation), as well as the presence of other users. As a result the actual rate might change from ON period to ON period (or OFF to OFF), because e.g. the newly encountered WiFi AP is more congested or is further away. Consequently the above nominal rates μ_c (μ_w) corresponds to the effective rate allocated by the AP or BS to the user (e.g. based on channel quality, other users' presence, and the respective MAC protocol) and is an *average* over all OFF (ON) periods (which might be known, for example, from general city-wide or time-of-day statistics). We validate this assumption against arbitrary, randomly generates rates as well in Section 4, and also generalize our analytical result to scenarios with more than 2 rates offered, in the next section.

Next, we will assume that while a user is moving between areas with and without WiFi, she also generates new "data requests". Each such request generates a *flow* of data, e.g. corresponding to a file upload, data synchronization (e.g. DropBox, Flickr, Google+), etc. in the uplink direction, or a web page access, file download, news feed, etc. in the downlink direction. We will use the terms "file", "flow", and "message" interchangeably, to refer to a concatenation of packets corresponding to the same application request (e.g., a downloaded file, a photo uploaded on FaceBook), and which must be uploaded (downloaded) in completion

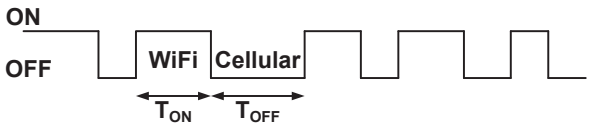

Fig. 1: The WiFi network availability model.

for the application request to be satisfied. Identifying and delimiting flows is a well researched problem and beyond the scope of this paper (see e.g. [12]). We assume the following simple model for the arrival and service of new flows at the user.

Definition 2. *[User Traffic and Service Model] Without loss of generality, we focus in one direction (uplink or downlink) and consider the following simple queueing model.*

- *New data requests arrive as a Poisson process with rate λ.*
- *The size of the flow/file corresponding to a data request is random and exponentially distributed.*
- *A flow arriving to find another flow currently in transmission will queue (the size of the queue assumed to be infinite), and will be served in a First Come First Served (FCFS) order.*
- *The service rate for the flow at the head of the queue will be equal to the WiFi rate (cellular rate), if the system is currently during an ON (OFF) period.*
- *A switch in connectivity might sometimes occur while a file transmission is ongoing. We assume that network transitions do not cause any interruptions to the traffic flow in service (i.e. the transmission continuous with the new rate immediately), and ignore vertical handover delays, if any. We discuss later how such interruptions can be included in the model.*

TABLE 1: Variables and Shorthand Notation.

Variable	Definition/Description
T_{ON}	Duration of ON (WiFi) periods
T_{OFF}	Duration of periods (OFF) without WiFi connectivity
λ	Average packet (file) arrival rate at the mobile user
$\pi_{i,c}$	Stationary probability of finding i files in cellular state
$\pi_{i,w}$	Stationary probability of finding i files in WiFi state
π_c	Probability of finding the system under cellular coverage only
π_w	Probability of finding the system under WiFi coverage
η_w	The rate of leaving the WiFi state
η_c	The rate of leaving the cellular state
μ_w	The service rate while in WiFi state
μ_c	The service rate while in cellular state
$E[S]$	The average service time
$E[T]$	The average system (transmission) time
$\rho = \lambda E[S]$	Average user utilization ratio
OE	The offloading efficiency
μ_k	The service rate while in network type k
η_k	The rate of leaving the type k network

The above two definitions capture our basic model for the offloading problem. We would like to stress that we do not claim that the actual availability periods or flow sizes are exponentially distributed (in fact, evidence for the contrary exists from measurement studies). The above assumptions are only made to keep our analysis tractable. In Section 2.6, we show how to extend our model to generic file size distributions. One could also try extend our framework to arbitrary ON and OFF distributions that can be approximated by Coxian distributions [13], fitting the first three moments to the real duration of the ON (OFF) period. In this case, there would be more states along one dimension in the Markov chain, and one could employ matrix-analytic methods [14]. However, these would offer little analytical insight. For this reason, we will test instead our model and its predictions against scenarios with realistic ON/OFF distributions in Section 4. Before proceeding further, we summarize in Table 1 some useful notation that will be used throughout the rest of the paper.

2.2 Delay Analysis

From the problem description, it is easy to see that the above setup corresponds to a single server queueing system whose rate varies according to an external (random) process. We are interested in the the total time a new user flow spends in the system (service+queueing) until it is served. It is a key metric of user experience that we would like to analyze. This is often referred to as the *system time* in queueing theory. However, we will also use the term *transmission delay* interchangeably.

Given the assumptions in the previous subsection, the on-the-spot offloading system can be modeled with a 2D Markov chain, as shown in Fig. 2.

$\pi_{i,w}$ denotes the stationary probability of having WiFi coverage *and* finding i flows in the system queue (i.e. $i-1$ waiting and one being transmitted over WiFi).

$\pi_{i,c}$ denotes the stationary probability of having only cellular coverage *and* i flows in the system.

Writing the balance equations for this chain gives

$$\pi_{0,c}(\lambda+\eta_c) = \pi_{1,c}\mu_c + \pi_{0,w}\eta_w \quad (1)$$
$$\pi_{0,w}(\lambda+\eta_w) = \pi_{1,w}\mu_w + \pi_{0,c}\eta_c \quad (2)$$
$$\pi_{i,c}(\lambda+\eta_c+\mu_c) = \pi_{i-1,c}\lambda + \pi_{i+1,c}\mu_c + \pi_{i,w}\eta_w, (i>0) \quad (3)$$
$$\pi_{i,w}(\lambda+\eta_w+\mu_w) = \pi_{i-1,w}\lambda + \pi_{i+1,w}\mu_w + \pi_{i,c}\eta_c, (i>0) \quad (4)$$

We define the probability generating functions for both the cellular and WiFi

$$G_c(z) = \sum_{i=0}^{\infty} \pi_{i,c} z^i, \text{and } G_w(z) = \sum_{i=0}^{\infty} \pi_{i,w} z^i, |z| \leq 1.$$

We can rewrite Eq.(1) and (3) as

$$\pi_{0,c}(\lambda+\eta_c+\mu_c) = \pi_{0,w}\eta_w + \pi_{1,c}\mu_c + \pi_{0,c}\mu_c$$
$$\pi_{i,c}(\lambda+\eta_c+\mu_c) = \pi_{i-1,c}\lambda + \pi_{i,w}\eta_w + \pi_{i+1,c}\mu_c, (i>0) \quad (5)$$

We multiply each of the equations from Eq.(5) by z^i and sum over all i's. After some calculus this yields

$$(\lambda+\eta_c+\mu_c)G_c(z) = \lambda z G_c(z) + \eta_w G_w(z) + \frac{\mu_c}{z}(G_c(z) - \pi_{0,c}) + \pi_{0,c}\mu_c. \quad (6)$$

By repeating the same process with Eq.(2) and (4), we get

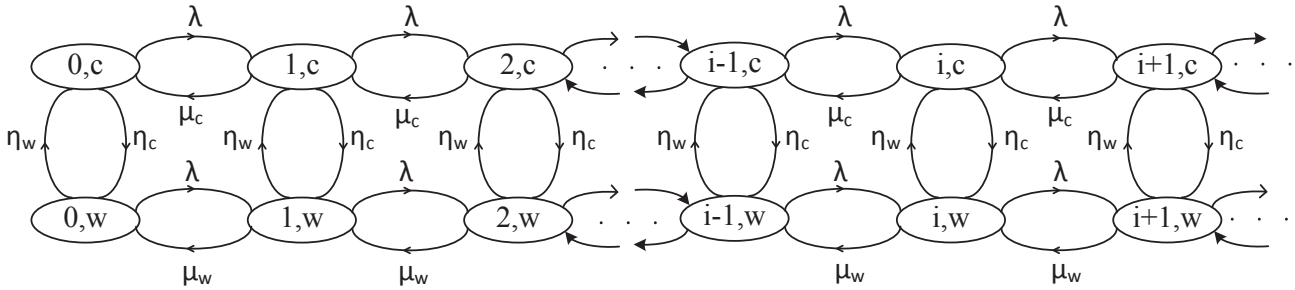

Fig. 2: The 2D Markov chain for the simple on-the-spot mobile data offloading scenario of Def. 1 and 2.

$$(\lambda+\eta_w+\mu_w)G_w(z) = \lambda z G_w(z)+\eta_c G_c(z) + \frac{\mu_w}{z}\left(G_w(z)-\pi_{0,w}\right)+\pi_{0,w}\mu_w. \quad (7)$$

Equations (6) and (7) define a system of equations in $G_c(z)$ and $G_w(z)$, from which we can get

$$f(z)G_c(z)=\pi_{0,w}\eta_w\mu_w z+\pi_{0,c}\mu_c\left[\eta_w z+(\lambda-z\mu_w)(1-z)\right], \quad (8)$$

where

$$f(z) = \lambda^2 z^3-\lambda(\eta_c+\eta_w+\lambda+\mu_w+\mu_c)z^2 + (\eta_c\mu_w+\eta_w\mu_c+\mu_c\mu_w+\lambda\mu_w+\lambda\mu_c)z-\mu_c\mu_w. \quad (9)$$

It can be proven that the polynom in Eq.(9) has only one root in the open interval $(0,1)$ [15]. This root is denoted as z_0. Setting $z=z_0$ into Eq.(8) gives

$$\pi_{0,w}\eta_w\mu_w z_0+\pi_{0,c}\mu_c\left[\eta_w z_0+\lambda z_0(1-z_0)-\mu_2(1-z_0)\right]=0.$$

After some algebraic manipulations with the last equation and Eq.(1), we obtain $\pi_{0,c}$ and $\pi_{0,w}$ as:

$$\pi_{0,c} = \frac{\eta_w(\mu-\lambda)z_0}{\mu_c(1-z_0)(\mu_w-\lambda z_0)}, \quad (10)$$

$$\pi_{0,w} = \frac{\eta_c(\mu-\lambda)z_0}{\mu_w(1-z_0)(\mu_c-\lambda z_0)}, \quad (11)$$

where $\mu=\pi_c\mu_c+\pi_w\mu_w$, and π_w is the steady-state probability of finding the system in some region with WiFi availability. Using standard Renewal theory [11]) we get $\pi_w=\frac{\eta_c}{\eta_c+\eta_w}$. Similarly, for the periods with only cellular access we have $\pi_c=\frac{\eta_w}{\eta_c+\eta_w}$[2].

Finally, for $G_c(z)$ and $G_c(z)$ we have from Eq.(8) and Eq.(6)

$$G_c(z)=\frac{\left[\eta_w(\mu-\lambda)z+\pi_{0,c}\mu_c(1-z)(\lambda z-\mu_w)\right]}{f(z)}, \quad (12)$$

$$G_w(z)=\frac{\left[\eta_c(\mu-\lambda)z+\pi_{0,w}\mu_w(1-z)(\lambda z-\mu_c)\right]}{f(z)}. \quad (13)$$

We define two new quantities $E[N_c]=\sum_{i=0}^{\infty} i\pi_{i,c}$ and $E[N_w]=\sum_{i=0}^{\infty} i\pi_{i,w}$. Hence, we have $E[N_c]=G_c^{'}(1)$ and $E[N_w]=G_w^{'}(1)$.

2. Note that μ is *not* the average experienced service rate, except for some limiting cases, e.g. when the system is always backlogged with traffic.

It is easy to see then that the average number of files in the system is $E[N]=E[N_c]+E[N_w]$. Replacing $z=1$ in Eq.(12)-(13), we get for the average number of files in the system

$$E[N]=\frac{\lambda}{\mu-\lambda}+\frac{\mu_c(\mu_w-\lambda)\pi_{0,c}+(\mu_c-\lambda)(\mu_w(\pi_{0,w}-1)+\lambda)}{(\eta_c+\eta_w)(\mu-\lambda)}.$$

Finally, using the Little's law $E[N]=\lambda E[T]$ [11], we obtain the average file delay in on-the-spot mobile data offloading:

Result 1. *The average file transmission delay in the data offloading scenario of Definition 1 and 2 is*

$$E[T]=\frac{1}{\mu-\lambda}+\frac{\mu_c(\mu_w-\lambda)\pi_{0,c}+(\mu_c-\lambda)(\mu_w(\pi_{0,w}-1)+\lambda)}{\lambda(\eta_c+\eta_w)(\mu-\lambda)}. \quad (14)$$

2.3 Low utilization approximation

In the previous subsection we derived a generic expression for the average delay of on-the-spot-offloading. However, the formula in Eq.(14) contains a root of a third order (cubic) equation, which while obtainable in closed-form, is quite complex. For this reason, in the remainder of this section we will consider simpler approximations for specific operation regimes. One such scenario of interest is when resources are underloaded (e.g. nighttime, rural areas) and/or traffic is relatively sparse (e.g. background traffic from social and mailing applications, messaging, Machine Type Communication, etc.).

For low utilization, there is almost no queueing so we can only consider the service time as an approximation of the total delay. We can thus use a fraction of the original Markov chain of Fig. 2 with only 4 states, as shown in Fig. 3 (i.e. number of jobs in the system ≤ 1). The service time will depend only on the probability of the flow arriving during a WiFi period (easily found to be $\frac{\eta_c}{\eta_c+\eta_w}$) or cellular only period ($\frac{\eta_w}{\eta_c+\eta_w}$)), and the amount of time to "hit" a 0 state (i.e. $\{0,c\}$ or $\{0,w\}$) from there. The latter can be derived in closed form by a simple application of first step analysis [16]. This gives us a first useful approximation, which becomes exact as $\lambda\to 0$ (the interested reader can find the detailed proof in [17]).

Low utilization approximation. *The average file transmission delay in the on-the-spot mobile data offloading for*

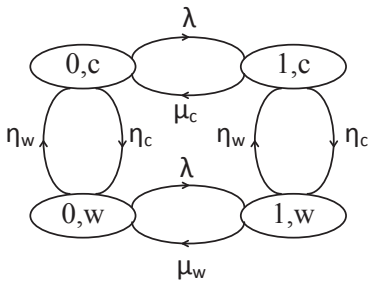

Fig. 3: The reduced Markov chain for $\rho \to 0$.

sparse traffic can be approximated by

$$E[T] = \frac{(\eta_w + \eta_c)^2 + \eta_c\mu_c + \eta_w\mu_w}{(\mu_c\mu_w + \mu_c\eta_w + \mu_w\eta_c)(\eta_c + \eta_w)}. \quad (15)$$

2.4 High utilization approximation

Another interesting regime is that of high utilization. As explained earlier, wireless resources are often heavily loaded, especially in urban centers, due to the increasing use of smart phones, tablets, and media-rich applications. Hence, it is of special interest to understand the average user performance in such scenarios. We provide an approximation that corresponds to the region of high utilization $(\rho \to 1)$, i.e. for which it holds that

$$\lambda \approx \frac{\eta_w}{\eta_c + \eta_w}\mu_c + \frac{\eta_c}{\eta_c + \eta_w}\mu_w.$$

Under this condition the polynomial of Eq.(9) becomes

$$f(z) = (z-1)[\lambda^2 z^2 - \lambda(\mu_c + \mu_w + \eta_c + \eta_w)z + \mu_c\mu_w]. \quad (16)$$

The root in the interval $(0, 1)$ of the function (16) is

$$z_0 = \frac{(\mu_c + \mu_w + \eta_c + \eta_w) - \sqrt{(\mu_c + \mu_w + \eta_c + \eta_w)^2 - 4\mu_c\mu_w}}{2\lambda},$$

since one other root is 1 and the third one is larger than 1. Hence, we get the following result:

High utilization approximation. *The average file transmission delay in the on-the-spot mobile data offloading for a user with heavy traffic can be approximated by*

$$E[T] = \frac{1}{\mu - \lambda}\left(1 - \frac{(\mu_c - \lambda)(\mu_w - \lambda)}{\lambda(\eta_c + \eta_w)}\right) + \frac{z_0}{\lambda(\eta_c + \eta_w)(1 - z_0)}\left(\frac{\mu_w - \lambda}{\mu_w - \lambda z_0}\eta_w + \frac{\mu_c - \lambda}{\mu_c - \lambda z_0}\eta_c\right) \quad (17)$$

2.5 Moderate utilization approximation

Finally, we can get an approximation for moderate utilization regimes by interpolating function $f(z)$. Due to space limitations, we only state here the result. Note, that unlike the low and high utilization approximations, which become tight in the limit, this is just a heuristic.

Moderate utilization approximation. *The average file transmission delay in the on-the-spot mobile data offloading for moderate traffic can be approximated by*

$$E[T] = \frac{1}{\mu - \lambda} + \frac{\mu_c(\mu_w - \lambda)\pi_{0,c} + (\mu_c - \lambda)(\mu_w(\pi_{0,w} - 1) + \lambda)}{\lambda(\eta_c + \eta_w)(\mu - \lambda)}, \quad (18)$$

where $\pi_{0,c}$ and $\pi_{0,w}$ are given by Eq.(10)-(11), and $z_0 = \frac{1}{\varepsilon}\frac{\mu_c\mu_w}{\eta_c\mu_w + \eta_w\mu_c - \lambda(\eta_c + \eta_w) + \mu_c\mu_w}$. ε is a fitting parameter that takes values in the range 1.4-1.6, and can be fine-tuned empirically for better results.

2.6 Generic file size distribution approximation

Our analysis so far considers exponentially distributed flow sizes. Yet, for some traffic types, heavy-tailed file sizes were reported [18]. Unfortunately, generalizing the above 2D chain analysis for generic files is rather hard, if not impossible. Nevertheless, we can use the M/G/1 P-K formula [11] as a guideline to introduce a similar "correction factor" related to smaller/higher file size variability. Due to space limitations, we state this here without proof. However, the equivalence with the M/G/1 vs. M/M/1 difference is easily evident, and the interested reader is referred to any queueing theory textbook.

Let c_v denote the coefficient of variation for the file size distribution, and $E[T]$ and $E[S]$ denote the system and service time, respectively, for exponentially distributed packet sizes (as derived before). The average system time for the generic packet size distributions in the ordinary M/G/1 can be written through the delay of the M/M/1 system as

$$E[T_g] = E[S] + E[T_Q] \cdot \frac{1 + c_v^2}{2}, \quad (19)$$

where $E[T_Q] = E[T] - E[S]$ is the average queueing time for the exponentially distributed packet sizes. After some very simple calculus steps, we obtain the following approximation for generic file sizes.

Result 2. *The average file transmission delay in the on-the-spot mobile data offloading for generic file size distributions can be approximated by*

$$E[T_g] = \frac{1 - c_v^2}{2}E[S] + \frac{1 + c_v^2}{2}E[T]. \quad (20)$$

2.7 Offloading efficiency

Finally, an important parameter that can quantitatively characterize data offloading is the *offloading efficiency* OE, defined as the ratio of the amount of transmitted data through WiFi against the total amount of transmitted data. Higher offloading efficiency means better performance for both client and operator. Knowing this parameter is especially important when it comes to calculating how much a user will have to pay, knowing that the charges for using Internet access are not the same for WiFi as are for cellular network. The derivation of this metric follows easily from the earlier discussions, and its expression is given below.

Result 3. *The offloading efficiency in an on-the-spot mobile data offloading system is*

$$OE = \frac{\mu_w}{\mu_w + \mu_c \frac{G_c(1)-G_c(0)}{G_w(1)-G_w(0)}}. \quad (21)$$

3 Analyzing Offloading in Heterogeneous Cellular Networks

In the previous section, we considered a simpler scenario with two networks (WiFi and Cellular) and respective rates. Nevertheless, in current and future networks different areas might be covered by different cellular network technologies (GPRS, EDGE, HSPA, LTE), and multiple "short-range" options might exist in addition to WiFi (e.g. femto- or other small-cell technologies). Coverage holes might also exist. Finally, even within the same technology (e.g. WiFi), the rate might be different across different access points. While such scenarios could still be emulated by the basic model, by "absorbing" these difference across ON and OFF periods into an average rate, as explained earlier, it would be interesting to see whether we can extend our basic model to better predict performance in more complex scenarios, where a mobile user might be switching between a number of different technologies and/or rates during a large time window.

3.1 The model

As earlier in Section 2.1, we again consider a mobile user moving between locations with different network characteristics. However, now there are N possible options a user can encounter, rather than just two 2. To keep discussion simple, in the remainder we will assume that these different options correspond to different technologies, such as WiFi, 2.5G, 3G, 3G+, 4G, etc., or even no network at all. Note however that the analysis to follow can be also applied to a scenario where, e.g. we have a number of different rates a user could experience within the same technology. This just requires increasing N, and updating the transition rates between different states accordingly (e.g. to some long-term statistics available).

We will assume that whenever there is WiFi coverage, all the traffic will be transmitted through the corresponding AP. When the WiFi connectivity is lost, the traffic will be sent (received) through the cellular technology available. In case there are multiple options in terms of the network coverage, we assume that the user will switch to the technology that offers the highest rates. Nevertheless, this could be a policy matter that will not affect our analysis. As before, we will assume that network transitions do not cause any impairments to the flow. At the end of this section, we briefly discuss how the generic model could be extended to include also possible interruptions caused from switching to a different technology.

We model the network availability with the multilevel scheme as shown in Fig. 4. The duration of each period is exponentially distributed with rate $\eta_i, i = 1, \ldots, N$. Each period T_i corresponds to the time duration during which the mobile user is communicating through the same access network technology without interruption. We call these periods *levels* or *phases*. The period durations of any level are mutually independent, and at the same time independent of the durations of periods of other levels. The data transmission rates during periods with different connectivities are denoted as $\mu_i, i = 1, \ldots, N$. As before, the traffic arrival process is Poisson and file sizes are exponentially distributed. The scheduling discipline is the same as before (FCFS).

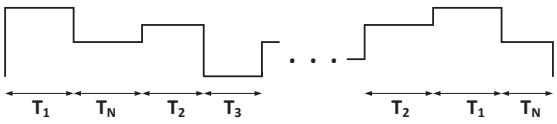

Fig. 4: The multi-network availability model.

3.2 Analysis

In this subsection we derive the expression for the average file delay in a multilevel on-the-spot offloading system. Based on the assumptions we have made our system can be modeled with a 2D Markov chain that is bounded in one dimension (the dimension that represents the number of levels). This is shown in Fig. 5[3]. The possible level transitions are shown only for the state $(1, 0)$ and partially for $(2, j)$ to avoid making the figure look more complex. $\pi_{i,k}$ denotes the stationary probability of being at level i (where i corresponds to areas served by a given technology with rate μ_i). There are a number of possible transitions now corresponding to the following events:

new arrival: From any state, the chain moves to the right (horizontally) with rate λ.

flow finishes transmission: From any state $\{i, k\}$ $(k > 0)$, the chain moves to the left (horizontally) with rate μ_i.

change in the network connectivity: The chain moves (vertically) from level i to another level j (transition to all the levels possible with no exception) with rate $\eta_{i,j}$.

If we denote with η_i the (total) rate at which a user leaves an area covered by technology (or rate) i, it holds that $\eta_i = \sum_{j=1}^{N} \eta_{i,j}$.

We start by writing the balance equations for this Markov chain as

$$(\lambda + \eta_i)\, \pi_{i,0} = \mu_i \pi_{i,1} + \sum_{j=1}^{N} \eta_{j,i} \pi_{j,0}, \quad (22)$$

for $i = 1, \ldots, n$, $j = 0$, and

$$\begin{aligned} &(\lambda + \mu_i + \eta_i)\, \pi_{i,k} \\ = \;& \lambda \pi_{i,k-1} + \mu_i \pi_{i,k+1} + \sum_{j=1}^{N} \eta_{j,i} \pi_{j,k}, \end{aligned} \quad (23)$$

3. It is worth mentioning that not all the possible transitions are depicted in Fig. 5.

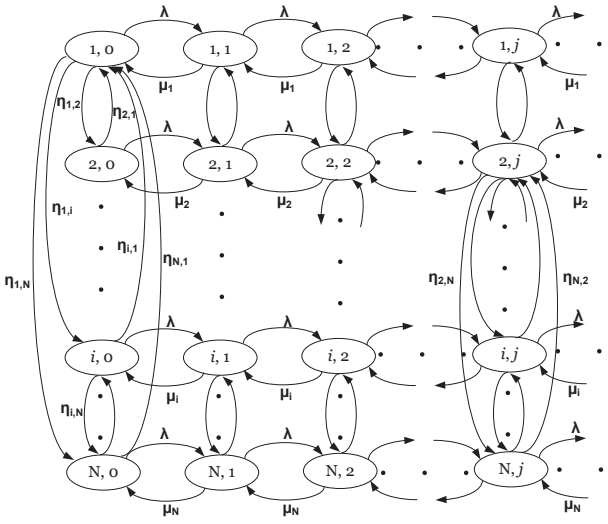

Fig. 5: The 2D Markov chain for multilevel on-the-spot mobile data offloading model.

for $i = 1, \ldots, n$ and $k > 0$. Summing Eq.(22) and Eq.(23) multiplied by z^k, and then summing up over all k, we get

$$\lambda \sum_{k=0}^{\infty} \pi_{i,k} z^k + \mu_i \sum_{k=1}^{\infty} \pi_{i,k} z^k + \sum_{k=0}^{\infty} \pi_{i,k} z^k \sum_{j=1}^{N} \eta_{i,j}$$
$$= \lambda \sum_{k=1}^{\infty} \pi_{i,k-1} z^k + \mu_i \sum_{k=1}^{\infty} \pi_{i,k} z^{k-1} + \sum_{k=0}^{\infty} \pi_{j,k} z^k \sum_{j=1}^{N} \eta_{j,i}. \quad (24)$$

We define the probability generating function for each level as

$$G_i(z) = \sum_{k=0}^{\infty} \pi_{k,i} z^k, \ |z| \leq 1 \ i = 1, \ldots, N. \quad (25)$$

Eq.(24) is now transformed to

$$\lambda G_i(z) + \mu_i \left[G_i(z) - \pi_{i,0} \right] + G_i(z) \sum_{j=1}^{N} \eta_{i,j}$$
$$= \lambda z G_i(z) + \frac{\mu_i}{z} \left[G_i(z) - \pi_{i,0} \right] + \sum_{j=1}^{N} \eta_{j,i} G_j(z). \quad (26)$$

After performing some algebra we have

$$\left[\lambda z(1-z) + \mu_i(z-1) + \eta_i z \right] G_i(z) - \sum_{j=1}^{N} \eta_{j,i} z G_j(z)$$
$$= \mu_i (z-1) \pi_{i,0}, i = 1, \ldots, N. \quad (27)$$

In Eq.(27), after introducing the substitution

$$f_i(z) = \lambda z(1-z) - \mu_i(1-z) + \eta_i z, \quad (28)$$

we obtain the following equation

$$\mathbf{F}(z)\mathbf{g}(z) = (z-1)\boldsymbol{\theta}, \quad (29)$$

where

$$\mathbf{F}(z) = \begin{bmatrix} f_1(z) & -\eta_{2,1}z & -\eta_{3,1}z & \ldots & -\eta_{N,1}z \\ -\eta_{1,2}z & f_2(z) & -\eta_{3,2}z & \ldots & -\eta_{N,2}z \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ -\eta_{1,N}z & -\eta_{2,N}z & -\eta_{3,N}z & \ldots & f_n(z) \end{bmatrix},$$

$$\mathbf{g}(z) = \begin{bmatrix} G_1(z) \\ G_2(z) \\ \vdots \\ G_N(z) \end{bmatrix}, \boldsymbol{\theta} = \begin{bmatrix} \mu_1 \pi_{1,0} \\ \mu_2 \pi_{2,0} \\ \vdots \\ \mu_N \pi_{N,0} \end{bmatrix}.$$

Aplying Cramer's rule to Eq.(29) we obtain

$$|\mathbf{F}(z)| \, G_i(z) = |\mathbf{F}_i(z)| \, (z-1). \quad (30)$$

$|\mathbf{F}_i(z)|$ is the determinant obtained after replacing the ith column of $|\mathbf{F}(z)|$ with $\boldsymbol{\theta}$. As can be observed from Eq.(28), at point $z = 1$, $f_1(1) = \eta_1$. Also, at this same point the sum of the elements in rows 2 to N of the first column $(-\eta_{1,2}z - \eta_{1,3}z - \ldots - \eta_{1,N}z)$ represents the sum of transition rates out of state 1 multiplied by -1. If we substract row 1 from the sum of the other rows, we have 0 at the element $\{1,1\}$ of the determinant $|\mathbf{F}_i(z)|$. Similar conclusions can be drawn for the other elements of the first row. Hence, we can obtain an equivalent determinant $|\mathbf{F}(z)|$ with all the elements of the first row equal to 0. So, $z = 1$ is one root of this determinant. Hence, we can write

$$|\mathbf{F}(z)| = (z-1)Q(z). \quad (31)$$

Replacing Eq.(31) into Eq.(30) we get

$$Q(z)G_i(z) = |\mathbf{F}_i(z)|. \quad (32)$$

In order to find the partial probability generating functions $G_i(z)$, we need first to find the zero probabilities $\pi_{1,0}, \pi_{2,0}, \ldots, \pi_{N,0}$. To do this, we proceed in the following way. First, we find the roots of $Q(z)$. Since our system is of order $N > 2$, these solutions can be obtained only numerically. The polynomial $Q(z)$ is of degree $2N-1$. However, only $N-1$ of its roots lie in the interval $(0,1)$ (which is our interval of interest)[4]. We denote these roots as $z_1, \ldots, z_{N-1}$. Since $G_i(z) \neq 0$ (All the probabilities $p_{k,i}$ are positive), then from Eq.(32), we have that $|\mathbf{F}_i(z_j)| = 0, i = 1, \ldots, N, j = 1, \ldots, N-1$. However, from Eq.(32) we can observe that, for each z_j, and any pair $1 \leq i, l \leq N$, $\frac{|\mathbf{F}_i(z_j)|}{|\mathbf{F}_l(z_j)|} = const$. This means that for each z_j we have N homogeneous linear equations that differ from each other only by a constant factor. Hence, $|\mathbf{F}_i(z_j)| = 0$ gives only one independent equation for each root z_j. Given that there are $N-1$ different roots z_j, it turns out that there are in total $N-1$ independent equations. Since we have N unknown probabilities $\pi_{1,0}, \pi_{2,0}, \ldots, \pi_{N,0}$, and only $N-1$ equations, we cannot obtain unique solutions for these

4. The proof to this claim is rather long and complicated, and due to space limitations we do not show it here. It was proven by Mitrani and Itzhak in [19].

probabilities. So, we need another condition that relates these zero probabilities, and that is independent of the other $N-1$ equations.

Let's consider the vertical cut between states k and $k+1$. The balance equation through this cut is

$$\lambda(\pi_{1,k}+\pi_{2,k}+\ldots+\pi_{N,k}) = \mu_1\pi_{1,k+1}+\ldots+\mu_N\pi_{N,k+1}. \quad (33)$$

Summing over all k yields

$$\lambda\sum_{i=1}^{N}\pi_i = \mu_1(\pi_1-\pi_{1,0})+\ldots+\mu_1(\pi_N-\pi_{N,0}). \quad (34)$$

$\pi_i = \sum_{k=0}^{\infty}\pi_{i,k}$ denote the percentage of time the system is in level i. Eq.(34) can be rewritten as

$$\mu-\lambda = \sum_{i=1}^{N}\mu_i\pi_{i,0}, \quad (35)$$

where $\mu = \sum_{i=1}^{N}\mu_i\pi_i$ is the average service rate of the system. Eq.(35) is the Nth equation of the system we need to solve in order to get the zero probabilities. However, we do need to determine the probabilities π_i first.

We can find π_i by following a standard embedded MC approach for the (collapsed) chain with only N states (corresponding to the N levels). If we define $q_{i,j}$, the transition probabilities in the embedded chain, as $q_{i,j} = \frac{\eta_{i,j}}{\eta_i}$, then

$$\pi_i = \frac{\frac{r_i}{\eta_i}}{\sum_{i}^{N}\frac{r_i}{\eta_i}}, \quad (36)$$

where r_i are the solutions to the global balance equations for the embedded DTMC: $\sum_{i=1}^{N} r_i = 1$, and $r_j = \sum_{i=1}^{N} r_i q_{i,j}$.

Replacing Eq.(36) into Eq.(35), we have the Nth equation of our system. Now, solving that system we get all the zero probabilities. The partial PGFs are found from Eq.(32) as

$$G_i(z) = \frac{|\mathbf{F}_i(z)|}{Q(z)}, i = 1,\ldots,N. \quad (37)$$

The average number of packets in the system is

$$E[N] = \sum_{i=1}^{N} G_i'(1). \quad (38)$$

Using Little's law $E[N] = \lambda E[T]$, we get the following result:

Result 4. *The average file delay in a multilevel on-the-spot offloading system is given by*

$$E[T] = \frac{1}{\lambda}\sum_{i=1}^{N}\left(\frac{|\mathbf{F}_i(z)|}{Q(z)}\right)'_{z=1}. \quad (39)$$

We conclude this section by discussing how the N-level model could be extended to handle interruptions of a flow, due to switching from one technology to another. If we assume that a flow is delayed a bit due to this interruption, but then resumes over the new network, this could be captured, for example, by introducing $2N$ levels, instead of N that we have now. Each of the current levels would have a corresponding quasi level, into which the state of system would switch first. These levels correspond to the time needed to resume transmission. After staying for some time in one of those levels, the state of the system would move to a "real" level, in which the communication would be reestablished. While being in the quasi-level state, the data rate is 0, and from it the system can only move to the corresponding level, attached to the quasi level state. The analysis would then be the same as that described before.

4 Simulation Results

4.1 Basic model validation

In this section we will validate our theory against simulations for a wide range of traffic patterns, different values of file sizes and different average WiFi availability periods and availability ratios. We define the WiFi availability ratio as $AR = \frac{E[T_{ON}]}{E[T_{ON}]+E[T_{OFF}]} = \frac{\eta_c}{\eta_w+\eta_c}$. Unless otherwise stated, the durations of WiFi availability and unavailability periods will be drawn from independent exponential distributions with rates η_w and η_c, respectively. We mainly focus on two scenarios, related to the user's mobility. The first one considers pedestrian users with data taken from [8]. Measurements in [8] report that the average duration of WiFi availability period is 122 min, while the average duration with only cellular network coverage is 41 min (we use these values to tune η_w and η_c). The availability ratio reported is 75 %. The second scenario corresponds to vehicular users, related to the measurement study of [9]. An availability ratio of 11 % has been reported in [9]. For more details about the measurements we refer the interested reader to [8] and [9]. Finally, unless otherwise stated, file/flow sizes are exponentially distributed, and file arrivals at the mobile user is a Poisson process with rate λ.

A1.Validation of the main delay result

We first validate our model and 2-level result (Eq.(14)) against simulations for the two mobility scenarios mentioned (pedestrian and vehicular). The data rate for WiFi is assumed to be 2 Mbps (this is close to the average data rate obtained from measurements with real traces in [20]), and we assume that the cellullar network is 3G, with rate 500 kbps. The mean flow size is assumed to be 125 kB[5].

Fig. 6 shows the average file transmission delay (i.e. queueing + transmission) for the pedestrian scenario, for different arrival rates. The range of arrival rates shown corresponds to a server utilization of 0-0.9. We can observe, in Fig. 6, that there is a good match between theory and simulations. Furthermore, the average file transmission delay is increased with the arrival rate, as expected, due to queueing effects. Fig. 7 further illustrates the average file transmission delay for the vehicular scenario. We can observe there that the average transmission time is larger

5. This value is normalized for the arrival rates considered, to correspond to the traffic intensities reported in [9]. We have also considered other values with similar conclusions drawn.

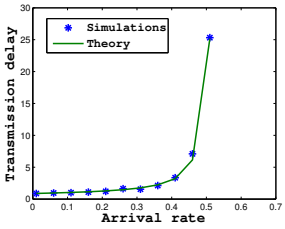

Fig. 6: Pedestrian user.

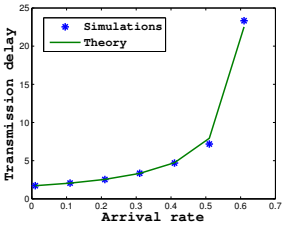

Fig. 7: Vehicular user.

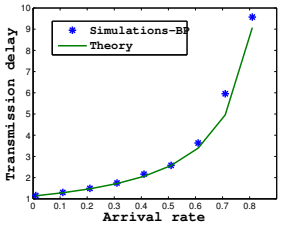

Fig. 8: BP vehicular periods.

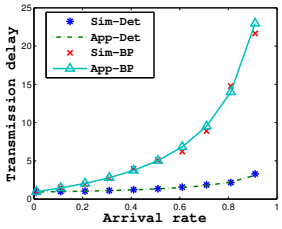

Fig. 9: Generic flow sizes.

than in Fig. 6. This is reasonable, due to the lower WiFi availability, resulting in most of the traffic being transmitted through the slower cellular network interface. Once more, we can observe a good match between theory and simulations.

A2.Validation against non-exponential ON-OFF periods

In the previous scenarios, we have used realistic values for the transmission rates and WiFi availabilities, but we have assumed exponential distributions for ON and OFF periods, according to our model. While the actual distributions are subject to the user mobility pattern, a topic of intense research recently, initial measurement studies ([8], [9]) suggest these distributions to be "heavy-tailed". It is thus interesting to consider how our model's predictions fare in this (usually difficult) case. To this end, we consider a scenario with "heavy-tailed" ON/OFF distributions (Bounded Pareto-BP). Due to space limitations, we focus on the vehicular scenario. The shape parameters for BP ON and OFF periods are $\alpha = 0.59$ and $\alpha = 0.64$, respectively. We consider a cellular rate of 800 kbps. We change the value of the data rate to see that our analysis holds for other values as well. Fig. 8 compares the average file delay for this scenario against our theoretical prediction. Interestingly, our theory still offers a reasonable prediction accuracy, despite the considerably higher variability of ON/OFF periods in this scenario[6]. While we cannot claim this to be a generic conclusion for any distribution and values, the results underline the utility of our model in practice.

A3.Validation of non-exp flow sizes

To conclude our validation, we finally drop the exponential flow assumption as well, and test our generic file size results of Eq.(20). Fig. 9 compares analytical and simulation results for deterministic, and Bounded Pareto distributed files sizes (shape parameter $\alpha = 1.2$ and $c_v = 3$). Mean file size is in both cases 125KB, and the rest of the parameters correspond to the vehicular scenario (exp. ON and OFF periods). We observe that higher size variability further increases delay, as expected. Somewhat more surprisingly, the observed accuracy in both cases is still significant, despite the heuristic nature of the approximation and the complexity of the queueing system.

A4. Validation of approximations

Having validated the main result of Eq.(14) we now proceed to validate the various simpler approximations we have proposed in Section 2. We begin with the low utilization approximation of Section 2.3 with $AR = 0.75$ (similar accuracy levels have been obtained with other values). Fig. 10 shows the flow delay for low arrival rates in the range $0.01 - 0.1$, which correspond to a maximum utilization of around 0.1. We can observe that the low utilization approximation provides a good match with the generic result and simulations. As λ increases, the difference between the approximated result and the actual value increases. For $\rho = 0.1$, the approximation error is around 5%. This is reasonable, as we have strictly assumed that there might be at most one file present in the system.

We next consider the high utilization regime and respective approximation (Eq.(17)). We consider utilization values of 0.8-0.95. Fig. 11 shows the delay for high values of λ, and $AR = 0.5$ (we have again tried different values). We can see there that our approximation is very close to the actual delay and should become exact as ρ goes to 1.

Finally, we consider approximation result (Eq.(18)) for moderate utilization values, in the range $0.3 - 0.7$ ($AR = 0.5$). Fig. 12 compares theory and simulations for the delay in this intermediate utilization regime. For moderate ρ, the value of the coefficient ε is 1.5. It is chosen empirically. Although this approximation is heuristic, and does not become exact for any utilization value (unlike the cases of the low/high utilization approximations), we can see that the accuracy is still satisfactory and improves for higher ρ.

4.2 Limitations of the 2-level model

So far, we have been assuming constant WiFi data rate in all the regions with WiFi coverage. While in theory it enables analytical tractability, this assumption is rather unrealistic, since the actual rate experienced in different APs will depend on AP load, distance, backhaul technology, etc. Therefore, it is particularly interesting to consider scenarios where the WiFi rate might be different at each connected AP. Specifically, we simulate a scenario where the average data rate over all APs is again 2 Mbps, but the actual rate for each ON (WiFi) period is selected uniformly in the interval 1-3 Mbps. The other parameters remain unchanged. In Fig. 13, we compare simulation results for this scenario against our theoretical result (which assumes a constant WiFi rate of 2 Mbps in every AP). From Fig. 13 it is evident that WiFi rate variability does not affect significantly the performance, making thus our results applicable in this case as well, despite the variable nature of the WiFi rate.

6. This has been the case with additional distributions and values we have tried. We have also observed that the error generally increases (decreases) when the difference between WiFi and cellular rates increases (decreases).

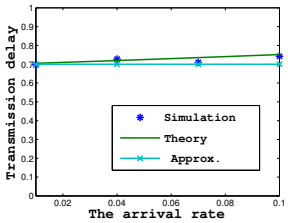

Fig. 10: The low utilization approx.

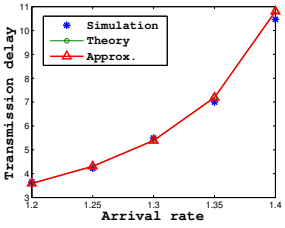

Fig. 11: The high utilization approx.

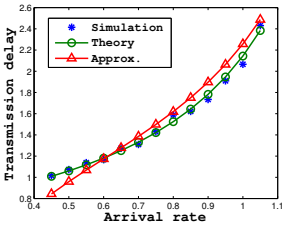

Fig. 12: The approx. for AR=0.5.

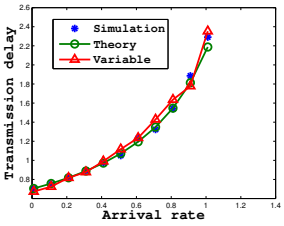

Fig. 13: Variable WiFi rates.

As we saw in Fig. 13, our model can predict with an excellent accuracy the delay even in a system in which the WiFi rate is not constant. Out of that one might infer that our two-level model is sufficient to provide a high-accuracy approximate analysis for a network technology with any discrete number of the data rates. It would be enough to lump all the levels into one level (phase) with a data rate equal to the average data rate of all the other levels. However, this turns out to be incorrect in the general case. The reason why this method gave good match in Fig. 13 is because the rates were chosen from a uniform distribution, i.e. the rates were close to each other. We investigate the effect of highly variable rates below.

We consider first the data rates to be drawn from an exponential distribution with the same average as in Fig. 13. The same holds for all the other system related parameters as well. Fig. 14 shows the average delay for this scenario. As can be seen from there, our theory cannot predict the delay correctly anymore. The discrepancy is even more pronounced when there is higher variability in the data rates. In the same plot we show the delay for data rates drawn from a Pareto distribution with shape parameter $\alpha = 1.2$, and the same average. The delay now is much higher, exceeding $5\times$ the predicted result by our theory.

Having established that the two level model fails in predicting the delay for variable rates that show a tendency of being quite dispersed from the mean, we move on with investigating the effect of availability ratio to the delay. Fig. 15 shows the average delay vs. arrival rate for $AR = 0.75$ and uniform, exponential and Pareto distributed data rates. The other parameters are the same as for Fig. 14. It is also shown the delay curve from our theory with constant rates. Here also, our theory can give accurate result when data rates underly uniform distribution, but fails when it comes to the higher variability data rates.

Finally, we consider the effect of larger difference between the WiFi and cellular rates. For that purpose we consider a scenario with cellular rate of 0.5 Mbps, instead of 1 Mbps, and with the same average WiFi rate (2 Mbps). The availability ratio is 0.5. Fig. 16 illustrates the average delay. In this case also, the same conclusions hold as before.

So, we can say that the accuracy of our model holds in scenarios where the data rates are relatively close to each other, and it does not depend heavily on the availability ratio nor on the cellular rate. When it comes to data rates that are subject to higher variability, we need the N level model of Section 3.

To further enhance our claims about the necessity of using the N-level model, we consider the scenarios with multiple access technologies (WiFi, 3G, HSPA, LTE). The corresponding parameter values are given in Table 2. The values taken belong to the range of intervals given in [21], [22]. The other system parameters are the same as in the previous considered scenarios. We lump the cellular network levels into one single level with average duration equal to the sum of their individual average durations (15 s), and average data rate equal to the weighted average of the corresponding rates (3.5 Mbps). Then, we use our 2-levels model to find the average delay. Fig. 17 illustrates the average delay vs. the arrival rate. On the same plot, we show the actual simulated delay. As can be seen from Fig. 17, the 2-levels model cannot capture the case with multiple heterogeneous networks, as the prediction is very far from the actual delay. Hence, for such cases we need the N-level model.

4.3 N-level model validation

Next, we consider the scenarios with multiple access technologies (WiFi, 3G, HSPA, LTE) or even without network coverage at all. Namely, there are operators that might offer 4G coverage only in some regions, while in the others they offer only 3G. There might also exist regions with little or no coverage at all (in sparse populated areas). In the following we will see how our multilevel theory of Section 3 will cope with the actual (simulated) scenarios. Unless otherwise stated, the data rates and average durations are given in Table 2.

First, we focus on the scenario when there are 3 possible network choices: WiFi, 3G and LTE. The policy here is that WiFi is the network with absolute priority. When there is no WiFi coverage, LTE has priority over 3G. We assume that there is always 3G network availability. There is an equal probability to move to any other access technology after leaving the current network. Since, there are only 3 possible levels, this probability is equal to 0.5. Flows are exponentially distributed with average size of 125 kB, and the arrival process is Poisson. The availability ratio of the WiFi network (Eq.(36)) is found to be 50%.

Fig.18 shows the average file delay vs. the arrival rate for this system. As can be seen our theory matches with simulations. As expected, the delay increases with increasing the traffic arrival rate, due to the queueing effect.

The second scenario shown in Fig. 18 corresponds to $N = 4$ possible levels: WiFi, 3G, HSPA, and LTE. The

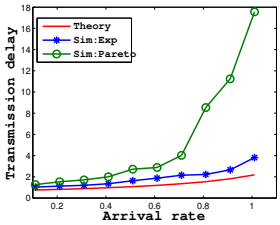

Fig. 14: Variable rates for $AR = 0.5$.

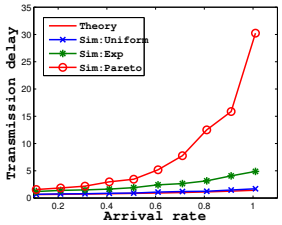

Fig. 15: Variable rate with $AR = 0.75$.

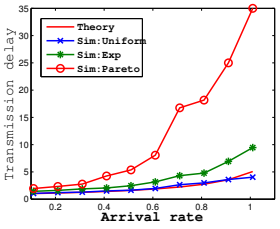

Fig. 16: Variable rate with lower cellular rate.

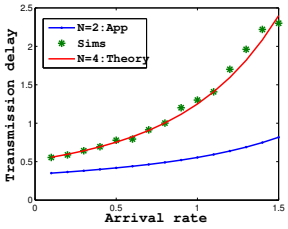

Fig. 17: The approximation with the 2 level model.

TABLE 2: The parameters for different access technologies [21], [22].

Technology	Data rate	Average duration
WiFi	2 Mbps	10 s
3G	1 Mbps	3 s
HSPA	1.5 Mbps	5 s
LTE	10 Mbps	5 s
No coverage	0	2 s

parameters are given in Table 2. The probability of moving to any specific level is 1/3 now. The availability ratio now turns out to be 43.5%. There is a nice fit with theory again. The average delay is a bit higher compared to the previous example, since the HSPA data rate is lower compared to WiFi, and the other networks' characteristics are the same.

We also consider the possibility of not having network coverage at all. Now, we have to consider 5 levels (Fig. 18). Given that the other 4 levels have the same parameters as above, for the no network availability we choose the average duration to be 2s. The probability of encountering a specific level after leaving the one in use is 0.25. The availability ratio is 40%. Again, there is a match between theory and simulations that shows that our theory is correct. As expected the delay is larger, because there are some time periods when there is no connectivity at all.

Finally, we consider scenarios with lower WiFi availability ratio, and higher LTE duration periods than before (10 s). The availability ratios for $N = 3, 4, 5$ are 40%, 36% and 33%, respectively. The other parameters are exactly the same as before. Fig. 19 illustrates the average delay. As can be seen, the delays are much lower now. This comes from the fact that there is a lower degree of WiFi connectivity, and higher degree of LTE coverage. Since the LTE data rates are much higher, the delay is significantly reduced. The delay reduction can exceed 20%. The only advantage in using WiFi offloading under these circumstances lies in the lower prices WiFi operators offer, as opposed to LTE charges.

4.4 Non-exponential assumptions

So far, we have validated our model for the exponentially distributed durations of the different levels, as well as for exponentially distributed flow sizes. Next, we drop these assumptions and see how our theory behaves under these more general conditions. First, we keep the exponential assumption on file sizes, and consider heavy-tailed distributions for the durations of the different levels. There are $N = 4$ possible levels. The average durations of the corresponding phases are identical to those of Fig. 18, only that now they are Bounded Pareto with shape parameter $\alpha = 1.2$. Fig. 20 shows the average file delay. Surprisingly enough, our theory that is valid only for exponentially distributed periods, is able to predict the delay even for heavy-tailed distributions with a remarkable accuracy.

Finally, we drop exponential assumptions for both the level durations and flow sizes, and see how our generic flow size distribution approximation (Eq.20) behaves. We keep other parameters unchanged. The phase durations are Bounded Pareto with identical shape parameter as before ($\alpha = 1.2$). We consider two scenarios in terms of the distribution of flow sizes. While in the first one, all the flows have constant size, in the second one the flows have sizes that are drawn from a Bounded Pareto distribution with parameters $L = 0.24, H = 93, \alpha = 1.2$. It should be mentioned that the average flow size remains unchanged. Fig. 21 illustrates the delay for both scenarios. As can be seen, our proposed approximation (although heuristic) can predict the average delay quite satisfactorily, despite the very complex system we are dealing with. This increases the usefulness of our model. Another outcome of the model is that the delay, as in all other queueing systems, is higher for packets with higher variability.

4.5 Offloading Gains

We have so far established that our analytical model offers considerable accuracy for scenarios commonly encountered in practice. In this last part, we will thus use our model to acquire some initial insight as to the actual offloading gains expected in different scenarios. The operator's main gain is some relief from heavy traffic loads leading to congestion. The gains for the users are the lower prices usually offered for traffic migrated to WiFi, as well as the potential higher data rates of WiFi connectivity. There are also reported energy benefits associated [23], but we do not consider them here. Specifically, we will investigate the actual gains from data offloading, in terms of average transmission delay (related to user performance) and offloading efficiency (% of total traffic actually sent over WiFi - of interest to both the operator and the user). We consider two key parameters of interest that can affect these metrics: availability ratio and WiFi/cellular rate difference.

We first consider how transmission delay changes as a function of availability ratio, for different traffic intensities:

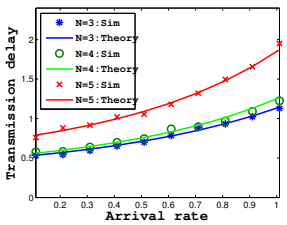

Fig. 18: The transmission delay.

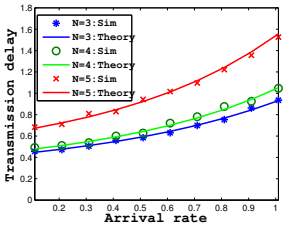

Fig. 19: The transmission delay.

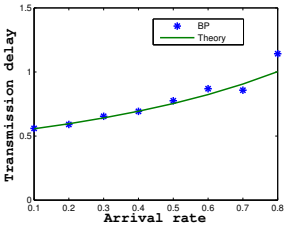

Fig. 20: Bounded Pareto periods.

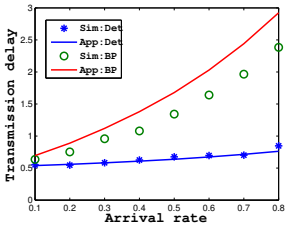

Fig. 21: Generic flow sizes.

very sparse, relatively sparse ($\rho = 0.15$) and medium ($\approx 40\%$). In the last scenario, however, when the user will be in zones in which it can connect only to lower rate access technologies the intensity would be much higher. Fig. 22 shows the average delay vs. AR for those traffic intensities. We can observe that the delay decreases as WiFi availability increases. More data are transmitted through the WiFi network, and hence the delay is lower since we have assumed that, on average, WiFi delivers better rates. A more interesting observation is that the delay improvement with higher WiFi availability values, is considerably more sharp, when the traffic load is higher. While for the arrival rate of $\lambda = 0.01$ the delay difference between the highest and the lowest availability ratios is less than 40%, this value exceeds $2\times$ for medium arrival rates. This seems to imply that denser WiFi deployments do not offer significant performance gains to users in low loaded regions, despite the higher rates offered, but could have a major impact on user experience, in heavily loaded areas.

As mentioned in Section 2, offloading efficiency is a very important quantity in characterizing mobile data offloading. Also, one might expect offloading efficiency to simply increase linearly with the availability ratio (i.e. % of data offloaded = % of time with WiFi connectivity). As it turns out, this is not the case. To better understand what affects this metric, we consider the impact of different cellular rates as well as different AR on the offloading efficiency. For the WiFi network we take the data rate to be 2 Mbps, and for the cellular we consider rates of 0.3 Mbps, 0.5 Mbps and 1 Mbps. Fig.23 illustrates the offloading efficiency vs. availability ratio for a moderate arrival rate of $\lambda = 0.3$. For comparison purposes we also depict the line $x = y$ (Offloading efficiency = AR). First, as expected, we can observe that offloading efficiency increases with availability ratio, in all scenarios. However, this increase is not linear. More interestingly, the actual offloading efficiencies are always higher than the respective availability ratio, and increase as the difference between the WiFi and the cellular rate increases. For $AR = 0.4$, 75% of the data are offloaded to WiFi when the ratio is 6.67 compared to 50% for a ratio of 2. The reason for this is that, due to the lower cellular rates, traffic arriving during the cellular (only) availability period ends up being transmitted during the next WiFi period due to queueing delays. This effect becomes more pronounced as the rate difference increases. Also, although not shown here, the respective offloading efficiency increases even further as traffic loads increase. Summarizing, these findings are particularly interesting to operators (and users), as they imply that high offloading efficiencies can be achieved for loaded regions, without necessarily providing almost full coverage with WiFi APs.

Finally, let us consider the impact of installing new LTE base stations and WiFi access points on the user experience. At first we assume that there is a coverage of 50 % (the same order as in [8]) with WiFi APs, and that there is no LTE. Next, in the regions with no WiFi coverage, the 3G base stations are being replaced successively with LTE base stations. We consider the impact of the deployment of LTE base stations on the sparse and medium-to-high traffic intensities ($\lambda = 0.1$ and $\lambda = 1$). The WiFi data rate is 2 Mbps, while for the 3G and LTE the data rate is 1 Mbps and 10 Mbps, respectively. Fig. 24 depicts the dependency of the average file delay on the % of the LTE coverage (not covered by WiFi AP) which have replaced the 3G base stations. On the x-axis the value 0 denotes that there is 50 % WiFi coverage, and the regions with no WiFi APs are covered with 3G base stations. The value of 50% on the x-axis refers to the complete replacement of 3G base stations with LTE base stations (at least in the regions with no WiFi). As can be seen from Fig. 24 when it comes to zones with sparse traffic, increasing the number of LTE base stations does not necessarily improve too much the performance of the mobile users. For example, if a mobile operator decides to completely replace the 3G base stations with LTE base stations, the delay will be reduced by less than $2\times$. Since the deployment and maintenance of the LTE base stations implies an increased cost for the mobile operator, it is not economical to upgrade the network to 4G in the regions with sparse traffic. As opposed to this, when it comes to zones with a large number and very active users, such as city centers, university campuses etc., the full deployment of LTE will reduce the delay for mobile users up to $6\times$ according to our scenario (see Fig. 24). Hence, in such regions it is beneficial for both the mobile operator and the users to upgrade the base stations.

Contrary to the previous case, now we decide to deploy additional APs and keep the actual 3G base stations. The other parameters remain the same as before. Fig. 25 shows the average delay vs. the WiFi availability ratio for the two traffic intensities (sparse and medium-to-high). We assume that at the beginning there is a coverage of 50% with WiFi APs. If we compare Fig. 25 and Fig. 24, for sparse traffic, we can notice that the difference in the delay is very low (less than 20%). On the other hand, for dense traffic if we

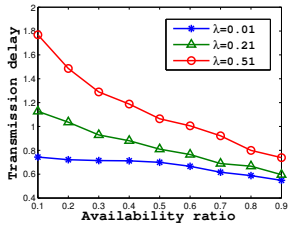

Fig. 22: Different traffic rates.

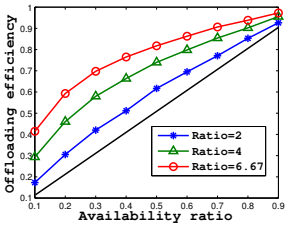

Fig. 23: Offloading efficiencies.

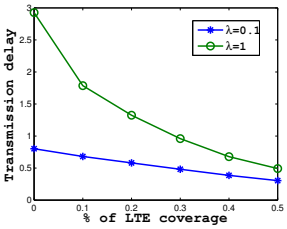

Fig. 24: The design problem with increasing LTE coverage.

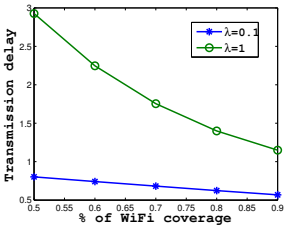

Fig. 25: The design problem with increasing WiFi coverage.

deploy LTE base stations instead of WiFi AP, the delay will be reduced further (reaching the maximum point of reduction of 50 %). Although the LTE's BTS coverage area is larger compared to the WiFi AP (more APs will be needed), still the incurred cost for the LTE base stations is much higher compared to the AP deployment, and operators should consider switching to 4G mostly in regions with very high traffic intensity.

5 Related Work

Authors in [24] propose to exploit opportunistic communications for information spreading in social networks. Their study is based on determining the minimum number of users that are able to reduce maximally the amount transmitted through the cellular network. A theoretical analysis with some optimization problems of the offloading for opportunistic and vehicular communication are given in [25] and [26]. The LTE offloading into WiFi direct is subject of study in [27]. The work in [28] is mainly concerned with studying the conditions under which rate coverage is maximized, for random deployment of APs belonging to different networks. Contrary to most of the other works, authors in [29] consider the situation in which cellular operators pay for using the AP from third parties. They use game theory to consider different issues, such as the amount of data and money a cellular operator should pay for utilizing the APs. In [30], a solution for mobile data offloading between 3GPP and non-3GPP access networks is presented. A WiFi based mobile data offloading architecture that targets the energy efficiency for smartphones was presented in [31]. An interesting work on determining the number of WiFi AP that need to be deployed in order to achieve a QoS is presented in [32].

As more related to mobile data offloading are the papers with measurements [8], [9]. Authors in [8] have tracked the behaviour of pedestrian users and their measurements suggest heavy-tailed periods of WiFi availability. The same holds for the time when there is no WiFi connectivity in the proximity of the mobile user. Similar conclusions for the availability periods are given in [9], where authors conduct measurements for vehicular users. These users are on metropolitan area buses. However, the mean duration of ON and OFF periods are different in the two scenarios of [8] and [9]. This is reasonable given the difference in speeds between vehicular and pedestrian users. The offloading efficiencies reported there are quite different, too. This result comes from different deadlines assumed in the two papers (related to delayed offloading). In addition to the two measurement-based studies [8], [9], already discussed in Section 4, there exists some additional interesting work in the area of offloading. Nevertheless, most related work does not deal with performance modeling and analysis of mobile data offloading. In [33], an integrated architecture has been proposed based on opportunistic networking to switch the data traffic from the cellular to WiFi networks. The results were obtained from real data traces.

In [34], the authors define a utility function related to delayed offloading to quantitatively describe the trade-offs between the user satisfaction in terms of the price that she has to pay and the experienced delay by waiting for WiFi connectivity. The authors use a semi-Markov process to determine the optimal handing-back point (deadline) for three scenarios. However, this analysis does not consider on-the-spot offloading, nor queueing effects. In our paper, we do take into account the queueing process of the packets at the user. The work in [35] considers the traffic flow characteristics when deciding when to offload some data to the WiFi. However, there is no delay-related performance analysis. A cost based analysis is provided in [36].

The approach we are using here is based on the probability generating functions and is motivated from [19], [37].

To our best knowledge, the closest work in spirit to ours is [20]. The results in [20] are the extension of the results in [8] containing the analysis for delayed offloading. The WiFi availability periods, as well as the periods of time when there is only cellular network coverage are modeled with exponential distributions. Also the packet sizes are exponentially distributed. Authors there also use 2D Markov chains to model the state of the system and use matrix-analytic methods to get a numerical solution for the offloading efficiency. However, their model does not apply directly to on-the-spot offloading. Also, they only provide numerical solutions. On the other hand, a performance analysis with closed form results for delayed offloading was provided in [38].

Summarizing, the novelty of our work is along the following dimensions: (i) we deal with on-the-spot offloading, (ii) we provide closed-form results and approximations, (iii) we provide an extension for generic packet size distributions, (iv) we validate our theory against realistic parameter values and distributions, (v) we provide some insight about the offloading gains that are of interest to both users and operators, (v) we generalize our analysis to capture any

number of possible network connections.

6 Conclusion

In this paper, we have proposed a queueing analytic model for the performance of on-the-spot mobile data offloading for generic number of access technologies, and we validated it against realistic WiFi network availability statistics. We have provided approximations for different utilization regions and have validated their accuracy compared to simulations and the exact theoretical results. We also showed that our model can be applied to a broader class of distributions for the durations of the periods between and with WiFi availability. Our model can provide insight on the offloading gains by using on-the-spot mobile data offloading in terms of both the offloading efficiency and delay. We have shown that the availability ratio of WiFi connectivity, in conjuction with the arrival rate plays a crucial role for the performance of offloading, as experienced by the user.

References

[1] "Cisco visual networking index: Global mobile data traffic forecast update, 2012-2017," Feb. 2013, http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/ white paper c11-520862.pdf.

[2] "Mobile data offloading through WiFi," 2010, proxim Wireless.

[3] T. Kaneshige, "iPhone users irate at idea of usage-based pricing," Dec. 2009, http://www.pcworld.com/article/184589/ATT IPhone Users Irate at Idea of Usage Based Pricing.html.

[4] "Growing data demands are trouble for Verizon, LTE capacity nearing limits," http://www.talkandroid.com/97125-growing-data-demands-are-trouble-for-verizon-lte-capacity-nearing-limits/, 2012.

[5] Http://www.3gpp1.eu/ftp/Specs/archive/23 series/23.829/.

[6] M. Qutqut, F.M.Al-Turjman, and H. Hassanein, "MWM: Mobile femtocells utilizing WiFi (a data offloading framework for cellular networks using mobile femtocells)," in *Proc. of IEEE ICC*, 2013.

[7] C. B. Sankaran, "Data offloading techniques in 3GPP rel-10 networks: A tutorial," *IEEE Communications Magazine*, June 2012.

[8] K. Lee, I. Rhee, J. Lee, S. Chong, and Y. Yi, "Mobile data offloading: How much can WiFi deliver," in *Proc. of ACM CoNEXT*, 2010.

[9] A. Balasubramanian, R. Mahajan, and A. Venkataramani, "Augmenting mobile 3G using WiFi," in *Proc. of ACM MobiSys*, 2010.

[10] J. Eriksson, H. Balakrishnan, and S. Madden, "Cabernet: Vehicular Content Delivery Using WiFi," in *14th ACM MOBICOM*, 2008.

[11] S. M. Ross, *Stochastic Processes*, 2nd ed. John Wiley & Sons, 1996.

[12] Y. Y. X. Meng, S. Wong and S. Lu, "Characterizing flows in large wireless data networks," in *Proceedings of ACM MOBICOM*, 2014.

[13] T. Osogami and M. Harchol-Balter, "Closed form solutions for mapping general distributions to minimal PH distributions," *Performance Evaluation*, 2003.

[14] M. F. Neuts, *Matrix Geometric Solutions in Stochastic Models: An Algorithmic Approach*. John Hopkins University Press, 1981.

[15] U. Yechiali and P. Naor, "Queueing problems with heterogeneous arrivals and service," *Operations Research*, vol. 19, no. 3, 1971.

[16] J. L. Snell and C. Grinsted, *Introduction to Probability*. American Mathematical Society, 2006.

[17] F. Mehmeti and T. Spyropoulos, "Performance analysis of on-the-spot mobile data offloading," in *in Proc. of IEEE Globecom*, 2013.

[18] A. Abhari and M. Soraya, "Workload generation for YouTube," *Multimedia Tools and Applications*, 2010.

[19] I. Mitrany and B.Avi-Itzhak, "A many-server queue with service interruptions," *Operations Research*, vol. 16, no. 3, 1968.

[20] K. Lee, J. Lee, Y. Yi, I. Rhee, and S. Chong, "Mobile data offloading: How much can WiFi deliver," *IEEE/ACM Trans. Netw.*, vol. 21, no. 2, 2013.

[21] R. Research, "Beyond LTE: Enabling the mobile broadband explosion," 2014.

[22] Http://www.swisscom.ch/en/residential/mobile/mobile-network.html.

[23] N. Balasubramanian, A. Balasubramanian, and A. Venkataramani, "Energy consumption in mobile phones: A measurement study and implications for network applications," in *Proc. of ACM IMC*, 2009.

[24] B. Han, P. Hui, A. Kumar, M. V. Marathe, and J. S. A. Srinivasan, "Mobile data offloading through opportunistic communications and social participation," *IEEE Tran. Mob. Computing*, vol. 11, no. 5, 2012.

[25] Y. Li, M. Qian, D. Jin, P. Hui, Z. Wang, and S. Chen, "Multiple mobile data offloading through delay tolerant networks," in *Proc. of ACM CHANTS*, 2011.

[26] Y. Li, D. Jin, Z. Wang, L. Zeng, and S. Chen, "Coding or not: Optimal mobile data offloading in opportunistic vehicular networks," *IEEE Tran. Intelligent Transportation Systems*, 2013.

[27] A. Pyattaev, K. Johnsson, S. Andreev, and Y. Koucheryavy, "3GPP LTE traffic offloading onto WiFi direct," in *Proc. of IEEE WCNC*, 2013.

[28] S. Singh, H. Dhillon, and J. Andrews, "Offloading in heterogeneous networks: Modeling, analysis, and design insights," *IEEE Tran. Wireless Communications*, vol. 12, no. 5, 2013.

[29] J. Lee, Y. Yi, S. Chong, and Y. Jin, "Economics of WiFi offloading: Trading delay for cellular capacity," in *Proc. of IEEE Infocom Workshop SDP*, 2013.

[30] D. Kim, Y. Noishiki, Y. Kitatsuji, and H. Yokota, "Efficient ANDSF-assisted Wi-Fi control for mobile data offloading," in *Proc. of IWCMC*, 2013.

[31] A. Y. Ding, B. Han, Y. Xiao, P. Hui, A. Srinivasan, M. Kojo, and S. Tarkoma, "Enabling energy-aware collaborative mobile data offloading for smartphones," in *Proc. of IEEE SECON*, 2013.

[32] J. Kim and N. Song, "Placement of WiFi access points for efficient WiFi offloading in an overlay network," in *Proc. of IEEE PIMRC*, 2013.

[33] S. Dimatteo, P. Hui, B. Han, and V. Li, "Cellular traffic offloading through WiFi networks," in *Proc. of IEEE MASS*, 2011.

[34] D. Zhang and C. K. Yeo, "Optimal handing-back point in mobile data offloading," in *Proc. of IEEE VNC*, 2012.

[35] S. Wietholter, M. Emmelmann, R. Andersson, and A. Wolisz, "Performance evaluation of selection schemes for offloading traffic to IEEE 802.11 hotspots," in *Proc. of IEEE ICC*, 2012.

[36] K. Berg and M. Katsigiannis, "Optimal cost-based strategies in mobile network offloading," in *Proc. of ICST CROWNCOM*, 2012.

[37] U. Yechiali, "A queueing-type birth-and-death process defined on a continous-time Markov chain," *Operations Research*, vol. 21, no. 2, 1973.

[38] F. Mehmeti and T. Spyropoulos, "Is it worth to be patient? Analysis and optimization of delayed mobile data offloading," in *Proc. of IEEE Infocom 2014*.

Fidan Mehmeti received the Bsc degree in Electronics in 2006 and Msc degree in Telecommunications in 2009, both from the University of Prishtina, Kosovo. Currently he is pursuing his PhD studies at Institute Eurecom in Sophia Antipolis, France. His main research interests are in performance modeling and analysis for cognitive radio networks and mobile data offloading.

Thrasyvoulos Spyropoulos received the Diploma in Electrical and Computer Engineering from the National Technical University of Athens, Greece, and a Ph.D degree in Electrical Engineering from the University of Southern California. He was a post-doctoral researcher at INRIA and then, a senior researcher with the Swiss Federal Institute of Technology (ETH) Zurich. He is currently an Assistant Professor at EURECOM, Sophia-Antipolis.