Performance Analysis of "On-the-spot" Mobile Data Offloading

Fidan Mehmeti
Mobile Communications Department
EURECOM, France
Email: mehmeti@eurecom.fr

Thrasyvoulos Spyropoulos
Mobile Communications Department
EURECOM, France
Email: spyropou@eurecom.fr

***Abstract*—An unprecedented increase in the mobile data traffic volume has been recently reported due to the extensive use of smartphones, tablets and laptops. Moreover, predictions say that this increase is going to be yet more pronounced in the next 3-4 years. 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 the benefits to operators have already been documented, with considerable amounts of traffic already switched over to WiFi, it is less clear how much and under what conditions the user gains as well. 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, and user mobility and traffic load. We validate our theory against simulations for realistic data and scenarios, and provide some initial insights as to the offloading gains expected in practice.**

***Keywords*-Mobile data offloading, Queueing, Markov chains.**

I. 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 degradation to 3G services. Possible solutions to this problem could be the 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]).

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 [4]), 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.

There exist two types of offloading: *on-the-spot* and *delayed* [5]. The usual way of offloading is on-the-spot offloading, where traffic is transmitted over the cellular network only when there is no WiFi availability. More recently, delayed offloading has been proposed: if there is currently no WiFi availability, (some) traffic can be delayed up to some chosen time threshold, instead of being sent immediately over the cellular interface. If up to that point, no AP is detected, the data are transmitted through the cellular network. At the moment, smart phones can only switch between interfaces. Using both interfaces in parallel, as well as per flow offloading (IFOP) are currently also being considered in 3GPP [4].

In this paper, our attention will be focused to on-the-spot offloading, since 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) [6]. Although on-the-spot offloading is already used and it does relieve the network, it is still a matter of debate if it offers any benefits to the user as well (in terms of performance, battery consumption, etc.) These benefits might depend on the availability and performance of WiFi networks and the cellular network, environment and type of user mobility [5], [7], etc.

To this end, in this paper we propose a queueing analytic model for performance analysis of on-the-spot mobile data offloading, that can be used to answer questions like the ones above. Our contributions can be summarized as following:

- We derive the expected delay of on-the-spot offloading as a function of WiFi availability, traffic intensity, and other key parameters (Section II-B).
- We propose simpler closed-form approximations for some interesting utilization regimes (Sections II-C, II-D, II-E).
- We validate our model using both synthetic, but also real data for most parameters of interest and demonstrate significant accuracy (Section III-A).
- We use our model to provide some preliminary answers to the questions of offloading efficiency and delay improve-

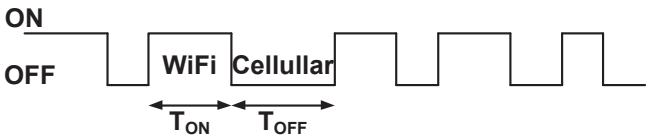

Fig. 1. The WiFi network availability model

ments through WiFi-based offloading (Section III-C).

II. Performance modeling

A. Problem setup

Consider a mobile user that enters and leaves zones with WiFi coverage (with a rate that depends on the user's mobility, e.g. pedestrian, vehicular, the environment in hand, e.g. rural, urban, etc.). Without loss of generality, we assume that there is always cellular network coverage. 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. This switch in connectivity might sometimes occur while some session/flow is running. Although it depends on the actual technology whether and how a vertical handover will occur in this case, we will assume that network transitions do not cause any interruptions to the traffic flow.

We will model the WiFi network availability as an ON-OFF alternating renewal process [8] $\left(T_{ON}^{(i)}, T_{OFF}^{(i)}\right), i \geq 1$, as shown in Fig. 1. ON periods represent the presence of the WiFi connectivity, while during the OFF periods data are transmitted only through the cellular network. i denotes the number of ON-OFF cycles elapsed until time t. The duration of any ON period $T_{ON}^{(i)}$ is assumed to be an exponentially distributed random variable with rate η_w, and is independent of the duration of any other ON or OFF period. The data transmission rate during WiFi connectivity periods is denoted with μ_w. Similarly, all OFF periods are assumed to be independent and exponentially distributed with rate η_c, and with data rate $\mu_c < \mu_w$. Finally, to simplify analysis, we assume that traffic arrives as a Poisson process with rate λ and file (or flow) sizes are random and exponentially distributed.

We also assume queueing occurs when a message (e.g. file, packet) arrives to find another message currently being queued or in transmission, and consider First Come First Served (FCFS) as the queuing discipline. The total time a file spends in the system (service+queueing) is referred to as the *system time*. We use also the term *transmission delay* interchangeably with system time.

We stress here that the above assumptions are only made for analytical tractability, and we do not claim the actual availability periods or packet sizes to be exponentially distributed. For this reason, we will further test our model and its predictions against real ON/OFF distributions in Section III. Furthermore, we could extend our framework to arbitrary ON and OFF distributions that can be approximated by Coxian distributions [9], fitting the first three moments to the real duration of the ON (OFF) period. In this case, there would just be more states along one dimension in the Markov chain, and we could use matrix-analytic methods [10]. But, closed-form approximations could also be pursued. Finally, we show in Section II-F how to extend our model to generic file size distributions. Before proceeding further, we summarize in Table I some useful notation that will be used throughout the rest of the paper.

TABLE I
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

B. Delay analysis of on-the-spot offloading

We will first use queueing analysis to derive a formula for the average transmission delay of a file in an on-the-spot data offloading scenario. Given the previously stated assumptions, our system can be modeled with a 2D Markov chain, as shown in Fig. 2. Our approach in this first step is similar to [11] with the difference that we have the same arrival rate during both periods. $\pi_{i,c}$ denotes the stationary probability of finding i files when there is only cellular network coverage, and $\pi_{i,w}$ is the stationary probability of finding i files during WiFi coverage. Writing the balance equations for this chain gives

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

$$\pi_{0,w}(\lambda + \eta_w) = \pi_{1,w}\mu_w + \pi_{0,c}\eta_c \quad (2)$$

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

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

The steady-state probability of finding the system in some region with WiFi availability is (from Renewal theory [8]) $\pi_w = \frac{\eta_c}{\eta_c+\eta_w}$. Similarly, for the periods with only cellular access we have $\pi_c = \frac{\eta_w}{\eta_c+\eta_w}$.

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

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

After solving the system of equations (6) and (7), we have

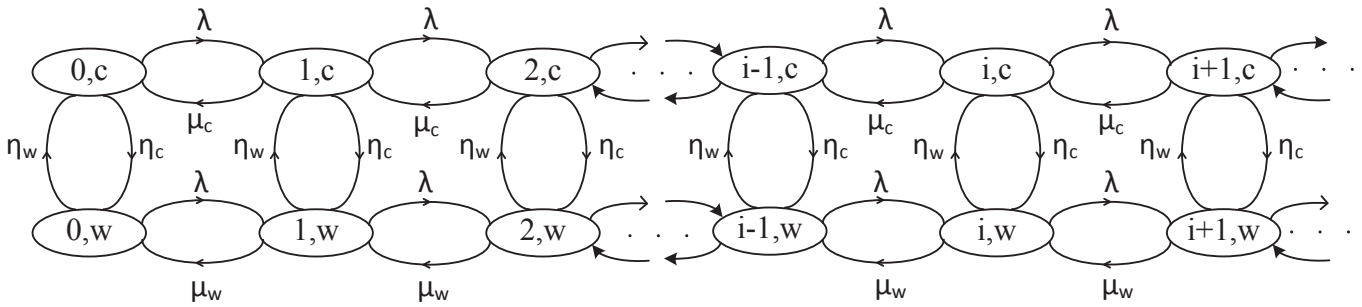

Fig. 2. The 2D Markov chain for on-the-spot Mobile data offloading model

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

where

$$\begin{aligned} f(z) \quad &= \lambda^2 z^3 - \lambda(\eta_c + \eta_w + \lambda + \mu_w + \mu_c)z^2 \\ &+ (\eta_c\mu_w + \eta_w\mu_c + \mu_c\mu_w + \lambda\mu_w + \lambda\mu_c)z - \mu_c\mu_w \end{aligned} \quad (9)$$

It can be proven that the polynom in Eq.(9) has only one root in the open interval $(0,1)$ [11]. This root is denoted as z_0. We omit this proof here due to space limitations. 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.(4), we obtain for $\pi_{0,c}$ and $\pi_{0,w}$

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

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

where $\mu = \pi_c\mu_c + \pi_w\mu_w$.

Finally, for $G_c(z)$ and $G_c(z)$ we have

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

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

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 $E[N_w] = G_w^{'}(1)$.

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

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

Finally, using the Little's law $E[N] = \lambda E[T]$ [8] , we obtain the average packet delay in on-the-spot mobile data offloading:

Result 1. *The average file transmission delay in the on-the-spot mobile data offloading is*

$$E[T] = \frac{1}{\mu-\lambda} + \frac{\mu_c(\mu_w - \lambda)\pi_{0,c} + (\mu_c - \lambda)(\mu_w(\pi_{0,w}-1) + \lambda)}{\lambda(\eta_c + \eta_w)(\mu - \lambda)} \quad (15)$$

C. Low utilization approximations

In the previous subsection we derived a generic expression for the average delay for on-the-spot-offloading. However, the formula in Eq.(15) contains a root of a third order (cubic) equation, and as such its solution is cumbersome, even if obtainable in closed-form. For this reason, in the remainder of this section we will consider simpler closed-form approximations for specific operation regimes. One such scenario of interest is when resources are underloaded (e.g. nighttime, rural areas, or mostly low traffic users, etc) and/or traffic is relatively sparse (some examples are, background traffic from social and mailing applications, messaging, Machine-to-Machine communication, etc.). We thus derive first a low utilization approximation for the average delay when $\rho \to 0$.

Under these assumptions the polynomial in Eq.(9) becomes a linear function with a zero at

$$z_0 = \frac{\mu_c\mu_w}{\eta_c\mu_w + \eta_w\mu_c + \mu_c\mu_w}. \quad (16)$$

Now, Eq.(10)-(11) reduce to $\pi_{0,c} = \frac{\eta_w}{\eta_c+\eta_w}$, and $\pi_{0,w} = \frac{\eta_c}{\eta_c+\eta_w}$, and Eq.(15) can be simplified as suggested in the following:

Low utilization approximation 1. *The average file transmission delay in the on-the-spot mobile data offloading for sparse traffic can be approximated by*

$$E[T] = \frac{\eta_c + \eta_w}{\eta_w\mu_c + \eta_c\mu_w}. \quad (17)$$

Another approach for finding an approximation for the case of sparse traffic is by only considering the service time. For very low utilization, there is no queueing and the service time corresponds in most cases to the total system time. To find the average service time, we use a fraction of the Markov chain from Fig. 2 with only 4 states, as shown in Fig. 3. This chain contains only 4 states, since we do not consider queueing. The system empties at either state $(0,c)$ or state $(0,w)$, since the packet transmission can be finished either during a cellular or a WiFi period.

The goal here is to find the average time until a packet arriving in a WiFi or cellular period finishes its service, i.e. the time until the system, starting from the state $(1,c)$ or $(1,w)$ first enters any of the states $(0,c)$ or $(0,w)$. Hence, the average service time is

$$E[S] = \frac{\eta_w}{\eta_c + \eta_w}E[T_c] + \frac{\eta_c}{\eta_c + \eta_w}E[T_w], \quad (18)$$

where $E[T_c]$ $(E[T_w])$ is the average time until a packet that enters service during a cellular (WiFi) network period finishes its transmission. This can occur during a different period.

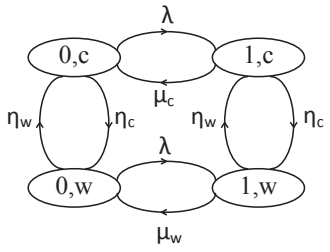

Fig. 3. The reduced Markov chain for $\rho \to 0$

The expression for $E[T_c]$ is equal to

$$E[T_c] = P[I_c = 1]E[T_c|I_c = 1] + P[I_c = 0]E[T_c|I_c = 0], \quad (19)$$

where I_c is an indicator random variable having value 1 if the first transition from state $(1, c)$ is to state $(0, c)$. This means that the packet is transmitted during the same cellular period. Otherwise, its value is 0. The probabilities of this random variables are $P[I_c = 1] = \frac{\mu_c}{\mu_c+\eta_c}$, and $P[I_c = 0] = \frac{\eta_c}{\mu_c+\eta_c}$,respectively. For the conditional expectations from Eq.(19), we have

$$E[T_c|I_c = 1] = \frac{1}{\mu_c + \eta_c}, \quad (20)$$

$$E[T_c|I_c = 0] = \frac{1}{\mu_c + \eta_c} + E[T_w]. \quad (21)$$

Eq.(20) is actually the expected value of the minimum of two exponentially distributed random varibles with rates μ_c and η_c. Replacing Eq.(20) and (21) into Eq.(19), we get

$$E[T_c] - \frac{\eta_c}{\mu_c + \eta_c}E[T_w] = \frac{1}{\mu_c + \eta_c}. \quad (22)$$

Following a similar procedure for $E[T_w]$ we obtain

$$E[T_w] - \frac{\eta_2}{\mu_w + \eta_w}E[T_c] = \frac{1}{\mu_w + \eta_w}. \quad (23)$$

After solving the system of equations Eq.(22)-(23), we have

$$E[T_w] = \frac{\mu_c + \eta_c + \eta_w}{\mu_c\mu_w + \mu_c\eta_w + \mu_w\eta_c}, \quad (24)$$

$$E[T_c] = \frac{\mu_w + \eta_c + \eta_w}{\mu_c\mu_w + \mu_c\eta_w + \mu_w\eta_c}. \quad (25)$$

Now, replacing Eq.(24)-(25) into Eq.(18), we have the average service time, and the second low utilization approximation is ($E[T] \approx E[S]$):

Low utilization approximation 2. *The average file transmission delay in the on-the-spot mobile data offloading for sparse traffic can be approximated by*

$$E[T] = \frac{(\eta_w + \eta_c)^2 + \eta_c\mu_c + \eta_w\mu_w}{(\mu_c\mu_w + \mu_c\eta_w + \mu_w\eta_c)(\eta_c + \eta_w)}. \quad (26)$$

D. 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

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

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

$$f(z) = (z - 1)[\lambda^2 z^2 - \lambda(\mu_c + \mu_w + \eta_c + \eta_w)z + \mu_c\mu_w]. \quad (27)$$

The root in the interval $(0, 1)$ of the function (27) is

$$z_0 = \frac{(\mu_c + \mu_w + \eta_c + \eta_w) - \sqrt{(\mu_c + \mu_w + \eta_c + \eta_w)^2 - 4\mu_c\mu_w}}{2\lambda}, \quad (28)$$

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*

$$E[T] = \frac{1}{\mu - \lambda}\left(1 - \frac{(\mu_c - \lambda)(\mu_w - \lambda)}{\lambda(\eta_c + \eta_w)}\right) + \frac{z_0}{\lambda(\eta_c + \eta_w)(1 - z_0)}\left(\frac{\mu_w - \lambda}{\mu_w - \lambda z_0}\eta_w + \frac{\mu_c - \lambda}{\mu_c - \lambda z_0}\eta_c\right) \quad (29)$$

E. Moderate utilization approximation

So far, we have proposed two approximations for the light and heavy traffic scenarios. These approximations are exact in the limits as $\rho \to 0$ and $\rho \to 1$, respectively. Finally, we provide a heuristic approximation for the average packet delay for intermediate utilization values in the range $0.2 - 0.8$. To do so, we perform linear interpolation of the function $f(z)$ in the range $(0, 1)$, and find the point z_{int} which is the root of the linear interpolated function. The values of the linear function at the end of the interval are $f(0) = -\mu_c\mu_w < 0$, and $f(1) = \eta_c\mu_w + \eta_w\mu_c - \lambda(\eta_c + \eta_w) > 0$. The interpolation function is thus $f(z) = f(0) + \frac{f(1)-f(0)}{1-0}(z - 0)$, with the zero at point

$$z_{int} = \frac{\mu_c\mu_w}{\eta_c\mu_w + \eta_w\mu_c - \lambda(\eta_c + \eta_w) + \mu_c\mu_w}. \quad (30)$$

The function $f(z)$ is concave on the interval $(0, 1)$ if the packet arrival rate satisfies the condition

$$\lambda < \eta_c + \eta_w. \quad (31)$$

In that case, we know that the root of the function $f(z)$ is lower than the zero of the interpolated function. Hence, for the moderate utilization region we propose the approximation $z_0 = \frac{z_{int}}{\varepsilon}$, with ε being in the range 1.4-1.6.

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

$$E[T] = \frac{1}{\mu - \lambda} + \frac{\mu_c(\mu_w - \lambda)\pi_{0,c} + (\mu_c - \lambda)(\mu_w(\pi_{0,w} - 1) + \lambda)}{\lambda(\eta_c + \eta_w)(\mu - \lambda)}, \quad (32)$$

where $\pi_{0,c}$ and $\pi_{0,w}$ are given by Eq.(10)-(11), $z_0 = \frac{z_{int}}{\epsilon}$, and z_{int} is given by Eq.(30).

F. Generic file size distribution approximation

Our analysis so far considers exponentially distributed file sizes. Yet, for some traffic types, heavy-tailed file sizes were reported [12]. 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 [8] as a guideline to introduce a similar "correction factor" related to smaller/higher file size variability[1]. 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,

[1]Due to space limitations, we state this here without further discussion. However, the equivalence with the M/G/1 vs. M/M/1 difference is easily evident, and the interested reader is referred to any standard queueing theory textbook.

for exponentially distributed packet sizes (as derived before). Then, the following approximation applies to 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*

$$E[T_g] = \frac{1-c_v^2}{2}E[S] + \frac{1+c_v^2}{2}E[T]. \quad (33)$$

III. Simulation results

A. 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 [5]. Measurements in [5] 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 [7]. An availability ratio of 11 % has been reported in [7]. For more details about the measurements we refer the interested reader to [5] and [7]. 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 here our model and main result (Eq.(15)) 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 [13]), and we assume that the cellullar network is 3G, with rate 500 kbps. The mean packet size is assumed to be 125 kB[2].

Fig. 4 shows the average file transmission delay (i.e. queueing + transmission) for the pedestrian scenario, for different arrival rates. The range of arrival rates shown correspond to a server utilization of 0-0.9. We can observe, in Fig. 4, that there is a good match between theory and simulations. Furthermore, the average file transmission delay is increased by increasing the arrival rate, as expected, due to queueing effects. Fig. 5 further illustrates the average file transmission delay for the vehicular scenario. We can observe there that the average transmission time is larger than in Fig. 4. 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 the theory and simulations.

In the previous scenarios, we have used realistic values for the transmission rates and WiFi availabilities, but we have so far 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 ([5], [7]) 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). Due to space limitations, we focus on the vehicular scenario. The shape parameters for the Bounded Pareto ON and OFF periods are $\alpha = 0.59$ and $\alpha = 0.64$, respectively and we now consider a cellular rate of 800 kbps. Figure 6 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[3]. 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.

A2. Validation of approximations

Having validated the main result of Eq.(15) we now proceed to validate the various simpler approximations we have proposed in Section II. We start with the low utilization approximations of Section II-B and consider the availability ratio to be 0.75 (similar accuracy levels have been obtained with other values). Fig. 7 shows the packet 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 both approximations show a good match with the generic result and with simulations, but the second approximation shows better performance than the first one. As λ increases, the difference between the approximated result and the actual value increases. For a utilization of 0.1, the first approximation error is around 5%. This is reasonable, as we have strictly assumed that $\lambda = 0$ for this approximation.

We next consider the high utilization regime and respective approximation (Eq.(29)). We consider utilization values of 0.8-0.95. Fig. 8 shows the delay for high values of λ, and an availability ratio of 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.(32)) for moderate utilization values, in the range $0.3 - 0.7$. The availability ratio is again 0.5. The ON and OFF periods are exponentially distributed with mean 1 (to satisfy the condition of Eq.(31)). Fig. 9 compares theory and simulations for the delay in this intermediate utilization regime. The value of the coefficient ε is 1.5. 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 utilization values.

B. Variable WiFi rates and non-exponential file sizes

In all of the above scenarios, we have been assuming constant data rate in the regions with WiFi connectivity. This assumption is unrealistic, since the actual rate experienced

[2]This value is normalized for the arrival rates considered, to correspond to the traffic intensities reported in [7]. We have also considered other values with similar conclusions drawn.

[3]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).

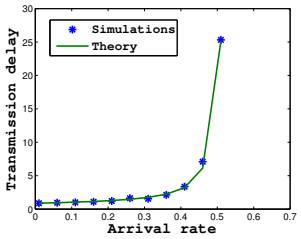

Fig. 4. Pedestrian user

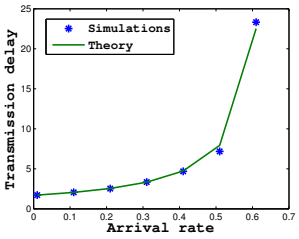

Fig. 5. Vehicular user

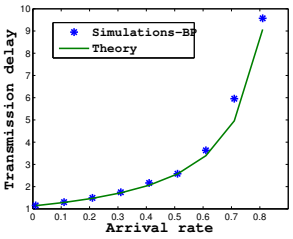

Fig. 6. BP vehicular periods

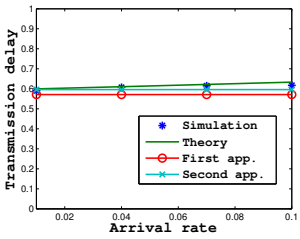

Fig. 7. The low utilization approx.

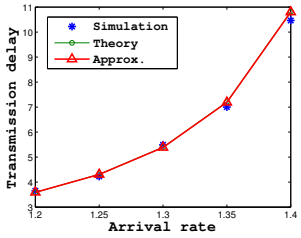

Fig. 8. The high utilization approx.

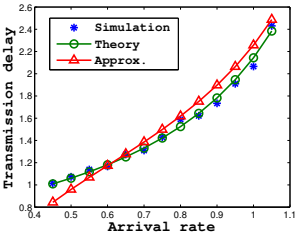

Fig. 9. The approx. for AR=0.5

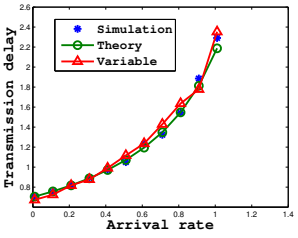

Fig. 10. Variable WiFi rates

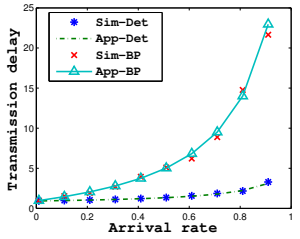

Fig. 11. Generic packet sizes

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 2Mbps, but the actual rate for each ON (WiFi) period is selected uniformly in the interval 1-3 Mbps. In Fig. 10 we compare simulation results for this scenario against our theory (which assumes a constant WiFi rate of 2 Mbps when connected). It is evident that rate variability does not affect performance significantly, making our results applicable in this case as well.

To conclude our validation, we finally drop the exponential packet assumption as well, and test our generic file size results of Eq.(33). Fig. 11 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.

C. 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 [14], 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 send over WiFi - of interest to both the operator and the user). We consider two key parameters of interest that can affect these metrics: WiFi availability ratio, and WiFi/cellular rate difference.

We first consider how transmission delay changes as a function of the availability ratio, for different traffic intensities. We have selected three traffic intensities: very sparse, relatively sparse ($\rho = 0.15$) and medium ($\approx 40\%$). Fig. 12 shows the dependence of the average delay to the availability ratio 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 for higher WiFi availability values, is considerably more sharp, when the average 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 don't 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.

Another parameter that can quantitatively characterize data offloading is the *offloading efficiency* 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. 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 WiFi availability ratios. We consider the impact of different rates of the cellular network 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.13 illustrates the offloading efficiency vs. availability ratio for a moderate arrival rate of $\lambda = 0.3$. For comparison purposes we

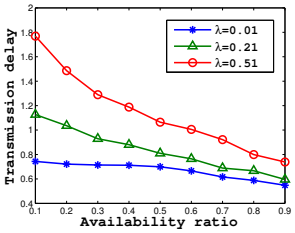

Fig. 12. Different traffic rates

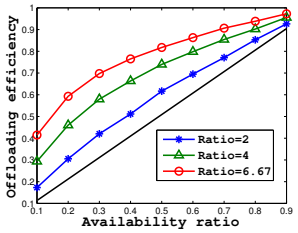

Fig. 13. Offloading efficiencies

also depict the line $x = y$ (Offloading efficiency = availability ratio). First, as expected, we can observe that offloading efficiency increases with AR, 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 an availability ratio of 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.

IV. Related Work

In addition to the two measurement-based studies [5][7], already discussed in Section III, 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 [15], 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 [16], 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 [17] 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 [18].

To our best knowledge, the closest work in spirit to ours is [13]. The results in [13] are the extension of the results in [5] containing the analysis for delayed offloading. 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.

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 out 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. Conclusion

In this paper, we have proposed a queueing analytic model for the performance of on-the-spot mobile data offloading, and we validated it against realistic WiFi network availability statistics. We have provided approximations for different utilization regions (low, moderate, and high utilization) 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 play a crucial role for the performance of offloading, as experienced by the user. In future work, we intend to extend our model to analyse delayed offloading.

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] http://www.3gpp1.eu/ftp/Specs/archive/23 series/23.829/.
[5] 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.
[6] J. Eriksson, H. Balakrishnan, and S. Madden, "Cabernet: Vehicular Content Delivery Using WiFi," in *14th ACM MOBICOM*, (San Francisco, CA), September 2008.
[7] A. Balasubramanian, R. Mahajan, and A. Venkataramani, "Augmenting mobile 3G using WiFi," in *Proc. of ACM MobiSys*, 2010.
[8] S. M. Ross, *Stochastic Processes*. John Wiley & Sons, 2 ed., 1996.
[9] T. Osogami and M. Harchol-Balter, "Closed form solutions for mapping general distributions to minimal PH distributions," *Performance Evaluation*, 2003.
[10] M. F. Neuts, *Matrix Geometric Solutions in Stochastic Models: An Algorithmic Approach*. John Hopkins University Press, 1981.
[11] U. Yechiali and P. Naor, "Queueing problems with heterogeneous arrivals and service," *Operations Research*, vol. 19, may-Jun. 1971.
[12] A. Abhari and M. Soraya, "Workload generation for YouTube," *Multimedia Tools and Applications*, 2010.
[13] K. Lee, J. Lee, Y. Yi, I. Rhee, and S. Chong, "Mobile data offloading: How much can WiFi deliver," *IEEE/ACM Trans. Netw.*, 2013.
[14] 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.
[15] S. Dimatteo, P. Hui, B. Han, and V. Li, "Cellular traffic offloading through WiFi networks," in *Proc. of IEEE MASS*, 2011.
[16] D. Zhang and C. K. Yeo, "Optimal handing-back point in mobile data offloading," in *Proc. of IEEE VNC*, 2012.
[17] 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.
[18] K. Berg and M. Katsigiannis, "Optimal cost-based strategies in mobile network offloading," in *Proc. of ICST CROWNCOM*, 2012.