Performance Modeling, Analysis, and Optimization of Delayed Mobile Data Offloading for Mobile Users

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

***Abstract*— Operators have recently resorted to Wi-Fi offloading to deal with increasing data demand and induced congestion. Researchers have further suggested the use of delayed offloading: if no Wi-Fi connection is available, (some) traffic can be delayed up to a given deadline or until WiFi becomes available. Nevertheless, there is no clear consensus as to the benefits of delayed offloading, with a couple of recent experimental studies largely diverging in their conclusions, nor is it clear how these benefits depend on network characteristics (e.g., Wi-Fi availability), user traffic load, and so on. In this paper, we propose a queueing analytic model for delayed offloading, and derive the mean delay, offloading efficiency, and other metrics of interest, as a function of the user's patience, and key network parameters for two different service disciplines (First Come First Served and Processor Sharing). We validate the accuracy of our results using a range of realistic scenarios and real data traces. Finally, we use these expressions to show how the user could optimally choose deadlines by solving the variations of a constrained optimization problem, in order to maximize her own benefits.**

***Index Terms*— Mobile data offloading, deadlines, queueing, 2D Markov chain, optimization.**

I. Introduction

LATELY, an enormous growth in the mobile data traffic has been reported. This increase is due to a significant penetration of smartphones and tablets in the market, as well as Web 2.0 and streaming applications which are very bandwidth-hungry. Furthermore, Cisco [1] reports that by 2019 the mobile data traffic will increase by 10 times compared to 2014, and will climb to 24.3 exabytes per month. Mobile video traffic will comprise 72% of the total traffic, compared to 55% in 2014 [1].

This increase in traffic demand is overloading cellular networks (especially in the dense areas), forcing them to operate close to their capacity limits, causing thus a significant degradation to 3G services. Upgrading to LTE or LTE-advanced, as well as the deployment of additional network infrastructure could help alleviate this capacity crunch [2], but reports already suggest that such solutions are bound to face the same problems [3]. Furthermore, these solutions may not be cost-effective from the operators' perspective: they imply an increased cost (for power, location rents, deployment and maintenance), without a similar increase in revenues [4], and because of the fact that a small number of users consume a large amount of traffic (3% of the users consume 40% of the traffic [4]).

A more cost-effective way to cope with the problem of highly congested mobile networks is by offloading some of the traffic through Femtocells (SIPTO, LIPA [5]), and the use of WiFi. In 2014, 46% of the total mobile data traffic was offloaded [1]. Projections say that this will increase to 54% by 2019 [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, such as AT&T or SFR, 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 WiFi offloading. The usual way of offloading is *on-the-spot offloading*: when there is WiFi available, all traffic is sent over the WiFi network; otherwise, *all* traffic is sent over the cellular interface. More recently, "delayed" offloading has been proposed: if there is currently no WiFi availability, (some) traffic can be delayed instead of being sent/received immediately over the cellular interface. In the simplest case, traffic is delayed until WiFi connectivity becomes available. This is already the case with current smartphones, where the user can select to send synchronization or backup traffic (e.g., Dropbox, Google+) only over WiFi. A more interesting case is when the user (or the device on her behalf) can choose a deadline (e.g., per application, per file, etc.). If up to that point no AP is detected, the data are transmitted over the cellular network [6], [7].

We have already analyzed the case of on-the-spot offloading in [8]. Delayed offloading offers additional flexibility and promises potential performance gains to both the operator and user. First, more traffic could be offloaded, further decongesting the cellular network. Second, if a user defers the transmission of less delay-sensitive traffic, this could lead to energy savings [9]. Finally, with more operators moving away from flat rate plans towards usage-based plans [10], users have incentives to delay "bulky" traffic to conserve their plan quotas or to receive better prices [11].

Nevertheless, there is no real consensus yet as to the added value of delayed offloading, if any. Recent experimental studies largely diverge in their conclusions about the gains of delayed offloading [6], [7]. Additionally, the exact amount of delay a flow can tolerate is expected to depend heavily

Manuscript received September 16, 2015; revised May 11, 2016; accepted July 1, 2016; approved by IEEE/ACM Transactions on Networking Editor T. Hou. Date of publication July 29, 2016; date of current version February 14, 2017. This work was supported by Intel Mobile Communications, Sophia Antipolis, France, through the project "WTFOM: Wireless Traffic Flow Optimization for Multicom." A subset of initial results has been presented at the IEEE INFOCOM 2014 main conference.

F. Mehmeti is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: fidan.mehmeti@uwaterloo.ca).

T. Spyropoulos is with the Department of Mobile Communications, EURECOM, Biot 06410, France (e-mail: spyropou@eurecom.fr).

Digital Object Identifier 10.1109/TNET.2016.2590320

on (a) the user, and (b) the application type. For example, a study performed in [12] suggests that "more than 50% of the interviewed users would wait up to 10 minutes to stream YouTube videos and 3-5 hours for file downloads". More importantly, *the amount of patience will also depend on the potential gains for the user*. As a result, two interesting questions arise in the context of delayed offloading:

- *If deadlines are externally defined (e.g., by the user or application), what kind of performance gains for the user/operator should one expect from delayed offloading and what parameters do these depend on?*
- *If an algorithm can choose the deadline(s) to achieve different performance-cost trade offs, how should these deadlines be optimally chosen?*

In this paper, we try to answer the two aforementioned questions. The main contributions of this paper can be summarized as follows:

- We propose a queueing analytic model for delayed offloading for two types of scheduling disciplines: First Come First Served (FCFS) and Processor Sharing (PS), based on 2D Markov chains, and derive expressions for the average delay, and other performance user related metrics as a function of deadlines, and key system parameters; we also give closed-form approximations for different regimes of interest;
- We validate our results extensively for both service disciplines, using also scenarios, parameters and data observed from real measurement traces, which depart from the assumptions made in our model;
- We formulate and solve basic cost-performance user-oriented optimization problems, and derive the achievable tradeoff regions as functions of the network parameters (WiFi availability, user load, etc.) in hand.

The paper is organized as follows. In the next section, we present our queueing analytic model and derive the average delay and probability of reneging for delayed offloading, together with the approximations for low and high utilization regimes for different service disciplines. We then validate our theory against simulations for a wide range of realistic scenarios in Section III. In Section IV, we solve different optimization problems having as outcome the optimal deadline. Then, in Section V, we discuss some related work. We conclude our work and provide some further discussion on potential future offloading-related work in Section VI.

II. Analysis of Delayed Offloading

In this section, we formulate the delayed offloading problem, and derive analytical expressions for key metrics of interest (e.g., mean flow delay). We 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) and the environment (e.g., rural, urban). Without loss of generality, we assume that there is always cellular network coverage. We also assume that the user generates flows over time (different sizes, different applications, etc.) that need to be transmitted (uploaded or downloaded) over the network.[1] Whenever there is coverage by some WiFi AP, all traffic will be transmitted through WiFi, assuming for simplicity, at first, FCFS queueing discipline. This scheduling discipline is realistic for some scenarios. But, there might be some other scenarios for which FCFS is not realistic. For that reason, we also consider the case with PS service discipline, as more viable.

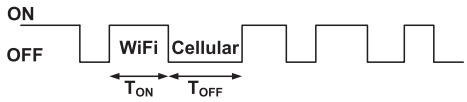

Fig. 1. The WiFi network availability model.

When the WiFi connectivity is lost, we assume that flows waiting in the queue and new flows arriving can be delayed until there is WiFi coverage again. However, each flow has a maximum delay it can wait for (a *deadline*), which might differ between flows and users [12]. If the deadline expires before the flow can be transmitted over some WiFi AP, then it is sent over the cellular network.[2]

To facilitate the analysis of the above system, we make the following assumptions. We model the WiFi network availability as an ON-OFF alternating renewal process [13] $\left(T_{ON}^{(i)}, T_{OFF}^{(i)}\right), i \geq 1$, as shown in Fig. 1. The duration of each ON period (WiFi connectivity), $T_{ON}^{(i)}$, is assumed to be an exponentially distributed random variable with rate η, and independent of the duration of other ON or OFF periods. During such ON periods data can be transmitted over the WiFi network with a data rate c_w. Similarly, all OFF periods (cellular connectivity only) are assumed to be independent and exponentially distributed with rate γ, and a data rate that is lower than the WiFi rate.[3] We further assume that traffic arrives as a Poisson process with rate λ, and that file sizes are exponentially distributed. Finally, to capture the fact that each file or flow may have a different deadline assigned to it, we assume that deadlines are also random with exponential distribution of rate ξ.

The above model is flexible enough to describe a large number of interesting settings: high vs. low WiFi availability (by manipulating $\frac{\gamma}{\gamma+\eta}$), low vs. high speed users (low γ, η vs. high γ, η, respectively), low utilization vs. congested scenarios (via λ and c_w), etc. However, the assumptions of exponentiality, while necessary to proceed with any meaningful analysis (as it will be soon made evident), might "hide" the effect of second order statistics (e.g., variability of ON/OFF periods, flow sizes, etc.). To address this, in Section III we relax most of these assumptions, and validate our results in scenarios with generic ON/OFF periods, generic flow size distributions, and non-exponential deadlines, taken among others, from real traces.

Our goal is to analyze this system to answer the following questions: (i) if the deadlines are given (e.g., defined "externally" by the user or application), what is the expected

[1] We will use the terms "flow", "file", and "packet" interchangeably throughout the paper, as the most appropriate term often depends on the application and the level at which offloading is implemented.

[2] In practice, the switch in connectivity might sometimes occur while some flow is running. Without loss of generality, we will assume that the transmission is resumed from the point it was interrupted when WiFi was lost. It might continue over the cellular network (vertical handover) or paused until WiFi becomes available again or the deadline expires.

[3] Although this might not *always* be the case, everyday experience as well as a number of measurements [6] suggest this to be the case, on average. In any case, our analysis and model hold for any WiFi and cellular rates.

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 a cellular state
$\pi_{i,w}$	Stationary probability of finding i files in a WiFi state
π_c	Probability of finding the system under cellular coverage only
π_w	Probability of finding the system under WiFi coverage
p_r	Probability of reneging
η	The rate of leaving the WiFi state
γ	The rate of leaving the cellular state
$E[\Gamma]$	The average file size
$\mu = \frac{c_w}{E[\Gamma]}$	The mean service transition rate while being in a WiFi state
ξ	The reneging rate
$E[S]$	The average service time
$E[T]$	The average system (transmission) time
T_d	The deadline time
$\rho = \lambda E[S]$	Average user utilization ratio

performance as a function of network parameters like WiFi availability statistics, and user traffic load? (ii) if the deadlines are "flexible", i.e., the user would like to choose these deadlines in order to optimize his overall performance (e.g., trading off some delay, waiting for WiFi, to avoid the often higher energy and monetary cost of cellular transmission), how should they be chosen?

We will answer the first question in the remainder of this section, and use the derived expressions to provide some answers to the second question, in Section IV. Before proceeding, we summarize in Table I some useful notation. The total time a file spends in the system (queueing+ service time) will be referred to as the *system time* or *transmission delay*.

A. Performance of WiFi Queue

All files arriving to the system are by default sent to the WiFi interface with a deadline assigned (drawn from an exponential distribution). Files are queued if there is another file already in service (i.e., being transmitted) or if there is no WiFi connectivity at the moment, until their deadline expires. If the deadline for a file expires (either while queued or while at the head of the queue, but waiting for WiFi), the file *abandons* the WiFi queue and is transmitted through the cellular network. These kind of systems are known as queueing systems with *impatient* customers [14] or with *reneging* [15]. Throughout our analysis, we will assume that files will abandon the queue only during periods without WiFi connectivity.[4] Nevertheless, in Section III we consider also deterministic deadlines. Our focus here will be on the WiFi queue for two reasons: First, this is the place where files accumulate most of the delay. Second, this is the point where a decision can be made, which will be relevant to the deadline optimization (Section IV). For the moment, we can assume that a file sent back to the cellular interface will incur a fixed delay (this might also include some mean queueing delay) that is larger, in general, than the service time over WiFi (i.e., $\frac{file\_size}{WiFi\_rate}$).

[4]In this manner, abandonments are plausibly associated with the accumulated "opportunity cost", i.e., the time spent waiting for WiFi connectivity (the "non-standard" option for transmission). Instead, if WiFi is available, but there are some files in front, it might make no sense to abandon, as queueing delays might also occur in the cellular interface.

Given the previously stated assumptions, the WiFi queue can be modeled with a 2D Markov chain, as shown in Fig. 2. States with WiFi connectivity are denoted with $\{i, w\}$, and states with only cellular connectivity with $\{i, c\}$. i corresponds to the number of customers in the system (service+queue). During WiFi states, the system empties at rate $\mu = \frac{c_w}{E[\Gamma]}$ (since files are transmitted 1-by-1), with $E[\Gamma]$ being the average file size. During cellular states the system empties at rate $i \cdot \xi$ since any of the i queued packets can renege. The following theorem uses probability generating functions (PGF) to derive the mean system time for this queue. The use of PGFs in 2D Markov chains is known for quite a long time [16]–[18].

Theorem 1: The mean system time for the WiFi queue under FCFS scheduling discipline, when delayed mobile data offloading is performed, is

$$E[T] = \frac{1}{\lambda}\left[\left(1+\frac{\gamma}{\eta}\right)\frac{\lambda-\mu(\pi_w-\pi_{0,w})}{\xi} + \frac{(\lambda-\mu)\pi_w + \mu\pi_{0,w}}{\eta}\right]. \quad (1)$$

Proof: Let $\pi_{i,c}$ and $\pi_{i,w}$ denote the stationary probabilities of finding i files when there is only cellular network coverage, or WiFi coverage, respectively.

Writing the balance equations for the cellular and WiFi state gives

$$(\lambda+\gamma)\pi_{0,c} = \eta\pi_{0,w} + \xi\pi_{1,c} \quad (2)$$

$$(\lambda+\gamma+i\xi)\pi_{i,c} = \eta\pi_{i,w} + (i+1)\xi\pi_{i+1,c} + \lambda\pi_{i-1,c} \quad (3)$$

$$(\lambda+\eta)\pi_{0,w} = \gamma\pi_{0,c} + \mu\pi_{1,w} \quad (4)$$

$$(\lambda+\eta+\mu)\pi_{i,w} = \gamma\pi_{i,c} + \mu\pi_{i+1,w} + \lambda\pi_{i-1,w} \quad (5)$$

The long term probabilities of finding the system in a cellular or WiFi state are $\pi_c = \frac{\eta}{\eta+\gamma}$ and $\pi_w = \frac{\gamma}{\eta+\gamma}$, respectively.

We define the PGFs for both the cellular and WiFi states as $G_c(z) = \sum_{i=0}^{\infty}\pi_{i,c}z^i$, and $G_w(z) = \sum_{i=0}^{\infty}\pi_{i,w}z^i, |z| \le 1$. After multiplying Eq.(3) with z^i and adding it to Eq.(2), we obtain

$$(\lambda+\gamma)G_c(z) + \xi\left(1-\frac{1}{z}\right)\sum_{i=1}^{\infty} i\pi_{i,c}z^i = \eta G_w(z) + \lambda z G_c(z). \quad (6)$$

The summation in the above equation gives $\sum_{i=1}^{\infty} i\pi_{i,c}z^i = zG_c'(z)$. Hence, after some rearrangements in Eq.(6), we obtain

$$\xi(1-z)G_c'(z) = (\lambda(1-z)+\gamma)\,G_c(z) - \eta G_w(z). \quad (7)$$

Repeating the same procedure for Eq.(4)-(5) we get

$$(\lambda+\eta)G_w(z) = \gamma G_c(z) + \lambda z G_w(z) + \mu\left(\frac{1}{z}-1\right)(G_w(z)-\pi_{0,w}),$$

which after some rearrangements leads to

$$((\lambda z-\mu)(1-z)+\eta z)\,G_w(z) = \gamma z G_c(z) - \mu(1-z)\pi_{0,w}.$$

Next, we make two replacements $\alpha(z) = \lambda(1-z)+\gamma$, and $\beta(z) = (\lambda z-\mu)(1-z)+\eta z$. Now, we have the system of equations

$$G_w(z) = \frac{\gamma z G_c(z) - \mu(1-z)\pi_{0,w}}{\beta(z)}, \quad (8)$$

$$G_c'(z) - \frac{\alpha(z)\beta(z) - \eta\gamma z}{\xi(1-z)\beta(z)}G_c(z) = \frac{\eta\mu\pi_{0,w}}{\xi\beta(z)}. \quad (9)$$

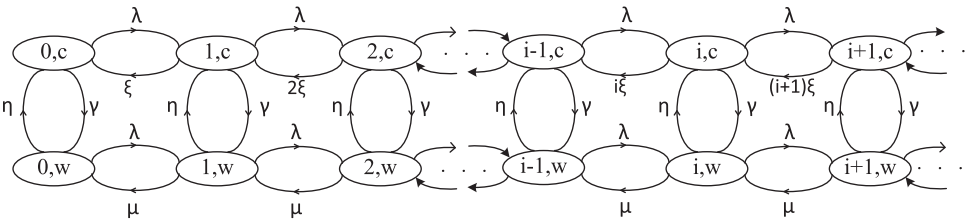

Fig. 2. The 2D Markov chain for the WiFi queue in delayed offloading.

The roots of $\beta(z)$ are

$$z_{1,2} = \frac{\lambda+\mu+\eta \mp \sqrt{(\lambda+\mu+\eta)^2 - 4\lambda\mu}}{2\lambda}. \quad (10)$$

It can be shown easily that these roots satisfy the relation $0 < z_1 < 1 < z_2$. We introduce the function $f(z) = -\frac{\alpha(z)\beta(z)-\eta\gamma z}{\xi(1-z)\beta(z)}$, as the multiplying factor of $G_c(z)$ in the differential equation Eq.(9). Performing some simple calculus operations, the above function transforms into

$$f(z) = -\frac{\lambda}{\xi} + \frac{\gamma}{\xi(1-z)}\left(\frac{\eta z}{\beta(z)} - 1\right). \quad (11)$$

After some algebra and applying the *partial fraction expansion*, the function $f(z)$ becomes

$$f(z) = -\frac{\lambda}{\xi} + \frac{\gamma}{\xi}\left(\frac{M}{z-z_1} + \frac{N}{z_2-z}\right). \quad (12)$$

We determine the coefficients M and N in the standard way as

$$M = \frac{\frac{\mu}{\lambda}-z}{z_2-z}\,|_{z=z_1} = \frac{\frac{\mu}{\lambda}-z_1}{z_2-z_1} = \frac{z_1z_2-z_1}{z_2-z_1} > 0,$$

and

$$N = \frac{\frac{\mu}{\lambda}-z}{z-z_1}\,|_{z=z_2} = \frac{\frac{\mu}{\lambda}-z_2}{z_2-z_1} < 0.$$

In order to solve the differential equation Eq.(9), we need to multiply it by $e^{\int f(z)dz}$. Hence, we get

$$G_c^{'}(z)e^{\int f(z)dz} + f(z)G_c(z)e^{\int f(z)dz} = \frac{\eta\mu\pi_{0,w}}{\xi\beta(z)}e^{\int f(z)dz}. \quad (13)$$

We thus need to integrate the function in Eq.(12):

$$\int f(z)dz = -\frac{\lambda}{\xi}z + \frac{\gamma M}{\xi}\ln|z-z_1| - \frac{\gamma N}{\xi}\ln(z_2-z). \quad (14)$$

The constant normally needed on the right-hand side of Eq.(14) can be ignored in our case. We next raise Eq.(14) to the power of e to get

$$e^{\int f(z)dz} = e^{-\frac{\lambda}{\xi}z}\,|z-z_1|^{\frac{\gamma M}{\xi}}\,(z_2-z)^{-\frac{\gamma N}{\xi}}. \quad (15)$$

Now, Eq.(13) is equivalent to

$$\frac{d}{dz}\left(e^{-\frac{\lambda}{\xi}z}\,|z-z_1|^{\frac{\gamma M}{\xi}}\,(z_2-z)^{-\frac{\gamma N}{\xi}}G_c(z)\right) = \frac{\eta\mu\pi_{0,w}}{\xi\beta(z)}e^{\int f(z)dz} \quad (16)$$

We define $k_1(z)$ and $k_2(z)$ as

$$k_1(z) = e^{-\frac{\lambda}{\xi}z}\,(z_1-z)^{\frac{\gamma M}{\xi}}\,(z_2-z)^{-\frac{\gamma N}{\xi}}, z \le z_1,$$
$$k_2(z) = e^{-\frac{\lambda}{\xi}z}\,(z-z_1)^{\frac{\gamma M}{\xi}}\,(z_2-z)^{-\frac{\gamma N}{\xi}}, z \ge z_1.$$

Eq.(16) now becomes

$$\frac{d}{dz}\left(k_1(z)G_c(z)\right) = \frac{\eta\mu\pi_{0,w}}{\xi\beta(z)}k_1(z), z \le z_1, \quad (17)$$
$$\frac{d}{dz}\left(k_2(z)G_c(z)\right) = \frac{\eta\mu\pi_{0,w}}{\xi\beta(z)}k_2(z), z \ge z_1, \quad (18)$$

and after integrating, we obtain

$$k_1(z)G_c(z) = \frac{\eta\mu\pi_{0,w}}{\xi}\int_0^z \frac{k_1(x)}{\beta(x)}dx + C_1, z \le z_1 \quad (19)$$
$$k_2(z)G_c(z) = \frac{\eta\mu\pi_{0,w}}{\xi}\int_{z_1}^z \frac{k_2(x)}{\beta(x)}dx + C_2, z \ge z_1. \quad (20)$$

The bounds of the integrals in Eq.(19) and Eq.(20) come from the defining region of z in Eq.(17)-(18). We need to determine the coefficients C_1 and C_2 in Eq.(19) and Eq.(20). We take $z = 0$ in Eq.(19). We have $k_1(0) = z_1^{\frac{\gamma M}{\xi}} z_2^{-\frac{\gamma N}{\xi}}$, and knowing that $G_c(0) = \pi_{0,c}$, we get $C_1 = \pi_{0,c}z_1^{\frac{\gamma M}{\xi}} z_2^{-\frac{\gamma N}{\xi}}$. In a similar fashion we get $C_2 = 0$.

Finally, for the PGF in the cellular state we have

$$G_c(z) = \frac{\eta\mu\pi_{0,w}\int_0^z \frac{k_1(x)}{\beta(x)}dx + \xi\pi_{0,c}z_1^{\frac{\gamma M}{\xi}} z_2^{-\frac{\gamma N}{\xi}}}{\xi k_1(z)}, \; z \le z_1, \quad (21)$$

$$G_c(z) = \frac{\eta\mu\pi_{0,w}\int_{z_1}^z \frac{k_2(x)}{\beta(x)}dx}{\xi k_2(z)}, \quad z \ge z_1. \quad (22)$$

In the last equation, the "zero probabilities" $\pi_{0,c}$ and $\pi_{0,w}$ are unknown. We can find them in the following way: we know that $\pi_c = \frac{\eta}{\eta+\gamma} = G_c(1) = \frac{\eta\mu\pi_{0,w}\int_{z_1}^1 \frac{k_2(x)}{\beta(x)}dx}{\xi k_2(1)}$. From this we have

$$\frac{\xi k_2(1)}{\eta+\gamma} = \mu\pi_{0,w}\int_{z_1}^1 \frac{k_2(x)}{\beta(x)}dx. \quad (23)$$

Similarly, from the boundary conditions in Eq.(21) for $z \le z_1$, we get

$$\eta\mu\pi_{0,w}\int_0^{z_1} \frac{k_1(x)}{\beta(x)}dx + \xi\pi_{0,c}z_1^{\frac{\gamma M}{\xi}} z_2^{-\frac{\gamma N}{\xi}} = 0. \quad (24)$$

After solving the system of equations Eq.(23) and Eq.(24), we obtain for the "zero probabilities":

$$\pi_{0,w} = \frac{\xi k_2(1)}{(\eta+\gamma)\mu}\frac{1}{\int_{z_1}^1 \frac{k_2(x)}{\beta(x)}dx}, \text{and} \quad (25)$$

$$\pi_{0,c} = -\frac{\eta k_2(1)\int_0^{z_1} \frac{k_1(x)}{\beta(x)}dx}{(\eta+\gamma)z_1^{\frac{\gamma M}{\xi}} z_2^{-\frac{\gamma N}{\xi}}\int_{z_1}^1 \frac{k_2(x)}{\beta(x)}dx}. \quad (26)$$

The value of the integral $\frac{k_1(x)}{\beta(x)}dx$ is always negative. Hence, $\pi_{0,c}$ is always positive.

Using a vertical cut between any two-pairs of neighboring states in Fig. 2, and writing balance equations we have

$$\lambda\pi_{i,c} + \lambda\pi_{i,w} = \mu\pi_{i+1,w} + (i+1)\xi\pi_{i+1,c}. \quad (27)$$

Summing over all i results in

$$\lambda(\pi_c + \pi_w) = \mu(\pi_w - \pi_{0,w}) + \xi\sum_{i=0}^{\infty}(i+1)\pi_{i+1,c}. \quad (28)$$

Obviously, the last equation reduces to

$$\lambda = \mu(\pi_w - \pi_{0,w}) + \xi E[N_c], \quad (29)$$

where $E[N_c] = G_c^{'}(1)$, and $E[N_w] = G_w^{'}(1)$. Eq.(29) yields

$$E[N_c] = \frac{\lambda - \mu(\pi_w - \pi_{0,w})}{\xi}. \quad (30)$$

So far, we have derived $E[N_c]$ as the first derivative at $z = 1$ of $G_c(z)$. In order to find the average number of files in the system, we need $E[N_w]$ as well. We can get it by differentiating Eq.(8):

$$G_w^{'}(z) = \frac{\beta(z)\left(\gamma G_c(z) + \gamma z G_c^{'}(z) + \mu\pi_{0,w}\right)}{\beta^2(z)} - \frac{\beta^{'}(z)\left(\gamma z G_c(z) - \mu(1-z)\pi_{0,w}\right)}{\beta^2(z)}, \quad (31)$$

and setting $z = 1$. After some calculus, we obtain

$$E[N_w] = \frac{\left(\gamma E[N_c] + \mu\pi_{0,c}\right)\eta - \gamma\pi_c(\mu - \lambda)}{\eta^2}. \quad (32)$$

Replacing Eq.(30) into Eq.(32) we get

$$E[N_w] = \frac{\gamma}{\eta}\frac{\lambda - \mu\left(\pi_w - \pi_{0,w}\right)}{\xi} + \frac{\mu\pi_{0,w}}{\eta} - \frac{\gamma\pi_c(\mu - \lambda)}{\eta^2}. \quad (33)$$

The average number of files in the system is

$$E[N] = E[N_c] + E[N_w]. \quad (34)$$

Finally, using the Little's law $E[N] = \lambda E[T]$ [13], we obtain the average packet delay in delayed offloading as in Eq.(1). □

The above result gives the total expected delay that incoming flows experience in the WiFi queue. For flows that do get transmitted over WiFi (i.e., whose deadline does not expire) this amounts to their total delay. Flows that end up reneging (deadline expires before transmission) must be transmitted through the cellular system and thus incur an additional delay Δ (related to their transmission time over the cellular link, i.e., $\frac{packet\_size}{cellular\_rate}$, and possibly some queueing delay as well). The following Corollary gives the probability of reneging for each flow.

Corollary 1: The probability that an arbitrary flow arriving to the WiFi queue will renege, i.e., its deadline will expire before it can be transmitted over a WiFi AP is

$$p_r = \frac{\lambda - \mu(\pi_w - \pi_{0,w})}{\lambda}. \quad (35)$$

The rate of flows sent back to the cellular network is given by $\lambda \cdot p_r$. This must be equal to $\xi \cdot E[N_c]$, which is the average abandonment rate in Fig. 2, i.e., $\lambda p_r = \xi E[N_c]$. Replacing $E[N_c]$ from Eq.(30) gives the above result. This also provides another important metric, the *offloading efficiency* of our system, OE, namely the percentage of data that get offloaded from cellular network. We find it in the following way. Observe a very long time interval t. During this time the total amount of data that has arrived into the system is $\lambda t E[\Gamma]$. As the system will be found in a WiFi period having a file to transmit with probability $\pi_w - \pi_{0,w}$, it follows that during $(\pi_w - \pi_{0,w})\,t$ time units it will be transmitting files with a data rate of $c_w = \mu E[\Gamma]$. So, the amount of offloaded data is $(\pi_w - \pi_{0,w})\,t\mu E[\Gamma]$, and the offloading efficiency is $\frac{(\pi_w - \pi_{0,w})t\mu E[\Gamma]}{\lambda t E[\Gamma]} = \frac{\mu(\pi_w - \pi_{0,w})}{\lambda}$. Given Eq.(35), we can write the following Corollary.

Corollary 2: The offloading efficiency of the delayed mobile data offloading system of Section II-A is given by

$$OE = 1 - p_r. \quad (36)$$

As a final note, it should be mentioned that the system is always stable. As such, there is no possibility of congestion. Namely, the system can be congested only during the ON periods, or equivalently, only from the files that do not renege, since during the OFF periods the rate at which files leave the system is proportional to the size of the queue. The arrival rate of the non-reneging files is $\lambda(1 - p_r) = \mu\left(\pi_w - \pi_{0,w}\right) < \mu$. Hence, the arrival rate of such files is always lower than the service rate during an ON period. So, the system is stable. Intuitively, the more files in the queue the more of them abandon the WiFi queue.

B. System Performance for Processor Sharing

The result in Eq.(1) was derived for the FCFS order of service. Under the same assumptions, Eq.(1) holds for the Processor Sharing (PS) policy as well. In the following, we prove this result.

Theorem 2: The mean system time for the WiFi queue under PS scheduling discipline, when delayed mobile data offloading is performed, is given by Eq.(1).

Proof: To prove this theorem, we need to show that the Markov chain of Fig. 2 remains exactly the same under the PS policy. As we have the same network setup, the parameters λ, η_w, and η_c remain unchanged. Now, let's consider the service rate. If there are i files in the system (in the WiFi state), each one of the i files shares $\frac{1}{i}$ of the resources, i.e., has a service rate of $\frac{1}{i}\mu$. Since there are i files (with identically exponentially distributed file sizes), the transition rate to move from state $\{i, w\}$ to $\{i-1, w\}$ is simply the rate of a minimum of i exponentially distributed random variables, i.e., $i \cdot \frac{1}{i}\mu = \mu$. So, for the WiFi states in a PS setup, we have the same rates going backwards as in the lower part of the Markov chain in Fig. 2.

When the system is in a cellular state, any of the queued files can renege independently with rate ξ. If there are i files waiting in the queue (during a cellular period), then the transmission rate of moving from the state $\{i, c\}$ to state $\{i-1, c\}$ is $i\xi$. This is the same transition rate as for the FCFS policy. Hence, we have shown that all the transition rates in Fig. 2 remain unchanged. As a consequence of that, Eq.(1) holds for the PS policy as well. □

Similarly, Eq.(35) holds for the PS policy too. In that direction, the previous two findings increase even further the value of our theoretical result.

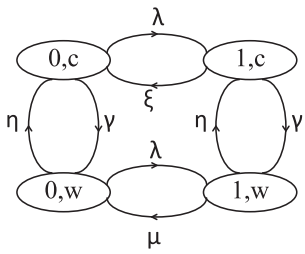

Fig. 3. The reduced Markov chain for $\rho \to 0$.

The above expressions can be used to predict the performance of a delayed offloading system, as a function of most parameters of interest, such as WiFi availability and performance, user traffic load, etc. As we shall see later, it does so with high accuracy even in scenarios where many of the assumptions do not hold. However, Eq.(1) cannot easily be used to solve optimization problems related to the deadline (ξ), analytically, as the parameters $\pi_{0,c}$ and $\pi_{0,w}$ involve ξ in a non-trivial way. To this end, we propose next some closed-form approximations for the low and high utilization regimes.

C. Low Utilization Approximation

One interesting scenario is when resources are underloaded (e.g., nighttime, rural areas, or mostly low traffic users) and/or traffic is relatively sparse (e.g., background traffic from social and mailing applications, messaging, Machine-to-Machine communication). For very low utilization, the total system time essentially consists of the service time, as there is almost no queueing. So, we can use a fraction of the Markov chain from Fig. 2 with only 4 states, as shown in Fig. 3, to derive $E[T]$ and p_r. The system empties at either state $\{0, w\}$, if the packet is transmitted while in WiFi connectivity period or state $\{0, c\}$, if the packet spends in queue more than the deadline it was assigned while waiting for WiFi availability.

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 state $\{1, c\}$ or $\{1, w\}$ first enters any of the states $\{0, c\}$ or $\{0, w\}$. Hence, the average service time is

$$E[S] = \frac{\eta}{\gamma+\eta}E[T_c] + \frac{\gamma}{\gamma+\eta}E[T_w], \tag{37}$$

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

$$E[T_c] = P[I_c = 1]E[T_c|I_c = 1] + P[I_c = 0]E[T_c|I_c = 0], \tag{38}$$

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 these random variables are $P[I_c = 1] = \frac{\xi}{\xi+\gamma}$, and $P[I_c = 0] = \frac{\gamma}{\xi+\gamma}$, respectively. For the conditional expectations from Eq.(38), we have

$$E[T_c|I_c = 1] = \frac{1}{\xi+\gamma}, \tag{39}$$

$$E[T_c|I_c = 0] = \frac{1}{\xi+\gamma} + E[T_w]. \tag{40}$$

Eq.(39) is actually the expected value of the minimum of two exponentially distributed random variables with rates ξ and γ. Substituting Eq.(39) and Eq.(40) into Eq.(38), we get

$$E[T_c] - \frac{\gamma}{\xi+\gamma}E[T_w] = \frac{1}{\xi+\gamma}. \tag{41}$$

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

$$E[T_w] - \frac{\eta}{\mu+\eta}E[T_c] = \frac{1}{\mu+\eta}. \tag{42}$$

After solving the system of equations Eq.(41)-(42), we have

$$E[T_w] = \frac{\xi+\gamma+\eta}{\xi\mu+\xi\eta+\mu\gamma}, \tag{43}$$

$$E[T_c] = \frac{\mu+\gamma+\eta}{\xi\mu+\xi\eta+\mu\gamma}. \tag{44}$$

Now, replacing Eq.(43)-(44) into Eq.(37), we have the average service time, and the low utilization approximation is ($E[T] \approx E[S]$):

$$E[T] = \frac{(\eta+\gamma)^2+\gamma\xi+\eta\mu}{(\xi\mu+\xi\eta+\mu\gamma)(\gamma+\eta)}. \tag{45}$$

To find the probability of reneging, we need to know $\pi_{0,c}$. We find it by solving the local balance equations for Fig. 3. After solving the system, we get

$$\pi_{0,c} = \frac{\eta}{\eta+\gamma}\frac{\xi(\mu+\lambda+\eta)+\mu\gamma}{\xi(\mu+\lambda+\eta)+\mu\gamma+\lambda^2+\lambda(\eta+\gamma+\mu)}. \tag{46}$$

Substituting Eq.(46), and $\pi_c = \frac{\eta}{\eta+\gamma}$ into Eq.(35), we get the probability of reneging for low utilization as

$$p_r = \frac{\theta_1\xi}{\theta_2\xi+\theta_3}, \tag{47}$$

where $\theta_1 = \frac{\eta(\lambda+\eta+\gamma+\mu)}{\eta+\gamma}$, $\theta_2 = \mu+\lambda+\eta$, and $\theta_3 = \mu\gamma+\lambda^2+\lambda(\eta+\gamma+\mu)$.

Hence, we can write the results for the low utilization regime for both the average delay and probability of reneging as following:

Low Utilization Approximation: The expected system time in the WiFi queue and the probability of reneging for sparse input traffic can be approximated by

$$E[T] = \frac{(\eta+\gamma)^2+\gamma\xi+\eta\mu}{(\xi\mu+\xi\eta+\mu\gamma)(\gamma+\eta)}, \tag{48}$$

$$p_r = \frac{\theta_1\xi}{\theta_2\xi+\theta_3}, \tag{49}$$

where $\theta_1 = \frac{\eta(\lambda+\eta+\gamma+\mu)}{\eta+\gamma}$, $\theta_2 = \mu+\lambda+\eta$, and $\theta_3 = \mu\gamma+\lambda^2+\lambda(\eta+\gamma+\mu)$.

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 here an approximation that corresponds to the region of high utilization ($\rho \to 1$).

High Utilization Approximation: The expected system time in the WiFi queue and the probability of reneging for a user with heavy traffic can be approximated by

$$E[T] = \frac{1}{\lambda}\left[\left(1+\frac{\gamma}{\eta}\right)\frac{\lambda-\mu\pi_w}{\xi}+\frac{(\lambda-\mu)\pi_w}{\eta}\right], \quad (50)$$

$$p_r = \frac{\lambda-\mu\pi_w}{\lambda}+\frac{\mu}{\lambda}\pi_{0,w}, \quad (51)$$

where $\pi_{0,w}$ is the first order Taylor series approximation of Eq.(25). The details of the derivation can be found in [19].

III. Performance Evaluation

A. Simulation Setup

In this section we will validate our theory against simulations for a wide range of traffic intensities, WiFi availability periods with different distributions, and different deadline times for two service disciplines (FCFS and PS). We perform the validations both for synthetic distributions (which may or may not be the same as the model assumptions) and for data obtained from real traces.

We consider a user moving around and entering zones with WiFi access (ON periods). While being in a region with WiFi access, the files in the user's queue are served with the WiFi rate the user receives. The WiFi rates are taken from a real trace [20]. The average data rate inferred from there is 1.28 Mbps.

Deadline Implementation: When the WiFi connectivity is lost (the OFF period), the user interrupts the transmission/reception process. During the time there is no WiFi access, for every file (the one in service and all the others that are queued) an independent deadline is set. The deadline timers start counting from the moment the WiFi connection is lost. There are two ways to implement the deadline in the simulator. We can either reset the deadline for a given file at the beginning of each OFF period, or we can set it once, and then, if that file is still in the queue, the renege clock will resume running in the next OFF period. For exponentially distributed deadlines, the results are the same in both approaches due to the memoryless property of the exponential distribution.

WiFi Queue Updates: In the meantime, while being in an OFF period, some other files might arrive. For them similarly, the deadline is set (independently) and the timer starts running immediately. If the time until the WiFi becomes available again is longer then the deadline for a file, that file reneges (receives service from the cellular network). During an OFF period multiple files can renege. For the files whose deadline doesn't expire by the time the WiFi access is available again (in the next ON period), the reneging clock is stopped. Then, either the file that was in service before entering the OFF period resumes its service from the point it stopped (if it hasn't reneged during the OFF period), or the new files starts the transmission (if the previous file in service has reneged). The procedure continues in the same way for each ON period. Unless otherwise stated, the usual service policy is FCFS. Finally, for each file we measure the time it spends in the WiFi queue, and take the average over all files. The simulation results are the averages over 100000 runs.

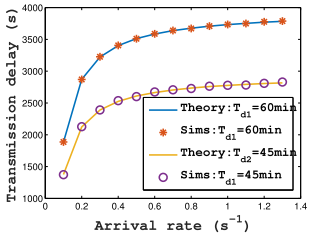

Fig. 4. The average delay for pedestrian users' scenarios.

WiFi Availability: We define the WiFi availability ratio as $AR = \frac{E[T_{ON}]}{E[T_{ON}]+E[T_{OFF}]} = \frac{\gamma}{\eta+\gamma}$. The WiFi availability depends on the mobility pattern of the mobile user. We focus on two scenarios. The first one considers mostly pedestrian users with statistics taken from [6]. Measurements in [6] report that the average duration of WiFi availability period is 122 min ($\eta = \frac{1}{122}min^{-1}$), while $E[T_{OFF}]$ is 41 min ($\gamma = \frac{1}{41}min^{-1}$), with AR=0.75. The second scenario corresponds to vehicular users, related to the measurement study of [7]. An AR of 11% has been reported in [7]. Unless otherwise stated, the durations of WiFi ON and OFF periods will be drawn from independent exponential distributions, where $E[T_{ON}]$ is taken from Table III, and for a given AR, $E[T_{OFF}]$ is determined. We also simulate a scenario where the ON and OFF periods are taken from the real trace of [20].

Other Parameters: The deadlines are exponentially distributed with rate ξ. We also simulate scenarios with deterministic deadlines. Unless otherwise stated, file sizes are assumed exponential. Nevertheless, we relax this condition and try other distributions for file sizes as well. Furthermore, we also use a trace of file sizes [21], with a mean of 2.27 Mb. We use this value as the average file size in all the simulation scenarios. Files are generated at the mobile user as a Poisson process with rate λ. See each scenario in the following sections for specific corresponding parameter values, and also Table III for traces-related statistics.

B. Validation of Main Delay Result

We first validate our model and main delay result (Eq.(1)) against simulations for the two mobility scenarios mentioned (pedestrian and vehicular). The data rate for WiFi is 1.28 Mbps. The mean packet size is 2.27 Mb for both scenarios.

Fig. 4 shows the average file transmission delay (i.e., queueing + service) for the pedestrian scenario, for two different average deadline times of $T_{d1} = 1$ hour ($\xi_1 = 1/3600\ s^{-1}$) and $T_{d2} = 45$ minutes ($\xi_2 = 1/2700\ s^{-1}$), respectively. The range of arrival rates shown corresponds to a server utilization of 0-0.9. We can observe from Fig. 4 that there is a good match between theory and simulations. Furthermore, the average file delay increases as arrival rate increases, as expected, due to queueing effects. On the other hand, the average delay increases for higher deadlines, since flows

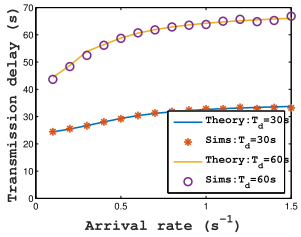

Fig. 5. The average delay for vehicular users' scenarios.

TABLE II
PROBABILITY OF RENEGING FOR PEDESTRIAN AND VEHICULAR SCENARIOS

Scenario	Deadline	$\lambda = 0.05$	$\lambda = 0.45$	$\lambda = 1.45$
Ped.(Theory)	1 hour	0.128	0.883	0.96
Ped.(Simulation)	1 hour	0.129	0.885	0.96
Veh.(Theory)	60 s	0.867	0.982	0.992
Veh.(Simulation)	60 s	0.867	0.98	0.991

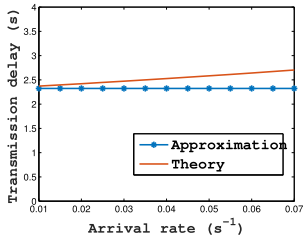

Fig. 6. Low utilization delay approximation for $AR = 0.8$.

with lower deadlines leave the WiFi queue earlier, leading to smaller queueing delays. Fig. 5 further illustrates the average file transmission delay for the vehicular scenario with average deadline times $T_{d1} = 30$ s ($\xi_1 = 1/30\ s^{-1}$) and $T_{d2} = 60$ s ($\xi_2 = 1/60\ s^{-1}$). Despite the differences with the vehicular scenario, similar conclusions can be drawn. Finally, Table II depicts the respective probabilities of reneging for the two scenarios. The percentage of flows that abandon the WiFi queue is higher in the vehicular case, since the availability ratio of the WiFi network is very low (11%), and deadlines are rather small. These observations agree with [7]. Our theory matches simulated results in all the scenarios.

C. Validation of Approximations

We next validate the approximations we have proposed in Section II. We start with the low utilization approximation of Section II-C with AR=0.8 (similar accuracy levels have been obtained with other values as well) and with a deadline of 20 s. Fig. 6 shows the packet delay for arrival rates in the range $0.01 - 0.07\ s^{-1}$, which correspond to a maximum utilization of up to 0.15. As λ increases, the difference between the approximated result and the actual value increases, since we have considered only the service time for this approximation. The same conclusion holds for the probability of reneging (Fig. 7).

Next, we consider the high utilization regime and respective approximation (Eq.(50)). We consider the value of ρ higher than 0.8. Fig. 8 shows the delay for high values of λ, AR=0.5, and an average deadline of 20 s. We can see that

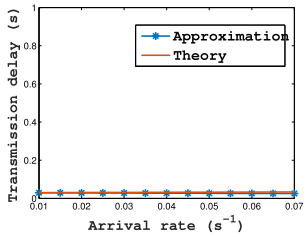

Fig. 7. Low utilization p_r approximation for $AR = 0.8$.

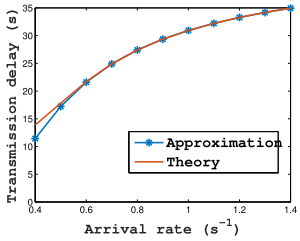

Fig. 8. High utilization delay approximation for $AR = 0.5$.

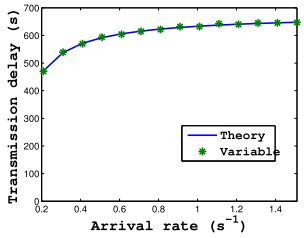

Fig. 9. Variable WiFi rates with the same average as theory.

our approximation is very close to the actual delay and should become exact as ρ gets larger.

D. Variable WiFi Rates and Non-Exponential Parameters

While in our model we consider a fixed transmission rate for all WiFi hotspots, this is not realistic in practice. Hence, we have also simulated scenarios where the WiFi rate varies uniformly in the range 0.5-2.06 Mbps. Fig. 9 shows the delay for the vehicular scenario (AR=0.11) with a deadline of 10 minutes. As can be seen from Fig. 9, even in this case, our theory can give accurate predictions for the incurred delay.

So far we have been assuming 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 ([6], [7]) suggest these distributions to be "heavy-tailed". To this end, we consider a scenario with "heavy-tailed" ON/OFF distributions (Bounded Pareto) for the vehicular case. The shape parameters for the Bounded Pareto ON and OFF periods are $\alpha = 0.59$ and $\alpha = 0.64$. The average deadline is 200 s. Fig. 10 compares the average file delay 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. 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.

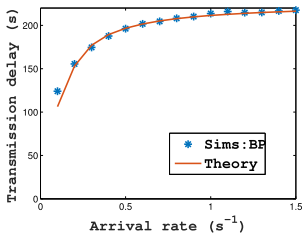

Fig. 10. The delay for BP ON-OFF periods vs. theory.

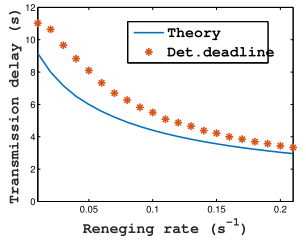

Fig. 11. The delay for deterministic deadlines vs. theory.

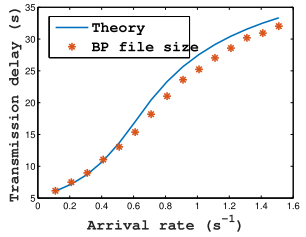

Fig. 12. The delay for BP packet sizes vs. theory.

In all of the above scenarios we have assumed variable deadlines for each file (drawn from an exponential distribution). Nonetheless, in some cases, the user might choose the same deadline for many (or most) flows that can be delayed, as a measure of her patience. To this end, we simulate a scenario where the deadline is fixed for an arrival rate of $0.1\,s^{-1}$. The other parameters are identical to the vehicular scenario. In Fig. 11 we compare simulation results for this scenario against our theory (that assumes exponential deadlines with same average). It is evident that even in this case there is a reasonable match with our theory, despite the different distributions for the deadline.

To conclude our validation with synthetic distributions, we finally drop the exponential file assumption as well, and test our theoretical result vs. generic file size results. Fig. 12 compares analytical and simulation results for Bounded Pareto files sizes with mean 2.27 Mb (shape parameter $\alpha = 1.2$ and coefficient of variation $c_v = 3$). The deadline is $T_d = 20$ s, $AR = 0.5$, and the other parameters correspond to the vehicular scenario. Our theoretical prediction remains reasonably close despite higher file size variability.

E. Validation Against Real Traces

Next, we go a step further and validate our theoretical result with simulations with data from real traces. In the first scenario, we keep the same parameters as in the scenario of Fig. 12, besides the file size. We take the actual file size values from a real trace [21]. The average file size, as reported before,

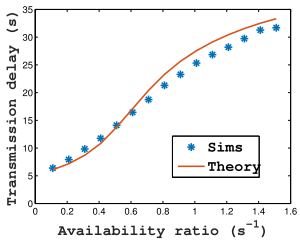

Fig. 13. Real trace files delay vs. theory.

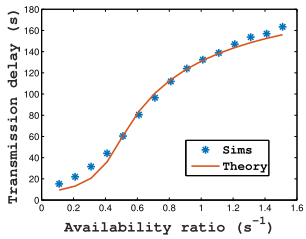

Fig. 14. The delay for trace file sizes and BP ON-OFF vs. theory.

is 2.27 Mb, and the coefficient of variation of the file size there is 2.5. Fig. 13 illustrates the average delay. As can be seen, the system performance can be predicted quite accurately with Eq.(1). In the second scenario with real traces, we keep the files from the same trace as before, but now we generate the ON and OFF periods according to the same Bounded Pareto distributions as in the scenario of Fig. 10. All the other parameters are the same. Fig. 14 shows the simulated average delay vs. the delay obtained from our theory (Eq.(1)). Again, we have shown the importance of our model and its accuracy in predicting the performance.

Finally, we relax all the conditions under which Eq.(1) was derived, and use real data. For that purpose, we have taken the trace containing trips of real buses of [7], that can be found in [20], and that includes: (a) the time of meetings of buses with WiFi points, (b) the duration of these meetings, and (from another related trace in [20]) (c) the amount of exchanged data as well as the duration of meeting times, from which we have obtained related WiFi rates of each meeting (i.e., the data rate in each ON period). We pick a subset of 10 buses and assume that users on the buses generate file requests randomly according to a Poisson process. The file to be downloaded (and the respective size) is chosen from a set of real file sizes corresponding to the trace in [21]. If a bus is within WiFi range during the request, the delivery is performed over WiFi, provided the contact duration is enough (the rate for this download is the one reported in the trace). Otherwise the download is stopped. In the OFF period, the renege clock runs. The average deadline time is 40 s, and is exponential. If the bus runs into a region with WiFi coverage before the deadline expires, the file in service will resume its transmission from the point it stopped, but with the new data rate. If the deadline is shorter than the duration of the OFF period, the file reneges and is served from the cellular network. We then calculate the mean file delay. For our theoretical plot, we look at the statistics from these traces to derive the quantities needed by our model, namely the average duration of ON and OFF periods, the average and the coefficient of variation of file size. The values of these quantities are shown in Table III. Based on

TABLE III

THE STATISTICS OF THE USED DATA TRACES

Variable	Average	c_v
ON periods	12.57 s	1.15
OFF periods	28.42 s	2.9
WiFi data rate	1.28 Mbps	0.98
File (flow) size	2.27 Mb	2.5

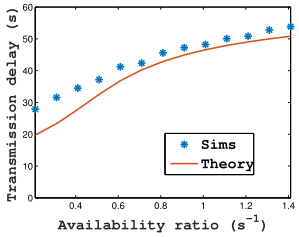

Fig. 15. Scenario with all parameters from real traces.

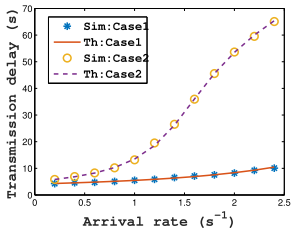

Fig. 16. The average delay for the PS policy for different deadlines.

a study of these traces, we observe that the ON/OFF durations are not exponential, the file sizes have $c_v = 2.5$, and the WiFi rates fluctuate with mean value 1.28 Mbps along different meetings. We then use these values with Eq.(1) to derive our predicted performance for this network. Fig. 15 illustrates the simulated vs. theoretical result (Eq.(1)). Although we have departed from all the theoretical assumptions, our model can still provide a good accuracy (the worst case error is around 20%) even in these extreme cases. This increases even further the value of our model.

F. PS Policy Validation

So far in this section, in all the scenarios, we were running simulations considering only the FCFS service policy. As we have shown, in all the cases, there is a match between simulations and the theoretical result (Eq.(1)). We have also justified, in Section II, that Eq.(1) should also hold for the PS policy. In the following scenarios, we validate that result for the PS policy. Towards that direction we consider a vehicular user having an average deadline of $T_{d1} = 30$ s, WiFi data rate of 10 Mbps, $AR = 0.5$, and the rest of the parameters identical to those of Fig. 5. Fig. 16 illustrates the average delay for that scenario. As we can see from the plot, the theory and simulations give the same result. We can also observe that a moderate arrival rate of 1 s^{-1} leads to an average delay of around 5 s. In the same figure we also show the delay curve for the vehicular scenario with parameters $T_{d2} = 60$ s, WiFi rate of 5 Mbps, and with the rest of parameters remaining unchanged. Again, we can see that there is a good fit between

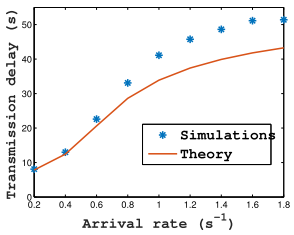

Fig. 17. The average delay (PS) for Bounded Pareto file sizes.

theoretical and simulated results. Due to the lower data rate in the second scenario the delays are much higher.

To conclude our validation for the PS part, we drop the exponential file size assumption, and test our theoretical result vs. heavy-tailed file size result. Fig. 17 compares analytical and simulation results for Bounded Pareto file sizes (shape parameter $\alpha = 1.8$ and $c_v = 2.2$). In this scenario the deadline is 25 s, the mean file size and WiFi rates as before (2.27 Mb and 1.28 Mbps), $AR = 0.5$. The other parameters correspond to the vehicular scenario. Although not to be expected, but our theoretical prediction remains reasonably close despite higher file size variability, with a maximum discrepancy of roughly 15%.

While in a "regular"[5] M/G/1/PS system the average system time is identical (for any packet size distribution) to the average system time of an M/M/1/PS system, and equal to [22]

$$E[T] = \frac{E[S]}{1 - \lambda E[S]}, \tag{52}$$

with $E[S]$ being the average packet size (in time units), this is not the case with our considered queueing system. Namely, our system is intermittent in its nature, and we have observed (simulation-wise) that for such systems Eq.(1) does not hold completely for generic packet sizes. Nevertheless, our result (that holds for exponentially distributed file sizes) can still predict with a good accuracy the performance of an M/G(intermittent)/1/PS system even for heavy-tailed file sizes, as seen in Fig. 17. Providing a more accurate theoretical result for generic file sizes remains part of the future work.

G. Delayed 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 [9], but we do not consider them here. In this last part, we will investigate the actual gains from data offloading, in terms of offloading efficiency. Higher offloading efficiency means better performance for both the client and operator. We compare the offloading efficiencies for on-the-spot offloading [8] vs. delayed offloading for different deadline times ($T_{d1} = 2$ min, $T_{d2} = 1$ min). For the cellular

[5]By a regular PS system we mean a system whose service characteristics do not change over time, i.e., the service rate remains always constant.

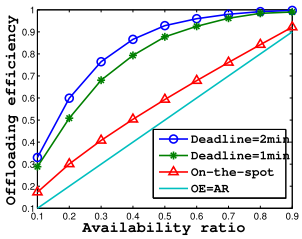

Fig. 18. Offloading gains for delayed vs. on-the-spot offloading.

network in the on-the-spot system, the data rate is 0.6 Mbps. The WiFi rate is 1.28 Mbps. Fig.18 illustrates the offloading efficiency vs. availability ratio for a moderate arrival rate of $\lambda = 0.2\ s^{-1}$. For comparison purposes we also depict the line $x = y$ ($OE = AR$). First, as expected, we can observe that offloading efficiency increases with availability ratio. However, this increase is not linear. More interestingly, the actual offloading efficiencies are always higher than the respective availability ratios. As expected, delayed offloading provides higher offloading efficiencies compared to on-the-spot offloading, with higher deadlines leading to higher offloading efficiencies. For the same AR, by doubling the deadline, the offloading efficiency increases by about 10%. Also, although not shown here, the respective OE increases even further as traffic load decreases. 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. Optimizing Delayed Offloading

The results considered so far allow us to predict the expected system delay when the deadlines are defined externally (e.g., by the user or the application). However, the user (or the device on her behalf) could choose the deadline in order to solve an optimization problem among additional (often conflicting) goals, such as the *monetary cost* for accessing the Internet and the *energy consumption* of the device. For example, the user might want to minimize the delay subject to a maximum (energy or monetary) cost, or to minimize the cost subject to a maximum delay she can tolerate.

To formulate and solve such optimization problems, we need analytical formulas for the average delay and the incurred cost. We already have such formulas for the delay of files sent over WiFi, where we will use the two approximations of Sections II-C and II-D. Furthermore, we can assume that files transmitted over the cellular network incur a fixed delay Δ, capturing both the service and queueing delays over the cellular interface.[6] To proceed, we need to also assume simple models for energy and cost, in order to get some initial intuition about the tradeoffs involved. We are aware that reality is more complex (for both energy and cost) and may differ based on technology (3G, LTE), provider, etc. We plan to extend our models in future work.

[6]We could also try to model the cellular queue as an M/M/1 or G/M/1 system, but we are more interested in the dynamics of the WiFi queue, since this is where the reneging decisions take place. To keep things simple, we defer this to future work. We provide a more detailed discussion on this topic in Section VI.

Assume a user has to download or upload a total amount of data equal to L. On average $p_r \cdot L$ data units will be transmitted over the cellular interface. Assume further that D_c and D_w denote the cost per transmitted data unit for a cellular and WiFi network, where $D_w < D_c$ (often $D_w = 0$). Finally, let c_c and c_w denote the transmission rates, and E_c and E_w energy spent per time unit during transmission over the cellular and WiFi network, respectively.[7] It is normally the case that $c_c < c_w$ as well as $E_c \approx E_w$ [24]. It follows then that the total monetary and energy costs, D, and E, could be approximated by

$$D = (D_c - D_w)p_r + D_w \text{ and } E = \left(\frac{E_c}{c_c} - \frac{E_w}{c_w}\right) p_r + \frac{E_w}{c_w}. \tag{53}$$

A. Optimization Problems

Optimization Problem 1: Eq.(53) suggests that both the average power consumption and cost depend linearly on the probability of reneging, p_r, which we have also derived in Section II, and which is a function of the system deadline $\frac{1}{\xi}$. The system delay is also a function of ξ. We can thus formulate optimization problems of the following form, for both the high and low utilization regimes, where ξ is the optimization parameter:

$$\begin{aligned} &\min_{\xi} \quad E[T] + p_r \Delta \\ &\text{s. t. } p_r \leq P_r^{max}, \end{aligned} \tag{54}$$

where $E[T]$ is given by Eq.(48), and p_r by Eq.(49), for low utilization, and Eq.(50) and Eq.(51), for high utilization, respectively. Due to the linearity of Eq.(53), we can express the constraint directly for p_r, where P_r^{max} depends on whether we consider monetary cost, energy or a weighted sum of both, and the respective parameters. Finally, we can also exchange the optimization function with the constraint to minimize the cost, subject to a maximum delay. This provides us with a large range of interesting optimization problems we can solve.

If we express the inequality constraint in Eq.(54) through ξ, we have the equivalent constraint $\xi \leq \frac{\theta_3 P_r^{max}}{\theta_1 - \theta_2 P_r^{max}}$. The probability of reneging from Eq.(49) is an increasing function of ξ, since $p_r^{'}(\xi) > 0$. This implies that maximum p_r corresponds to maximum ξ. We denote by $f(\xi)$ the total average delay of Eq.(54) (delay function from now on). Hence, we have

$$f(\xi) = \frac{A_1\xi + A_2}{B_1\xi + B_2} + \frac{\theta_1\xi}{\theta_2\xi + \theta_3}\Delta, \tag{55}$$

where $A_1 = \gamma$, $A_2 = (\eta+\gamma)^2 + \eta\mu$, $B_1 = (\mu+\eta)(\gamma+\eta)$, and $B_2 = \mu\gamma(\gamma+\eta)$. In order to solve the optimization problem given by Eq.(54), we need to know the behavior of the delay function. For that purpose, we analyze the monotonicity and convexity of Eq.(55). To do that, we need the first and second derivatives, which are

$$f^{'}(\xi) = \frac{A_1B_2 - A_2B_1}{(B_1\xi + B_2)^2} + \frac{\theta_1\theta_3\Delta}{(\theta_2\xi + \theta_3)^2}, \quad \text{and}$$
$$f^{''}(\xi) = \frac{2(A_2B_1 - A_1B_2)}{(B_1\xi + B_2)^3} - \frac{\theta_1\theta_2\theta_3\Delta}{(\theta_2\xi + \theta_3)^3}.$$

[7]The chosen energy model is clearly an oversimplified one, and is only used to derive some initial insights on the tradeoffs involved. Nevertheless, our model can be extended to include more realistic energy models, such as the one in [23]. Due to space limitations, we refer the interested reader to our tech report [19].

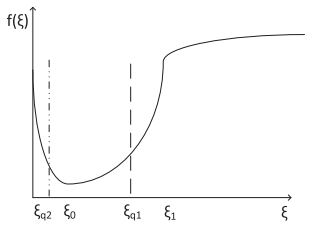

Fig. 19. The delay function for the optimization problem.

It is worth noting that $A_1B_2 < A_2B_1$. This prevents the delay function from being always concave. The delay function is decreasing in the interval for which $f^{'}(\xi) \leq 0$. This happens when

$$\xi \leq \xi_0 = \frac{\theta_3\sqrt{\frac{A_2B_1-A_1B_2}{\theta_1\theta_3\Delta}} - B_2}{B_1 - \theta_2\sqrt{\frac{A_2B_1-A_1B_2}{\theta_1\theta_3\Delta}}}.$$

Hence, the delay function is decreasing in the interval $(0, \xi_0)$, and increasing in the rest, with ξ_0 being a minimum. Further, the solution of $f^{''}(\xi) > 0$ gives the interval where the function is convex. This happens when

$$\xi \leq \xi_1 = \frac{\theta_3\sqrt[3]{2\frac{A_2B_1-A_1B_2}{\theta_1\theta_2\theta_3\Delta}} - B_2}{B_1 - \theta_2\sqrt[3]{2\frac{A_2B_1-A_1B_2}{\theta_1\theta_2\theta_3\Delta}}}. \quad (56)$$

It can be easily proven that $\xi_0 < \xi_1$.

Such constrained-optimization problems are often solved with the Lagrangian method and Karush-Kuhn-Tucker (KKT) conditions. However, the optimal solution for our problem can be found more easily. The delay function is shown in Fig. 19. The optimal deadline depends on the maximum cost, that is proportional to the probability of reneging. So, we can determine the optimal deadline based on the value of P_r^{max}. If this value of P_r^{max} is quite high, the corresponding reneging rate ξ_{q1} (dashed line in Fig. 19) will be higher than the global minimum ξ_0. Consequently, the global minimum of Eq.(56) is also the optimal reneging rate. On the other hand, if the maximum cost is quite low (low P_r^{max}), the maximum reneging rate $\xi_{q,2}$ (dotted line in Fig. 19) is lower than the global minimum. This implies that the minimum delay will be achieved for the maximum reneging rate of $\xi_{q,2} = \frac{\theta_3 P_r^{max}}{\theta_1 - \theta_2 P_r^{max}}$. In other words, the average deadline time that minimizes the delay for a given maximum cost is

$$T_{d,opt} = \frac{1}{\xi_{opt}} = \frac{1}{\min\left(\xi_0, \frac{\theta_3 P_r^{max}}{\theta_1 - \theta_2 P_r^{max}}\right)}. \quad (57)$$

Optimization Problem 2: After minimizing the transmission delay subject to a maximum reneging rate (cost, energy), our next goal is to minimize the reneging probability subject to a maximum transmission delay, which can be for example due to QoS requirements. Hence, the optimization problem in this case would be

$$\begin{aligned} \min_{\xi} \quad & p_r = \frac{\theta_1\xi}{\theta_2\xi + \theta_3} \\ \text{s. t.} \quad & E[T] + p_r\Delta \leq T_{max}. \end{aligned} \quad (58)$$

Just as in Optimization problem 1, we study the monotonicity and convexity of the delay function, which now is the constraint function. For the probability of reneging, we already know that it is an increasing function in ξ. Following a similar procedure as in the previous problem, we get for the optimum value of the deadline (from a quadratic constraint)

$$T_{d,opt} = \frac{1}{\max\left(0, \frac{K_2 - \sqrt{K_2^2 - 4K_1K_3}}{2K_1}\right)}, \quad (59)$$

where $K_1 = A_1\theta_2 + \theta_1\Delta B_1 - T_{max}B_1\theta_2$, $K_2 = T_{max}B_1\theta_3 + T_{max}B_2\theta_2 - A_1\theta_3 - A_2\theta_2 - \theta_1\Delta B_2$, $K_3 = A_2\theta_3 - T_{max}B_2\theta_3$.

We can observe from Eq.(59) that there are two possible scenarios in terms of the optimal deadline. In the first case, an optimal deadline will not exist (its value will be equal to ∞), meaning that the optimal thing to do is to always wait for the WiFi to become available. In the second scenario, a finite non-zero optimal deadline will exist.

Next, we give the solutions to optimization problems for high utilization regime (Optimization problems 3 and 4), where the expressions for $E[T]$ and p_r are given by Eq.(50) and Eq.(51), respectively. The procedure to follow for their solution is similar to the two previous optimization problems. So, we will show only the final results.

Optimization Problem 3: In the first optimization problem for the high utilization regime, our objective function is the transmission delay, and the constraint function is the probability of reneging. So, we have the following problem

$$\begin{aligned} \min_{\xi} \quad & E[T] + p_r\Delta \\ \text{s. t.} \quad & p_r \leq P_r^{max}. \end{aligned} \quad (60)$$

Using the same methodology as before, we get the optimal value of the deadline time that minimizes the average delay, given a maximum cost. That value is

$$T_{d,opt} = \frac{1}{\min\left(\sqrt{\frac{C_1}{D_1\Delta}}, \frac{P_r^{max} - D_2}{D_1}\right)}, \quad (61)$$

where $C_1 = \frac{1}{\lambda}\left(1 + \frac{\gamma}{\eta}\right)(\lambda - \mu\pi_w)$, $C_2 = \frac{(\lambda-\mu)\pi_w}{\lambda\eta}$, $D_1 = \frac{\lambda - \mu\left[\pi_w - \pi_{0,w}(1) + \pi_{0,w}^{'}(1)\right]}{\lambda}$, and $D_2 = \frac{\mu}{\lambda}\pi_{0,w}^{'}(1)$.

Optimization Problem 4: Finally, in the last optimization problem we want to minimize the probability of reneging subject to a maximum delay a packet should experience in the system. The corresponding optimization problem is

$$\begin{aligned} \min_{\xi} \quad & p_r \\ \text{s. t.} \quad & E[T] + p_r\Delta \leq T_{max}, \end{aligned} \quad (62)$$

and its solution is

$$T_{d,opt} = \frac{2D_1\Delta}{T_{max} - C_2 - D_2\Delta - \sqrt{(T_{max} - C_2 - D_2\Delta)^2 - 4C_1D_1\Delta}}. \quad (63)$$

B. Practical Implementation

Our assumption is that the proposed scheme is implemented on the UE side. A detailed architectural description of our approach is beyond the scope of this paper. Nevertheless,

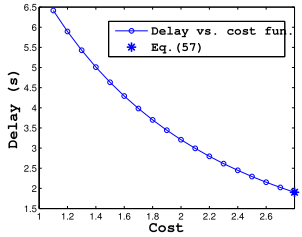

Fig. 20. The delay vs. cost curve for high cellular rate.

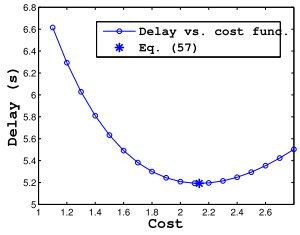

Fig. 21. The delay vs. cost curve for low cellular rate.

we present here for completeness some discussion about how some of the key parameters can be obtained in practice, to implement the algorithm.

WiFi/Cellular Data Rates: The UE can conduct passive rate measurements of both interfaces, e.g., it can maintain a moving average based on the size and delay of earlier file downloads. This way, an estimate of the "effective" rate is available that also considers the impact of other users, and not just the transmission rate related to the channel and rate adaptation. An alternative option is to get them from the 3GPP network entity known as ANDSF [25].

File Sizes: In many cases, files sizes are known in advance (in case of some web page/file/video download). Otherwise, the user can send an initial HTTP query to the server about the file size, and get a response. We have tested this option with a number of URLs, and the majority of servers is ready to respond to such queries.

Arrival Rate: The file generation rate can be easily estimated by the user, similarly to the data rates, by keeping running estimates over longer periods of the time between arrivals.

Availability Ratio: For AR, that depends on $E[T_{ON}]$ and $E[T_{OFF}]$, a number of options are available. Again, the UE could maintain statistics based on connectivity events (to BSs or APs) that are anyway captured by a UE. Alternatively, as in [26], the user can send its GPS data to BS, and the later one in return can estimate the availability ratio and $E[T_{ON}]$ for that user, based on its perfect knowledge of the APs deployment. The user can then compute $E[T_{OFF}]$. We assume that the cellular operator has perfect knowledge of AP positions, due to the fact that big operators own a large number of APs nowadays.

Some additional details about such a UE side implementation (e.g., convergence properties), and a potential BS side implementation can be found in [19].

C. Optimization Evaluation

We will now validate the solutions of the previous optimization problems for two different cases (vehicular user's scenario). In both of them the arrival rate is $0.1\,s^{-1}$, and the maximum cost per data unit one can afford is 2.8 monetary units. The transmission of a data unit through WiFi costs 1, and through cellular interface 5 units. The choice of these values is simply to better visualize the results; different values yield similar conclusions. The WiFi rate is 1 Mbps, and $E[\Gamma] = 1$ Mb. Fig. 20 shows the delay vs. cost curve for cellular rate being $2\times$ lower than WiFi rate. First thing to observe is that the minimum delay is achieved for the highest possible cost (2.8). The optimal average deadline is $T_d = 1$ s. This is in agreement with the optimal value predicted using Eq.(57), and shown with an asterisk in Fig. 20. We replace Eq.(49) into Eq.(53) to get the relationship between the cost and the renege rate. We have shown in Eq.(53) that the cost is directly proportional to p_r, and the later one is an increasing function of ξ. This implies that the maximum cost is in fact the maximum ξ (minimum deadline). This practically means that in Eq.(54) Δ is small and that the delay in the WiFi queue represents the largest component of the delay. As a consequence of that, it is better to redirect the files through the cellular interface as soon as possible. Hence, in these cases (when cellular rate is comparable to the WiFi), the optimum is to assign the shortest possible deadline constrained by the monetary cost.

TABLE IV
OPTIMAL DETERMINISTIC DEADLINE TIMES VS. THEORY

Sc.	Constraint	T_d (theory)	T_d (deterministic)	Relative error
1	$D \leq 2.8$	2.71	2.2	18%
2	$T_{max} = 6.6$	1.56	1.22	22%
3	$D \leq 3.8$	1.73	1.55	11%
4	$T_{max} = 15$	7.97	6.82	15%

Fig. 21 corresponds to a scenario with the same parameters as Fig. 20, except that now the cellular rate is much lower ($10\times$) than the WiFi rate. In that case, Δ is high, and $p_r\Delta$ is the largest component of the delay function. As can be seen from Fig. 21, leaving the WiFi queue immediately is not the best option. The optimum delay is achieved for $T_d = 5$ s. This corresponds to an average cost of $D = 2.1$. This is very close to the theoretical solution of the problem. This is reasonable since for a large difference between the WiFi and cellular rates it is better to wait and then (possibly) be served with higher rate, than to move to a much slower interface (cellular).

Next, we use the solutions of the four optimization problems (for exponentially distributed deadline times) to see how accurately our theory can predict the optimal deadline times, but for deterministic deadlines. The optimal policy essentially finds the optimal value for the *average* deadline (assuming these are exponential). In practice, the chosen deadline will be assigned to all files, and will be deterministic. We consider four scenarios, one for each optimization problem. The costs are the same as before. The arrival rate for low utilization scenarios is $0.1\ s^{-1}$, while for the high ones it is $1.5\ s^{-1}$; $c_w = 1$ Mbps, $E[\Gamma] = 1$ Mb. In Table IV we show the optimal deadlines by using our model (e.g., Eq.(57)), and the optimal deterministic deadlines by using simulations (delay vs. cost plots) with the same parameters as in theory. As can be seen from Table IV, the error in determining the optimal deadline decreases for higher arrival rates. The error is in the range of 10%-20%. This is reasonable since the simulated scenarios are with deterministic deadlines and in our theory

we assume exponential deadlines. Another reason is that in optimization problems we are only using the low and high utilization approximations, and not the exact result (Eq.(1)).

V. Related Work

Authors in [27] 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 of traffic transmitted over the cellular network. A theoretical analysis with some optimization problems related to offloading for opportunistic and vehicular communications is given in [28]. The LTE offloading into WiFi direct is the subject of study in [29]. The work in [30] 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 other works, authors in [31] consider the situation in which cellular operators pay for using the APs from third parties. They use game theory for that purpose. In [25], 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 [32]. In [33], an end-to-end system for adaptive traffic offloading for WiFi-LTE deployment is designed and implemented. Other architectures for implementing offloading are presented in [26], and [34]. Some interesting works on determining the number and position of WiFi APs to be deployed in order to achieve a QoS are [35]–[37].

Some recent influential work in offloading relates to measurements of WiFi availability [6], [7]. Authors in [6] have tracked the behavior of 100 users (most of which were pedestrians) and their measurements reveal that during 75% of the time there is WiFi connectivity. In [7], measurements were conducted on users riding metropolitan area buses. In contrast to the previous study, the WiFi availability reported there is only around 10%. The mean duration of WiFi availability and non-availability periods is also different in the two studies, due to the difference in speeds between vehicular and pedestrian users. The most important difference between the two studies relates to the reported offloading efficiency, with [7] reporting values in the range from 20%-33% for different deadlines, and [6] reporting that offloading does not exceed 3%. We believe this is due to the different deadlines and availabilities considered.

The authors in [38] 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. However, their analysis does not consider queueing effects. Such queueing effects may affect the performance significantly, especially in loaded systems (which are of most interest) or with long periods without WiFi. The work in [39] considers the traffic flow characteristics when deciding when to offload some data to the WiFi. However, there is no delay-related performance analysis. A WiFi offloading system that takes into account a user's throughput-delay tradeoff and cellular budget constraints is proposed in [40]. However, only heuristic algorithms are proposed, and queueing effects are ignored. Summarizing, in contrast to our work, these papers either perform no analysis or use simple models that ignore key system effects such as queueing.

The approach we are using is based on the probability generating functions and is motivated from [16]–[18].

To our best knowledge, the closest work in spirit to ours is [41]. The results in [41] are an extension of results in [6] containing the analysis for delayed offloading. Authors there also use 2D Markov chains to model the state of the system. However, they use matrix-analytic methods to obtain a numerical solution for the offloading efficiency. Such numerical solutions unfortunately do not provide insights on the dependencies between different key parameters, and cannot be used to formulate and analytically solve optimization problems that include multiple metrics.

As a final note, in [8], we have proposed a queueing analytic model for on-the-spot mobile data offloading, and a closed form solution was derived for the average delay. While the model we propose here shares some similarities (ON/OFF availabilities, 2D Markov chain approach) with the basic model in [8], it is in fact considerably more difficult to solve.

VI. Conclusion and Future Work

In this paper we have proposed a queueing analytic model for the performance of delayed mobile data offloading, and have validated it against realistic WiFi network availability statistics. We have shown that our model holds for different scheduling disciplines. We have also considered a number of scenarios where one or more of our model's assumptions do not hold, and have observed acceptable accuracy in terms of predicting the system delay as a function of the user's patience. Finally, we have also shown how to manipulate the maximum deadlines, in order to solve various optimization problems involving the system delay, monetary costs, and power consumption. In the following, we discuss some possible interesting extensions of our model, that remain part of the future work.

Throughout this paper, we were assuming that deadlines are chosen randomly for each file, and do not depend on the actual file size. So, it may happen that a very short file can end up having a very large deadline and vice versa. In general, this might not be very realistic. In practice, a user would set a larger deadline for a larger file (e.g., when downloading a movie). Although capturing the dependency between the deadline and the file size would make our model more realistic, it is beyond the scope of this paper. It remains part of the future work.

A similar problem to setting the deadlines be proportional to file sizes is to introduce the notion of *slowdown* [42], and then optimize it. The slowdown is defined as the ratio of the total system time and the file size, i.e., the average waiting time per file size unit. This is also a very interesting problem, and we will consider it in the future.

As already mentioned in Section IV, the files that are transmitted over the cellular interface incur a fixed delay that captures both the service and queueing time. Despite the fact that we are not considering queueing at the cellular interface, our modeling approach would still be valid in the low utilization regime, where we could easily neglect the queueing delay in the cellular queue, and take the extra delay contribution from the reneging packets simply as $\frac{packet\_size}{cellular\_rate}$. On the other hand, when it comes to moderate and high utilization regimes the situation becomes more complex. In that case, the arrival process at the cellular queue is not Poisson, and

as a best case scenario a G/M/1 queue should be considered. Nevertheless, in order to solve such a system either some other approximations need to be considered (e.g., model the arrival process with a hyperexponential distribution, and consider the queueing system $H_2/M/1$), or only a numerical solution can be obtained. We defer this to future work, too.

REFERENCES

[1] (Feb. 2015). *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2014–2019*. [Online]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.pdf

[2] *Mobile Data Offloading Through WiFi*, Proxim Wireless, Fremont, CA, USA, 2010.

[3] (2012). *Growing Data Demands Are Trouble for Verizon, LTE Capacity Nearing Limits*. [Online]. Available: http://www.talkandroid.com/97125-growing-data-demands-are-trouble-for-verizon-lte-capacity-nearing-limits/

[4] T. Kaneshige. (Dec. 2009). *AT&T iPhone Users Irate at Idea of Usage-Based Pricing*. [Online]. Available: http://www.pcworld.com/article/184589/ATT_IPhone_Users_Irate_at_Idea_of_Usage_Based_Pricing.html

[5] *SIPTO and LIPA*, accessed on 2012. [Online]. Available: http://www.3gpp1.eu/ftp/Specs/archive/23series/23.829/

[6] K. Lee, J. Lee, Y. Yi, I. Rhee, and S. Chong, "Mobile data offloading: How much can WiFi deliver?" in *Proc. ACM Co-NEXT*, 2010, Art. no. 26.

[7] A. Balasubramanian, R. Mahajan, and A. Venkataramani, "Augmenting mobile 3G using WiFi," in *Proc. ACM MobiSys*, 2010, pp. 209–222.

[8] F. Mehmeti and T. Spyropoulos, "Performance analysis of 'on-the-spot' mobile data offloading," in *Proc. IEEE GLOBECOM*, Dec. 2013, pp. 1577–1583.

[9] N. Balasubramanian, A. Balasubramanian, and A. Venkataramani, "Energy consumption in mobile phones: A measurement study and implications for network applications," in *Proc. ACM IMC*, 2009, pp. 280–293.

[10] S. Sen, C. Joe-Wong, S. Ha, and M. Chiang, "Incentivizing time-shifting of data: A survey of time-dependent pricing for Internet access," *IEEE Commun. Mag.*, vol. 50, no. 11, pp. 91–99, Nov. 2012.

[11] S. Ha, S. Sen, C. Joe-Wong, Y. Im, and M. Chiang, "Tube: Time-dependent pricing for mobile data," in *Proc. ACM SIGCOMM*, 2012, pp. 247–258.

[12] S. Sen, C. Joe-Wong, S. Ha, J. Bawa, and M. Chiang, "When the price is right: Enabling time-dependent pricing of broadband data," in *Proc. ACM SIGCHI Conf. Human Factors Comput. Syst.*, 2013, pp. 2477–2486.

[13] S. M. Ross, *Stochastic Processes*. New York, NY, USA: Wiley, 1996.

[14] D. Y. Barrer, "Queuing with impatient customers and ordered service," *Oper. Res.*, vol. 5, no. 5, pp. 650–656, 1957.

[15] R. E. Stanford, "Reneging phenomena in single channel queues," *Math. Oper. Res.*, vol. 4, no. 2, pp. 162–178, 1979.

[16] U. Yechiali and P. Naor, "Queuing problems with heterogeneous arrivals and service," *Oper. Res.*, vol. 19, no. 3, pp. 722–734, 1971.

[17] N. Perel and U. Yechiali, "Queues with slow servers and impatient customers," *Eur. J. Oper. Res.*, vol. 201, no. 1, pp. 247–258, 2010.

[18] E. Altman and U. Yechiali, "Analysis of customers' impatience in queues with server vacations," *Queueing Syst.*, vol. 52, no. 4, pp. 261–279, 2006.

[19] F. Mehmeti and T. Spyropoulos, "Optimization of delayed mobile data offloading," EURECOM, Biot, France, Res. Rep. RR-13-286, 2013. [Online]. Available: http://www.eurecom.fr/en/publication/4097

[20] *WiFi Availability Trace*, accessed on 2011. [Online]. Available: http://traces.cs.umass.edu/index.php/Network/Network1

[21] *Packet Size Trace*, accessed on 2013. [Online]. Available: http://www.wsdream.com/dataset.html

[22] L. Kleinrock, *Queueing Systems: Computer Applications*, vol. 2. New York, NY, USA: Wiley, 1976.

[23] Y. Xiao, P. Savolainen, A. Karppanen, M. Siekkinen, and A. Ylä-Jääski, "Practical power modeling of data transmission over 802.11g for wireless applications," in *Proc. ACM e-Energy*, 2010, pp. 75–84.

[24] A. Sharma, V. Navda, R. Ramjee, V. N. Padmanabhan, and E. M. Belding, "Cool-Tether: Energy efficient on-the-fly WiFi hot-spots using mobile phones," in *Proc. ACM CoNEXT*, 2009, pp. 109–120.

[25] D. S. Kim, Y. Noishiki, Y. Kitatsuji, and H. Yokota, "Efficient ANDSF-assisted Wi-Fi control for mobile data offloading," in *Proc. IWCMC*, 2013, pp. 343–348.

[26] S. Dimatteo, P. Hui, B. Han, and V. O. K. Li, "Cellular traffic offloading through WiFi networks," in *Proc. IEEE MASS*, Oct. 2011, pp. 192–201.

[27] B. Han *et al.*, "Mobile data offloading through opportunistic communications and social participation," *IEEE Trans. Mobile Comput.*, vol. 11, no. 5, pp. 821–834, May 2012.

[28] Y. Li *et al.*, "Multiple mobile data offloading through delay tolerant networks," in *Proc. ACM CHANTS*, 2011, pp. 43–48.

[29] A. Pyattaev, K. Johnsson, S. Andreev, and Y. Koucheryavy, "3GPP LTE traffic offloading onto WiFi direct," in *Proc. IEEE WCNC*, Apr. 2013, pp. 135–140.

[30] S. Singh, H. S. Dhillon, and J. G. Andrews, "Offloading in heterogeneous networks: Modeling, analysis, and design insights," *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, pp. 2484–2497, May 2013.

[31] J. Lee, Y. Yi, S. Chong, and Y. Jin, "Economics of WiFi offloading: Trading delay for cellular capacity," in *Proc. IEEE INFOCOM Workshop SDP*, Apr. 2013, pp. 357–362.

[32] A. Y. Ding *et al.*, "Enabling energy-aware collaborative mobile data offloading for smartphones," in *Proc. IEEE SECON*, Jun. 2013, pp. 487–495.

[33] R. Mahindra, H. Viswanathan, K. Sundaresan, M. Y. Arslan, and S. Rangarajan, "A practical traffic management system for integrated LTE-WiFi networks," in *Proc. ACM MobiCom*, 2014, pp. 189–200.

[34] A. Y. Ding *et al.*, "Vision: Augmenting WiFi offloading with an open-source collaborative platform," in *Proc. MCS*, 2015, pp. 44–48.

[35] J. Kim, N.-O. Song, B. H. Jung, H. Leem, and D. K. Sung, "Placement of WiFi access points for efficient WiFi offloading in an overlay network," in *Proc. IEEE PIMRC*, Sep. 2013, pp. 3066–3070.

[36] A. Abdrabou and W. Zhuang, "Probabilistic delay control and road side unit placement for vehicular ad hoc networks with disrupted connectivity," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 1, pp. 129–139, Jan. 2011.

[37] K. Berg and M. Katsigiannis, "Optimal cost-based strategies in mobile network offloading," in *Proc. ICST CROWNCOM*, 2012, pp. 95–100.

[38] D. Zhang and C. K. Yeo, "Optimal handing-back point in mobile data offloading," in *Proc. IEEE VNC*, Nov. 2012, pp. 219–225.

[39] S. Wiethölter, M. Emmelmann, R. Andersson, and A. Wolisz, "Performance evaluation of selection schemes for offloading traffic to IEEE 802.11 hotspots," in *Proc. IEEE ICC*, Jun. 2012, pp. 5423–5428.

[40] Y. Im *et al.*, "AMUSE: Empowering users for cost-aware offloading with throughput-delay tradeoffs," in *Proc. IEEE INFOCOM*, Apr. 2013, pp. 435–439.

[41] 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, pp. 536–550, Apr. 2013.

[42] M. Harchol-Balter, *Performance Modeling and Design of Computer Systems: Queueing Theory in Action*. Cambridge, U.K.: Cambridge Univ. Press, 2013.

Fidan Mehmeti received the B.Sc. degree in electronics and the M.Sc. degree in telecommunications from the University of Prishtina, Kosovo, in 2006 and 2009, respectively, and the Ph.D degree from the Institute Eurecom, Sophia Antipolis, France, in 2015. He is currently a Post-Doctoral Fellow with the University of Waterloo, Canada. His main research interests are in performance modeling and analysis for cognitive radio networks, mobile data offloading, and heterogeneous networks.

Thrasyvoulos Spyropoulos received the Diploma degree in electrical and computer engineering from the National Technical University of Athens, Greece, and the Ph.D. degree in electrical engineering from the University of Southern California. He was a Post-Doctoral Researcher with INRIA and also a Senior Researcher with the Swiss Federal Institute of Technology Zurich. He is currently an Assistant Professor with EURECOM, Sophia Antipolis.