EURECOM
Department of Mobile Communications
Campus Sophia Tech
Les Templiers
450 route des Chappes
B.P. 193
06410 Biot
FRANCE

Research Report RR-13-286

Optimization of Delayed Mobile Data Offloading

July 25th, 2013
Last update March 26th, 2016

Fidan Mehmeti, Thrasyvoulos Spyropoulos

[1]EURECOM's research is partially supported by its industrial members: BMW Group, Cisco, Monaco Telecom, Orange, SAP, SFR, STEricsson, Swisscom, Symantec.

Tel : (+33) 4 93 00 81 00
Fax : (+33) 4 93 00 82 00
Email : {mehmeti,spyropou}@eurecom.fr

Optimization of Delayed Mobile Data Offloading

Fidan Mehmeti, Thrasyvoulos Spyropoulos

Abstract

Operators have recently resorted to WiFi offloading to deal with increasing data demand and induced congestion. Researchers have further suggested the use of "delayed offloading": if no WiFi 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. WiFi availability), user traffic load, etc. 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. We validate the accuracy of our results using a range of realistic scenarios, and use these expressions to show how to optimally choose deadlines.

Index Terms

Mobile data offloading, Deadlines, Queueing, 2D Markov chain, Optimization.

Contents

List of Figures

1 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 have high-bandwidth requirements. Cisco [2] reports that by 2017 the mobile data traffic will increase by 13 times, and will climb to 13.2 exabytes per month. Mobile video traffic will comprise 66% of the total traffic, compared to 51% in 2012 [2].

This increase in traffic demand is overloading cellular networks, forcing them to operate close to (and often beyond) their capacity limits. Upgrading to LTE or LTE-advanced, as well as the deployment of additional network infrastructure could help alleviate this capacity crunch [3], but reports already suggest that such solutions are bound to face the same problems [4]. 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 [5].

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 [6]), and the use of WiFi. In 2012, 33% of the total mobile data traffic was offloaded [2]. Projections say that this will increase to 46% by 2017 [2]. 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 [3], etc. Also, wireless operators have already deployed or bought a large number of WiFi access points (AP) [3].

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 through the cellular network [7, 8].

We have already analyzed the case of on-the-spot offloading in [9]. 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 [10]. Finally, with more operators moving away from flat rate plans towards usage-based plans [11], users have incentives to delay "bulky" traffic to conserve their plan quotas or to receive better prices [12].

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 [7, 8]. Additionally, the exact amount of delay a flow can tolerate is expected to depend heavily on (a) the user, and (b) the application type. For example, a study performed in [13] 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?*

The main contributions of this paper can be summarized as follows: (i) We propose a queueing analytic model for the problem of delayed offloading, based on two-dimensional Markov chains, and derive expressions for the average delay, and other performance metrics as a function of the deadlines, and key system parameters; we also give closed-form approximations for different regimes of interest; (*Section 2*) (ii) We validate our results extensively, using also scenarios and parameters observed in real measurement traces that depart from the assumptions made in our model; (*Section 3*) (iv) We formulate and solve basic cost-performance optimization problems, and derive the achievable tradeoff regions as a function of the network parameters (WiFi availability, user load, etc.) in hand (*Section 4*).

2 Analysis of Delayed Offloading

In this section, we formulate the delayed offloading problem, and derive analytical expressions for key metrics (e.g. mean per 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 a First Come First Served (FCFS) queueing discipline. When the WiFi connectivity is lost, we assume that flows waiting in the queue and new flows arriving can be delayed until

[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.

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 [13]. 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 [14] $\left(T_{ON}^{(i)}, T_{OFF}^{(i)}\right), i \geq 1$, 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 rate equal to μ. 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 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 variables that are exponentially distributed with 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 μ), etc. However, the assumptions of exponentiallity, 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 3 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.[4]

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 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 4.

[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 [7] suggest this to be the case, on average.

[4]We could further extend our framework to arbitrary ON and OFF distributions using Coxian distributions and matrix-analytic methods [15]. However, the latter are only numerical, reducing their utility. We defer such scenarios and potential closed-form approximations to future work.

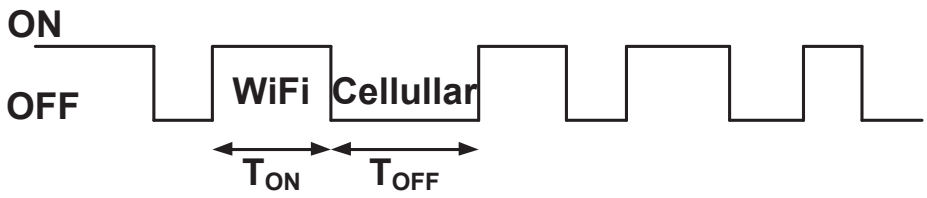

Figure 1: The WiFi network availability model.

Before proceeding, we summarize in Table 1 some useful notation. Also, the total time a file spends in the system (queueing+ service time) will be referred to as the *system time* or *transmission delay*.

Table 1: Variables and Shorthand Notation.

Variable	Definition/Description
T_{ON}	Duration of ON (WiFi) periods
T_{OFF}	Duration of periods (OFF) without WiFi connectivity
λ	Average packet (file) arrival rate at the mobile user
$\pi_{i,c}$	Stationary probability of finding i files in cellular state
$\pi_{i,w}$	Stationary probability of finding i files in WiFi state
π_c	Probability of finding the system under cellular coverage only
π_w	Probability of finding the system under WiFi coverage
p_r	Probability of reneging
η	The rate of leaving the WiFi state
γ	The rate of leaving the cellular state
μ	The service rate while in 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

2.1 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 (in FCFS order) 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 [16] or with *reneging* [17]. Throughout our analysis, we'll

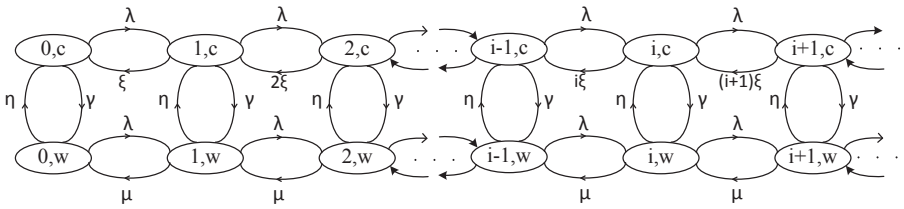

Figure 2: The 2D Markov chain for the WiFi queue in delayed offloading.

assume that files will abandon the queue only during periods without WiFi connectivity[5]. Nevertheless, in Section 3 we consider also the deterministic deadlines, i.e. a file will be sent over the cellular network. 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 4). 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. $\frac{file\_size}{\mu}$).

Given the previously stated assumptions, the WiFi queue can be modeled with a 2D Markov chain, as shown in Figure 2. States with WiFi connectivity are denoted with $\{i, w\}$, and states with cellular connectivity only with $\{i, c\}$. i corresponds to the number of customers in the system (service+queue). During WiFi states, the system empties at rate μ (since files are transmitted 1-by-1) and 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 [18], [19], [20].

Theorem 1. *The mean system time for the WiFi queue 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 probability of finding i files when there is only cellular network coverage, or WiFi coverage, respectively.

[5]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.

Writing the balance equations for the cellular and WiFi states gives

$$
\begin{aligned}
(\lambda+\gamma)\pi_{0,c} &= \eta\pi_{0,w} + \xi\pi_{1,c} && (2)\\
(\lambda+\gamma+i\xi)\pi_{i,c} &= \eta\pi_{i,w} + (i+1)\xi\pi_{i+1,c} + \lambda\pi_{i-1,c} && (3)\\
(\lambda+\eta)\pi_{0,w} &= \gamma\pi_{0,c} + \mu\pi_{1,w} && (4)\\
(\lambda+\eta+\mu)\pi_{i,w} &= \gamma\pi_{i,c} + \mu\pi_{i+1,w} + \lambda\pi_{i-1,w} && (5)
\end{aligned}
$$

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

We define the probability generating functions for both the cellular and WiFi states as $G_c(z) = \sum_{i=0}^{\infty} \pi_{i,c} z^i$, and $G_w(z) = \sum_{i=0}^{\infty} \pi_{i,w} z^i, |z| \leq 1$. After multiplying Eq.(3) with z^i and adding 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

$$\xi(1-z)G_c^{'}(z) = \left(\lambda(1-z)+\gamma\right)G_c(z) - \eta G_w(z). \quad (7)$$

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 yields to

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

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

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

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

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 easily be shown 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 of Eq.(9). Performing some simple calculus operations, the above function transforms into

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

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

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

We determine the coefficients M and N in the standard way as $M = \frac{\frac{\mu}{\lambda} - z}{z_2 - z} \mid_{z=z_1} = \frac{\frac{\mu}{\lambda} - z_1}{z_2 - z_1} = \frac{z_1 z_2 - z_1}{z_2 - z_1} > 0$, and $N = \frac{\frac{\mu}{\lambda} - z}{z - z_1} \mid_{z=z_2} = \frac{\frac{\mu}{\lambda} - z_2}{z_2 - z_1} < 0$.

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

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

We thus need to integrate the function in Eq.(15):

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

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

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

Now, Eq.(13) is equivalent to

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

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

$$k_1(z) = e^{-\frac{\lambda}{\xi}z}\left(z_1 - z\right)^{\frac{\gamma M}{\xi}}\left(z_2 - z\right)^{-\frac{\gamma N}{\xi}}, z \le z_1, \tag{17}$$

$$k_2(z) = e^{-\frac{\lambda}{\xi}z}\left(z - z_1\right)^{\frac{\gamma M}{\xi}}\left(z_2 - z\right)^{-\frac{\gamma N}{\xi}}, z \ge z_1. \tag{18}$$

Eq.(16) now becomes

$$\frac{d}{dz}\left(k_1(z)G_c(z)\right) = \frac{\eta\mu\pi_{0,w}}{\xi\beta(z)}k_1(z), z \le z_1, \tag{19}$$

$$\frac{d}{dz}\left(k_2(z)G_c(z)\right) = \frac{\eta\mu\pi_{0,w}}{\xi\beta(z)}k_2(z), z \ge z_1, \tag{20}$$

and after integrating we obtain

$$k_1(z)G_c(z) = \frac{\eta\mu\pi_{0,w}}{\xi}\int_0^z \frac{k_1(x)}{\beta(x)}dx + C_1, z \leq z_1 \tag{21}$$

$$k_2(z)G_c(z) = \frac{\eta\mu\pi_{0,w}}{\xi}\int_{z_1}^z \frac{k_2(x)}{\beta(x)}dx + C_2, z \geq z_1. \tag{22}$$

The bounds of the integrals in Eq.(21) and Eq.(22) come from the defining region of z in Eq.(19)-(20). We need to determine the coefficients C_1 and C_2 in Eq.(21) and Eq.(22). We take $z = 0$ in Eq.(21). 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 for $C_1 = \pi_{0,c} z_1^{\frac{\gamma M}{\xi}} z_2^{-\frac{\gamma N}{\xi}}$. In a similar fashion we get for $C_2 = 0$.

Finally, for the PGF in the cellular state we have

$$G_c(z) = \frac{\eta\mu\pi_{0,w}\int_0^z \frac{k_1(x)}{\beta(x)}dx + \xi\pi_{0,c} z_1^{\frac{\gamma M}{\xi}} z_2^{-\frac{\gamma N}{\xi}}}{\xi k_1(z)}, z \leq z_1, \tag{23}$$

$$G_c(z) = \frac{\eta\mu\pi_{0,w}\int_{z_1}^z \frac{k_2(x)}{\beta(x)}dx}{\xi k_2(z)}, z \geq z_1. \tag{24}$$

In the last two equations, 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

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

Similarly, from the boundary conditions in Eq.(23) for $z \leq z_1$, we get

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

After solving the system of equations Eq.(25) and Eq.(26), for the 'zero probabilities' we obtain

$$\pi_{0,w} = \frac{\xi k_2(1)}{(\eta+\gamma)\mu}\frac{1}{\int_{z_1}^1 \frac{k_2(x)}{\beta(x)}dx}, \text{and} \tag{27}$$

$$\pi_{0,c} = -\frac{\eta k_2(1)\int_0^{z_1} \frac{k_1(x)}{\beta(x)}dx}{(\eta+\gamma) z_1^{\frac{\gamma M}{\xi}} z_2^{-\frac{\gamma N}{\xi}} \int_{z_1}^1 \frac{k_2(x)}{\beta(x)}dx}. \tag{28}$$

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

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

$$\lambda\pi_{i,c} + \lambda\pi_{i,w} = \mu\pi_{i+1,w} + (i+1)\xi\pi_{i+1,c}. \tag{29}$$

Summing over all i yields to

$$\lambda(\pi_c + \pi_w) = \mu(\pi_w - \pi_{0,w}) + \xi\sum_{i=0}^{\infty}(i+1)\pi_{i+1,c}. \tag{30}$$

The last equation, obviously reduces to

$$\lambda = \mu(\pi_w - \pi_{0,w}) + \xi E[N_c], \tag{31}$$

where $E[N_c] = G_c^{'}(1)$, and $E[N_w] = G_w^{'}(1)$. Eq.(31) yields

$$E[N_c] = \frac{\lambda - \mu(\pi_w - \pi_{0,w})}{\xi}. \tag{32}$$

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)

$$\begin{aligned} G_w^{'}(z) &= \frac{\beta(z)\left(\gamma G_c(z) + \gamma z G_c^{'}(z) + \mu\pi_{0,w}\right)}{\beta^2(z)} \\ &- \frac{\beta^{'}(z)\left(\gamma z G_c(z) - \mu(1-z)\pi_{0,w}\right)}{\beta^2(z)}, \end{aligned} \tag{33}$$

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

$$E[N_w] = \frac{\left(\gamma E[N_c] + \mu\pi_{0,c}\right)\eta - \gamma\pi_c(\mu - \lambda)}{\eta^2}. \tag{34}$$

Replacing Eq.(32) into Eq.(34) we get

$$E[N_w] = \frac{\gamma}{\eta}\frac{\lambda - \mu\left(\pi_w - \pi_{0,w}\right)}{\xi} + \frac{\mu\pi_{0,w}}{\eta} - \frac{\gamma\pi_c(\mu - \lambda)}{\eta^2}. \tag{35}$$

The average number of files in the system is

$$E[N] = E[N_c] + E[N_w]. \tag{36}$$

Finally, using the Little's law $E[N] = \lambda E[T]$ [14], we obtain the average packet delay in delayed data 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 incure 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.

Corollary 2. *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*

$$p_r = \frac{\lambda - \mu(\pi_w - \pi_{0,w})}{\lambda}. \tag{37}$$

In other words, 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.(32) gives us the above result. This also gives us another important metric, the *offloading efficiency* of our system, namely the percentage of flows that get offloaded over some WiFi network, as $E_{off} = 1 - p_r$.

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 remarkable accuracy even in scenarios where many of the assumptions don't 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.

2.2 Low utilization approximation

One interesting scenario 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.). 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 the state $(1, c)$ or $(1, w)$ first enters any of the states $(0, c)$ or $(0, w)$. Hence, the average service time is

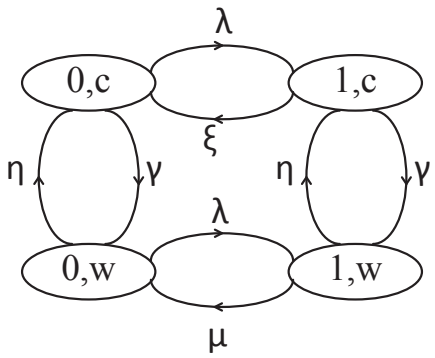

Figure 3: The reduced Markov chain for $\rho \to 0$.

$$E[S] = \frac{\eta}{\gamma+\eta}E[T_c] + \frac{\gamma}{\gamma+\eta}E[T_w], \tag{38}$$

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

$$E[T_c] = P[I_c = 1]E[T_c|I_c = 1] + P[I_c = 0]E[T_c|I_c = 0], \tag{39}$$

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.(39), we have

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

$$E[T_c|I_c = 0] = \frac{1}{\xi+\gamma} + E[T_w]. \tag{41}$$

Eq.(40) is actually the expected value of the minimum of two exponentially distributed random varibles with rates ξ and γ. Replacing Eq.(40) and (41) into Eq.(39), we get

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

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

$$E[T_w] - \frac{\eta}{\mu+\eta}E[T_c] = \frac{1}{\mu+\eta}. \tag{43}$$

After solving the system of equations Eq.(42)-(43), we have

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

$$E[T_c] = \frac{\mu + \gamma + \eta}{\xi\mu + \xi\eta + \mu\gamma}. \tag{45}$$

Now, replacing Eq.(44)-(45) into Eq.(38), we have the average service time, and the low utilization approximation is ($E[T] \approx E[S]$)

$$E[T] = \frac{(\eta + \gamma)^2 + \gamma\xi + \eta\mu}{(\xi\mu + \xi\eta + \mu\gamma)(\gamma + \eta)}. \tag{46}$$

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

$$\pi_{0,c} = \frac{\eta}{\eta + \gamma} \frac{\xi(\mu + \lambda + \eta) + \mu\gamma}{\xi(\mu + \lambda + \eta) + \mu\gamma + \lambda^2 + \lambda(\eta + \gamma + \mu)}. \tag{47}$$

Replacing Eq.(47) and $\pi_c = \frac{\eta}{\eta+\gamma}$ into Eq.(37), we get the probability of reneging for low utilization as

$$p_r = \frac{\theta_1 \xi}{\theta_2 \xi + \theta_3}, \tag{48}$$

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

2.3 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. Here, we provide an approximation that corresponds to the region of high utilization ($\rho \to 1$).

The expected system time in the WiFi queue for a user with heavy traffic can be approximated by

$$E[T] = \frac{1}{\lambda}\left[\left(1 + \frac{\gamma}{\eta}\right)\frac{\lambda - \mu\pi_w}{\eta} + \frac{(\lambda - \mu)\pi_w}{\eta}\right]. \tag{49}$$

The approximation Eq.(52) comes from Eq.(1) by replacing $\pi_{0,w} = 0$.

To find the approximation for the probability of reneging in the high utilization regime we proceed as follows. Since from Eq.(37) the only term that depends on ξ is $\pi_{0,w}$ (we will need it later to solve optimization problems), we will not take it equal to 0. We will approximate it by a first order Taylor approximation at $\xi = 1$. For that purpose, we will denote $\pi_{0,w}$ as $\pi_{0,w}(\xi)$. So, we write

$$\pi_{0,w}(\xi) = \pi_{0,w}(1) + (\xi - 1)\pi'_{0,w}(1), \tag{50}$$

where $\pi_{0,w}(1) = \frac{A(1)}{(\eta+\gamma)\mu \int_{z_1}^{1} \frac{A(x)}{\beta(x)} dx}$, and $\pi'_{0,w}(1) = \frac{A(1)}{(\eta+\gamma)\mu} \frac{\ln(A(1)e) \int_{z_1}^{1} \frac{A(x)}{\beta(x)} dx - \int_{z_1}^{1} \frac{A(x) \ln A(x)}{\beta(x)} dx}{\left(\int_{z_1}^{1} \frac{A(x)}{\beta(x)} dx\right)^2}$,

where $A(x) = \frac{(x-z_1)^{\gamma M}}{(z_2-x)^{\gamma N}} e^{-\lambda x}$.

Hence, the probability of reneging in the high utilization regime can be approximated by

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

where $\pi_{0,w}$ is given by Eq.(50).

Now, we can write jointly the expressions for the average file delay and probability of reneging in a delayed offloading system for high utilization.

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*

$$E[T] = \frac{1}{\lambda}\left[\left(1 + \frac{\gamma}{\eta}\right)\frac{\lambda - \mu\pi_w}{\xi} + \frac{(\lambda-\mu)\pi_w}{\eta}\right], \tag{52}$$

$$p_r = \frac{\lambda - \mu\pi_w}{\lambda} + \frac{\mu}{\lambda}\pi_{0,w}, \tag{53}$$

where $\pi_{0,w}$ is the first order Taylor series approximation of Eq.(27).

3 Performance evaluation

In this section we will validate our theory against simulations for a wide range of traffic intensities, different values of file sizes, WiFi availability periods with different distributions, and different deadline times. We define the WiFi availability ratio as $AR = \frac{E[T_{ON}]}{E[T_{ON}]+E[T_{OFF}]} = \frac{\gamma}{\eta+\gamma}$. Unless otherwise stated the durations of WiFi availability and unavailability periods will be drawn from independent exponential distributions with rates η and γ, respectively. The deadlines are exponentially distributed with rate ξ, although we will simulate scenarios with deterministic deadlines as well. We mainly focus on two scenarios, related to the user's mobility. The first one considers mostly pedestrian users with data taken from [7]. Measurements in [7] 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 and γ). The availability ratio is thus 75 %. The second scenario corresponds to vehicular users, related to the measurement study of [8]. An availability ratio of 11 % has been reported in [8], although not all the details are mentioned there. For more details about the measurements we refer the interested reader to [7] and [8]. Finally, unless otherwise stated, file/flow sizes are exponentially distributed, and file arrival at the mobile user is a Poisson process with rate λ.

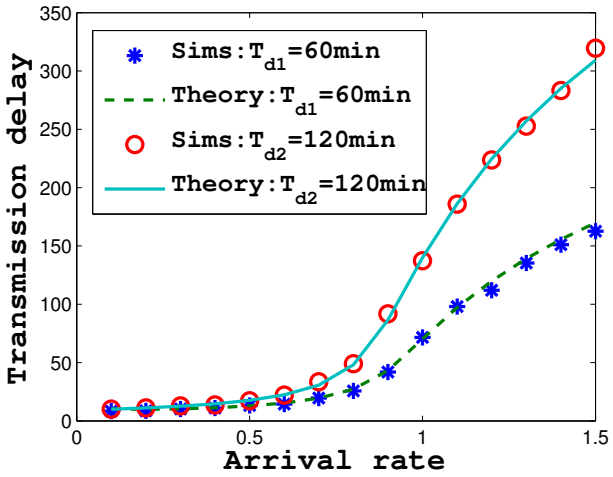

Figure 4: The average delay for pedestrian users' scenarios.

3.1 Validation of main delay result

We first validate here 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 assumed to be 1 Mbps. The mean packet size is assumed to be 7.5 MB for the pedestrian scenario and 125 kB for the vehicular scenario.

Fig. 4 shows the average file transmission delay (i.e. queueing + transmission) for the pedestrian scenario, for two different average deadline times of $T_{d1} = 1$ hour ($\xi_1 = 1/3600s^{-1}$) and $T_{d2} = 2$ hours ($\xi_2 = 1/7200s^{-1}$), 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 transmission delay is increased by increasing the arrival rate, as expected, due to queueing effects. On the other hand, the average delay increases for higher deadlines, since flows 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 ($\xi_1 = 1/30s^{-1}$) and $T_{d2} = 60$s ($\xi_2 = 1/60s^{-1}$). Despite the differences of the vehicular scenario, similar conclusions can be drawn.

Finally, Table 2 depicts the respective probabilities of reneging for the two scenarios. The percentage of flows that abandon the WiFi queue is higher in the vehicular scenario, since the availability ratio of the WiFi network is very small (11%), and deadlines are rather small. These observations agree with [8]. Nevertheless, our theory matches simulation values in all scenarios.

So far, we have assumed exponential distributions for ON and OFF periods, according to our model. While the actual distributions are subject to the user mobility pattern, a topic of intense research recently, initial measurement studies ([7, 8])

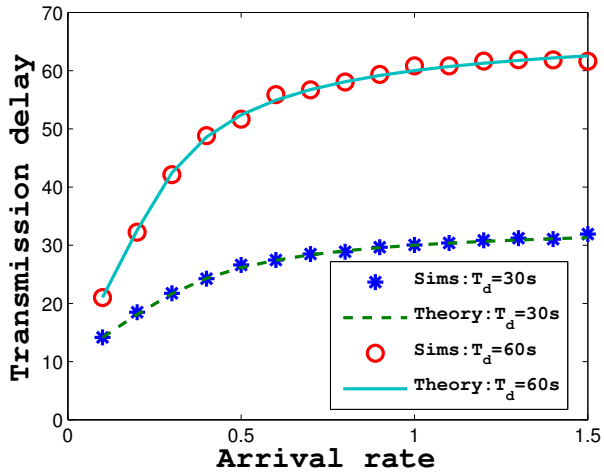

Figure 5: The average delay for vehicular users' scenarios.

Table 2: Probability of reneging for pedestrian and vehicular scenarios.

Scenario	Deadline	$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 1$	$\lambda = 1.5$
Pedestrian(Theory)	1 hour	0.103	0.109	0.252	0.501
Pedestrian(Simulation)	1 hour	0.1	0.117	0.239	0.508
Vehicular(Theory)	60 s	0.32	0.778	0.889	0.926
Vehicular(Simulation)	60 s	0.32	0.776	0.891	0.925

suggest these distributions to be "heavy-tailed". 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 only. The shape parameters for the Bounded Pareto ON and OFF periods are $\alpha = 0.59$ and $\alpha = 0.64$, respectively. The average deadline is 100s. Fig. 6 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.

3.2 Validation of approximations

We next validate the approximations we have proposed in Section 2. We start with the low utilization approximation of Section 2.2 and consider the availability ratio to be 0.75 (similar accuracy levels have been obtained with other values) and with a deadline of 2 min. Fig. 7 shows the packet delay for low arrival rates in the range $0.01 - 0.11$, which correspond to a maximum utilization of up to 0.2. As λ increases, the difference between the approximated result and the actual value

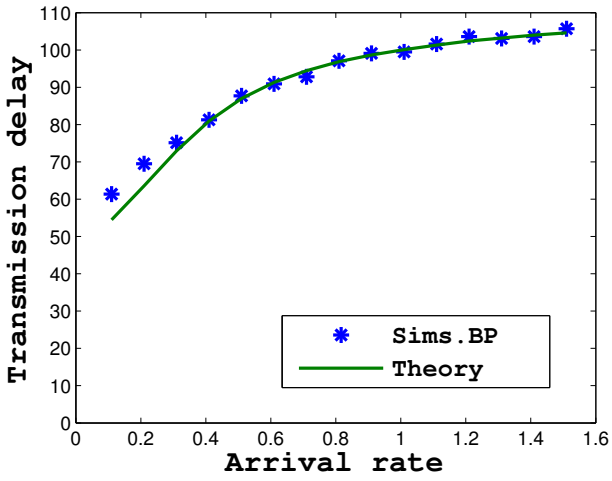

Figure 6: The delay for BP ON-OFF periods vs. theory.

increases, since we have considered only the service time for this approximation. The same conclusion holds for the probability of reneging (Fig. 8).

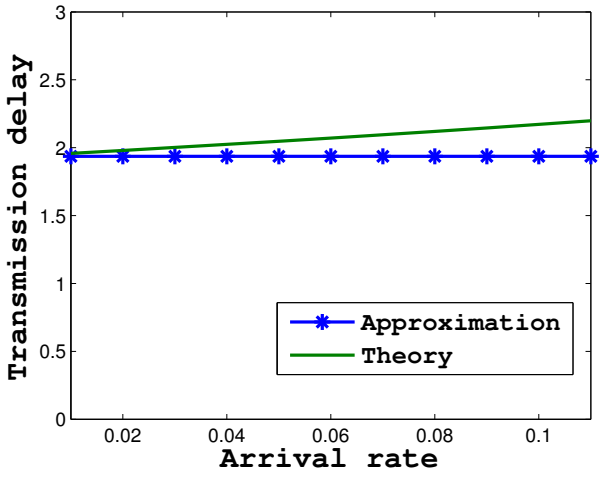

Figure 7: Low utilization delay approximation for $AR = 0.75$.

Next, we consider the high utilization regime and respective approximation (Eq.(52)). We consider utilization values around 0.8. Fig. 9 shows the delay for high values of λ, and an availability ratio of 0.5. We can see there that our approximation is very close to the actual delay and should become exact as ρ gets larger.

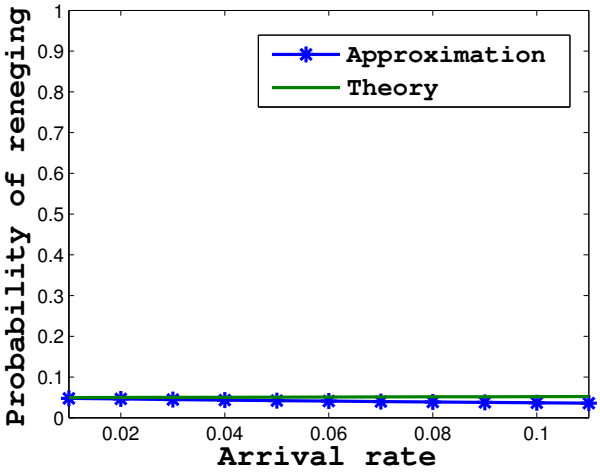

Figure 8: Low utilization p_r approx. for $AR = 0.75$

3.3 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. For this reason, we have also simulated scenarios where the WiFi rate varies uniformly in the range 0.4-1.6 Mbps. Fig. 10 shows the delay for for the vehicular scenario with a deadline of 10 minutes. Even in this case, our theory can give solid predictions for the incurred delay.

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

To conclude our validation, we finally drop the exponential packet assumption as well, and test our theoretical result vs. generic file size results. Fig. 12 compares analytical and simulation results for deterministic (10 s deadline), and Fig. 13 does it for Bounded Pareto distributed files sizes (shape parameter $\alpha = 1.2$ and $c_v = 3$), where the deadline is $T_d = 20$s. Mean file size is in both cases 125KB, the availability ratio is 0.5, and the rest of the parameters correspond to the vehicular scenario. Our theoretical prediction remains reasonably close, despite higher size variability.

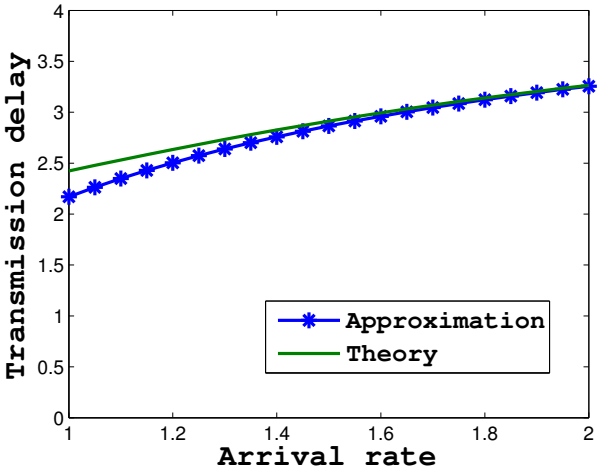

Figure 9: High utilization delay approximation for $AR = 0.5$.

3.4 Delayed offloading gains

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 client and operator. We compare the offloading efficiencies for on-the-spot offloading [9] vs. delayed offloading for different deadline times ($T_{d1} = 2min, T_{d2} = 1min$). Fig.14 illustrates the offloading efficiency vs. availability ratio for a moderate arrival rate of $\lambda = 0.2$. For comparison purposes we also depict the line $x = y$ (offloading efficiency = availability ratio). First, as expected, we can observe that offloading efficiency increases with AR. However, this increase is not linear. More interestingly, the actual offloading efficiencies are always higher than the respective availability ratios. As expected, the delayed offloading provides higher offloading efficiencies compared to on-the-spot offloading, with higher deadlines leading to higher offloading efficiencies.

4 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 the user can tolerate.

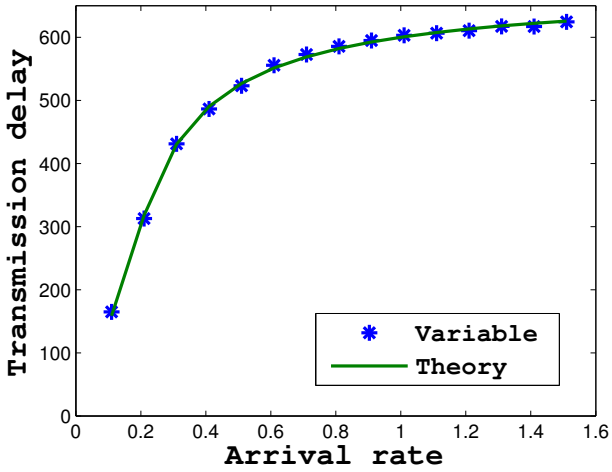

Figure 10: Variable WiFi rates with the same average as theory.

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 Section 2.2 and 2.3. 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.

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 through the cellular interface. Assume further that D_c and D_w denote the costs 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. It is normally the case that $c_c < c_w$ as well as $E_c \approx E_w$ [21].

It follows then that the total monetary and energy costs, D, and E, could be approximated by

$$D = (D_c - D_w)p_r + D_w \quad \text{and} \quad E = \left(\frac{E_c}{c_c} - \frac{E_w}{c_w}\right) p_r + \frac{E_w}{c_w}. \tag{54}$$

[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.

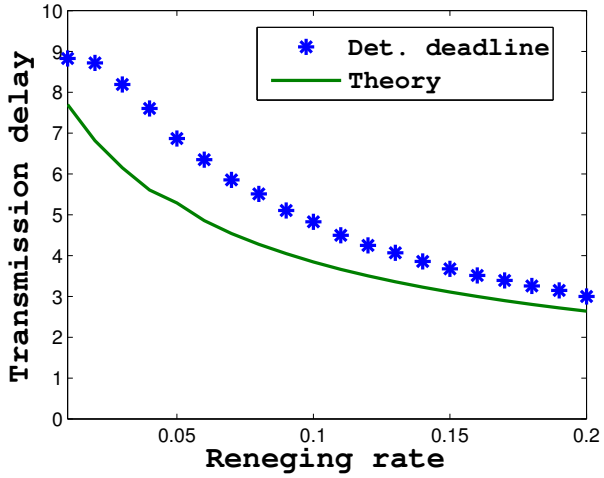

Figure 11: The delay for deterministic deadlines vs. theory.

4.1 Optimization problems

Eq.(54) 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 2, 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

$$\begin{aligned} \min_{\xi} \quad & E[T] + p_r \Delta \\ \text{s. t.} \quad & p_r \leq P_r^{max}, \end{aligned} \tag{55}$$

where $E[T]$ is given by Eq.(46), and p_r by Eq.(48), for low utilization, and Eq.(52) and Eq.(53), for high utilization, respectively. Due to the linearity of Eq.(54), 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 constrain, 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.(55) 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.(48) 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.(55) (delay function from now on). Hence we have

$$f(\xi) = \frac{A_1 \xi + A_2}{B_1 \xi + B_2} + \frac{\theta_1 \xi}{\theta_2 \xi + \theta_3} \Delta, \tag{56}$$

where $A_1 = \gamma$,$A_2 = (\eta + \gamma)^2 + \eta\mu$,$B_1 = (\mu + \eta)(\gamma + \eta)$, $B_2 = \mu\gamma(\gamma + \eta)$. In order to solve the optimization problem given by Eq.(55), we need to know the

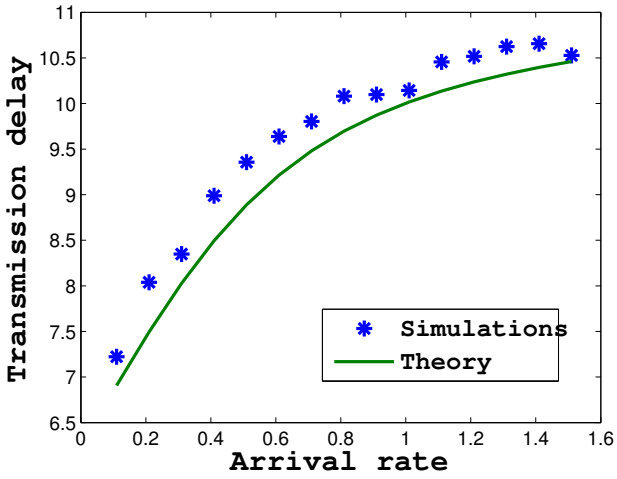

Figure 12: Deterministic packets

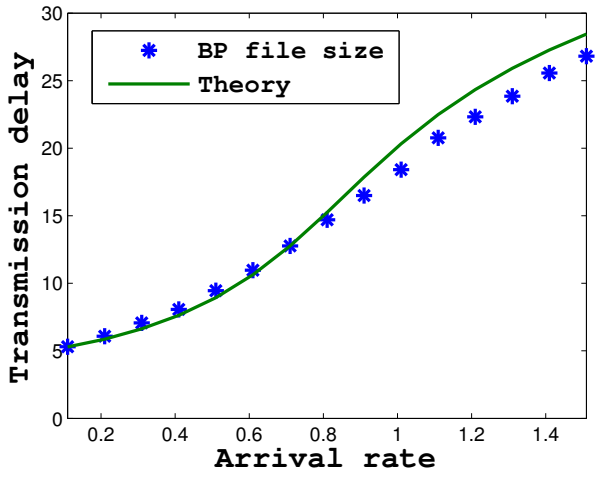

Figure 13: The delay for BP packet sizes vs. theory.

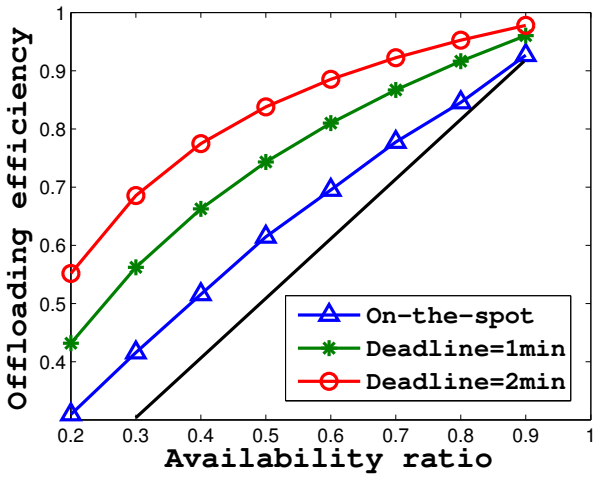

Figure 14: Offloading gains for delayed vs. on-the-spot offloading.

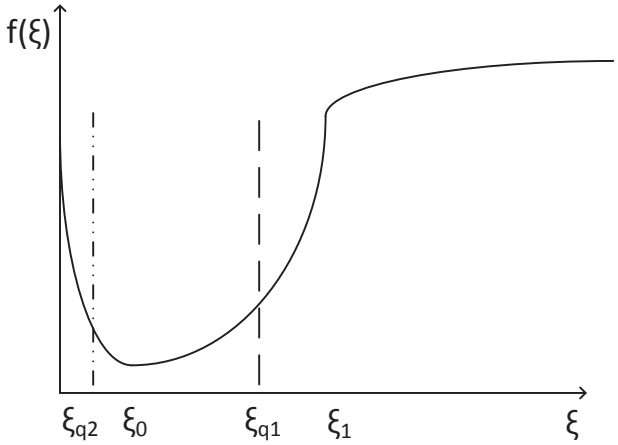

Figure 15: The delay function for the optimization problem.

behavior of the delay function. For that purpose, we analyze the monotonicity and convexity of Eq.(56). To do that we need the first and second derivatives, which are

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

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

It is worth noting that $A_1B_2 < A_2B_1$. This prevents delay function 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

$$\xi \leq \xi_1 = \frac{\theta_3\sqrt[3]{2\frac{A_2B_1 - A_1B_2}{\theta_1\theta_2\theta_3\Delta}} - B_2}{B_1 - \theta_2\sqrt[3]{2\frac{A_2B_1 - A_1B_2}{\theta_1\theta_2\theta_3\Delta}}}. \tag{57}$$

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

Such constrained-optimization problems are often solved with the Lagrangian method and KKT conditions. However, the optimal solution for our problem can be found more easily. The delay function looks like in Fig. 15. 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. 15) will be higher than the global minimum ξ_0. Consequently, the global minimum of Eq.(57) 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. 15) 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

$$T_{d,opt} = \frac{1}{\xi_{opt}} = \frac{1}{\min\left(\xi_0, \frac{\theta_3 P_r^{max}}{\theta_1 - \theta_2 P_r^{max}}\right)}. \tag{58}$$

Similar steps can be followed to solve the same optimization problem for high utilization, as well as other problems.

Optimization problem 2: After minimizing the transmission delay subject to a maximum reneging rate (cost, energy), our next goal now 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

$$\begin{aligned} \min_{\xi} \quad & p_r = \frac{\theta_1 \xi}{\theta_2 \xi + \theta_3} \\ \text{s. t.} \quad & E[T] + p_r \Delta \leq T_{max}. \end{aligned} \tag{59}$$

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

$$T_{d,opt} = \frac{1}{\max\left(0, \frac{K_2 - \sqrt{K_2^2 - 4K_1K_3}}{2K_1}\right)}, \tag{60}$$

where $K_1 = A_1\theta_2 + \theta_1 \Delta B_1 - T_{max} B_1 \theta_2$, $K_2 = T_{max} B_1 \theta_3 + T_{max} B_2 \theta_2 - A_1 \theta_3 - A_2 \theta_2 - \theta_1 \Delta B_2$, $K_3 = A_2 \theta_3 - T_{max} B_2 \theta_3$, $C_1 = \frac{1}{\lambda}\left(1 + \frac{\gamma}{\eta}\right)(\lambda - \mu \pi_w)$.

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.(52) and Eq.(53), respectively.

Optimization problem 3: Having solved the optimization problem for low utilization, we move on to the high utilization regime, and use the approximation for the average file delay to solve two optimization problems. In the first one, our objective function is the transmission delay, and the constrain function is the probability of reneging. So, we have the following problem

$$\begin{aligned} \min_{\xi} \quad & E[T] + p_r \Delta \\ \text{s. t.} \quad & p_r \leq P_r^{max}, \end{aligned} \tag{61}$$

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

$$T_{d,opt} = \frac{1}{\min\left(\sqrt{\frac{C_1}{D_1 \Delta}}, \frac{P_r^{max} - D_2}{D_1}\right)}, \tag{62}$$

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

$$\begin{aligned} \min_{\xi} \quad & p_r \\ \text{s. t.} \quad & E[T] + p_r \Delta \leq T_{max}, \end{aligned} \tag{63}$$

that gives as a solution

$$T_{d,opt} = \frac{2D_1\Delta}{T_{max} - C_2 - D_2\Delta - \sqrt{(T_{max} - C_2 - D_2\Delta)^2 - 4C_1D_1\Delta}}. \quad (64)$$

4.2 A more realistic energy model

It should be mentioned that the energy model proposed in this report is a simple one, and we have clearly indicated that we are doing so to get some initial intuition about the tradeoffs involved. It should be mentioned that extending the model or considering more realistic models is part of the future work. Nevertheless, we can show that our model can be adapted to capture the results of [1]. Our model is exact when we consider a single large burst, i.e., the whole data is given as a single burst of size L (see Fig. 16 below). In that case, it makes sense to consider a static energy consumption over the considered time interval. In [1], authors consider two power saving modes for 802.11 communications. The first one is *Continuously Active Mode* (CAM), in which the user can be either in TRANSMIT/RECEIVE mode or IDLE mode. The second one is the *Power Saving Mode* (PSM), when the user can be in the TRANSMIT/RECEIVE, IDLE, or SLEEP state. The last state starts if the time after the last bit was transmitted/received is larger than a given TIMEOUT time.

For the CAM mode, the authors give the following formula for the consumed energy:

$$E = P_R T_B + P_I T_I, \quad (65)$$

where P_R is the power when downloading data, $P_I < P_R$ is the power in the idle mode, while T_B and T_I are the durations of time the user stays in the receive and idle mode, respectively. The same holds when the user is transmitting (only the power now is P_T). For the CAM mode presented in [1], Fig. 17 depicts the spread of data over time.

In Fig. 17, T_B is the sum of burst durations, and T_I is the sum of the durations of burst intervals. Let's denote with $T = T_B + T_I$ the total duration of the observed interval of time during which the user transmits/receives L data units. We should note that we use the same terminology here as the one presented in [1]. If we need to adapt our model to hold for the scenario illustrated in Fig. 17 we need to proceed as following.

We assume again that the total amount of data to be downloaded/uploaded is L bits. We can again use our model, but now we introduce the notion of "virtual power". We lump the two power levels (P_R and P_I) into a single power level P, and assume that the user consumes energy only during the burst durations, such that the total consumed energy is the same. However, this power is not P_R anymore. The total consumed energy in this case is $\frac{L}{c}P$, where c is the data rate. The energy spent following the model in [1] is $P_R\frac{L}{c} + \left(T - \frac{L}{c}\right)P_I$. From these two expressions, we have

$$\frac{L}{c}P = P_R\frac{L}{c} + \left(T - \frac{L}{c}\right)P_I,$$

that yields to

$$P = P_R + \frac{c}{L}\left(T - \frac{L}{c}\right)P_I = P_R - P_I + \frac{c}{L}TP_I = P_R - P_I + \frac{T}{T_B}P_I. \quad (66)$$

The values for P_I and P_R are given in [1]. For example, for the mobile phone model Nokia N810, these values are $P_I = 0.884$ W and $P_R = 1.181$ W. The values of T_B for different phone models are provided in [1], too.

If we observe a longer time interval, we can approximate the expression $\frac{T_B}{T}$ with the utilization ratio ρ, and Eq.(66) leads to

$$P = P_R - P_I + \frac{1}{\rho}P_I = P_R + P_I \cdot \frac{(1-\rho)}{\rho}. \quad (67)$$

In Eq.(54), following the above discussion, we have $E_w = P_{R,w} + P_{I,w}\frac{1-\rho_w}{\rho_w}$. For the cellular network we would have $E_c = P_{R,c} + P_{I,c}\frac{1-\rho_c}{\rho_c}$. The utilization ratios mentioned here are $\rho_w = \pi_w - \pi_{0,w}$ and $\rho_c = \pi_c - \pi_{0,c}$, and these values are derived in our paper. The values for the cellular interface can be taken in any of the papers bearing with some measurement campaign in cellular networks, e.g., [22].

We have shown that our model presented in Section 4 can be used to capture a more general case for the CAM mode.

For the more involved case of PSM, following a similar approach as for the CAM mode, we can show that our model can be adapted to capture that mode as well. The additional parameters related to the PSM mode are the sleep power P_S, and the duration of the idle period T_{to}, also known as the *timeout time*. Following the same logic as for CAM, the virtual power consumption in the PSM mode can be given as

$$\begin{aligned} E = PT_B = P_R T_B + P\left(T_I \le T_{to}\right) \cdot P_I E\left[T_I | T_I \le T_{to}\right] + \\ P\left(T_I > T_{to}\right)\left(P_I T_{to} + P_S\left(E\left[T_I | T_I > T_{to}\right] - T_{t0}\right)\right). \end{aligned} \quad (68)$$

The term $P\left(T_I \le T_{to}\right) \cdot P_I E\left[T_I | T_I \le T_{to}\right]$ in Eq.(66) corresponds to the average energy consumed during burst intervals smaller than timeout time, i.e., when SLEEP state is not activated. The third term in Eq.(66) corresponds to the average consumed energy in burst intervals longer than TIMEOUT time.

After performing some calculus operations on Eq.(66), we can find that the expressions for power consumption are $E_w = P_{R,w} + P_{S,w}\left(\frac{1}{\rho_w - 1}\right) + \frac{T_{to}}{T_B}(P_{I,w} - P_{S,w})$ for the WiFi interface, and $E_c = P_{R,c} + P_{S,c}\left(\frac{1}{\rho_c - 1}\right) + \frac{T_{to}}{T_B}(P_{I,c} - P_{S,c})$ for the cellular interface. Again, the values of the related parameters can be taken from [1] for the WiFi, and [22] for the cellular network.

Finally, we are going to show simulation-wise that lumping all the possible power levels into a single level (when transmitting/receiving), and considering zero energy consumption when IDLE/SLEEP, as shown above, can provide very accurate results. For that purpose we compare the total (simulated) consumed energy

Figure 16: The burst of data in the model proposed in this report.

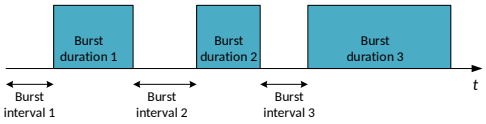

Figure 17: The burst of data appearance in [1].

during an observed time period with the consumed energy as obtained from Eq.(66) for virtual power consumption. The power while being in the RECEIVE state is $P_R = 1.181$ W, and when being idle it is $P_I = 0.884$ W. The burst intervals are uniformly distributed in the range 0-10 s, while the burst size is exponentially distributed with the average values in the range 1-10 Mbits. The WiFi data rate is 2 Mbps. Fig. 18 illustrates the consumed energy during the observed time interval. As can be seen from there, our prediction based on Eq.(66) provides a very high accuracy, and as such we can say that Eq.(66) can be used even in the cases when the transmitted/received data is not just a single burst of data with size L.

The same conclusions can be obtained when considering the PSM mode (with the introduction of SLEEP periods) as well. Fig. 19 illustrates the consumed energy by a mobile phone in the PSM mode. The timeout interval considered here is 3 s. The power in the sleep mode is $P_S = 0.042$ W. The other parameters remain unchanged compared to the previous scenario of Fig. 18.

4.3 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 report. Nevertheless, 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, not just the transmission rate related to the channel and rate adaptation. In fact, we have been working on a prototype (as part of a different work) of such an implementation. An alternative option is to get them from the 3GPP network entity called Access Network Discovery and Selection Function (ANDSF), or even its IEEE alternative solution 802.21 [23].

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

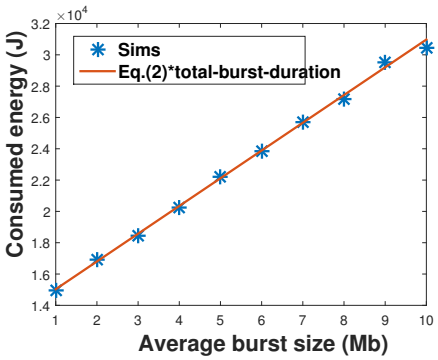

Figure 18: The total consumed energy by a mobile user in the CAM mode.

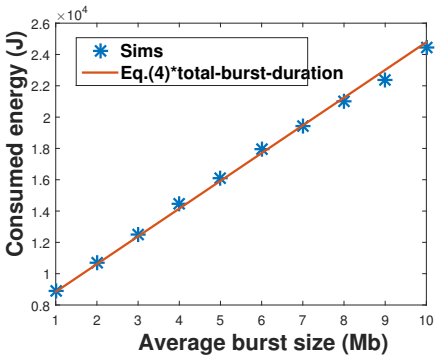

Figure 19: The total consumed energy by a mobile user in the PSM mode.

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 are ready to respond to such queries.

For the remaining parameters, lets consider e.g., Optimization problem 1 (Eq.(58)). For the average file delay $E[T]$ (for any utilization region) of Eq.(1), the user needs to know also: the arrival rate (λ), the availability ratio (AR), that depends on $E[T_{ON}]$ and $E[T_{OFF}]$, and $\pi_{0,w}$. The file arrival (generation) rate can be easily estimated by the user, similarly to the data rates, by keeping running estimates over longer periods for the time between arrivals. For the availability ratio, 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 [24], 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 AP 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 (which would be the ones used for offloading). $\pi_{0,w}$ could also be computed by the UE. Finally, the user knows the maximum amount she is willing to pay per data unit (P_r^{max}).

Observing the solution to the optimization problem 1 (Eq.(58)), the optimal deadline depends, among others, on the parameters $\xi_0, \theta_1, \theta_2, \theta_3$. However, we can observe in the report that these 4 variables depend on the parameters discussed earlier that we already explained how the UE can infer.

As a final note, although not the original intention of this work, the algorithm could also be adapted to a BS-based implementation, where the operator makes the offloading decisions, e.g., an operator owing BSs and APs. In that case, the BS collects the necessary statistics, has knowledge of the WiFi and cellular rates, and can implement the policy. What is more, such a centralized implementation avoids some potential coordination/convergence issues when multiple UEs implement this policy aggressively.

Namely, if the policy is implemented in a naive way, suboptimal behavior and convergence issues could be observed under some scenarios. E.g., if there are a number of UEs who see a high rate, underloaded WiFi AP, they might all switch most of their traffic to that AP. This in turn leads to this AP getting congested and offering very low rates to each individual UE, which in turn will be detected through a running average estimate of the rate, in which case all UEs switch most of their traffic back to the cellular interface. And so on. However, this is an extreme scenario, which could be avoided by a more careful implementation of our policy. For example, the UE could avoid changing abruptly the deadline time (and the amount of traffic offloaded) to the new optimum in case the network conditions have changed considerably. Instead, she should change the deadline gradually by increments. E.g., the new deadline is also implemented as a low-pass filter (some type of running average) to force the deadline to move to the right direction, when network conditions change, but avoid rapid oscillations. This approach avoids the potential convergence issues.

If the policy is implemented at the BS side, it provides the operator with a centralized control of all UEs. Specifically, the operator could choose the deadlines accordingly for all UEs covered, in order to solve a network-wide optimization problem, still subject to delay constraints or offloaded data constraints per terminal, using our paper, but now aiming at some globally optimal objective. The operator would then communicate, through the network entity known as ANDSF, the optimal deadlines to the users. The UE cannot implement aggressively our policy. This way, the convergence issue is overcome by BS side implementation.

4.4 Optimization evaluation

We will now validate the solutions of the previous optimization problem for two different cases. In both of them the arrival rate is 0.1, 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 5 units. The choice of these values is simply for better visualizing the results; different values yield similar conclusions. Fig. 20 shows the delay vs. cost curve for cellular rate being $2\times$ lower than WiFi rate. First thing that we can observe is that the minimum delay is achieved for

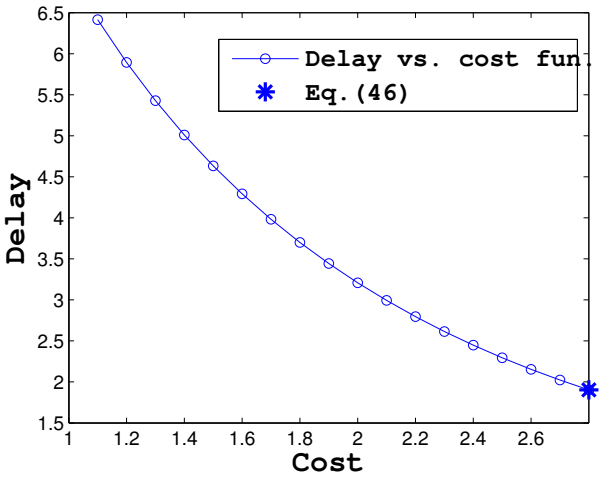

Figure 20: The delay vs. cost curve for high cellular rate.

the highest possible cost (2.8). The optimal average deadline is $T_d = 1$s. This is in agreement with the optimal value predicted from Eq.(58), and shown with an asterisk in Fig. 20. We replace Eq.(48) into Eq.(54) to get the relationship between the cost and the renege rate. We have shown in Eq.(54) 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.(55), 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.

Fig. 21 corresponds to a scenario with the same parameters as in Fig. 20, except that now the cellular rate is much lower ($10\times$). 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, it is not the best option to leave as soon as possible from the WiFi queue, i.e. choose the smallest possible deadline. The optimum delay is achieved for $T_d = 5$s. This corresponds to an average cost of $D = 2.1$ which is also 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).

Further, 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 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

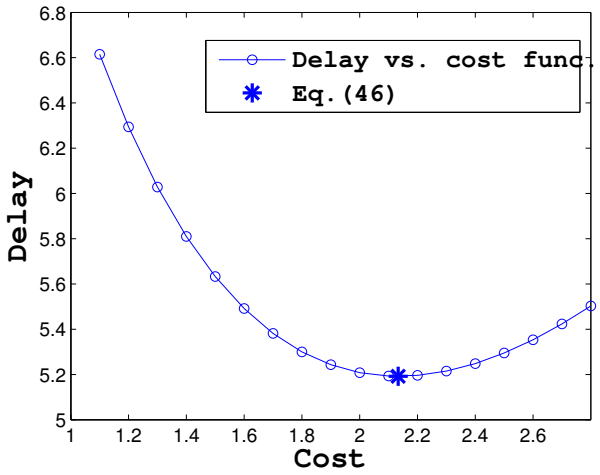

Figure 21: The delay vs. cost curve for low cellular rate.

the same as before. The arrival rate for low utilization scenarios is 0.1, while for the high ones, 1.5. In Table 3, we show the optimal deadlines by using our model (e.g. Eq.(58)), 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 3, the error in determining the optimal deadline decreases for higher arrival rates. The error is in the range 10%-20%. This is reasonable, since the simulated scenarios are with deterministic deadlines and in our theory we use 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)).

Table 3: 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%

5 Related Work

Some recent influential work in offloading relates to measurements of WiFi availability [7, 8]. Authors in [7] 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 [8], 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 [8] reporting values in the range from 20%-33% for different deadlines, and [7] reporting that offloading does not exceed 3%. We believe this is due to the different deadlines assumed together with the different availabilities.

The authors in [25] 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 [26] considers the traffic flow characteristics when deciding when to offload some data to the WiFi. However, there is no delay-related performance analysis. Modeling the cost factors is the focus of [27], which also shows where the offloading APs should be installed. A WiFi offloading system that takes into account a user's throughput-delay tradeoffs and cellular budget constraints is proposed in [28]. However, only heuristic algorithms are proposed, and queueing effects are ignored. Finally, in [24], an integrated architecture has been proposed based on opportunistic networking to switch the data traffic from the cellular to WiFi networks. Summarizing, in contrast to our work, these papers either perform no analysis or use simple models that ignore key system effects such as queueing.

To our best knowledge, the closest work in spirit to ours is [29]. The results in [29] are the extension of the results in [7] 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 [9], 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 [9], it is in fact considerably more difficult to solve.

6 Conclusion

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 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 energy costs. In future work, we intend to consider more complex models for both the WiFi and cellular queues, as well as per-flow scheduling and dispatch policies.

References

[1] 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. of ACM e-Energy*, 2010.

[2] "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.

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

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

[5] 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.

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

[7] 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.

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

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

[10] 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.

[11] 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 Communications Magazine*, vol. 50, no. 11, 2012.

[12] S. Ha, S. Sen, C. Joe-Wong, Y. Im, and M. Chiang, "Tube: time-dependent pricing for mobile data," in *Proc. of ACM SIGCOMM*, 2012.

[13] C. Joe-Wong, S. Ha, and J. Bawa, "When the price is right: Enabling time-dependent pricing for mobile data," *ACM SIGCHI 2013*.

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

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

[16] D.Y.Barrer, "Queuing with impatient customers and ordered service," *Operations Research*, 1957.

[17] R.E.Stanford, "Reneging phenomena in single channel queues," *Mathematics of Operations Research*, 1979.

[18] U. Yechiali and P. Naor, "Queueing problems with heterogeneous arrivals and service," *Operations Research*, 1971.

[19] N. Perel and U. Yechiali, "Queues with slow servers and impatient customers," *Europearn Journal of Operations Research*, no. 201, 2010.

[20] E. Altman and U. Yechiali, "Analysis of customers' impatience in queues with server vacations," *Queueing Systems*, vol. 52, 2006.

[21] 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. of ACM CoNEXT*, 2009.

[22] N. Ding, D. Wagner, X. Chen, A. Pathak, Y. C. Hu, and A. Rice, "Characterizing and modeling the impact of wireless signal strength on smartphone battery drain," in *Proc. of ACM SIGMETRICS*, 2013.

[23] K. Taniuchi, Y. Ohba, V. Fajardo, S. Das, M. Tauil, Y. Cheng, A. Dutta, D. Baker, M. Yajnik, and D. Famolari, "IEEE 802.21: Media independent handover: Features, applicability, realization," *IEEE Communic. Mag.*, vol. 47, no. 1, 2009.

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

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

[26] 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.

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

[28] Y. Im, C. J. Wong, S. Ha, S. Sen, T. Kwon, and M. Chiang, "AMUSE: Empowering users for cost-aware offloading with throughput-delay tradeoffs," in *Proc. of IEEE Infocom Mini-conference*, 2013.

[29] K. Lee, J. Lee, Y. Yi, I. Rhee, and S. Chong, "Mobile data offloading: How much can WiFi deliver," *IEEE/ACM Trans. Netw.*, 2013.