The Effect of Mobility on Delayed Data Offloading

1st Xiaoyi Zhou
School of Electronic Information and Electrical Engineering
Shanghai Jiao Tong University
Shanghai, China
zhouxiaoyi@sjtu.edu.cn

2nd Tong Ye
School of Electronic Information and Electrical Engineering
Shanghai Jiao Tong University
Shanghai, China
yetong@sjtu.edu.cn

3rd Tony T. Lee
School of Science and Engineering
The Chinese University of Hong Kong, Shenzhen
Guangdong, China
tonylee@cuhk.edu.cn

***Abstract*—Delayed offloading is a widely accepted solution for mobile users to offload their traffic through Wi-Fi when they are moving in urban areas. However, delayed offloading enhances offloading efficiency at the expense of delay performance. Previous works mainly focus on the improvement of offloading efficiency while keeping delay performance in an acceptable region. In this paper, we study the impact of the user mobility on delayed data offloading in respect to the tradeoff between offloading efficiency and delay performance. We model a mobile terminal with delayed data offloading as an *M/MMSP/1* queuing system with three service states. To be practical, we consider the feature of currently commercial mobile terminals in our analysis. Our analytical result shows that the mobility of the users can reduce the queueing delay incurred by the delayed offloading, and suggests that delayed offloading strategies can be optimized according to the mobility of the terminals once the delay requirement is given.**

***Index Terms*—Mobile data offloading, cellular network, Wi-Fi, mobility, *M/MMSP/1* queue**

I. Introduction

In recent years, the urban areas have witnessed the surge in mobile data traffic. A Ciscos survey shows that the global mobile traffic has grown 17-fold over the past 5 years [1]. The explosion of mobile traffic has led to the cellular network overloaded problem that causes the ruins of users satisfaction [2]. Even though 5G will bring new spectrum to market, but the proliferating of devices and connections will demand additional bandwidth [3]. A widely accepted solution is to offload part of the mobile traffic through Wi-Fi interface with cheaper bandwidth [4]. This solution currently becomes more and more attractive to both mobile network operators and mobile users.

Nowadays, cellular network coverage is nearly ubiquitous in urban areas. To provide offloading service, many public places, such as residential areas, commercial districts and transportation hubs, in urban areas are installed with Wi-Fi hotspots [5]. When the mobile users are moving in these places, they pass through Wi-Fi network coverages and cellular network coverages alternately. After losing Wi-Fi signal, if users can tolerate certain delay to wait for the next Wi-Fi access point, they will be able to offload more traffic through Wi-Fi, such that they can keep their communication costs as low as possible.

Based on such idea, several kinds of delayed offloading strategies have been proposed in [6]–[11]. The goal is to promote offloading efficiency, while keeping delay performance in an acceptable region. Herein, the offloading efficiency is defined as the ratio of the data offloaded via Wi-Fi to the total transmitted data. In [6], service requests enter a Wi-Fi buffer when the buffer is not full; otherwise, the requests will be transmitted through the cellular network. In [7] and [8], each file, e.g., emails or pictures, is assigned with a timer when it enters the Wi-Fi buffer. If this file is still waiting in the Wi-Fi queue when the timer reaches a preset deadline, it will be sent via the cellular network. Ref. [9] proposed to decide whether a newly arrived packet is directly transmitted via the cellular network, according to the Wi-Fi buffer length and the network connection. Clearly, these strategies implies that the mobile terminals are able to send the traffic through cellular network and Wi-Fi at the same time. However, such kind of concurrent transmission [12] is not supported by most of currently commercial mobile terminals [13].

In this paper, we study the data offloading problem of currently commercial mobile terminals, which can use only one kind of wireless channel to transmit traffic at the same time. Our goal is to find out if there is any factor that may affect the tradeoff between the offloading efficiency and the delay performance. We analyze the delayed data offloading by using an *M/MMSP/1* queuing model with three service states. Our analytical results show that: though the offloading efficiency is enhanced at the expense of queuing delay, the moving speed of mobile users in motion is helpful to reduce the queuing delay incurred by the delayed offloading. This indicates that the deadline of delayed offloading strategies can be optimized according to the mobility of the terminals once the delay requirement is given.

The rest of this paper is organized as follows. In section II, we describe the delayed offloading strategy in detail, and show that the offloading procedure is essentially a three-state *M/MMSP/1* queueing system. In section III, we establish a

978-1-7281-1856-7/19/$31.00 ©2019 IEEE

hybrid embedded Markov chain to derive the mean delay and the offloading efficiency. In section IV, we show that the moving speed of mobile users in motion can reduce the queuing delay incurred by the delayed offloading. Section V concludes this paper.

II. Delayed Offloading of Terminals Without Concurrent Transmission

When people are moving in the urban area, they pass through Wi-Fi coverages and cellular network coverages alternately. For a mobile terminal without concurrent transmission capability, it perceives the wireless channel switching between two states over time, as illustrated in Fig. 1, where C denotes the state that there is only cellular signal while F represents the state that the Wi-Fi signal is available.

Assume that the duration times of wireless channel states F and C are exponentially distributed with parameters f_F and f_C, respectively. The wireless channel perceived by the mobile terminal in motion is a kind of Markov channel [14]. Clearly, given the deployment of Wi-Fi hotspots, the faster the user moves, the larger f_F and f_C are. Thus, we use $f = \frac{1}{1/f_F + 1/f_C}$ to delineate the mobility of the user.

A. Delayed Offloading Procedure

In the face of such wireless environment, as described by the Markov channel in Fig. 1, we consider the delayed offloading procedure for currently commercial terminals as follows. When the Wi-Fi signal is available, the terminal transmits the traffic through Wi-Fi. Once the Wi-Fi connection is lost, the terminal pauses the traffic transmission to wait for the next Wi-Fi hotspot and randomly selects a deadline at the same time. If a Wi-Fi signal is available before the deadline expires, the terminal will recover the transmission through Wi-Fi; otherwise, it will go on with the transmission via the cellular network.

Hence, during the whole delayed offloading procedure, the terminal has three transmission (or service) states: (1) delayed state (or state 0), transmission is delayed, (2) cellular state (or state 1), transmission via the cellular network, and (3) Wi-Fi state (or state 2), transmission via Wi-Fi. The transitions among these service states are plotted in Fig. 2.

B. Three-state Markov Modulated Service Process

Suppose that the deadline set for the delayed state is an exponential random variable with parameter f_D. The data transmission process of mobile terminals can be considered as a three-state Markov modulated service process (MMSP) [15]. Let $f_{i,j}$ be the transition rate from state i to state j in Fig. 2, where i,$j = 0, 1, 2$. According to Fig. 1 and 2, $f_{i,j}$ is given by

$$f_{2,0} = f_F \tag{1a}$$

$$f_{0,1} = f_D \tag{1b}$$

$$f_{0,2} = f_{1,2} = f_C. \tag{1c}$$

It follows that the steady-state probabilities of each service state in Fig. 2 are given by

$$\pi_0 = (1-R)\,\frac{\tau f}{\tau f + 1 - R} \tag{2a}$$

$$\pi_1 = (1-R)\,\frac{1-R}{\tau f + 1 - R} \tag{2b}$$

$$\pi_2 = R, \tag{2c}$$

where $\tau = 1/f_D$ is the expectation of the deadline and $R = \frac{1/f_F}{1/f_C + 1/f_F} = \frac{f_C}{f_C + f_F}$ is referred to as Wi-Fi available ratio in this paper since it actually indicates the ratio of the time that the terminal can perceive Wi-Fi signals.

Let μ_j be the transmission rate of state j in Fig. 2. Clearly, the transmission rate of the delayed state is $\mu_0 = 0$. It follows that the average transmission rate that a mobile terminal with the delayed offloading strategy can provide is given by:

$$\hat{\mu} = \pi_1 \mu_1 + \pi_2 \mu_2 = (1-R)\,\frac{1-R}{\tau f + 1 - R}\mu_1 + R\mu_2. \tag{3}$$

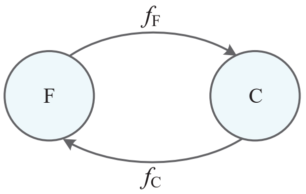


Fig. 1. Transition of wireless environment state.

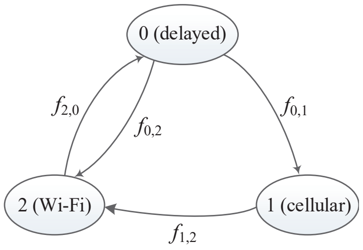


Fig. 2. State transition of the data transmission.

III. Analysis of Mean Delay and Offloading Efficiency

In this section, we analyze the performance of the data offloading of the terminals without concurrent transmission capability. Suppose the input traffic is a Poisson process with rate λ. The data transmission of the offloading procedure can be delineated as an *M/MMSP/1* queueing system with three service states. The difficulty of the analysis of the *M/MMSP/1* queue lies in the fact that the service time of a file is related to the service state when its service starts. To cope with this problem, we use the hybrid embedded Markov chain developed in [16].

A. Embedded Points

Two types of time points are embedded into the data offloading process. We consider the epoch when a file starts its service, since the service time of files depends on the service state at this epoch. We also observe the epoch at the transition of service states, since the dependency of the service time is essentially caused by service state transitions during the service of a file. We thus define the two types of embedded points as follows:

1. State-transition point Φ_j: epoch when the service state transits to state j;
2. Start-service point S_j: epoch when a file starts its service while the service state is j.

where $j = 0, 1, 2$. Clearly, the time interval between two adjacent embedded points is exponentially distributed.

Suppose the current epoch is an embedded point of which the service state is the delayed state $j = 0$, as Fig. 3(a) shows. Since the service is suspended at current epoch, the next event may be a state transition from service state 0 to service state i after time I_i which is an exponential random variable with parameter $f_{0,i}$, where $i = 1, 2$. Thus, the type of the next embedded point is determined by which kind of service state transition happens first. It follows that the distribution of the interval $I = \min_i I_i$ from current point to the next point is exponentially distributed with parameter $\sum_{i=1}^{2} f_{0,i}$ and the next embedded point is Φ_i with probability $f_{0,i} / \sum_{i=1}^{2} f_{0,i}$, where $i = 1, 2$.

Similarly, when the current epoch is an embedded point of which the transmission state is state $j > 0$, the next embedded point will be $\Phi_{\overline{j}}$ with probability $f_{j,\overline{j}} / \left(f_{j,\overline{j}} + \mu_j\right)$ or S_j with probability $\mu_j / \left(f_{j,\overline{j}} + \mu_j\right)$, where $\overline{j} \triangleq (j + 1)\, mod\, 3$, as shown in Fig. 3(b) and (c). Also, the distribution of the interval from current point to the next point is exponentially distributed with parameter $f_{j,\overline{j}} + \mu_j$.

B. Start Service Probability

The start service probability $\hat{\pi}_j$ is defined as the probability that a data file starts its service in state j. Consider a newly arrived file, which sees n files in the buffer. These files are labeled according to their sequence in queue. The head-of-line (HOL) file is labeled with 0 and the newly arrived file is labeled with n. We define two types of conditional probabilities corresponding to the embedded points:

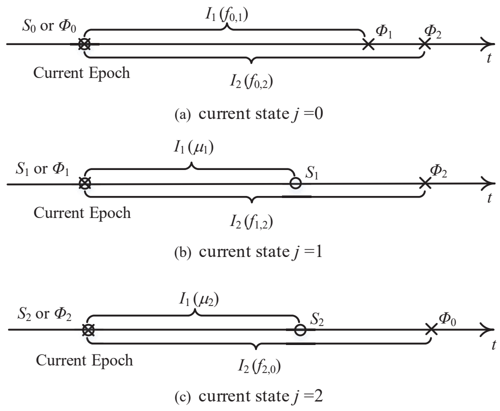
1. $\hat{\pi}_{n,j}(m) = P\{$the m^{th} data file starts its service in service state j | the newly arrived file sees n files in buffer$\}$
2. $\hat{\varphi}_{n,j}(m) = P\{$the service state transits to state j when the m^{th} data file is in service | the newly arrived file sees n? files in buffer$\}$.

$\hat{\pi}_{n,j}(m)$ is defined on the embedded point S_j, at which the $(m-1)^{th}$ file finishes its service when the service state is j. Thus, the last event may be that the $(m-1)^{th}$ file starts its service when the service state is j or that the service state transits to state j when the $(m-1)^{th}$ file is in service. Therefore, for $1 \leq m \leq n$, the equations of $\hat{\pi}_{n,j}(m)$ in each state are obtained:

$$\hat{\pi}_{n,0}(m) = 0 \tag{4a}$$

$$\hat{\pi}_{n,1}(m) = \frac{\mu_1}{\mu_1 + f_{1,2}} \Big(\hat{\pi}_{n,1}(m-1) + \hat{\varphi}_{n,1}(m-1) \Big) \tag{4b}$$

$$\hat{\pi}_{n,2}(m) = \frac{\mu_2}{\mu_2 + f_{2,0}} \Big(\hat{\pi}_{n,2}(m-1) + \hat{\varphi}_{n,2}(m-1) \Big). \tag{4c}$$



(a) current state j =0

(b) current state j =1

(c) current state j =2

Fig. 3. Relationship between two kinds of embedded points.

Similarly, the equations of $\hat{\varphi}_{n,j}(m)$ are given by:

$$\hat{\varphi}_{n,0}(m) = \frac{f_{2,0}}{\mu_2 + f_{2,0}} \Big(\hat{\pi}_{n,2}(m) + \hat{\varphi}_{n,2}(m) \Big) \tag{5a}$$

$$\hat{\varphi}_{n,1}(m) = \frac{f_{0,1}}{f_{0,1} + f_{0,2}} \Big(\hat{\pi}_{n,0}(m) + \hat{\varphi}_{n,0}(m) \Big) \tag{5b}$$

$$\begin{aligned} \hat{\varphi}_{n,2}(m) = & \frac{f_{0,2}}{f_{0,1} + f_{0,2}} \Big(\hat{\pi}_{n,0}(m) + \hat{\varphi}_{n,0}(m) \Big) + \\ & \frac{f_{1,2}}{\mu_1 + f_{1,2}} \Big(\hat{\pi}_{n,1}(m) + \hat{\varphi}_{n,1}(m) \Big). \end{aligned} \tag{5c}$$

Combing (4) and (5), we have the relations between $\hat{\pi}_{n,j}(m)$ and $\hat{\pi}_{n,j}(m-1)$:

$$\begin{pmatrix} \hat{\pi}_{n,0}(m) \\ \hat{\pi}_{n,1}(m) \\ \hat{\pi}_{n,2}(m) \end{pmatrix} = \hat{Q} \begin{pmatrix} \hat{\pi}_{n,0}(m-1) \\ \hat{\pi}_{n,1}(m-1) \\ \hat{\pi}_{n,2}(m-1) \end{pmatrix} = \hat{Q}^m \begin{pmatrix} \hat{\pi}_{n,0}(0) \\ \hat{\pi}_{n,1}(0) \\ \hat{\pi}_{n,2}(0) \end{pmatrix}, \tag{6}$$

where the coefficient matrix $\hat{Q}$ is

$$\hat{Q} = \begin{pmatrix} 0 & 0 & 0 \\ \beta \frac{f_{0,1}}{f_{0,1}+f_{0,2}} \Big(1+\frac{f_{2,0}}{\mu_2}\Big) & \beta \Big(1+\frac{f_{0,1}}{f_{0,1}+f_{0,2}} \frac{f_{2,0}}{\mu_2}\Big) & \beta \frac{f_{0,1}}{f_{0,1}+f_{0,2}} \frac{f_{2,0}}{\mu_2} \\ \beta \Big(\frac{f_{1,2}}{\mu_1} + \frac{f_{0,2}}{f_{0,1}+f_{0,2}}\Big) & \beta \frac{f_{1,2}}{\mu_1} & \beta \Big(1+\frac{f_{1,2}}{\mu_1}\Big) \end{pmatrix}, \tag{7}$$

and

$$\beta = \frac{R(1-R)(1-R+\tau f)\mu_1\mu_2}{(1-R)^2 f\mu_1 + R(1-R+\tau f) f\mu_2 + R(1-R)(1-R+\tau f)\mu_1\mu_2}.$$

We solve (6) and obtain

$$\hat{\pi}_{n,0}(m) = 0 \tag{8a}$$

$$\begin{aligned} \hat{\pi}_{n,1}(m) = & \theta_1 + \Big(\theta_2 - \frac{f_{0,2}}{f_{0,1} + f_{0,2}}\Big) \beta^m \hat{\pi}_{n,0}(0) + \\ & \theta_2 \beta^m \hat{\pi}_{n,1}(0) - \theta_1 \beta^m \hat{\pi}_{n,2}(0) \end{aligned} \tag{8b}$$

$$\begin{aligned} \hat{\pi}_{n,2}(m) = & \theta_2 - \Big(\theta_2 - \frac{f_{0,2}}{f_{0,1} + f_{0,2}}\Big) \beta^m \hat{\pi}_{n,0}(0) - \\ & \theta_2 \beta^m \hat{\pi}_{n,1}(0) + \theta_1 \beta^m \hat{\pi}_{n,2}(0), \end{aligned} \tag{8c}$$

where $1 \leq m \leq n$ and $\theta_j = \frac{\pi_j \mu_j}{\sum_{j=0}^{2} \pi_j \mu_j}$. When $m = 0$, $\hat{\pi}_{n,j}(0)$ is the probability that the HOL file starts its service when the service state is j, given that the newly arrived file sees n

files in the buffer. Thus, $\hat{\pi}_{n,j}(0) = p_{n,j}/p_n$, where $p_{n,j}$ is the stationary probability that there are n files in the buffer and the service state is j, and $p_n = \sum_{j=0}^{2} p_{n,j}$ is the stationary probability that there are n files in the buffer.

By definition, a newly arrived file that sees n files in buffer upon its arrival starts its service in state j is $\hat{\pi}_{n,j}(n)$. Thus, the start service probability $\hat{\pi}_j$ is

$$\hat{\pi}_j = \sum_{n=0}^{\infty} p_n \hat{\pi}_{n,j}(n). \tag{9}$$

Combining (8) and (9), we have:

$$\hat{\pi}_0 = p_{0,0} \tag{10a}$$

$$\begin{aligned}\hat{\pi}_1 = \theta_1 + &\Big(\theta_2 - \frac{\tau f}{\tau f + 1 - R}\Big) G_0(\beta) + \\ &\theta_2 G_1(\beta) - \theta_1 G_2(\beta) - \frac{1-R}{\tau f + 1 - R} p_{0,0}\end{aligned} \tag{10b}$$

$$\begin{aligned}\hat{\pi}_2 = \theta_2 - &\Big(\theta_2 - \frac{\tau f}{\tau f + 1 - R}\Big) G_0(\beta) - \\ &\theta_2 G_1(\beta) + \theta_1 G_2(\beta) - \frac{1-R}{\tau f + 1 - R} p_{0,0},\end{aligned} \tag{10c}$$

where $G_j(z) = \sum_{n=0}^{\infty} p_{n,j} z^n$. The numerical solutions of $G_j(z)$ and $p_{0,0}$ can be derived by establishing a two-dimensional continuous time Markov chain, which is not given in detail due to space limitation.

C. Mean Service Time

Let T_j be the time needed to serve a file if the file starts its service in state j. Consider a file that the system is empty and in the delayed state when it arrives at the system. We say this file start its service in delayed state $j = 0$. The service state changes to the state i in the next embedded point with probability $f_{0,i}/\sum_{i=1}^{2} f_{0,i}$, where $i = 1, 2$. After that, the time this file still needed to finish the service is T_i. Considering that the time from current point to the next embedded point is $1/\sum_{i=1}^{2} f_{0,i}$, the expectation of T_0 is given by:

$$\begin{aligned}E[T_0] = &\frac{f_{0,1}}{f_{0,1}+f_{0,2}}\left(\frac{1}{f_{0,1}+f_{0,2}} + E[T_1]\right) + \\ &\frac{f_{0,2}}{f_{0,1}+f_{0,2}}\left(\frac{1}{f_{0,1}+f_{0,2}} + E[T_2]\right).\end{aligned} \tag{11a}$$

Similarly, we obtain $E[T_1]$ in (11b) and $E[T_2]$ in (11c).

$$E[T_1] = \frac{\mu_1}{\mu_1+f_{1,2}}\frac{1}{\mu_1+f_{1,2}} + \frac{f_{1,2}}{\mu_1+f_{1,2}}\left(\frac{1}{\mu_1+f_{1,2}} + E[T_2]\right) \tag{11b}$$

$$E[T_2] = \frac{\mu_2}{\mu_2+f_{2,0}}\frac{1}{\mu_2+f_{2,0}} + \frac{f_{2,0}}{\mu_2+f_{2,0}}\left(\frac{1}{\mu_2+f_{2,0}} + E[T_0]\right). \tag{11c}$$

Solving (11), we can derive $E[T_j]$ in (12a)-(12c). And thus the mean service time:

$$E[T] = \sum_{j=0}^{2} \hat{\pi}_j E[T_j]. \tag{13}$$

D. Mean Waiting Time and Mean Delay

The waiting time of a file is the duration that from the time it arrives at the system to the time it becomes the HOL file. It also equals to the sum of the residual service time of the HOL file and the service time of all the data files before this file. For the k^{th} file in queue ($0 \le k \le n$), we define two types of conditional elapse time:

1. $W_{n,k}(k)$: the expected time from the epoch when a newly arrived file becomes the k^{th} file in queue while the service state is j to the epoch when it becomes the HOL file, given that it sees n files in the buffer when it arrives;
2. $V_{n,j}(k)$: the expected time from the epoch when the service state transits to state j while the newly arrived file is now the k^{th} file in queue to the epoch when it becomes the HOL file, given that it sees n files in the buffer when it arrives.

Following the similar arguments used to derive $\hat{\pi}_{n,j}(m)$, we have

$$\begin{pmatrix} V_{n,0}(k) \\ W_{n,1}(k) \\ W_{n,2}(k) \end{pmatrix} = \sum_{i=1}^{k-1} \left(\hat{Q}^T\right)^i \begin{pmatrix} E[T_0] \\ E[T_1] \\ E[T_2] \end{pmatrix} + \begin{pmatrix} E[T_0] \\ E[T_1] \\ E[T_2] \end{pmatrix}. \tag{14}$$

$$E[T_0] = \frac{\tau f^2 + (1-R) f + (1-R)\tau f \mu_1 + R(1-R)(\tau f + 1 - R)\mu_2 + R(1-R)^2 \tau \mu_1 \mu_2}{(1-R)^2 f \mu_1 + R(\tau f + 1 - R) f \mu_2 + R(1-R)(\tau f + 1 - R)\mu_1 \mu_2} \tag{12a}$$

$$E[T_1] = \frac{\tau f^2 + (1-R) f + R(1-R)(\tau f + 1 - R)\mu_2}{(1-R)^2 f \mu_1 + R(\tau f + 1 - R) f \mu_2 + R(1-R)(\tau f + 1 - R)\mu_1 \mu_2} \tag{12b}$$

$$E[T_2] = \frac{\tau f^2 + (1-R) f + R(1-R)^2 \mu_1 + (1-R)\tau f \mu_1}{(1-R)^2 f \mu_1 + R(\tau f + 1 - R) f \mu_2 + R(1-R)(\tau f + 1 - R)\mu_1 \mu_2} \tag{12c}$$

Let W be the mean waiting time. Solving (14) and using the relation $W = \sum_{j=0}^{2} \sum_{n=0}^{\infty} (n)\, p_{n,j} W_{n,j}$, we obtain:

$$W = \frac{1}{1-\frac{\lambda}{\hat{\mu}}} \left[\frac{\lambda}{\hat{\mu}} E[T] + \frac{1}{1-\beta} \sum_{j=0}^{2} E[T] (\pi_j - \hat{\pi}_j) - \frac{\beta}{1-\beta} \frac{(1-R)\tau}{1-R+\tau f} (\pi_0 - \hat{\pi}_0) \right]. \tag{15}$$

Also, we have the mean delay as follows:

$$D = W + E[T]. \tag{16}$$

E. Offloading Efficiency

Recall that the offloading efficiency, denoted by η, is defined as the ratio of the traffic transmitted via Wi-Fi to the total traffic. We consider a very long time period $[0, T]$. In this period, the total input traffic is λT. On the other hand, $\pi_2 - p_{0,2}$ is the probability that server is transmitting traffic while the service state is $j = 2$, and thus the part of traffic served by Wi-Fi is $(\pi_2 - p_{0,2})\, \mu_2 T$. It follows that

$$\eta = \frac{\mu_2}{\lambda} (\pi_2 - p_{0,2}). \tag{17}$$

IV. Performance Evaluations

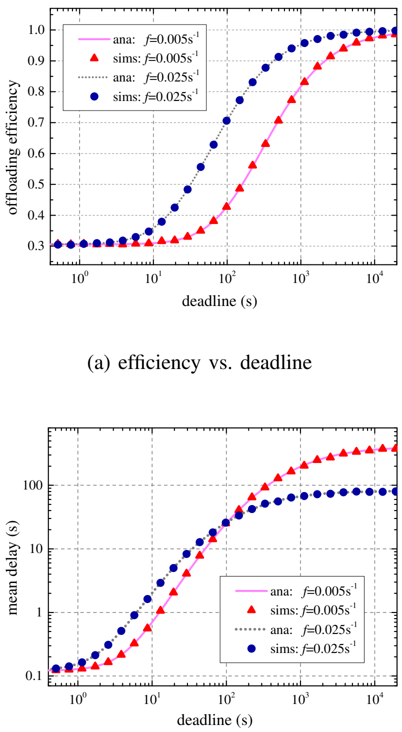
In reality, the deployment of Wi-Fi hotspots in a city is fixed. The factors that may affect the performance of the offloading procedure are the deadline set for the delayed service state and the moving speed of the users. Based on the analytical results in Section III, we study how the deadline and the moving speed of users affect the performances such as mean delay and offloading efficiency. We also conduct simulation experiments in this section of which the settings are the same with that in modeling.

We consider two application scenarios in this section. One is that the terminals are carried by pedestrians, of which the channel transition rate $f_C = 0.007s^{-1}$, $f_F = 0.016s^{-1}$, and thus $f = 0.005s^{-1}$. The other one is that the terminals are carried by vehicles. In this case, the channel transition rate $f_C = 0.035s^{-1}$, $f_F = 0.079s^{-1}$, and thus $f = 0.025s^{-1}$ [8].

A. $\mu_1 = \mu_2 = \mu$

To facilitate discussion, we first consider the case where the transmission rate of the Wi-Fi connection and the cellular network are the same $\mu_1 = \mu_2 = \mu$.

We plot the offloading efficiency η and the mean delay D versus the expectation of the deadline τ in Fig. 4, where $\lambda = 2.5$ file/s and $\mu = 10.75$ files/s. It can be seen from Fig. 4 that both η and D monotonously increases with τ no matter what f is, which means the offloading efficiency is improved at the expense of the mean delay. However, it is very interesting to see that, in the whole range of τ, the increments of η are the same while those of D are different under two application scenarios. For example, when the increment of D is 395s when $f = 0.005s^{-1}$, and that is 81s when $f = 0.025s^{-1}$, though the increments of η in both cases are 0.7. This implies that the terminal with a higher mobility or a larger f tends to pay a smaller cost in the mean delay to obtain the same increment of the offloading efficiency, which is visualized by Fig. 5 where D is plotted as a function of η.



(a) efficiency vs. deadline

(b) delay vs. deadline

Fig. 4. Performance of the delayed offloading when $\mu_1 = \mu_2$.

Based on the analytical results in Section III, we explain this point by considering two extreme cases of τ as follows. When τ is very small, the delayed service state in Fig. 2 disappears and the terminal will transmit the file immediately after it losses the Wi-Fi signal. In this case, though there are Wi-Fi state and cellular state as well as the state transitions, the transmission rate μ keeps unchanged over time, which implies that the service process and thus the queue length is independent of the service state transitions. In other words, the *M/MMSP/1* queue now reduces to an *M/M/1* queue with service rate μ. It follows that $p_{0,2} = \pi_2 p_0 = \pi_2 (1 - \lambda/\mu)$, and thus η in (17) is now equal to π_2, no matter what f is.

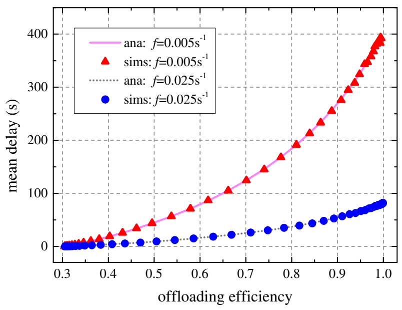
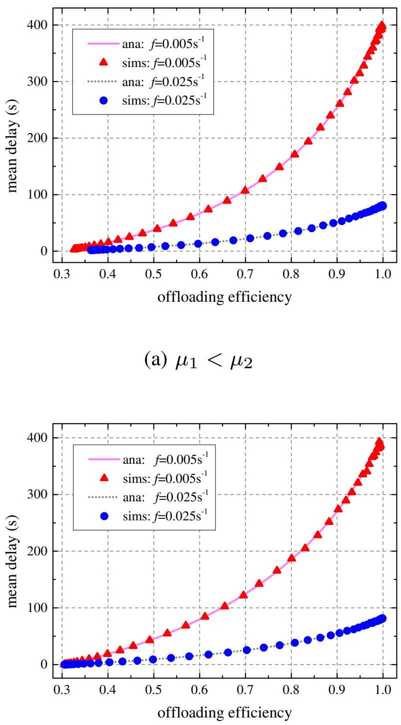


Fig. 5. Delay vs. efficiency when $\mu_1 = \mu_2$.



(a) $\mu_1 < \mu_2$

(b) $\mu_1 > \mu_2$

Fig. 6. Delay vs. efficiency when (a) $\mu_1 < \mu_2$ (b) $\mu_1 > \mu_2$.

Also, the mean delay in (16) changes to $D = 1/\left(\mu - \lambda\right)$ for all fs. On the other hand, when τ is extremely large, the cellular state in Fig. 2 disappears, and $\pi_1 = \hat{\pi}_1 = 0$ accordingly. In this case, the offloading process is actually an *M/MMSP/1* with two service states, Wi-Fi state and delayed state, and the terminal only transmits traffic via Wi-Fi. Thus, η in (17) increases to 1, for all fs. Moreover, it is easy to show that D in (16) goes to

$$D^* = \frac{1}{R\mu - \lambda}\left[1 + \frac{R\left(1-R\right)^2 \mu}{f}\right], \tag{18}$$

if τ approaches to infinity. It is obvious that D^* decreases with the mobility f.

B. $\mu_1 \neq \mu_2$

We also study the relationship between the mean delay and the efficiency when $\mu_1 \neq \mu_2$ in Fig. 6, where all the parameters except μ_1 and μ_2 are all the same with those in Fig. 4. Especially, $\mu_1 = 2.64$ files/s and $\mu_2 = 10.75$ files/s in Fig. 6(a) and $\mu_1 = 13.21$ files/s and $\mu_2 = 10.75$ files/s in Fig. 6(b). The results in Fig. 6 show that the mobility of the terminal is helpful to reduce the expense of mean delay during the increase of offloading efficiency.

We thus conclude from the above discussion that the mobility of mobile users can reduce the delay incurred by the delay offloading. Based on this conclusion, mobile users can optimize the system performance according to their mobility. For example, a user has a delay requirement based on the delay tolerance. Our conclusion shows that the cost of delay to enhance the offloading efficiency is small when the mobility of the user is high. In this case, the user can greatly increase the deadline to improve the offloading efficiency while ensuring the delay requirement.

V. Conclusion

In this paper, we analyze the delayed offloading problem of currently commercial mobile terminals. We develop a three-state *M/MMSP/1* queueing model to derive the offloading efficiency and the mean delay. Through the analysis, we find that the mobility of the users plays an important role in the tradeoff between the offloading efficiency and the mean delay. In particular, the mobility of mobile users can reduces the delay incurred by the delayed offloading.

References

[1] (2019, February) Cisco visual networking index: Global mobile data traffic forecast update, 2017-2022. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-738429.pdf

[2] O. B. Yetim and M. Martonosi, "Dynamic adaptive techniques for learning application delay tolerance for mobile data offloading," in *Proc. IEEE INFOCOM*, 2015, pp. 1885–1893.

[3] (2019, February) The 5g paradox: The need for more offloading options in the next-generation wireless era. [Online]. Available: https://wia.org/wp-content/uploads/WIA_Offload-web.pdf

[4] (2010) White paper - mobile data offloading through wi-fi. [Online]. Available: https://www.sourcesecurity.com/docs/moredocs/proximmicrosite/Mobile-Data-Offloading-Through-WiFi-V1.2.pdf

[5] M. Afanasyev, T. Chen, G. M. Voelker, and A. C. Snoeren, "Analysis of a mixed-use urban wifi network: when metropolitan becomes neapolitan," in *Proceedings of the 8th ACM SIGCOMM conference on Internet measurement*, 2008, pp. 85–98.

[6] N. Cheng, N. Lu, N. Zhang, X. S. Shen, and J. W. Mark, "Opportunistic wifi offloading in vehicular environment: A queueing analysis," in *2014 IEEE Global Communications Conference*, 2014, pp. 211–216.

[7] K. Lee, J. Lee, Y. Yi, I. Rhee, and S. Chong, "Mobile data offloading: How much can wifi deliver?" *IEEE/ACM Transactions on Networking (ToN)*, vol. 21, no. 2, pp. 536–550, 2013.

[8] F. Mehmeti and T. Spyropoulos, "Performance modeling, analysis, and optimization of delayed mobile data offloading for mobile users," *IEEE/ACM Transactions on Networking (TON)*, vol. 25, no. 1, pp. 550–564, 2017.

[9] A. Ajith and T. Venkatesh, "Qoe enhanced mobile data offloading with balking," *IEEE Communications Letters*, vol. 21, no. 5, pp. 1143–1146, 2017.

[10] N. Wang and J. Wu, "Opportunistic wifi offloading in a vehicular environment: Waiting or downloading now?" in *Proc. IEEE INFOCOM*, 2016, pp. 1–9.

[11] 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*, 2012, pp. 5423–5428.

[12] C. Hua, H. Yu, R. Zheng, J. Li, and R. Ni, "Online packet dispatching for delay optimal concurrent transmissions in heterogeneous multi-rat networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 7, pp. 5076–5086, 2016.

[13] C. Zhang, B. Gu, Z. Liu, K. Yamori, and Y. Tanaka, "Cost- and energy-aware multi-flow mobile data offloading using markov decision process," *IEICE Transactions on Communications*, vol. advpub, 2017.

[14] L. Huang and T. T. Lee, "Generalized pollaczek-khinchin formula for markov channels," *IEEE Transactions on Communications*, vol. 61, no. 8, pp. 3530–3540, 2013.

[15] L. Huang and T. T. Lee, "Queueing behavior of hybrid arq wireless system with finite buffer capacity," in *2012 21st Annual Wireless and Optical Communications Conference (WOCC)*, 2012, pp. 32–36.

[16] J. Zhang, Z. Zhou, T. T. Lee, and T. Ye, "Delay analysis of three-state markov channels," in *12th Int. Conf. Queueing Theory Network Applications*, vol. 61, August 2017, pp. 101–117.