An Analytical Framework to Characterize the Efficiency and Delay in a Mobile Data Offloading System

Yoora Kim
Dept. Math., Univ. of Ulsan
93 Daehak-ro, Nam-gu
Ulsan, South Korea
yrkim@ulsan.ac.kr

Kyunghan Lee*
ECE, UNIST
UNIST-gil 50
Ulsan, South Korea
khlee@unist.ac.kr

Ness B. Shroff
ECE and CSE, OSU
2015 Neil Ave.
Columbus, OH 43210
shroff.11@osu.edu

ABSTRACT

Smart mobile devices are generating a tremendous amount of data traffic that is putting stress on even the most advanced cellular networks. Delayed offloading has recently been proposed as an efficient mechanism to substantially alleviate this stress. The idea is simple. It allows a mobile device to delay transmission of data packets for a certain amount of time, while it searches WiFi (or similarly femtocell) networks to offload the data during the time. When the time expires, it completes the remaining portion of the delayed transmission through the cellular network that is available at the moment. In this paper, we develop an analytical framework using an embedded Markov process for the delayed offloading system. We provide a closed-form expression for estimating how much data generated by the users can be offloaded to WiFi networks from cellular networks even when there are non-Markovian data arrivals and service interruptions. We conduct extensive numerical studies with various ranges of delay, service interruption time, arrived data, and service rate. These numerical studies show that the current deployment of WiFi networks measured from a metropolitan city is capable of offloading about 80% of the generated data with 30 minutes of delay and 1 Mbps of WiFi data rate, but increasing the data rate does not help improve the amount of offloading. Further studies using this framework on two new deployment strategies of WiFi networks give guidance on how to upgrade WiFi networks by revealing that the amount of offloading for 30 minutes of delay and 1 Mbps of data rate can be drastically improved to about 90% or 98% according to the strategy.

Categories and Subject Descriptors

C.2.1 [**Network Architecture and Design**]: Wireless communication; D.4.8 [**Performance**]: Queueing theory

Keywords

Mobile data offloading, WiFi networks, queueing, reneging

* indicates corresponding author.

MobiHoc'14, August 11–14, 2014, Philadelphia, PA, USA.

http://dx.doi.org/10.1145/2632951.2632991.

1. INTRODUCTION

Recent advances in mobile devices (e.g., smartphones, tablets) are fueling the dramatic growth of mobile data traffic all over the world. A recent report has predicted that the total number of mobile Internet users is projected to surpass desktop Internet users by 2014 [19]. In order to support ubiquitous connectivity from such mobile devices to the Internet, cellular network providers are struggling to increase the capacity of their cellular networks by reducing cell sizes, widening the wireless channel bandwidth and upgrading the communication standards. As a result, in many countries, LTE and LTE-A (Long Term Evolution Advanced) networks supporting up to 150 Mbps are being operated with 20 to 40 MHz of channel bandwidth. However, even with these efforts, it is expected that the deluge of mobile data traffic will soon catch up and saturate the capacity of cellular networks [1]. This continued growth of mobile data traffic is creating challenges to the business model of cellular network providers who need to incur prohibitively large expenses for upgrading their networks, while the earnings from their subscribers remain relatively stagnant (at least in developed countries).

To alleviate this problem and aid cellular networks, it has been proposed that data can be offloaded on available WiFi networks (or on femtocells in the near future) [4, 12, 15, 17, 25]. This is often referred to as *mobile data offloading*. The idea is that mobile devices can offload their data traffic from cellular networks onto WiFi networks whenever WiFi networks are available[1]. Indeed, even now, most smartphones and tablets use mobile data offloading, by prioritizing WiFi networks over cellular networks. According to recent observations from Lee *et al.* [17], mobile data offloading already cuts about 65% of the data traffic burden to cellular networks by letting the traffic flow through WiFi networks. This 65% reduction seems surprising but the authors have suggested that the number can grow as high as 82% when allowing one hour delay for delay-insensitive data traffic and have named this technique *delayed mobile data offloading*.

The conceptual operation of delayed mobile data offloading (i.e., delayed offloading) is quite simple. Briefly, delayed offloading sets delay deadlines for data transmission that can be predetermined or chosen by a user and lets the mobile device find opportunities to offload the data to WiFi networks until the assigned deadline expires. To enable delayed offloading in real mobile devices, there have been efforts to redesign a transmission protocol which can allow a delay. The efforts include the bundle protocol [11, 20, 22] that keeps a session virtually connected under intermittent connectivity until its completion.

[1]Note that the central tenet of mobile data offloading can be directly applied to femtocell or picocell networks instead of WiFi networks.

After its proposal in [3, 17], delayed mobile data offloading has attracted many researchers, and various extensions [2] have been pursued including multi-hop offloading [18], social offloading [5], and offloading with ICN (Information Centric Networking) [9]. However, as of yet, we do not have a rigorous understanding of the performance of the delayed offloading system. This is because developing a rigorous analytical framework for delayed offloading system is quite challenging because as pointed out in [17], the system involves "queueing with reneging and service interruptions." In queueing, *reneging* means that a customer will leave the queue when her patience exceeds a certain limit. This resembles the behavior of a packet for its chosen deadline in the delayed offloading system. Service interruption literally means unwilling discontinuity of service in the queue, and this models connection and disconnection periods of a mobile device to WiFi networks in the system. In [17], 15 days of experimental study with about 100 iPhone users mostly in Seoul, Korea revealed that connection and disconnection periods showed heavy-tail distributions, especially truncated Pareto distributions with averages of 50 minutes and 25 minutes, respectively. Given these heavy-tail distributions of connection and disconnection periods, mathematical challenges arise in deriving a closed-form equation from the queueing process (which is obviously non-Markovian) on how much mobile data can be offloaded to WiFi networks for a chosen deadline. We propose a general analytical framework that handles such challenges. Our analytical results are validated by verifying the agreement of the amount of offloading obtained from our results and that shown in [17] for the same input traffic and system parameters. Through extensive numerical studies under various environments of WiFi networks including those of today and of projections made for the near future, we show that our framework is useful in designing a new radical deployment strategy of WiFi networks that achieves a dramatic improvement in the amount of offloading. Using our framework, we further clarify which parameter of WiFi networks most affects the efficiency of offloading.

Our contributions in this paper are as follows: (i) We develop an analytical framework for analyzing a queueing system with deterministic reneging and service interruptions. (ii) From our framework, we obtain closed-form formulas for key performance metrics in the delayed offloading system such as *offloading efficiency* and mean *packet delay*. (iii) Our framework is quite general in that it is capable of handling generally distributed system parameters including connection and disconnection periods, packet inter-arrival time, WiFi data rate, and packet size. (iv) We provide a guidance on how to upgrade WiFi networks to obtain much higher efficiency in offloading through extensive numerical studies using our analytical framework. We believe that our framework can be a stepping stone towards analyzing the delayed offloading system and its variants. Our analysis is applicable to generalized queueing systems with service interruptions and reneging, which can also be extended to explain the behavior of complex mobile systems.

2. DELAYED MOBILE DATA OFFLOADING SYSTEM

In this section, we overview the delayed mobile data offloading system proposed in [17]. In this paper, we describe the system from the viewpoint of data uploading from a mobile device to the Internet, but the same framework can be readily used for downloading. The only difference between delayed offloading for uploading and downloading is at the network operation in a mobile device. Uploading works as pushing while downloading works as pulling. In both cases, delayed offloading necessitates a bundle protocol[2] at the network layer which handles the responsibility of fulfilling a data request even when the connection to the Internet is intermittent. Thus, in delayed offloading, the session initiated for a data request is kept virtually alive by the bundle protocol until the requested data transfer is finished.

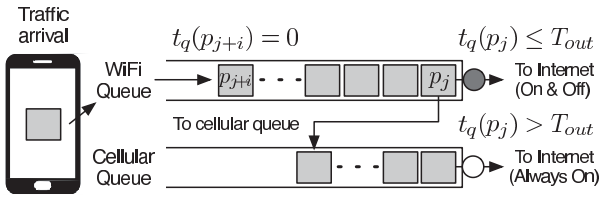


Figure 1: Delayed offloading mechanism in a mobile device.

Fig. 1 shows a simplified architecture of the delayed offloading system for the uploading scenario. There are two coupled queues for data upload at the MAC layer of a mobile device, called WiFi queue and cellular queue, and both queues are served by a FIFO (first-in-first-out) discipline. When the user of a mobile device requests to upload data (e.g., photos and videos taken at a park which need to be synchronized to a cloud backup service like iCloud) with a certain amount of allowed delay, denoted by T_{out}, the traffic request is first inserted into the WiFi queue and waits for the device to be connected to any WiFi network for up to a maximum of T_{out} units of time. If data in the WiFi queue is completely transmitted through WiFi networks before the time-out T_{out} has expired, we say that the data is successfully offloaded. If offloading of data fails, the system lets the data leave the WiFi queue and be relocated to the cellular queue in the mobile device for immediate transmission through 3G or 4G networks. We call such an event a *reneging* event.

In this paper, we analyze the ratio of the amount of successfully offloaded data over the amount of total data requested, which we call the *offloading efficiency* of the system. Since a mobile device is intermittently connected to WiFi networks by the mobility pattern of its owner (i.e., a human or sometimes a vehicle driven by a human) and the data rate for each connection is random, analyzing the offloading efficiency is a challenging problem. Further, given that empirical observations indicate that both the connection and the disconnection processes to any WiFi network of a mobile device follow heavy-tail distributions [17], it becomes even harder to analytically characterize the offloading efficiency for a given T_{out}. The detailed technical challenges from a queueing theory perspective and our proposed approach will be discussed in detail in Section 4.

3. MODEL DESCRIPTION

3.1 Queueing System Model

We assume that time axis is divided into unit intervals referred to as slots, and the slots are indexed by $t\,(t = 1, 2, \ldots)$. We set the system to start at time 1 so that time slot t covers the time interval $[t, t + 1)$. In the following, we describe our discrete-time queueing system in detail. We begin by describing the service process of the WiFi queue.

Service process of the WiFi queue. Let $C(t)$ be a random variable that denotes the connectivity status of the mobile device to a WiFi network. We define $C(t) := 1$ when the mobile device is

[2]See [11] for a candidate implementation in a mobile device.

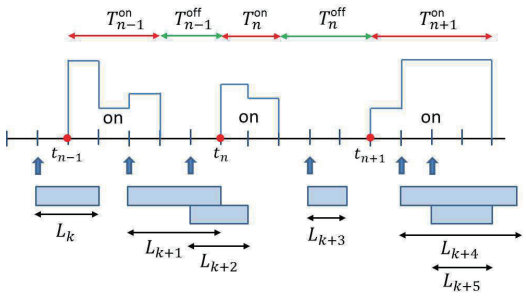


Figure 2: System model.

connected to a WiFi network during time slot t and $C(t) := 0$ otherwise. Then, the connectivity process $\{C(t)\}_{t=1}^{\infty}$ is modeled by an on-off process, where the on-state corresponds to being in a WiFi zone and the off-state corresponds to being out of the WiFi zones. Let T_n^{on} and T_n^{off} denote the nth $(n = 1, 2, \ldots)$ on-period and off-period of the process $\{C(t)\}_{t=1}^{\infty}$, respectively (See Fig. 2). We assume that the pair $(T_n^{\text{on}}, T_n^{\text{off}})$ is i.i.d. (independent and identically distributed) across n. However, for a fixed n, *we allow* T_n^{on} *and* T_n^{off} *to be dependent.* Therefore, the process $\{(T_n^{\text{on}}, T_n^{\text{off}})\}_{n=1}^{\infty}$ becomes an alternating renewal sequence [21]. Let $(T_{\text{on}}, T_{\text{off}})$ be the generic random vector for $(T_n^{\text{on}}, T_n^{\text{off}})$. Empirical observations in [17] show that T_{on} and T_{off} follow truncated heavy-tail distributions. Our model assumes general distributions having finite mean $\mathrm{E}[T_{\text{on}}] < \infty$ and $\mathrm{E}[T_{\text{off}}] < \infty$, and includes such truncated heavy-tail distributions.

Let $S(t)$ denote the transmission rate provided to the WiFi queue during time slot t. Then, we have

$$S(t) = \begin{cases} 0 & \text{if } C(t) = 0, \\ S_{\text{WiFi}}(t) & \text{if } C(t) = 1, \end{cases} \tag{1}$$

where $S_{\text{WiFi}}(t)$ denotes the transmission rate of the WiFi network to which the mobile device is connected during time slot t. For simplicity, we assume that $S_{\text{WiFi}}(t)$ is i.i.d. across t (with a generic random variable R), but our analysis is readily extensible to the case where $\{S_{\text{WiFi}}(t)\}_{t=1}^{\infty}$ is a Markov process.

Arrival process into the WiFi queue. We next describe the packet arrival process into the WiFi queue. Let $A(t)$ be a random variable that denotes the number of packets that are generated by the mobile device at time slot t. Then, the packet generation process $\{A(t)\}_{t=1}^{\infty}$ becomes the arrival process into the WiFi queue. Let A_k denote the kth $(k = 1, 2, \ldots)$ inter-arrival time. We assume that A_k is i.i.d. across k (with a generic random variable A) and follows a general distribution having finite mean $\mu := \mathrm{E}[A] < \infty$, i.e., the arrival process $\{A(t)\}_{t=1}^{\infty}$ is assumed to be a renewal process.

Packet size. In this paper, we allow the size of a packet to be variable which includes the fixed (deterministic) size as a special case. Let L be a random variable that denotes the size of a generic packet. In practice, packet size is bounded by the maximum transmission unit. Hence, we assume that the packet size distribution has finite support on the range $[1, L_{\max}]$.

Queue evolution equation. We now describe the queueing dynamics at the WiFi queue. Let $Q(t)$ be a random variable that denotes the number of packets in the WiFi queue at the beginning of time slot t. Then, the queueing process $\{Q(t)\}_{t=1}^{\infty}$ evolves according to the following recursion:

$$Q(t+1) = Q(t) + A(t) - D_{\text{WiFi}}(t) - D_{\text{Renege}}(t).$$

Here, $D_{\text{WiFi}}(t)$ is the number of packets that is completely transmitted through a WiFi network during time slot t, and $D_{\text{Renege}}(t)$ is the number of packets that renege during time slot t. Note that $D_{\text{WiFi}}(t)$ is given by

$$D_{\text{WiFi}}(t) = \begin{cases} 0, & \text{if } Q(t) + A(t) = 0 \text{ or } S(t) = 0, \\ \sup\{k : \sum_{i=1}^{k} L_i(t) \le S(t)\}, & \text{otherwise,} \end{cases}$$

where $L_i(t)$ $(i = 1, 2, \ldots, Q(t) + A(t))$ denotes the size of the ith packet in the WiFi queue at time t. Also, note that $D_{\text{Renege}}(t) \le Q(t) + A(t) - D_{\text{WiFi}}(t)$. That is, reneging (if any) occurs among $Q(t) + A(t) - D_{\text{WiFi}}(t)$ number of packets in the WiFi queue. Even for the case $S(t) > 0$ (i.e., when the service is being carried out during time slot t), reneging could occur for the head-of-line packet in the WiFi queue when the transmission rate is not sufficiently fast to complete the service of the head-of-line packet, i.e., $S_{\text{WiFi}}(t) < L_1(t)$.

3.2 Performance Metrics

In this section, we provide formal definitions of the performance metrics in the delayed offloading system. The primary quantity of our interest is the offloading efficiency.

DEFINITION 1 (OFFLOADING EFFICIENCY ρ). *The offloading efficiency is defined as the probability that a packet in the WiFi queue is completely served by WiFi networks before being reneged to cellular networks.*

Note that the offloading efficiency defined here does not take the packet loss in WiFi networks into consideration, and is statistically the same with the amount of packets served by WiFi networks divided by the amount of packets inserted in the WiFi queue. The second quantity of our interest is the mean packet delay of the delayed offloading system.

DEFINITION 2 (MEAN PACKET DELAY $\overline{W}$). *The mean packet delay is defined as the average duration of time that a packet generated by a mobile device stays in the WiFi queue before it is served or reneged.*

The offloading efficiency ρ is determined by the probability for a packet being reneged P_{Renege} (called the reneging probability), by the relation $\rho = 1 - P_{\text{Renege}}$. When we set a larger T_{out}, then ρ increases and $\overline{W}$ also increases. Hence, the offloading efficiency and the mean packet delay are in a trade-off relationship whose control knob is T_{out}.

4. TECHNICAL CHALLENGES AND PROPOSED APPROACH

4.1 Technical Challenges

Our queueing system is characterized by three factors: (i) server vacation with non-exhaustive service, (ii) heavy-tailed vacation and non-vacation periods, and (iii) impatient customers with deterministic reneging times. Here, the server vacation system with non-exhaustive services refers to a queueing system in which the server stops service and can have a vacation even when there is a customer in the queue. These three factors cause significant technical challenges, as described below.

First, the heavy-tail distributions and the server vacation make the queueing system non-Markovian as well as non-work-conserving. Furthermore, from the viewpoint of Kendall's notation, the resultant service times also follow a heavy-tail distribution and are correlated across each packet. Hence, factors (i) and (ii) yield $G/G/1$

queue with correlated and heavy-tailed service time. The third factor requires us to track the waiting time of each packet in the queue. Moreover, due to reneging, the queue length is affected by the waiting time, while, in turn, the waiting time is affected by the queue length. Thus, both the queue length and the waiting time need to be jointly investigated. In conclusion, we need a mathematical technique that enables us to handle the joint queue length and waiting time simultaneously for non-Markovian queueing system having correlated and heavy-tailed service times.

There have been several queueing papers that consider reneging customers or server vacation exclusively. Regarding reneging (without server vacation), there have been studies on queues with deterministic impatience time D (See [24] and references therein). In [6] and [7], Barrer studied $M/M/1+D$ and $M/M/n+D$ queues and obtained the customer loss probability. In [24], Xiong *et al.* studied the $M/G/1+D$ queue using level crossing analysis, but analytical solution is given only for $M/H_2/1+D$ queue having two-stage hyper-exponential service times, and a numerical approach is presented for the more general $M/G/1+D$ queue. Recently, Kim *et al.* [13] studied $M/PH/1+D$ queue having phase-type distributed service times. Regarding the server vacation (without the reneging), there also have been extensive studies, e.g., [10,23]. In [10], Doshi provided a survey for queueing systems with vacations. In [23], Takagi worked on mathematical modeling and analysis for a broad class of server vacation systems. However, to the best of our knowledge, there was no analytic work on the queueing system that is addressed in this paper.

4.2 Proposed Approach

We take a three-step approach to obtain the offloading efficiency and the mean queueing delay.

Step 1 (Analysis at an embedded point). First, by noting that the sequence of on/off periods $\{(T_n^{\text{on}}, T_n^{\text{off}})\}_{n=1}^{\infty}$ is a renewal sequence, we observe the system only at a subset of time slots when the on-period begins. At such time instants, called embedded points, we define a state vector in such a way that it is capable of capturing queue length, waiting time of each packet, and residual inter-arrival time with a minimum number of states. The details are provided in Section 5.1.

Step 2 (Analysis at an arbitrary time). Next, we derive the limiting distribution at an arbitrary point in time using the result on the embedded process. The key idea of our derivation is motivated by the Renewal Reward Theorem [8, 21] (which considers i.i.d. rewards across cycles). In our analysis, rewards are not necessarily i.i.d. across cycles, so we introduce a notion of conditional rewards to extract an i.i.d. subsequence of rewards. The details are provided in Section 5.2.

Step 3 (Derivation of performance metrics). Finally, from the limiting distribution at an arbitrary point in time, we can derive the analytic formulas for the offloading efficiency and the mean packet delay in the WiFi queue. The details are provided in Section 5.3.

Our contributions in the proposed analytical technique are as follows. We provide a mathematical approach to analyze a queueing system with deterministic reneging and heavy-tailed service interruption. In particular, our approach incorporates discrete-time Markov chain theory and a version of the Renewal Reward Theorem. In non-Markovian queueing systems, a judicious selection of embedded points makes the analysis tractable at times (often with Markov chain theory). In our case, in addition to the use of embedded points, we also strategically define a set of state variables in order to build up a foundation to the next step analysis using renewal reward type of arguments. Also, we develop an extended version of the Renewal Reward Theorem that is applicable even when the reward is not *i.i.d.* but correlated across cycles. Our framework explicitly derives the offloading efficiency by solving a system of matrix equations obtained from Markov chain theory. While beyond the scope of the current paper, our exact analysis could also be used to derive simple analytical approximations that might be tight in certain asymptotic regimes.

5. ANALYTICAL FRAMEWORK

5.1 Analysis at an Embedded Point

We observe the system only at the time instants when the mobile device that has been out of the WiFi coverage is switched to be connected to any WiFi network. We take such a time instant as our embedded point. That is, the nth ($n = 1, 2, \ldots$) embedded point corresponds to the beginning of the nth on-period. Let t_n denote the time at which the nth embedded point is located. Refer to Fig. 2 for a depiction of the embedded point. We assume that initially the mobile device is connected to a WiFi network, i.e., $C(1) = 1$. Then, from Fig. 2, we have

$$t_n = \begin{cases} 1 & n = 1, \\ t_{n-1} + T_{n-1}^{\text{on}} + T_{n-1}^{\text{off}} & n \geq 2. \end{cases}$$

At each embedded point (i.e., at time t_n), we observe the following four state variables: (i) Let $Q_n := Q(t_n)$ be the number of packets in the WiFi queue at time t_n. (ii) Let $W_n := W(t_n)$, where $W(t)$ for $t = 1, 2, \ldots$ is the waiting time of the head-of-line packet in the WiFi queue at time t. If $Q_n = 0$, then $W_n = 0$ holds. Under the FIFO policy, W_n represents the age of the oldest packet among the packets present in the WiFi queue at time t_n. (iii) Let $U_n := U(t_n)$, where $U(t)$ for $t = 1, 2, \ldots$ is the size of the head-of-line packet in the WiFi queue at time t. If $Q_n = 0$, then $U_n = 0$ holds. Thus, U_n represents the amount of unfinished work on the oldest packet at time t_n. (iv) Let $E_n := E(t_n)$, where $E(t)$ for $t = 1, 2, \ldots$ is the elapsed time from the moment when the last packet in the WiFi queue arrives to time t. When $Q_n = 1$, we have $E_n = W_n$. Finally, we let $\mathcal{S}$ be the set of states which can be expressed by the tuple (Q_n, W_n, U_n, E_n). Using the tuple, we define a state vector at the nth embedded point as follows:[3]

$$\boldsymbol{X}_n := (Q_n, W_n, U_n, E_n).$$

We call the process $\{\boldsymbol{X}_n\}_{n=1}^{\infty}$ the *embedded process*. Note that the embedded process can be viewed as a sampling of the process $\{\boldsymbol{X}(t)\}_{t=1}^{\infty} := \{(Q(t), W(t), U(t), E(t))\}_{t=1}^{\infty}$ at every $t = t_n$, giving the relation $\{\boldsymbol{X}(t_n)\}_{n=1}^{\infty} = \{\boldsymbol{X}_n\}_{n=1}^{\infty}$. Our sampling of embedded points results in a nice analytical property, as shown in Lemma 1.

LEMMA 1. *The embedded process $\{\boldsymbol{X}_n\}_{n=1}^{\infty}$ is a discrete-time Markov chain with state space $\mathcal{S}$.*

PROOF. Due to space limitation, we provide an outline of the proof, and the detailed proof is given in Appendix B in our technical report [14]. For ease of understanding, we suppose that $Q_n \geq 1$. Then, W_n and E_n represent the ages of the oldest packet and the youngest packet, respectively, among the Q_n packets in the WiFi queue. Hence, given the pair (W_n, E_n), we can infer the age of each packet in the WiFi queue from the inter-arrival time distribution $\mathrm{P}(A \leq x)$. In addition to the age distribution, we also need information on the size of each packet in the WiFi queue. Since only the head-of-line packet has a different packet-size distribution from

[3]Throughout this paper, we use a bold font for a vector notation.

the other packets (since it may be partially served), we observe the information via U_n. Finally, in order to predict the possibility of packet generation at time t_n, we need the elapsed time E_n from the latest packet generation. Therefore, using the triple information on ages, packet sizes, and probability of packet generation at a given time n, we can analyze how $\boldsymbol{X}_{\hat{n}}$ will behave in the future for $\hat{n} > n$, i.e., the process $\{\boldsymbol{X}_n\}_{n=1}^{\infty}$ becomes a Markov chain. □

If the packet size follows a geometric distribution, then U_n also follows the geometric distribution due to the memoryless nature. Hence, if we set the state vector as

$$\boldsymbol{X}'_n := (Q_n, W_n, E_n),$$

then $\{\boldsymbol{X}'_n\}_{n=1}^{\infty}$ is a discrete-time Markov chain. Furthermore, if the arrival process is Bernoulli, then by the memoryless nature again, we can further shorten the state vector as

$$\boldsymbol{X}''_n := (Q_n, W_n).$$

Then, $\{\boldsymbol{X}''_n\}_{n=1}^{\infty}$ is a discrete-time Markov chain.

In the following, we derive the transition probability matrix of the Markov chain $\{\boldsymbol{X}_n\}_{n=1}^{\infty}$. Let the ith and the jth elements of state space $\mathcal{S}$ be denoted by $\boldsymbol{i}$ and $\boldsymbol{j}$, respectively. We define

$$\begin{aligned} P^{\text{on}}_{\boldsymbol{i},\boldsymbol{j}} &:= \mathrm{P}(\boldsymbol{X}(t+1) = \boldsymbol{j} \mid \boldsymbol{X}(t) = \boldsymbol{i}, C(t) = 1), \\ P^{\text{off}}_{\boldsymbol{i},\boldsymbol{j}} &:= \mathrm{P}(\boldsymbol{X}(t+1) = \boldsymbol{j} \mid \boldsymbol{X}(t) = \boldsymbol{i}, C(t) = 0). \end{aligned} \tag{2}$$

Note that the probabilities $P^{\text{on}}_{\boldsymbol{i},\boldsymbol{j}}$ and $P^{\text{off}}_{\boldsymbol{i},\boldsymbol{j}}$ are concerned with transition across adjacent slots: $P^{\text{on}}_{\boldsymbol{i},\boldsymbol{j}}$ is the probability of being in state $\boldsymbol{j}$ at the beginning of time slot $(t+1)$, provided that the system is in state $\boldsymbol{i}$ at the beginning of time slot t and a WiFi network is available during time slot t. A similar interpretation applies to the probability $P^{\text{off}}_{\boldsymbol{i},\boldsymbol{j}}$. The formulas for $P^{\text{on}}_{\boldsymbol{i},\boldsymbol{j}}$ and $P^{\text{off}}_{\boldsymbol{i},\boldsymbol{j}}$ are given in Appendix. We define

$$\hat{P}_{\boldsymbol{i},\boldsymbol{j}} := \mathrm{P}(\boldsymbol{X}_{n+1} = \boldsymbol{j} \mid \boldsymbol{X}_n = \boldsymbol{i}).$$

Note that now $\hat{P}_{\boldsymbol{i},\boldsymbol{j}}$ is concerned with transition across adjacent embedded points. Let $M := [\hat{P}_{\boldsymbol{i},\boldsymbol{j}}]$ denote the one-step transition probability matrix of the Markov chain $\{\boldsymbol{X}_n\}_{n=1}^{\infty}$. Since $\hat{P}_{\boldsymbol{i},\boldsymbol{j}}$ is rewritten as $\hat{P}_{\boldsymbol{i},\boldsymbol{j}} = \mathrm{P}(\boldsymbol{X}(t_{n+1}) = \boldsymbol{j} \mid \boldsymbol{X}(t_n) = \boldsymbol{i})$, by incorporating the duration between the points t_n and t_{n+1}, we can compute the matrix $M = [\hat{P}_{\boldsymbol{i},\boldsymbol{j}}]$ from the probabilities $P^{\text{on}}_{\boldsymbol{i},\boldsymbol{j}}$ and $P^{\text{off}}_{\boldsymbol{i},\boldsymbol{j}}$, as shown below in Lemma 2.

LEMMA 2. *Let $M_{on} := [P^{on}_{\boldsymbol{i},\boldsymbol{j}}]$ and $M_{off} := [P^{off}_{\boldsymbol{i},\boldsymbol{j}}]$. Then, the transition probability matrix $M = [\hat{P}_{\boldsymbol{i},\boldsymbol{j}}]$ is obtained by*

$$M = \sum_{a \geq 1, b \geq 1} (M_{on})^a (M_{off})^b P(T_{on} = a, T_{off} = b).$$

PROOF. The key idea of our derivation relies on the Chapman-Kolmogorov's Theorem [16] that the k-step ($k = 1, 2, \ldots$) transition probability matrix of a discrete-time Markov chain is the kth power of the one-step transition probability matrix; and the k-step we consider corresponds to the length of the interval $[t_n, t_{n+1})$. Due to space limitation, we omit the details. For the detailed proof, please refer to Appendix D in our technical report [14]. □

From the transition probability matrix M, we can analyze the limiting behavior of the Markov chain $\{\boldsymbol{X}_n\}_{n=1}^{\infty}$ as follows. For each $\boldsymbol{i} \in \mathcal{S}$, we define

$$\pi_{\boldsymbol{i}} := \lim_{n \to \infty} \mathrm{P}(\boldsymbol{X}_n = \boldsymbol{i}) = \mathrm{P}(\boldsymbol{X}_\infty = \boldsymbol{i}),$$

which denotes the probability of the system being in state $\boldsymbol{i}$ at an arbitrary embedded point. Let $\boldsymbol{\pi} := (\pi_{\boldsymbol{i}})$ denote the limiting distribution vector. If the Markov chain $\{\boldsymbol{X}_n\}_{n=1}^{\infty}$ is ergodic, then the limiting distribution $\boldsymbol{\pi}$ exists and is obtained by solving the following system of matrix equations:

$$\boldsymbol{\pi} M = \boldsymbol{\pi}, \quad \boldsymbol{\pi} \boldsymbol{e} = 1,$$

where $\boldsymbol{e}$ is a vector of ones. A sufficient condition for ergodicity is existence of a constant $A_{\max}$ such that $\mathrm{P}(A \leq A_{\max}) = 1$. Under this condition, the elapsed time $E(\cdot)$ is bounded by $A_{\max}$, and thus the space $\mathcal{S}$ becomes finite. Since an irreducible Markov chain with finite state space is positive recurrent [16], we can show ergodicity. The condition $\mathrm{P}(A \leq A_{\max}) = 1$ is reasonable in the delayed mobile data offloading system since it is targeted for users who actively generate data.

5.2 Analysis at an Arbitrary Time

In the previous section, we have analyzed the distribution of the random vector $\boldsymbol{X}(t_n)$ at an arbitrary embedded point t_n as $n \to \infty$, i.e., $\boldsymbol{X}_\infty = \boldsymbol{X}(t_\infty)$. In this section, we analyze the distribution of $\boldsymbol{X}(t)$ at an arbitrary time t as $t \to \infty$, i.e., $\boldsymbol{X}(\infty)$. To this end, we derive formulas for the following probabilities for $\boldsymbol{i} \in \mathcal{S}$:

$$\begin{aligned} \xi^{\text{on}}_{\boldsymbol{i}} &:= \lim_{t \to \infty} \mathrm{P}(\boldsymbol{X}(t) = \boldsymbol{i}, C(t) = 1), \\ \xi^{\text{off}}_{\boldsymbol{i}} &:= \lim_{t \to \infty} \mathrm{P}(\boldsymbol{X}(t) = \boldsymbol{i}, C(t) = 0). \end{aligned} \tag{3}$$

Then, the limiting probability $\xi_{\boldsymbol{i}} := \lim_{t \to \infty} \mathrm{P}(\boldsymbol{X}(t) = \boldsymbol{i})$ is obtained by $\xi_{\boldsymbol{i}} = \xi^{\text{on}}_{\boldsymbol{i}} + \xi^{\text{off}}_{\boldsymbol{i}}$.

The key idea of our derivation is motivated by the Renewal Reward Theorem [8], which states the following: consider a renewal process $\{N(t)\}_{t \geq 0}$ having cycle lengths $\{C_k\}_{k=1}^{\infty}$. Suppose that a reward R_k is earned during the kth cycle, and assume that the pair (C_k, R_k) is i.i.d. across k (with a generic random vector (C, R)). If we let $R(t) = R_1 + \ldots + R_{N(t)}$ denote the total reward earned by time t, then with probability 1, we obtain

$$\frac{R(t)}{t} \to \frac{\mathrm{E}[R]}{\mathrm{E}[C]} \quad \text{as} \quad t \to \infty, \tag{4}$$

provided that $\mathrm{E}[R] < \infty$ and $\mathrm{E}[C] < \infty$. The theorem says that the long-term average reward is just the expected reward earned during a cycle, divided by the expected time of a cycle.

In our queueing model, the sequence $\{(T^{\text{on}}_n, T^{\text{off}}_n)\}_{n=1}^{\infty}$ is an alternating renewal sequence. Hence, we can view the joint nth on/off-periods as the nth cycle having length $T^{\text{on}}_n + T^{\text{off}}_n$. To compute the distribution $\xi^{\text{on}}_{\boldsymbol{i}}$, we count the total number of slots with state $(\boldsymbol{X}(t), C(t)) = (\boldsymbol{i}, 1)$ during the nth cycle and set the result as our reward $R^{\text{on}}_{n,\boldsymbol{i}}$. Then, the long-term average reward becomes the time average probability of being in state $(\boldsymbol{X}(t), C(t)) = (\boldsymbol{i}, 1)$, which is equal to the ensemble average probability under ergodicity condition. *Note that $R^{on}_{n,\boldsymbol{i}}$ in this setting is not necessarily independent, but identically distributed across n for a fixed $\boldsymbol{i} \in \mathcal{S}$.* Therefore, we cannot apply the formula (4) directly so we take a detour to analyze the long-term behavior. To our surprise, we reach the same conclusion that the long-term average reward is just the expected reward earned during a cycle, divided by the expected time of a cycle, as shown in Lemma 3.

LEMMA 3. *Suppose that the embedded process $\{\boldsymbol{X}_n\}_{n=1}^{\infty}$ is ergodic. Then, the limiting distributions $\xi^{on}_{\boldsymbol{i}}$ and $\xi^{off}_{\boldsymbol{i}}$ defined in (3) exist and are obtained by*

$$\xi^{on}_{\boldsymbol{i}} = \frac{\sum_{\boldsymbol{j} \in \mathcal{S}} \pi_{\boldsymbol{j}} \mu^{on}_{\boldsymbol{j},\boldsymbol{i}}}{E[T_{on} + T_{off}]}, \quad \xi^{off}_{\boldsymbol{i}} = \frac{\sum_{\boldsymbol{j} \in \mathcal{S}} \pi_{\boldsymbol{j}} \mu^{off}_{\boldsymbol{j},\boldsymbol{i}}}{E[T_{on} + T_{off}]}, \tag{5}$$

where

$$\mu^{on}_{\boldsymbol{j},\boldsymbol{i}} := E\Big[\sum_{s=t_n}^{t_{n+1}-1} 1_{\{\boldsymbol{X}(s)=\boldsymbol{i},C(s)=1\}} \,\Big|\, \boldsymbol{X}_n=\boldsymbol{j}\Big],$$

$$\mu^{off}_{\boldsymbol{j},\boldsymbol{i}} := E\Big[\sum_{s=t_n}^{t_{n+1}-1} 1_{\{\boldsymbol{X}(s)=\boldsymbol{i},C(s)=0\}} \,\Big|\, \boldsymbol{X}_n=\boldsymbol{j}\Big],$$

and where $1_{\{\cdot\}}$ *denotes the indicator function.*

PROOF. The proof consists of two parts. In the first part, we analyze the time average probability that the system stays in state $\boldsymbol{i}$ ($\boldsymbol{i} \in \mathcal{S}$) with WiFi being available. We define

$$R^{\text{on}}_{\boldsymbol{i}}(t) := \sum_{s=1}^{t} 1_{\{\boldsymbol{X}(s)=\boldsymbol{i},C(s)=1\}},$$

which denotes the total amount of rewards earned by time t. Then, the time average $\frac{1}{t}R^{\text{on}}_{\boldsymbol{i}}(t)$ converges with probability 1 as follows:

$$\frac{R^{\text{on}}_{\boldsymbol{i}}(t)}{t} \to \frac{1}{\mathrm{E}[T_{\text{on}}+T_{\text{off}}]}\sum_{\boldsymbol{j}\in\mathcal{S}}\pi_{\boldsymbol{j}}\mu^{\text{on}}_{\boldsymbol{j},\boldsymbol{i}} \quad \text{as} \quad t\to\infty. \tag{6}$$

In the second part, we show that the time average probability on the right-hand side of (6) is equal to the ensemble average probability due to the ergodicity of the process $\{\boldsymbol{X}_n\}_{n=1}^{\infty}$. Hence, we obtain

$$\lim_{t\to\infty}\mathrm{P}(\boldsymbol{X}(t)=\boldsymbol{i},C(t)=1) = \frac{1}{\mathrm{E}[T_{\text{on}}+T_{\text{off}}]}\sum_{\boldsymbol{j}\in\mathcal{S}}\pi_{\boldsymbol{j}}\mu^{\text{on}}_{\boldsymbol{j},\boldsymbol{i}}.$$

For details, refer to Appendix E in [14]. □

Note that in Lemma 3, $\mu^{\text{on}}_{\boldsymbol{j},\boldsymbol{i}}$ and $\mu^{\text{off}}_{\boldsymbol{j},\boldsymbol{i}}$ represent the expected reward earned during a cycle, conditioned that the cycle starts from state $\boldsymbol{j}$. We call $\mu^{\text{on}}_{\boldsymbol{j},\boldsymbol{i}}$ and $\mu^{\text{off}}_{\boldsymbol{j},\boldsymbol{i}}$ the *conditional rewards*. Since $\pi_{\boldsymbol{j}}$ represents the probability that a cycle starts from state $\boldsymbol{j}$, the numerator in (5) results in the expected reward earned during a cycle. The denominator in (5) represents the expected time of a cycle.

The formulas for $\mu^{\text{on}}_{\boldsymbol{j},\boldsymbol{i}}$ and $\mu^{\text{off}}_{\boldsymbol{j},\boldsymbol{i}}$ remain to be derived. For a matrix M, let $[M]_{j,i}$ denote the (j,i)th element of M. Then, the conditional rewards can be obtained from the matrices $M_{\text{on}} = [P^{\text{on}}_{\boldsymbol{i},\boldsymbol{j}}]$ and $M_{\text{off}} = [P^{\text{off}}_{\boldsymbol{i},\boldsymbol{j}}]$ as follows:

$$\begin{aligned}\mu^{\text{on}}_{\boldsymbol{j},\boldsymbol{i}} &= \mathrm{E}\Big[\sum_{k=1}^{T_{\text{on}}}[(M_{\text{on}})^{k-1}]_{j,i}\Big],\\ \mu^{\text{off}}_{\boldsymbol{j},\boldsymbol{i}} &= \mathrm{E}\Big[\sum_{k=1}^{T_{\text{off}}}[(M_{\text{on}})^{T_{\text{on}}}(M_{\text{off}})^{k-1}]_{j,i}\Big].\end{aligned} \tag{7}$$

The proof of (7) is given in Appendix F in our technical report [14]. Combining (7) and Lemma 3 in a matrix form yields the following:

$$\boldsymbol{\xi}_{\text{on}} := (\xi^{\text{on}}_{\boldsymbol{i}}) = \frac{\boldsymbol{\pi}\mathrm{E}\big[\sum_{k=1}^{T_{\text{on}}}(M_{\text{on}})^{k-1}\big]}{\mathrm{E}[T_{\text{on}}+T_{\text{off}}]},$$

$$\boldsymbol{\xi}_{\text{off}} := (\xi^{\text{off}}_{\boldsymbol{i}}) = \frac{\boldsymbol{\pi}\mathrm{E}\big[(M_{\text{on}})^{T_{\text{on}}}\sum_{k=1}^{T_{\text{off}}}(M_{\text{off}})^{k-1}\big]}{\mathrm{E}[T_{\text{on}}+T_{\text{off}}]},$$

from which we finally obtain the limiting distribution of the process $\{\boldsymbol{X}(t)\}_{t=1}^{\infty}$ as

$$\boldsymbol{\xi} := \big(\mathrm{P}(\boldsymbol{X}(\infty)=\boldsymbol{i})\big) = \boldsymbol{\xi}_{\text{on}} + \boldsymbol{\xi}_{\text{off}}. \tag{8}$$

5.3 Offloading Efficiency and Mean Packet Delay

By using the limiting distribution $\boldsymbol{\xi} = (\xi_{(q,w,u,e)})$ in (8), we can derive the formulas for our performance metrics ρ and $\overline{W}$. We begin by providing the formula for ρ.

THEOREM 1. *The offloading efficiency ρ under the delayed mobile data offloading system is obtained by*

$$\rho = 1-\lambda^{-1}\sum_{q\geq1,u\geq1,e\geq1}\big(\xi^{off}_{(q,T_{out},u,e)} + \xi^{on}_{(q,T_{out},u,e)}P(R<u)\big).$$

PROOF. Here, once again, we provide an outline of the proof due to space limitation. For a given time t, a packet reneges (if any) due to one of the following two reasons: (i) Suppose that there is a packet in the WiFi queue at time t, and that the mobile device is not connected to any WiFi network, i.e., $Q(t)\geq 1$ and $C(t)=0$. In this case, a reneging event occurs for the head-of-line packet in the WiFi queue if its age becomes T_{out}, i.e., $W(t)=T_{\text{out}}$. (ii) Now suppose that there is a packet in the WiFi queue at time t, and that the mobile device is connected to a WiFi network, i.e., $Q(t)\geq 1$ and $C(t)=1$. In this case, a reneging event occurs for the head-of-line packet in the WiFi queue if its age is T_{out} and the service rate of the WiFi network is less than the size of the packet, i.e., $W(t)=T_{\text{out}}$ and $S(t)=S_{\text{WiFi}}(t)<U(t)$.

Since the above two cases are exclusive, summing up the reneging probabilities occurred by reasons (i) and (ii) leads to P_{Renege}. Two reasons correspond to the first and the second term in Theorem 1, respectively. Please refer to Appendix G in our technical report [14] for the detailed proof. □

Now we provide the formula for $\overline{W}$.

THEOREM 2. *The mean packet delay $\overline{W}$ under the delayed mobile data offloading system is obtained by*

$$\overline{W} = \lambda^{-1}\sum_{q\geq1,w\geq1,u\geq1,e\geq1} q\cdot\xi_{(q,w,u,e)}.$$

PROOF. The average number of packets in the WiFi queue, denoted by $\overline{Q}$, is obtained as

$$\overline{Q} = \sum_{q=1}^{\infty} q\mathrm{P}(Q(\infty)=q) = \sum_{q\geq1,w\geq1,u\geq1,e\geq1} q\cdot\xi_{(q,w,u,e)}.$$

Hence, by applying Little's Law [21], we have $\overline{W}=\lambda^{-1}\overline{Q}$. □

6. NUMERICAL STUDY

In this section, we verify the correctness and accuracy of our analytical framework by performing extensive numerical studies, which also provide us some useful insights. For the verification, we develop an event-driven simulator that mimics the behavior of the delayed offloading system shown in Fig. 1.

6.1 System Setup and Parameters

Our analytical framework and simulator are capable of taking generalized input parameters. In the huge space of combinations of such system parameters, we focus on practical values observed in [1, 17] and candidate values expected for near future as shown below.

User Traffic. Based on [1], we assume that a mobile device will roughly generate 7 GB of data per month in the near future, whose inter-arrival process follows a truncated Pareto distribution with 3 minutes and 6 hours for its lower and upper truncation. When multimedia recording and viewing are frequently performed in a device, 7 GB/month is a reasonable number given that the full HD video is recorded and played at up to 30 Mbps. To capture wider range of user behaviors including behaviors of near future, we also consider 3.5 GB/month and 14 GB/month of data generation. Since it is unclear how much portion of the total data volume is subject to realtime demand in the near future, we assumed that the entire

traffic is delay tolerant. Extension of our framework that can accommodate various deadlines for different traffic types including zero delay tolerance is left as future work.

WiFi Data Rate. From the measurement study on WiFi data rates in [17] showing that 1.2 Mbps on average is empirically achievable, we set the capacity of WiFi networks as either of 0.1 Mbps, 1 Mbps, 2 Mbps, 10 Mbps, and 50 Mbps. Given three factors: (i) the average data rate of WiFi networks in a region may highly vary (e.g., by the factor of development of a city), (ii) there exist high-speed WiFi standards in the market (such as 802.11ac providing up to 600 Mbps as its nominal speed), and (iii) WiFi APs are mostly located in crowded areas, considering up to 50 Mbps for individual WiFi data rate is rational (even when considering near future).

WiFi Deployment. The availability of WiFi networks in a city can be captured by the distributions of disconnected periods (i.e., inter-connection time) and connected periods (i.e., connection time) to any WiFi AP. In the measurement study of [17], the average inter-connection and connection times are shown to be (25, 50) minutes respectively in Seoul, Korea. Based on this status quo of WiFi deployment, we vary the inter-connection time and the connection time to have (15, 45) minutes and (4, 12) minutes on average for modeling possible WiFi environments in the near future (with more WiFi APs). We assume that both the connection time and the inter-connection time conform to a truncated Pareto distribution.

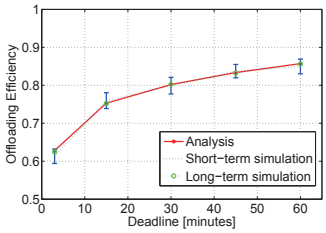
Deadline. We vary the deadline for offloading from 3 minutes to 1 hour and observe the impact of the system parameters on major performance metrics.

6.2 Numerical Analysis vs. Simulation

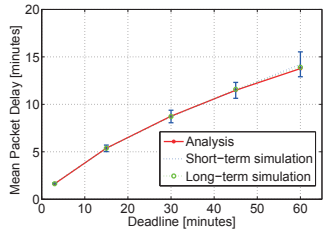
We validate our analytical framework by comparing its results for various parameter settings with those from short-term and long-term simulations. Figs. 3 (a) and (b) show the offloading efficiency and the queueing delay for various deadlines with 1 Mbps of WiFi data rate, 7 GB/month of input traffic, 25 minutes of inter-connection time, and 50 minutes of connection time. As expected, our framework provides almost the same results with those from long-term simulation and the average of a number of short-term simulations. Given that long-term simulations require much longer simulation time as well as more memory for their computations, our framework has its unique value. Note that the offloading efficiency from our framework for the deadline of 1 hour coincides with the result shown in [17] for the same parameter setting.

6.3 Offloading Scenarios and Lessons

Using our framework, we analyze diverse offloading scenarios that reflect the current network situation and possible near future environments through combinations of average inter-connection and connections times of (25, 50), (15, 45), and (4, 12) minutes. (25, 50) setting is considered as *current deployment* of WiFi networks while (15, 45) setting is chosen as a *wider deployment* of WiFi networks with substantially reduced amount of inter-connection time and slightly decreased amount of connection time. (4,12) setting is selected to model a completely different type of WiFi deployment in which WiFi APs are imagined to be installed in the middle of paths of movement such as on traffic lights or street lights that will result in much shorter inter-connection and connection time. We call it *prevalent deployment*. Note that (4,12) and (15,45) settings have the same connection ratio. We compare the wider deployment and the prevalent deployment over the current deployment to obtain guidance on how to upgrade WiFi networks in terms of achieved offloading efficiency, and the discussion on the mean packet delay is omitted due to space limitation.



(a) Offloading efficiency



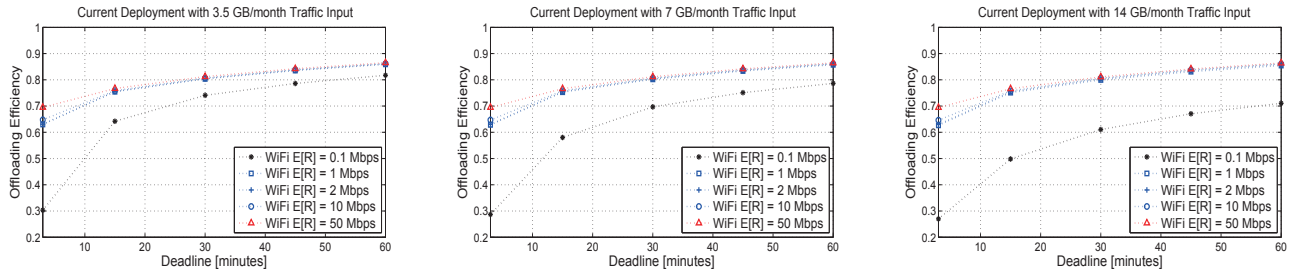
(b) Mean packet delay

Figure 3: The results from our framework compared with those from short-term and long-term simulations under data rate of 1 Mbps and input traffic of 7 GB/month. Short-term simulations are performed 20 times and their 95% confidence intervals are depicted.

For a comprehensive comparison, we vary the average data rate of WiFi networks, total amount of generated data per month per user, and deadline for offloading. Figs. 4 (a), (b), and (c) show the offloading efficiency from the current deployment of WiFi network for input traffic of 3.5 GB/month, 7 GB/month, and 14 GB/month, respectively, for various deadlines. Fig. 4 (a) shows that the offloading efficiency varies from 0.3 to 0.87 according to the deadline. It is very interesting to see that the average data rate of WiFi networks has almost no effect to the offloading efficiency when the data rate is beyond 1 Mbps. It is mainly because the arriving traffic does not come in large enough chunks to exploit data rates of greater than 1Mbps. This observation implies that upgrading WiFi APs for a newer standard provides diminishing returns from the point of view of the offloading efficiency. Figs. 4 (b) and (c) reconfirm that the effect of increased data rate is negligible. Figs. 4 (a), (b) and (c) commonly show that the offloading efficiency with the deadline of 1 hour is bounded by 87% under the current deployment.

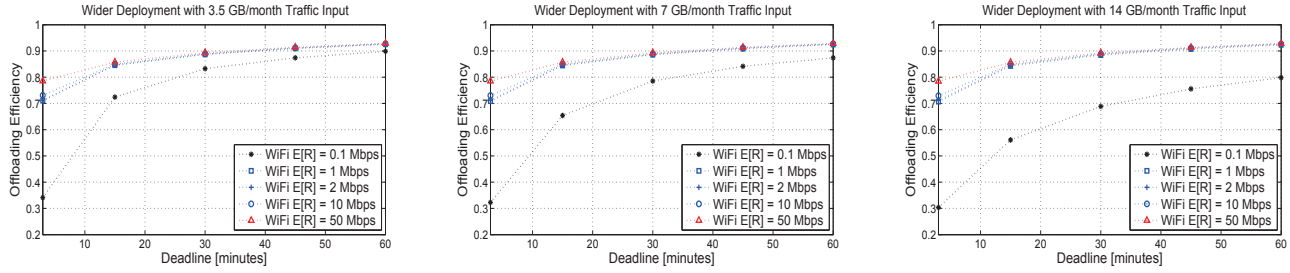
The results from the wider deployment depicted in Figs. 5 (a), (b), and (c) show that the offloading efficiency can be significantly improved by altering the deployment. For instance, with a deadline of 30 minutes, the wider deployment achieves about 89% of offloading efficiency which is nearly 10% higher than that of the current deployment. To our surprise, 89% is the level of offloading efficiency that was not achievable even with 60 minutes of deadline and higher data rate up to 50 Mbps. This implies that installing more WiFi APs, resulting in reducing the average inter-connection time gives far greater gains in offloading efficiency, as compared to increasing the data rate carried on the existing WiFi APs. Even in the wider deployment, data rates beyond 1 Mbps result only in a small improvement, except for the case of extremely short deadlines (e.g., 3 minute).

To purse higher offloading efficiency, we further test the prevalent deployment which may necessitate a larger number of WiFi APs than in the wider deployment. However, note that the installation cost of the prevalent deployment can be similar to that of the



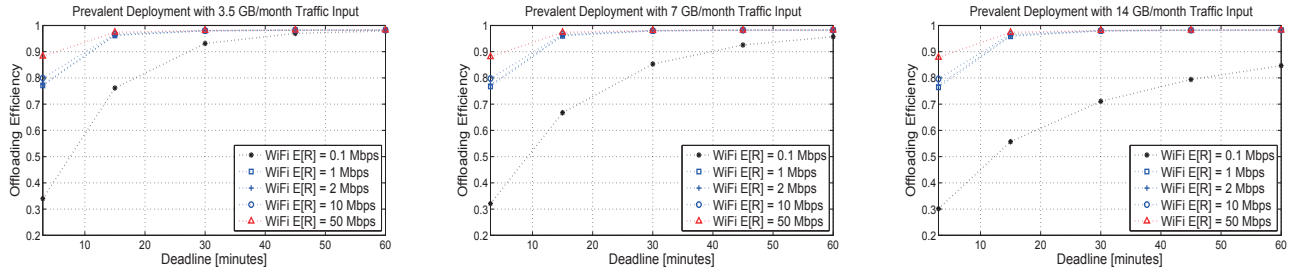
(a) 3.5 GB/month (b) 7 GB/month (c) 14 GB/month

Figure 4: Offloading efficiency obtained from the current deployment setting for various data rates and deadlines.



(a) 3.5 GB/month (b) 7 GB/month (c) 14 GB/month

Figure 5: Offloading efficiency obtained from the wider deployment setting for various data rates and deadlines.



(a) 3.5GB/month (b) 7 GB/month (c) 14 GB/month

Figure 6: Offloading efficiency obtained from the prevalent deployment setting for various data rates and deadlines.

wider deployment if we manage installation only with low-grade WiFi APs of slower data rate and low-speed wireless backhaul that substantially cuts construction cost. With the wireless backhaul, it is possible to deploy a number of WiFi APs at traffic lights or street lights on frequently visited roads at a manageable cost. Figs. 6 (a), (b), and (c) show impressive offloading performance as high as 98% even with 1 Mbps at 30 minutes of deadline. This extreme efficiency is not achievable in the current deployment and in the wider deployment no matter how fast the WiFi data rate is. Fig. 7 emphasizes the benefit of the prevalent deployment by directly comparing the offloading efficiency to other deployments under 15 minutes of deadline with various WiFi data rates. The prevalent deployment strategy dominates the wider deployment strategy for the whole range of WiFi data rates. Also, the offloading efficiency of the prevalent deployment with 0.3 Mbps already exceeds that of the wider deployment with 50 Mbps (or even higher).

The overall observations made from our framework imply that changing the paradigm of choosing installation locations of WiFi APs is worth considering. For instance, additional installation of WiFi APs can be recommended in the new locations that break down long inter-connection times into shorter ones such as road sides rather than in the conventional crowded areas such as department stores.

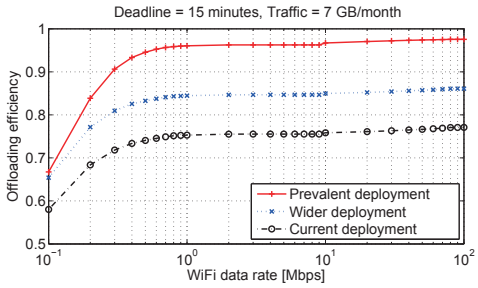


Figure 7: Offloading efficiency from three different deployment settings for the same deadline of 15 minutes and input traffic of 7 GB/month.

7. CONCLUDING REMARKS

In this paper, we provide an analytical framework that closely captures the realistic behavior of the delayed mobile data offloading system which can offload mobile traffic from cellular networks to WiFi networks (or similarly to femto/picocell networks). The analysis of the offloading system involves technical challenges due to its non-Markovian characteristics mainly resulting from the alternating heavy-tailed service and interruption periods as well as the deterministic reneging behavior. We address the challenges by uniquely designating the service activation events as embedded points, which transforms the system to a Markovian one. Using the redefined embedded process, we rigorously develop closed-form equations for the performance metrics of interest in the delayed offloading system, including offloading efficiency and mean packet delay. We expect that our analytical framework and the closed-form equations shown to almost perfectly predict the behaviors of the system, would significantly advance understandings on the offloading system and its variants. We also anticipate that the reverse engineering of the analytical framework would lead to optimization on the system as well as its control.

8. ACKNOWLEDGMENTS

This work was supported in part by the 2014 Research Fund of University of Ulsan and the Future Strategic Fund(1.14003.01) of UNIST (Ulsan National Institute of Science and Technology), and by Army Research Grant W911NF-12-1-0385 and National Science Foundation CNS-1012700.

9. REFERENCES

[1] Cisco visual networking index: Global mobile data traffic forecast update, 2012-2017, March 2013. http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-481360_ns827_Networking_Solutions_White_Paper.html.

[2] A. Aijaz, H. Aghvami, and M. Amani. A survey on mobile data offloading: technical and business perspectives. *IEEE Wireless Communications*, 20(2):104–112, 2013.

[3] A. Balasubramanian, R. Mahajan, and A. Venkataramani. Augmenting mobile 3G using WiFi. In *Proceedings of ACM MobiSys*, 2010.

[4] X. Bao, Y. Lin, U. Lee, I. Rimac, and R. Choudhury. Dataspotting: Exploiting naturally clustered mobile devices to offload cellular traffic. In *INFOCOM, 2013 Proceedings IEEE*, pages 420–424, April 2013.

[5] M. Barbera, J. Stefa, A. Viana, M. Dias de Amorim, and M. Boc. Vip delegation: Enabling vips to offload data in wireless social mobile networks. In *Proceedings of International Conference on Distributed Computing in Sensor Systems (DCOSS)*, 2011.

[6] D. Y. Barrer. Queuing with impatient customers and indifferent clerks. *Operations Research*, 5(5):644–649, 1957.

[7] D. Y. Barrer. Queuing with impatient customers and ordered service. *Operations Research*, 5(5):650–656, 1957.

[8] P. Bremaud. *Markov Chains*. Springer, 2008.

[9] A. Detti, M. Pomposini, N. Blefari-Melazzi, S. Salsano, and A. Bragagnini. Offloading cellular networks with information-centric networking: The case of video streaming. In *Proceedings of IEEE WoWMoM*, 2012.

[10] B. Doshi. Queueing systems with vacations: A survey. *Queueing systems*, 1(1):29–66, 1986.

[11] Y. Go, Y. Moon, G. Nam, and K. Park. A disruption-tolerant transmission protocol for practical mobile data offloading. In *Proceedings of ACM MobiOpp*, 2012.

[12] T. Han, N. Ansari, M. Wu, and H. Yu. On accelerating content delivery in mobile networks. *Communications Surveys Tutorials, IEEE*, 15(3):1314–1333, Third 2013.

[13] J. Kim and J. Kim. M/PH/1 queue with deterministic impatience time. *Commun. Korean Math. Soc.*, 28(2):383–396, 2013.

[14] Y. Kim, K. Lee, and N. B. Shroff. An analytical framework to characterize the efficiency and delay in a mobile data offloading system. Technical report, January 2014. available at http://msn.unist.ac.kr/papers/MobiHoc_2014_KLS_TR.pdf.

[15] V. Kone, H. Zheng, A. Rowstron, G. O'Shea, and B. Zhao. Measurement-based design of roadside content delivery systems. *Mobile Computing, IEEE Transactions on*, 12(6):1160–1173, June 2013.

[16] V. G. Kulkarni. *Modeling and Analysis of Stochastic Systems*. Chapman & Hall, Ltd., London, UK, UK, 1995.

[17] K. Lee, J. Lee, I. Rhee, Y. Yi, and S. Chong. Mobile data offloading : How much can WiFi deliver? In *Proceedings of ACM CoNEXT*, 2010.

[18] Y. Li, G. Su, P. Hui, D. Jin, L. Su, and L. Zeng. Multiple mobile data offloading through delay tolerant networks. In *Proceedings of ACM Workshop on Challenged Networks (CHANTS)*, 2011.

[19] M. Meeker. Internet trends - morgan stanley, April 2010. http://www.businessinsider.com/mary-meeker-mobile-internet?op=1.

[20] W.-B. Pöttner, J. Morgenroth, S. Schildt, and L. Wolf. Performance comparison of dtn bundle protocol implementations. In *Proceedings of ACM Workshop on Challenged Networks (CHANTS)*, 2011.

[21] S. Ross. *Stochastic Processes*. John Wiley & Sons, New York, 1996.

[22] K. L. Scott and S. Burleigh. Bundle protocol specification, rfc 5050, November 2007. http://tools.ietf.org/html/rfc5050.

[23] H. Takagi. *Queueing Analysis: A Foundation of Performance Evaluation (Volume 1: Vacation and priority Systems)*. Amsterdam: North Holland, 1991.

[24] W. Xiong, D. Jagerman, and T. Altiok. M/G/1 queue with deterministic reneging times. *Perform. Eval.*, 65(3-4):308–316, 2008.

[25] X. Zhuo, W. Gao, G. Cao, and S. Hua. An incentive framework for cellular traffic offloading. *Mobile Computing, IEEE Transactions on*, 13(3):541–555, March 2014.

APPENDIX

In Tables 1 and 2, we summarize the formulas for the transition probabilities $P^{\text{off}}_{\boldsymbol{i},\boldsymbol{j}}$ and $P^{\text{on}}_{\boldsymbol{i},\boldsymbol{j}}$, respectively. In the table, we use the following notations. First, $f_X(x) := \mathrm{P}(X = x)$ denotes the probability mass function of a discrete random variable X. Remind that L and R are random variables that represent the packet sizes and the WiFi data rates, respectively. Next, the functions $g_1(\cdot,\cdot)$, $g_2(\cdot,\cdot,\cdot)$, and $g_3(\cdot,\cdot)$ are concerned with work load in the WiFi

Table 1: Probability of $\boldsymbol{X}(t+1)=\boldsymbol{j}$ conditioned that $\boldsymbol{X}(t)=\boldsymbol{i}$ and $C(t)=0$

	Case I $q_1=0$ $\boldsymbol{i}=(0,0,0,e_1)$	Case II $q_1=1, w_1<T_{\text{out}}$ $\boldsymbol{i}=(1,w_1,u_1,w_1)$	Case III $q_1=1, w_1=T_{\text{out}}$ $\boldsymbol{i}=(1,T_{\text{out}},u_1,T_{\text{out}})$	Case IV $q_1\geq 2, w_1<T_{\text{out}}$ $\boldsymbol{i}=(q_1,w_1,u_1,e_1)$	Case V $q_1\geq 2, w_1=T_{\text{out}}$ $\boldsymbol{i}=(q_1,T_{\text{out}},u_1,e_1)$
$\boldsymbol{j}=(0,0,0,e_1+1)$	$1-f_{\tilde{A}}(e_1)$	0	0	0	0
$\boldsymbol{j}=(1,1,u_2,1)$	$f_{\tilde{A}}(e_1)f_L(u_2)$	0	$f_{\tilde{A}}(T_{\text{out}})f_L(u_2)$	0	0
$\boldsymbol{j}=(0,0,0,T_{\text{out}}+1)$	0	0	$1-f_{\tilde{A}}(T_{\text{out}})$	0	0
$\boldsymbol{j}=(1,w_1+1,u_1,w_1+1)$	0	$1-f_{\tilde{A}}(w_1)$	0	0	0
$\boldsymbol{j}=(2,w_1+1,u_1,1)$	0	$f_{\tilde{A}}(w_1)$	0	0	0
$\boldsymbol{j}=(q_1,w_1+1,u_1,e_1+1)$	0	0	0	$1-f_{\tilde{A}}(e_1)$	0
$\boldsymbol{j}=(q_1+1,w_1+1,u_1,1)$	0	0	0	$f_{\tilde{A}}(e_1)$	0
$\boldsymbol{j}=(q_1-1,w_2,u_2,e_1+1)$	0	0	0	0	$(1-f_{\tilde{A}}(e_1))$ $\cdot f_L(u_2)f_{V_2(\boldsymbol{i})}(w_2-1)$
$\boldsymbol{j}=(q_1,w_2,u_2,1)$	0	0	0	0	$f_{\tilde{A}}(e_1)f_L(u_2)$ $\cdot f_{V_2(\boldsymbol{i})}(w_2-1)$

Table 2: Probability of $\boldsymbol{X}(t+1)=\boldsymbol{j}$ conditioned that $\boldsymbol{X}(t)=\boldsymbol{i}$ and $C(t)=1$

	Case I $q_1=0$ $\boldsymbol{i}=(0,0,0,e_1)$	Case II $q_1=1, w_1<T_{\text{out}}$ $\boldsymbol{i}=(1,w_1,u_1,w_1)$	Case III $q_1=1, w_1=T_{\text{out}}$ $\boldsymbol{i}=(1,T_{\text{out}},u_1,T_{\text{out}})$
$\boldsymbol{j}=(0,0,0,e_1+1)$	$1-f_{\tilde{A}}(e_1)$	0	0
$\boldsymbol{j}=(1,1,u_2,1)$	$f_{\tilde{A}}(e_1)f_{L-R}(u_2)$	0	0
$\boldsymbol{j}=(0,0,0,1)$	$f_{\tilde{A}}(e_1)\mathrm{P}(L\leq R)$	0	0
$\boldsymbol{j}=(1,w_1+1,u_2,w_1+1)$	0	$(1-f_{\tilde{A}}(w_1))f_R(u_1-u_2)$	0
$\boldsymbol{j}=(0,0,0,w_1+1)$	0	$(1-f_{\tilde{A}}(w_1))\mathrm{P}(R\geq u_1)$	0
$\boldsymbol{j}=(2,w_1+1,u_2,1)$	0	$f_{\tilde{A}}(w_1)f_R(u_1-u_2)$	0
$\boldsymbol{j}=(1,1,u_2,1)$	0	$f_{\tilde{A}}(w_1)g_2(u_1,u_2,1)$	0
$\boldsymbol{j}=(0,0,0,1)$	0	$f_{\tilde{A}}(w_1)\mathrm{P}(R-L\geq u_1)$	0
$\boldsymbol{j}=(0,0,0,T_{\text{out}}+1)$	0	0	$1-f_{\tilde{A}}(T_{\text{out}})$
$\boldsymbol{j}=(1,1,u_2,1)$	0	0	$f_{\tilde{A}}(T_{\text{out}})\mathrm{P}(R-L<u_1)$
$\boldsymbol{j}=(0,0,0,1)$	0	0	$f_{\tilde{A}}(T_{\text{out}})\mathrm{P}(R-L\geq u_1)$

	Case IV $q_1\geq 2, w_1<T_{\text{out}}$ $\boldsymbol{i}=(q_1,w_1,u_1,e_1)$	Case V $q_1\geq 2, w_1=T_{\text{out}}$ $\boldsymbol{i}=(q_1,T_{\text{out}},u_1,e_1)$
$\boldsymbol{j}=(q_1,w_1+1,u_2,e_1+1)$	$(1-f_{\tilde{A}}(e_1))f_R(u_1-u_2)$	0
$\boldsymbol{j}=(q_2,w_2,u_2,e_1+1)$	$(1-f_{\tilde{A}}(e_1))g_2(u_1,u_2,q_1-q_2)f_{V_{q_1-q_2+1}(\boldsymbol{i})}(w_2-1)$	0
$\boldsymbol{j}=(0,0,0,e_1+1)$	$(1-f_{\tilde{A}}(e_1))g_3(q_1,u_1)$	0
$\boldsymbol{j}=(q_1+1,w_1+1,u_2,1)$	$f_{\tilde{A}}(e_1)f_R(u_1-u_2)$	0
$\boldsymbol{j}=(q_2,w_2,u_2,1)$	$f_{\tilde{A}}(e_1)g_2(u_1,u_2,q_1-q_2+1)f_{V_{q_1-q_2+2}(\boldsymbol{i})}(w_2-1)$	0
$\boldsymbol{j}=(1,1,u_2,1)$	$f_{\tilde{A}}(e_1)g_2(u_1,u_2,q_1)$	0
$\boldsymbol{j}=(0,0,0,1)$	$f_{\tilde{A}}(e_1)g_3(q_1+1,u_1)$	0
$\boldsymbol{j}=(q_1-1,w_2,u_2,e_1+1)$	0	$(1-f_{\tilde{A}}(e_1))g_1(u_1,u_2)f_{V_2(\boldsymbol{i})}(w_2-1)$
$\boldsymbol{j}=(q_2,w_2,u_2,e_1+1)$	0	$(1-f_{\tilde{A}}(e_1))g_2(u_1,u_2,q_1-q_2)f_{V_{q_1-q_2+1}(\boldsymbol{i})}(w_2-1)$
$\boldsymbol{j}=(0,0,0,e_1+1)$	0	$(1-f_{\tilde{A}}(e_1))g_3(q_1,u_1)$
$\boldsymbol{j}=(q_1,w_2,u_2,1)$	0	$f_{\tilde{A}}(e_1)g_1(u_1,u_2)f_{V_2(\boldsymbol{i})}(w_2-1)$
$\boldsymbol{j}=(q_2,w_2,u_2,1)$	0	$f_{\tilde{A}}(e_1)g_2(u_1,u_2,q_1-q_2+1)f_{V_{q_1-q_2+2}(\boldsymbol{i})}(w_2-1)$
$\boldsymbol{j}=(1,1,u_2,1)$	0	$f_{\tilde{A}}(e_1)g_2(u_1,u_2,q_1)$
$\boldsymbol{j}=(0,0,0,1)$	0	$f_{\tilde{A}}(e_1)g_3(q_1+1,u_1)$

queue and the data rates, and are defined as

$$\begin{aligned} g_1(u_1,u_2) &:= \mathrm{P}(R<u_1)f_L(u_2),\\ g_2(u_1,u_2,m) &:= \mathrm{P}(u_1+L_{m-1}^{\text{agg}}+L-R=u_2, L\geq u_2), \\ g_3(q,u) &:= \mathrm{P}(u+L_{q-1}^{\text{agg}}\leq R), \end{aligned} \tag{9}$$

where $L_m^{\text{agg}} := R_1+\ldots+R_m$ and R_k are i.i.d. copies of R. The closed-form formulas for the functions in (9) are determined by the distributions of L and R. Finally, the random variables $\tilde{A}$ and $V_s(\boldsymbol{i})=V_s((q,w,u,e))$ are related to the packet arrival time and are distributed as follows:

$$\begin{aligned} f_{\tilde{A}}(e) &= \mathrm{P}(A=e\,|\,A\geq e),\\ f_{V_s(\boldsymbol{i})}(v) &= \frac{\mathrm{P}(A_{s-1}^{\text{agg}}=w-v)\mathrm{P}(A_{q-s}^{\text{agg}}=v-e)}{\mathrm{P}(A_{q-1}^{\text{agg}}=w-e)}, \end{aligned}$$

where $A_m^{\text{agg}} := A_1+\ldots+A_m$ and A_k are i.i.d. copies of A. The closed-form formulas for $f_{\tilde{A}}(\cdot)$ and $f_{V_s(\boldsymbol{i})}(\cdot)$ are determined by the packet inter-arrival time distribution. For more details, please refer to Appendix C in our technical report [14].