Received: 12 March 2020 | Revised: 11 November 2020 | Accepted: 12 December 2020

DOI: 10.1049/ntw2.12012

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH PAPER

Delayed mobile data offloading scheme for quality of service traffic: Design and analysis

Anusree Ajith | **Tiruchirai Gopalakrishnan Venkatesh**

Department of Electrical Engineering, Indian Institute of Technology Madras, Chennai, India

Correspondence

Anusree Ajith, Department of Electrical Engineering, Indian Institute of Technology Madras, Sardar Patel Road, Adyar, Chennai, 600036, India.
Email: anusree.ajith@gmail.com

Abstract

As the world is entering the era of 5G, we are witnessing exponential growth in the volume of data traffic generated and consumed by mobile devices. Mobile data offloading handles the surge in mobile data traffic by offloading part of the traffic onto the Wi-Fi network. In this context, we design and analyse a quality of service (QoS) enhancement scheme for delayed mobile data offloading. We first consider prioritised queuing of traffic as a scheme for QoS provisioning. Using the concepts of virtual waiting time, renewal process, and level crossing arguments, we derive the average transmission delay of the offloaded and reneged packets of high priority. Next, we mathematically validate the benefit of balking all such packets whose expected virtual waiting time exceeds their deadline. Using a three-dimensional Markov chain model, we derive the required balking probability. Moreover, we merge the technique of balking with prioritisation for improving QoS in delayed offloading. Through an extensive simulation, we validate our analysis and demonstrate how the proposed scheme reduces the transmission delay without sacrificing offloading efficiency. Our investigation has the potential to be adopted in future mobile data offloading standards for improving QoS.

1 | INTRODUCTION

In recent years, there has been exponential growth in the use of wireless devices such as smartphones and tablets. The ever-increasing use of the Internet over these handheld devices has created a surge in data traffic. Global mobile data traffic was 4.4 exabytes per month in 2015; it grew at the rate of 63% and reached 7.2 exabytes per month at the end of 2016. Furthermore, it was expected to grow to 49 exabytes per month by 2021, a sevenfold increase over 2016 [1].

To meet the surge in the mobile data demand, use of Wi-Fi as a supplementary network to offload cellular traffic has been found to be a promising approach. Wi-Fi advocates for an unlicensed spectrum, which makes Wi-Fi attractive to telecom operators. Dual-mode phones use cellular radio, which contain global system for mobile communications as well as IEEE 802.11 (Wi-Fi) radio. The main advantages of mobile data offloading are that from the user's perspective, data offloading can be seen as a viable way to have a higher data rate and also being cost-effective. Apart from the high data rate, for all transfer sizes, data transfer in WLAN is significantly more power efficient than 3G [2]. This reduces the transmission time and hence the transmission power.

Two types of mobile data offloading have been discussed in the literature: on-the-spot and delayed offloading [3,4]. On-the-spot offloading takes place only when the user is within the Wi-Fi coverage area. When the user moves out of the Wi-Fi coverage area, offloading is stopped. In delayed mobile data offloading, offloading occurs whenever the Wi-Fi is available. When the user moves out of the Wi-Fi coverage area, the packets are not immediately routed through the cellular network. Transmission of data packets through the cellular network occurs only if the user does not reenter the Wi-Fi coverage area within a deadline. An efficient mobile data offloading scheme should have a low transmission delay and high offloading efficiency.

Offloading data traffic via an long-term evolution wireless local area network (LTE-WLAN) interworking has been specified for 3GPP release 10 [5]. Until that release, LTE-WLAN interworking was supported only in the core network. Because there was no tight integration between LTE and WLAN, offloading was mainly mobile-controlled. However, this helps the operator to reduce congestion in the backbone network. In 3GPP release 12 [6], Radio Access Network

(RAN)-assisted offloading was proposed. The focus of 3GPP release 13 [7] is operator-controlled mobile data offloading. LTE-WLAN Aggregation (LWA) and LTE-WLAN Radio Level Integration with IPsec Tunnel (LWIP) [8], as specified in release 13, allows the simultaneous combination of both cellular and Wi-Fi technologies to increase user equipment throughput. LWA standardises the aggregation of RAN and WLAN at the Packet Data Convergence Protocol(PDCP) layer, whereas in LWIP, the aggregation is done above the PDCP layer, at the internet protocol layer.

Motivation for the proposed work is as follows. Although much standardisation has been done in LTE-WLAN integration, the choice of the policy regarding which offloading decision has to be made is operator-dependent. Therefore, many efficient offloading schemes have been presented in the literature, and the current discussion falls into this category. To the best of our knowledge, the prioritisation and advantage of balking in delayed mobile data offloading have not been fully investigated in the literature.

In the present work we have designed and analysed a quality of service (QoS) enhancement scheme for delayed mobile data offloading. Our scheme is based on (1) prioritising QoS traffic, and (2) an intelligent balking decision, both geared towards reducing packet latency.

Major contributions are that:

1. We have developed a preemptive priority queuing model for a mobile data offloading scheme catering to traffic with different QoS requirements.
2. By analysing the stochastic process characterising the virtual waiting time $V_1(t)$ of the packets, we have derived a Volterra integral equation for the probability density function of $V_1(t)$.
3. Using the Laplace Stieltjes transform technique, we then solved the integral equation and arrived at the average transmission time of the offloaded packets and the reneged packets. Using these results we motivate the advantage for balking.
4. We derive the optimal balking probability of a packet by setting it to be equal to the probability that its virtual waiting time exceeds its deadline.
5. We develop a three-dimensional (3D) Markov chain model for delayed mobile data offloading with balking and priority. Solving the model, we derive the offloading efficiency of packets belonging to different priority classes.
6. Through extensive simulation, we validate our analysis and demonstrate how the proposed scheme reduces transmission delay without sacrificing much offloading efficiency.
7. The effect of finite cellular delay on our model is also analysed.

The rest of the article is organised as follows: In Section 2, we present a brief literature survey of analytical work related to delayed mobile data offloading and then highlight the unique contributions of our research. In Section 3, we validate the importance of balking for QoS enhancement in delayed offloading with multiple traffic classes. We study delayed offloading with balking for multiple traffic classes in Section 4. Results and discussions are presented in Section 5. Conclusions are drawn in Section 6.

2 | LITERATURE SURVEY

Insight into the potential benefits of mobile data offloading using trace-driven simulation was given by Lee et al. [9]. Traces show that 65% of traffic can be offloaded by incorporating on-the-spot offloading. Further, an additional 29% of traffic can be offloaded for a deadline of one hour. Sok-Ian Sou developed an analytical model to quantify the amount of 3G resources saved by offloading and deadline assurance for measuring the quality of user experience with policy and charging control support [10]. Mobile data offloading for LTE in the unlicensed spectrum has also been proposed [11]. A detailed survey on mobile data offloading can be found in Rebecchi et al. [3] and Zhou et al. [12] and references therein. Literature on the problem of mobile data offloading can be classified based on (1) the system model, (2) system parameters, (3) performance metrics, and (4) the analytical framework adopted, such as game theory [13–15], optimization [16–19], artificial intelligence/learning-based methods [20,21], and stochastic/queuing theory-based methods [9,23-25]. Table 1 provides the classification of some of the literature in the area of mobile data offloading.

Abbreviations: APs, Access points; BS, Base station; CSMA/CA, Carrier-sense multiple access with collision avoidance; MU, Moblie user; QoE, Quality of experience; SINR, Signal-to-interference-plus-noise ratio;

We now confine our literature survey to analytical models dealing with mobile data offloading. Analytical models based on queueing theory are prevalent in the literature. Lee et al. [9] developed a queueing model for delayed mobile data offloading along with service interruption and reneging. They derived offloading efficiency using matrix-geometric techniques [22]. Mehmeti et al. analysed delayed mobile data offloading using Z-transforms and derived optimal values of the deadline [23]. Ajith et al. [24] derived offloading efficiency and packet transmission delay for a delayed mobile data offloading scheme using the concept of virtual waiting time.

Yu et al. [4] explained operator-initiated offloading and user-initiated offloading schemes. The deployment of Wi-Fi can be treated as an independent Poisson point process and the user's movement as a semi-Markov process. Effective data offloading based on these processes was modelled in Hu et al. [25]. Xu et al. [26] analysed a delayed offloading scenario with multiple priority classes and derived packet delay and offloading efficiency. Cheung [16] posed single-user user-initiated network selection as a finite-horizon sequential decision problem to minimise cost for the case of a delayed Wi-Fi offloading scheme.

Sok-Ian Sou and Yi-Ting Peng showed that multipath Wi-Fi offloading has better resource use compared with opportunistic offloading [27]. Although delayed mobile data offloading has been widely investigated, the literature still lacks design and performance analysis of an offloading scheme that

TABLE 1 Classification of the literature on mobile data offloading

Theoretical Framework	Reference	System model	System parameters varied	Performance metrics studied	Remark
Game theory based	[13]	Mobile data operator (MDO),third-party APs, QoE aware mobile users	Price, number of APs	MDO profit, offloaded traffic, motivation for APs, social welfare	Contract problem, Stackelberg game
	[14]	Mobile network operator,small cell service providers	Uniform price, traffic volumes to offload	Average service prices	Nash equilibrium, topological infrastructure placement problem
	[15]	Vehicular ad hoc networks,optimal pricing strategy	Minimum unit price, data rate	Downloading time, utility of client and helper, knock down unit price	Stackelberg game
Optimization based	[17]	Macrocell base station and small-cell AP system parameters: traffic demand	Traffic demand	Optimal system cost, BS bandwidth usage, computational time	Joint optimization of BS bandwidth allocation, MUs traffic scheduling and power allocation
	[18]	Digital twin edge networks	Base station density, number of users, movement speed.	Average offloading latency, offloading failure rate, and service migration rate	Multiobjective dynamic optimization problem
	[19]	Cellular network with mobile nodes, and APs	Deadline	User preference of monetary cost, energy consumption	Discrete time optimal control problem, CEO
Artificial intelligence/ learning	[20]	Mobile edge computing environment, unmanned aerial vehicles (UAVs)	Number of iterations	UAV action probability, average offloaded data	Stochastic learning automata, reinforcement learning
	[21]	Internet of Things, mobile edge computing	Arrival rate, average cumulative reward	Total cost of consumed energy, bandwidth use, and buffer length	Computational offloading, deep reinforcement learning
Stochastic queuing/ theory	[26]	Mobile users under cellular and Wi-Fi coverage	Wi-Fi availability, deadlines	Transmission delay and offloading efficiency	Preemptive priority queuing analytic model, multilevel priority traffic
	[23]	Mobile users under cellular network and Wi-Fi coverage	Arrival rate. availability ratio, reneging rate	Transmission delay	Markov chain model
	[24]	Cellular network with Wi-Fi hot spots, mobile users	Wi-Fi availability ratio, queue length, packet deadline	Transmission delay, offloading efficiency	Balking scheme, QoE
	[25]	Poisson point process and CSMA/CA	SINR, intensity of AP	Expected offloaded traffic	Stochastic geometry

ensures QOS for different types of traffic. The purpose of this report is to fill this gap.

The unique contribution of our work and how it differs from existing related literature are as that:

1. Mehmeti et al. [23] analysed delayed mobile data offloading, but they do not consider the case of multiple-class traffic. We deal with multiclass traffic and show its benefit for QoS enhancement.
2. We derive the average transmission time of offloaded and reneged packets and mathematical validation of balking compared with the reference [24]. Furthermore, Ajith Venkatesh [24] deal with only single-class packets with no concept of priority. In the current work using Markov chain and priority queuing theory, we show how prioritisation along with balking improves QoS.
3. Our proposed idea of balking appropriate packet is equivalent to the optimal solution derived for the optimization problem in Cheung and Huang [16].
4. Our work differs from that of Xu et al. [26] because we consider multiple-class traffic in delayed offloading scenario and also show that balking can further improve QoS.
5. Unlike previous work in the literature, we derive the transmission delay of both offloaded and reneged packets. For the first time, the concept of a virtual waiting time and level crossing technique is used to analyse the formation of a queue in mobile data offloading.

3 | DELAYED OFFLOADING

In this section, we consider prioritising the traffic of a user as a technique to improve QoS in delayed offloading. We first derive and critically analyse the factors contributing to the average transmission delay of high-priority packets. This analysis should help us arrive at a QoS enhancement scheme that reduces transmission delay.

3.1 | System model

We consider a mobile node that roams randomly and therefore enters and leaves zones with Wi-Fi coverage area randomly. We consider that the network traffic can be categorised into two classes based on QoS requirements. The packet generation process of class 1 and class 2 traffic is modelled as a Poisson process with rate λ_1 and λ_2, respectively. Class 1 is given preemptive priority over class 2. We assume packet size to be exponentially distributed with mean κ for both the classes. Each packet has a deadline that is exponentially distributed with parameter γ_1 and γ_2 for class 1 and class 2 packets, respectively. All packets wait in the queue for transmission over WLAN. If transmission is not started before the expiry of deadline, the packet gets reneged and it is transmitted over the cellular network. Furthermore, we assume there is no queuing delay for transmission over the cellular network because it is usually assumed in the literature [3]. The effect of a finite cellular delay will be probed in Subsection 5.2. Let the Wi-Fi data rate be denoted as R_W. The availability of the Wi-Fi network is modelled as an ON-OFF alternating renewal process with ON and OFF periods being exponentially distributed with parameters α and β, respectively [28].

The whole system can be modelled as a queue with the following characteristics: (1) preemptive priority for class 1 over class 2, (2) packets renege on deadline expiration, (3) the server breakdown is governed by the ON-OFF process, and (4) server services are at rate $\mu = \frac{R_W}{\kappa}$ for both class 1 and class 2 when Wi-Fi is ON. Transmission delay and offloading efficiency are the metrics in which we are interested in delayed offloading. Transmission delay (T) is defined as the average queueing delay before packet transmission begins via either the cellular network or WLAN. Offloading efficiency (η) is the fraction of packets that are transmitted over WLAN.

3.2 | Derivation of transmission delay

In this section, we analyse the stochastic process characterising the virtual waiting time of class 1 packets using the level crossing method [29]. The analysis will help us to derive formulas for (1) the average transmission delay of offloaded class 1 packets, T_{S_1}, and (2) the average transmission delay of reneged class 1 packets, T_{R_1}.

Virtual waiting time [30] is defined as the time for which a hypothetical packet arriving at time t would have to wait before commencing service. Figure 1 shows typical realisations of $V_1(t)$, the virtual waiting time of class 1 packets, and $K(t)$, the Wi-Fi availability status. $V_1(t)$ has a value of 0 when queue is empty of class 1 packets and the server status is ON. The time interval during which $V_1(t)$ has a value of 0 is called the idle period (IP). While the queue is empty of class 1 packets ($V_1(t) = 0$), an initial jump in $V_1(t)$ can be due to either server breakdown (time instant m of Figure 1) or a packet arrival while the server is ON (time instant a_1 of Figure 1). The time interval when $V_1(t)$ has a value greater than 0 is called the busy period (BP). After the initial jump, the virtual waiting time decreases at rate $V_1'(t) = -1$ until it reaches 0 if there is no packet arriving within that duration. If a new packet arrives before $V_1(t)$ reaches 0, the virtual waiting time jumps by a value corresponding to the completion time of the new packet. The completion time of the packet, C, is defined as the time the instant service of a packet is started until the completion of the service, including all service unavailability periods as a result of server breakdown (denoted as C_i for ith class 1 packet). Because we follow preemptive priority for class 1 over class 2, the Laplace Stieltjes transform (LST) of density function of C has been derived in Ajith and Venkatesh [24] as:

$$L\{f_C(x)\}(s) = \frac{\mu}{s+\mu}\left[\frac{1-\frac{\alpha\beta}{(\mu+\alpha)(\mu+\beta)}}{1-\frac{\alpha\beta^2}{(\mu+\alpha)(\mu+\beta)(s+\beta)}}\right] \quad (1)$$

TABLE 2 Comparison with related work

	Our proposed model	Xu et al. [26]	Mehmeti and Spyropoulos [23]
System parameter used	$k_1 = k_2 = 2.5$ MB, $R_w = 2$ Mbps, $d_1 = 30$s, $d_2 = 5$ min, AR = 0.75 $\lambda_1 = 0.09$ packet/s $\lambda_2 = 0.1$	**Pedestrian case** $k_1 = k_2 = 2$ MB $R_w = 1.97$ Mbps $d_1 = 60$ min $d_2 = 30$ min AR = 0.75 $\lambda_1 = 0.5$, $\lambda_2 = 0.1$	**Vehicular case** $k_1 = k_2 = 2.27$ MB $R_w = 1.28$ Mbps $d_1 = 30$s, 60s AR = 0.11. Single class arrival rate is 0.1
Average transmission delay	From Figure 4a,b of current paper: delay of high-priority data is 10.8 s and delay of low-priority data is 75–150 s	From Figure 4a,b of Xu et al. [26]: delay of high-priority data and low-priority data is around 25 s (for $\lambda_2 = 0.1$)	From Figure 5 of Mehmeti and Spyropoulos [23]: for $d_1 = 30$ s, delay of data is 23 s. For $d_1 = 60$ s, delay of data is 42 s.
Offloading efficiency	From Figure 5b of current paper: offloading efficiency of class 1 packet is 44% and class 2 is 36% at $\lambda_2 = 0.1$	From Figure 9 of Xu et al. [26]: offloading efficiency of class 2 is 18% for AR = 0.1 and $\lambda_1 = 0.1$, deadline = 30–90 min.	
Reneging probability	From Figure 6a of current paper: probability of reneging is 0.23 for class 1 and 0.4 for class 2 at $\lambda_2 = 0.1$		From Table II of Mehmeti and Spyropoulos [23]: probability of reneging is 0.867 for deadline = 60 s

Because the completion time includes service unavailability periods, the virtual waiting time does not jump for OFF duration. However, the initial jump in virtual waiting time can be the result of service unavailability periods. The virtual waiting time of class 1 does not jump for the arrival of class 2 packets (b_1, b_2 and b_3) because class 1 packets have preemptive priority over class 2. Also, the virtual waiting time does not jump for class 1 packets that renege before the BP (a_2 and a_7), because they do not contribute to the virtual waiting time of future packet arrivals.

Figure 1 shows that the whole system can be modelled as an alternating regenerative process that alternates between *BP* and *IP*. The probability density function of the virtual waiting time of the class 1 packet, $f_{V_1}(x)$, can be written as:

$$f_{V_1}(x) = f_{V_1}(x|IP)P(IP) + f_{V_1}(x|BP)P(BP) \quad (2)$$

based on whether a packet arrival event occurs during *BP* or *IP*. Here, $P(IP)$ and $P(BP)$ are the probability that an arrival takes place in a IP or BP, respectively. According to the Poisson arrivals see time averages property [31], $P(IP)$ and $P(BP)$ can be formulated as:

$$P(IP) = \frac{E[IP]}{E[IP] + E[BP]}. \quad (3)$$

$$P(BP) = 1 - P(IP) \quad (4)$$

Because IP ends owing to a packet arrival or service interruption, $E[IP] = \frac{1}{\lambda_1 + \alpha}$. Here, $E[BP]$ has been derived in Equation (3) of Ajith Venkatesh [24]. Taking the Laplace transform on both side of Equation (2), the LST of $f_{V_1}(x)$ in Equation (2) is given by:

$$L\{f_{V_1}(x)\}(s) = L\{f_{V_1}(x \mid IP)\}(s)P(IP) + L\{f_{V_1}(x \mid BP)\}(s)P(BP) \quad (5)$$

Because any packets arriving during *IP* are served immediately,

$$L\{f_{V_1}(x \mid IP)\}(s) = 1. \quad (6)$$

To find the LST of $f_{V_1}(x)$, we have to derive the LST of $f_{V_1}(x|BP)$.

For further analysis, we use the level crossing method [29]. Consider the particular value $V_1(t) = x$ as a level as shown in Figure 1. A downcrossing of level x is a left-continuous hit of level x from above [29]. An upcrossingof level x is defined as the jump, which starts at a value below level x and ends at a value above level x. For example, an instant of downcrossing and upcrossing of level x can be seen at time h and time a_1, respectively in Figure 1. Downcrossing rate of level x is the number of downcrossings of level x per unit time. Similarly, the upcrossing rate of level x is the number of upcrossings of level x per unit time. The basic level crossing argument states that for every level x and every sample path, in the long run, the total downcrossing rate is equal to the total upcrossing rate [29].

The downcrossing rate of level x during *BP* is given by the probability density function (PDF) of V_1 conditioned on *BP*, $f_{V_1}(x|BP)$. The upcrossing rate of level x during *BP* can be calculated as the sum of the upcrossing rate of level x from level 0 and the upcrossing rate of level x from any level $y \in (0, x)$. Let *IJ* correspond to a random variable representing the initial jump of virtual waiting time from *IP*, and $F_{IJ}(x)$ be its distribution function. While queue is empty, an initial jump in $V_1(t)$ can either result from server breakdown or be due to a packet arrival while the server is ON. Therefore, $F_{IJ}(x)$ can be formulated as:

$$F_{IJ}(x) = \frac{\alpha}{\alpha+\lambda_1} F_Y(x) + \frac{\lambda_1}{\alpha+\lambda_1} F_C(x) \tag{7}$$

The upcrossing rate of level x during BP from level 0 is the rate at which the initial jump exceeds x. $\bar{F}_{IJ}(x)$ is the probability with which the initial jump exceeds level x. Because the initial jump occurs only once during BP, the upcrossing rate of level x during BP from level 0 is given by $\frac{\bar{F}_{IJ}(x)}{E[BP]}$.

In BP, the virtual waiting time jumps by a value corresponding to the completion time of the packet. Given v as the initial virtual waiting time (the starting point of the jump), for the jump to cross level x, the completion time must be greater than $x - v$. The probability of upcrossing level x from level v owing to a packet arrival is given by $P(C \geq x - v)$, which is the complementary CDF of C evaluated at $x - v$, $\bar{F}_C(x - v)$. A packet will stay in queue if its deadline, d, is greater than the virtual waiting time. $P(d \geq v)$ is the probability with which the packet will not renege, which is the complementary CDF of the deadline distribution evaluated at v, that is, $e^{-\gamma_1 v}$. Only packets that are not reneged change $V_1(t)$ upon their arrival. Therefore, $\bar{F}_C(x - v)e^{-\gamma_1 v}$ is the probability that upcrossing from level v crosses level x. $f_{V_1}(v|BP)$ is the probability that the virtual waiting time has value v over BP. Integrating v from 0 to x gives the probability of crossing level x from level $(0, x)$ over BP. Multiplying by the arrival rate gives the required upcrossing rate of level x during BP from any level $y \in (0, x)$.

Using the level crossing method [29], on equating the downcrossing rate to upcrossing rate, we get:

$$f_{V_1}(x|BP) = \frac{\bar{F}_{IJ}(x)}{E[BP]} + \lambda_1 \int_0^x \bar{F}_C(x - v)e^{-\gamma_1 v} f_{V_1}(v|BP)dv \tag{8}$$

This is a Volterra integral equation of the second kind [32] which can be solved by recursion. Similar to the analysis in Iravani and Balcıoglu [33], we get the LST of $f_{V_1}(x|BP)$ as:

$$L\{f_{V_1}(x|BP)\}(s) = \frac{L\{\bar{F}_{IJ}(x)\}(s)}{E[BP]} + \frac{\lambda_1 L\{\bar{F}_C(x)\}(s)\Phi_{m-1}(s+\gamma_1)}{E[BP]} \tag{9}$$

where $L\{\bar{F}_{IJ}(x)\}(s)$ and $L\{\bar{F}_C(x)\}(s)$ are the LST of complementary distribution of the initial jump and completion time, respectively, and m is the number of recursions. $\Phi_i(s)$ can be written as:

$$\Phi_1(s) = L\{\bar{F}_{IJ}(x)\}(s)$$
$$\Phi_m(s) = L\{\bar{F}_{IJ}(x)\}(s) + \lambda_1 L\{\bar{F}_C(x)\}(s)\Phi_{m-1}(s+\gamma_1) \tag{10}$$

Substituting Equations (3), (4), (6) and (9) into Equation (5) yields the LST of $f_{V_1}(x)$ as:

$$L\{f_{V_1}(x)\}(s) = \frac{(\lambda_1+\alpha)}{1+(\lambda_1+\alpha)E[BP]}\left[\frac{1}{\lambda_1+\alpha} + L\{\bar{F}_{IJ}(x)\}(s) + \lambda_1 L\{\bar{F}_C(x)\}(s)\Phi_{m-1}(s+\gamma_1)\right] \tag{11}$$

The LST of the distribution function of the transmission delay of the offloaded/served packet according to Stanford [34] can be derived as:

$$L\{F_{T_{S_1}}(x)\}(s) = \frac{L\{f_{V_1}(x)\}(s)\,|_{s=s+\gamma_1}}{s\eta_1} \tag{12}$$

where η_1 is the offloading efficiency of class 1 packets. According to Ajith and Venkatesh [24], this can be derived as:

$$\eta_1 = \frac{\mu}{\lambda_1}\frac{E[BP]}{E[BP]+E[IP]}\frac{\beta}{\beta+\alpha} \tag{13}$$

The average transmission delay of the offloaded/served packet, T_{S_1}, can be calculated by:

$$T_{S_1} = -L\{f_{T_{S_1}}(x)\}'(s)\,|_{s=0} \tag{14}$$

where:

$$L\{f_{T_{S_1}}(x)\}(s) = sL\{F_{T_{S_1}}(x)\}(s) + f_{T_{S_1}}(0).$$

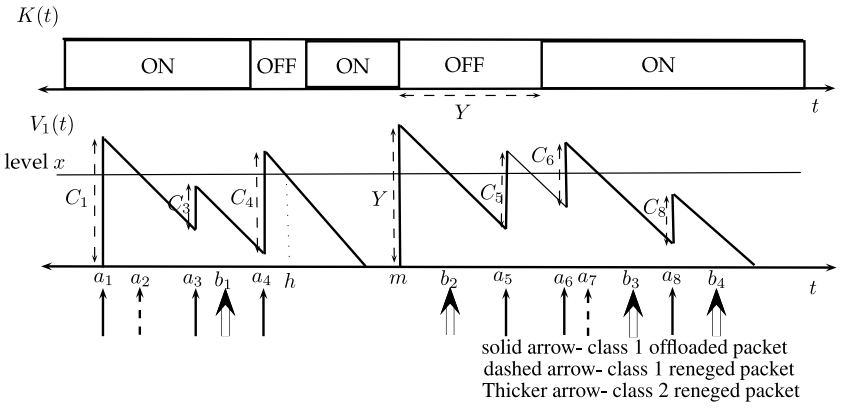

FIGURE 1 Sample realization of Wi-Fi availability, $K(t)$, and virtual waiting time of class 1, $V_1(t)$

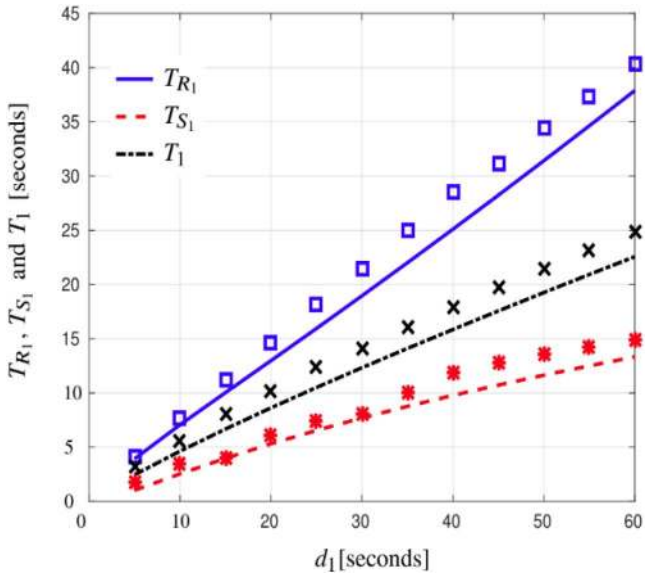

FIGURE 2 Transmission delay of class 1 reneged packets, T_{R_1}, transmission delay of class 1 served or offloaded packets, T_{S_1} and overall transmission delay of class 1 packets, T_1, for different class 1 deadline, d_1 (*Params*: $\lambda_1 = 0.09$/sec, $\varkappa_1 = \kappa_2 = 2.5$ MB, $R_W = 2Mbps$, $AR = 0.75$). Lines represent analytical results and the corresponding markers denote simulation results

The average transmission delay of class 1 packets in the mobile data offloading scheme, T_1, can be calculated as [34]:

$$T_1 = \frac{1-\eta_1}{\gamma_1}. \tag{15}$$

The relationship between T_{S_1},T_{R_1} and T_1 can be written as:

$$T_1 = T_{S_1}\eta_1 + T_{R_1}(1-\eta_1) \tag{16}$$

Therefore, the average transmission delay of the reneged packet, T_{R_1}, is:

$$T_{R_1} = \frac{T_1 - T_{S_1}\eta_1}{1-\eta_1} \tag{17}$$

Equations (14) and (17), which give the average transmission delay of offloaded and reneged packets, respectively, are the final results of the section.

3.3 | Results

To validate the analytical model, extensive simulations were carried out in MATLAB [35]. A node is simulated by generating Poisson traffic of class 1 and class 2. The Wi-Fi data rate is assumed to be 2 Mbps. Because delayed offloading could afford to transmit a larger packet than on-the-spot offloading, a packet length of 2.5 MB is assumed in the simulation. Wi-Fi connectivity is assumed to be available three-fourths of the time ($AR = 0.75$), as obtained in measurement studies in [9]. Because class 1 has preemptive priority over class 2, the arrival process of class 2 is immaterial in the analysis of transmission delay for class 1. In Figures 2–6, lines represent analytical results and the corresponding markers denote simulation results.

The transmission delay of offloaded class 1 packets, T_{S_1}, transmission delay of reneged class 1 packets, T_{R_1}, and transmission delay of class 1 packets, T_1 are plotted in Figure 2 for different deadlines of class 1. The transmission delay of reneged class 1 packets is larger than the transmission delay of offloaded class 1 packets. T_1 is a weighted sum of T_{R_1} and T_{S_1}. Thus, to reduce T_1 when T_{R_1} is higher than T_{S_1}, there is a need to balk appropriate packets. The reneged packets do not contribute to offloading efficiency. However, they increase the average queuing delay. If we could identify packets that are likely to be reneged, and if they are made to balk and routed to cellular network directly, the performance of the delayed offloading scheme can be improved.

4 | DELAYED OFFLOADING WITH BALKING AND PRIORITY

In this section, we analyse the system that includes prioritising traffic and appropriate balking to improve QoS in am delayed offloading scenario.

4.1 | System model

The system model is the same as the one in Section 3.1. The only change is that not all packets wait in the queue for Wi-Fi transmission; instead, the class j packet joins the queue with probability $1 - b^{(j)}_{(n_1,n_2,k)}$, where n_1 and n_2 are the number of class 1 and class 2 packets in the queue, and k is the Wi-Fi service status at the time of the arrival of the corresponding packet. $b^{(j)}_{(n_1,n_2,k)}$ is the balking probability of the packet, which is the probability with which the packet upon arrival has to leave the queue without waiting in the queue.

4.2 | Derivation of balking probability of a packet

To derive the balking probability, $b^{(j)}_{(n_1,n_2,k)}$, we first derive the distribution function of the virtual waiting time, $V_1(t)$ and $V_2(t)$, of class 1 and class 2, respectively.

Consider a packet whose arrival time is t' with deadline d. Let the tuple $N(t') = [n_1\ n_2]$, denotes the status of the queue, where n_1 and n_2 indicate the number of class 1 and class 2 packets in the queue at t'. Let $K(t') = k \in \{c, w\}$ be the Wi-Fi service status, where w denotes the time for which Wi-Fi and cellular connectivity are available, and c denotes the time for which only cellular connectivity is available. Let $V_1(t')$ and $V_2(t')$ be the virtual waiting time of class 1 and class 2 traffic, respectively.

Let $T_0^{(j)}$ ($j \in \{1, 2\}$ denotes the class) be the time from instant t' until one of the packets in the queue enters the

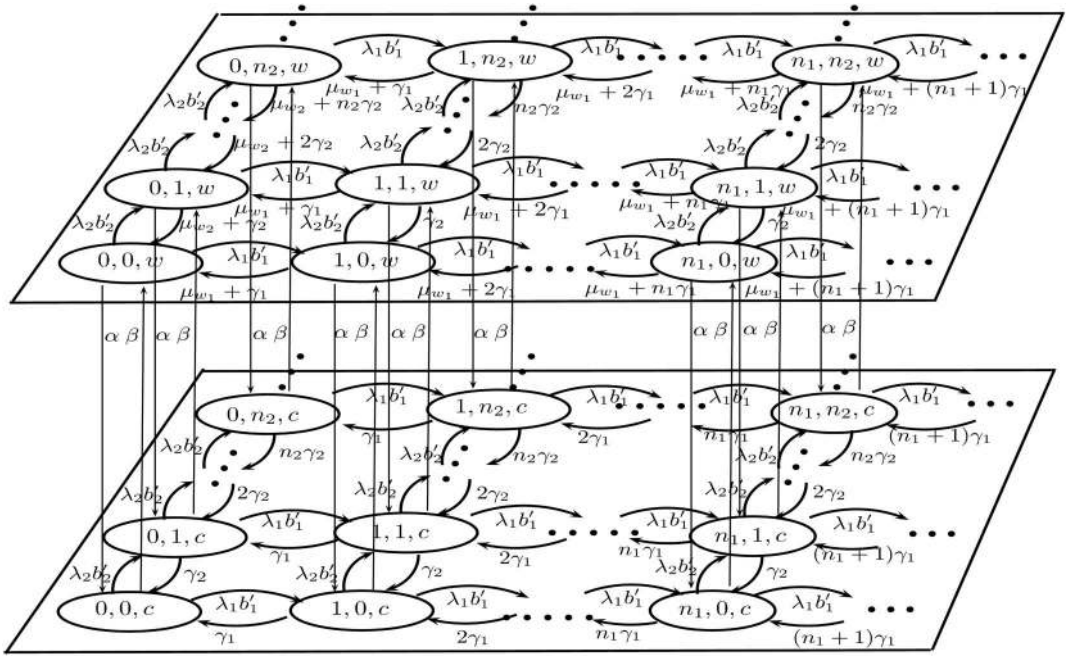

FIGURE 3 3D Markov chain model for delayed mobile data offloading with balking and priority. b_j' corresponds to $b_{(j,n_1,n_2,k)}$

service by reducing number of packets in the queue by 1, if queue is not empty. In this case, $T_0^{(j)}$ is equal to Y, if $k = c$, and equal to C, if $k = w$. If the queue is empty, $T_0^{(j)}$ is the time from instant t' until the packet that has arrived enters into service. $T_0^{(j)}$ is then equal to Y, if $k = c$, and equal to 0, if $k = w$. From this discussion, $T_0^{(j)}$ can be written as:

$$T_0^{(j)} = \begin{cases} Y, & \text{if } k = c \\ C, & \text{if } k = w, \quad \textit{queue non - empty} \\ 0, & \text{if } k = w, \quad \textit{queue empty} \end{cases}$$

$T_i^{(1)}$ is the time from the instant the previous packet leaves the system until the ith class 1 packet leaves the system, $1 \le i \le n_1$. Similarly, $T_i^{(2)}$ is the time from the instant the previous packet leaves the system until the ith class 2 packet leaves the system, $1 \le i \le n_2$:

$$T_i^{(j)} = \begin{cases} C & w.p. \quad P(H_{j,i} > C) \\ 0 & w.p. \quad P(H_{j,i} < C) \end{cases}$$

Here, $H_{j,i} \sim \text{Exp}(i\gamma_j)$, is the reneging rate of class j when the number of class j packets waiting in the queue is i. Also,

$$P(H_{j,i} < C) = \int_0^\infty f_C(x) F_{H_{j,i}}(x) dx$$

and:

$$P(H_{j,i} > C) = 1 - P(H_{j,i} < C)$$

Here $f_C(x)$ is the PDF of C and $F_{H_{j,i}}(x)$ is the cumulative distribution function of $H_{j,i}$.

Given $N(t') = [n_1\ n_2]$ and $K(t') = k$, we get:

$$V_1(t') = T_0^{(1)} + \sum_{i=1}^{n_1} T_i^{(1)} \tag{18}$$

$$V_2(t') = T_0^{(2)} + \sum_{i=1}^{n_2} T_i^{(2)} + \sum_{i=1}^{n_1} T_i^{(1)} \tag{19}$$

Taking the Laplace transform on both side of Equations (18) and (19), the LST of the virtual waiting time, V_j of class j packet, $j \in \{1, \text{two}\}$ given $N(t') = [n_1\ n_2]$ and $K(t') = k$, is:

$$\begin{aligned} &L_{V_1}(s \mid (N(t') = [n_1\ \ n_2], K(t') = k)) \\ &\quad = L_{T_0}^{(1)}(s) \prod_{i=1}^{n_1} L_{T_i}^{(1)}(s) \end{aligned} \tag{20}$$

$$\begin{aligned} &L_{V_2}(s \mid (N(t') = [n_1\ \ n_2],\ K(t') = k)) \\ &\quad = L_{T_0}^{(2)}(s) \prod_{i=1}^{n_2} L_{T_i}^{(2)}(s) \prod_{i=1}^{n_1} L_{T_i}^{(1)}(s) \end{aligned} \tag{21}$$

where $L_{T_i}^{(j)}(s)$ is the LST of $T_i^{(j)}$:

$$L_{T_0}^{(j)}(s) = \begin{cases} \dfrac{\beta}{s+\beta}, & \text{if } k = 0 \\ L_C(s), & \text{if } k = 1 \quad \textit{queue non - empty} \\ 1, & \text{if } k = 1, \quad \textit{queue empty} \end{cases}$$

$$L_{T_i}^{(j)}(s) = L_C(s)P(H_{j,i} > C) + 1P(H_{j,i} < C) \quad (22)$$

The conditional CDF of the virtual waiting time, V_j of the class j packet, can be derived as:

$$F_{V_j}(v_j \mid (N(t') = [n_1 n_2], K(t') = k)) = L^{-1}\left\{\frac{L_{V_j}(s \mid (N(t') = [n_1 n_2], K(t\prime) = k))}{s}\right\}(x)\,|_{x=v_j} \quad (23)$$

Because the number of class 1 and class 2 packets in the queue are n_1 and n_2, respectively, and the Wi-Fi service status is k, the probability that V_j is less than d' is given by:

$$F_{V_j}(v_j \mid (N(t') = [n_1 n_2], K(t') = k))|_{v_j = d'}$$

Algorithm 1 Balking

Variables:
- Class $j \in \{1,2\}$; ▹ 1-high priority, 2- low priority
- Status $K = k \in \{c, w\}$; ▹ c-Wifi OFF, w-Wifi ON
- Deadline d;
- Queue length1 n_1; ▹ Length of Class 1 queue
- Queue Length2 n_2; ▹ Length of Class 2 queue
- b; ▹ Balking Probability

procedure $Balk\_OR\_NOT((j, n_1, n_2, k, d))$
  $N(t') = [n_1\ n_2]$
  $K(t') = k$
  Switch(j)**do**
    Case1 :
      $V_1(t') = T_0^{(1)} + \sum_{i=1}^{n_1} T_i^{(1)}$
      $L_{V_1}(s \mid (N(t'), K(t'))) = L_{T_0}^{(1)}(s) \prod_{i=1}^{n_1} L_{T_i}^{(1)}(s)$

$$F_{V_1}(v_1 \mid (N(t'), K(t'))) = L^{-1}\left\{\frac{L_{V_1}(s \mid (N(t'), K(t')))}{s}\right\}(x)\,|_{x=v_1}$$

      If $(k = w)$
        $b = 1 - F_{v_1}(d \mid N(t') = [n_1, n_2], k = w)$
      elseif $k = c$
        $b = 1 - F_{v_1}(d \mid N(t') = [n_1, n_2], k = c)$
      endif
    Case2 :
      $V_2(t') = T_0^{(2)} + \sum_{i=1}^{n_2} T_i^{(2)} + \sum_{i=1}^{n_1} T_i^{(1)}$

$$L_{V_2}(s \mid (N(t'), K(t'))) = L_{T_0}^{(2)}(s) \prod_{i=1}^{n_2} L_{T_i}^{(2)}(s) \prod_{i=1}^{n_1} L_{T_i}^{(1)}(s)$$

$$F_{V_1}(v_1 \mid (N(t'), K(t'))) = L^{-1}\left\{\frac{L_{V_1}(s \mid (N(t'), K(t')))}{s}\right\}(x)\,|_{x=v_1}$$

      If $(k = w)$
        $b = 1 - F_{v_2}(d \mid N(t') = [n_1, n_2], k = w)$
      elseif $k = c$
        $b = 1 - F_{v_2}(d \mid N(t') = [n_1, n_2], k = c)$
      endif
  endSwitch
  if$(b < rand(0,1))$
    if$(j = 1)$
      Join Class 1 queue;
    else
      Join Class 2 queue;
    endif
  else
    Balk;
  endif
endProcedure

Then, the balking probability of the class j packet, $b_{(j,n_1,n_2,k)}$, can be formulated as:

$$b_{(j,n_1,n_2,k)} = 1 - F_{V_j}(d' \mid (N(t\prime) = [n_1 n_2], K(t') = k)). \quad (24)$$

Algorithm 1, Balking, is the algorithm that every mobile node executes to make a balking decision. The algorithm takes as input (1) the class of the packet, (2) the deadline of the packet, (3) the queue length of the high- and low-priority queues, and (4) the Wi-Fi status at the time of arrival. Upon execution of the algorithm, the packet either joins the queue corresponding to its class or balks.

4.3 | Derivation of transmission delay and offloading efficiency

Based on the system model in Section 4.1, delayed offloading with balking and priority can be modelled as a 3D Markov chain with states (n_1, n_2, k), as shown in Figure 3. Here, $n_1 \in \{0, 1, \ldots\}$ denotes the number of class 1 packets waiting in the queue, $n_2 \in \{0, 1, \ldots\}$ denotes the number of class 2 packets waiting in the queue, and $k \in \{c, w\}$ denotes status of the Wi-Fi service. When the system is in state (n_1, n_2, k), class j packet arrival to the Wi-Fi queue happens with the rate, $\lambda(1 - b_{(j,n_1,n_2,k)})$, where $b_{(j,n_1,n_2,k)}$ is as given in Equation (24).

Let $\pi_{(n_1,n_2,k)}$ denote the steady-state probability of state (n_1, n_2, k). Values of $\pi_{(n_1,n_2,k)}$ can be found thus: Let R be the rate matrix corresponding to the Markov chain, and π be the steady-state vector for R with elements as $\pi_{(n_1,n_2,k)}$. To find the steady-state vector, notice that $\pi R = 0$ [28]. Therefore, $R^T \pi^T = 0$, and to solve for π, we have to obtain the null space of R^T. After finding a vector in null space of R, a scalar multiple of that, which satisfies $\sum_{k=c,w} \sum_{n_1=1}^{\infty} \sum_{n_2=1}^{\infty} \pi_{(n_1,n_2,k)} = 1$, gives π. Given the steady-state probabilities of states, we can determine the average number of class j packets in the queue as:

$$E[N_j] = \sum_{k=c,w} \sum_{n_1=1}^{\infty} \sum_{n_2=1}^{\infty} n_j \pi_{(n_1,n_2,k)} \quad (25)$$

The reneging rate of class j packets is the mean number of class j packets that get reneged per unit time, which is given by

$$R_j = \gamma_j E[N_j] \quad (26)$$

The balking rate of class j packets is the mean number of class j packets that do not join the queue per unit time:

$$B_j = \lambda \sum_{k=c,w} \sum_{n_1=1}^{\infty} \sum_{n_2=1}^{\infty} b_{(j,n_1,n_2,k)} \pi_{(n_1,n_2,k)} \quad (27)$$

We can calculate the reneging and balking probability of class j packet as:

$$P_{R_j} = \frac{R_j}{\lambda_j} \tag{28}$$

$$P_{B_j} = \frac{B_j}{\lambda_j} \tag{29}$$

Finally, the offloading efficiency of class j is given by:

$$\eta_j = 1 - P_{R_j} - P_{B_j} \tag{30}$$

and the transmission delay of class j is given by:

$$T_j = \frac{E[N_{Q_j}]}{(\lambda - B_j)}(1 - P_{B_j}) \tag{31}$$

where we assume balked packets contribute zero to transmission delay and use Little's law [36] with effective arrival rate $\lambda - B_j$.

4.4 | Exact analysis to virtual waiting time of class 2

In deriving Equations (18) and (19), we made an exponential approximation for the PDF of the completion time. $V_2(t')$ of the tagged class 2 packet as given in Equation (19) must include a term that accounts for the arrival of a high-priority class 1 packet that arrives after the arrival of a tagged class 2 packet but is served before the tagged packet. It can be accounted for by introducing a correction term, W_{add}, whose derivation is as follows.

$V_2(t')$ of the tagged class 2 packet as given in Equation (21) must include a term that accounts for the arrival of a high-priority class 1 packet that arrives after the arrival of a tagged packet but is served before the tagged packet. It can be accounted for by the term G, which is the BP generated by the class 1 packet during $V_2(t')$. Let $V_2(t') = x$. The Laplace transform of G given x can be derived as follows: Let N be the number of class 1 packets that arrive during time duration x and still present until the end of x (the number of packets arrived minus the number of packets reneged). According to Stanford et al. [34], it can be considered as an M/M/∞ queue with arrivals occurring at rate λ_1, according to a Poisson process, and move the process from state i to $i + 1$. Service times have an exponential distribution with parameter γ_1, and there are always sufficient servers such that every arriving job is served immediately. Transitions from state i to $i - 1$ are at rate $i\gamma_1$. Then, the PMF of N given x can be derived as [28]:

$$P(N = j \mid x) = e^{-\frac{\lambda_1}{\gamma_1}(1-\gamma_1 x)} \frac{\left[\frac{\lambda_1}{\gamma_1}(1 - \gamma_1 x)\right]^j}{j!} \tag{32}$$

Then, the Laplace transform of G given x is given by:

$$L_G(s \mid x) = \sum_{j=0}^{\infty} L_j(s) P(N = j \mid x) \tag{33}$$

where $L_j(s)$ is the BP with j class 1 packets and is as given by Equation (2) of Iravani and Balcıoglu [33]. Therefore, the Laplace transform of G is given by:

$$L_G(s) = \int_0^{\infty} L_G(s \mid x) v_2(x) dx \tag{34}$$

where:

$$v_2(x) = L^{-1}\left\{L_{V_2}(s \mid (N(t\prime) = [n_1 \;\; n_2], K(t\prime) = k)) = L^{(2)}_{T_0}(s)\right\}(x)$$

Because calculation of Equation (34) is computationally complex, a correction term can be found as follows. Let W_{add} be the average time required to serve class 1 packets that arrive after the arrival of the tagged class 2 packet but are served before class 1 owing to their preemptive priority. Let n_1 and n_2 be the number of packets in the queue at the time of the tagged class 2 arrival. Then, W_{add} can be approximated as (without considering reneging of class 1 packets):

$$W_{add} = \left(\frac{n_1}{\mu_1} + \frac{n_2}{\mu_2}\right)\frac{\lambda_1}{\mu_1} + W_{add}\frac{\lambda_1}{\mu_1} \tag{35}$$

Therefore,

$$W_{add} = \frac{\rho_1}{1 - \rho_1}\left(\frac{n_1}{\mu_1} + \frac{n_2}{\mu_2}\right) \tag{36}$$

We then introduce the correction term to deadline d as $d' = d - W_{add}$, where W_{add} is the correction term added to incorporate the class 1 packet that arrived after the tagged class 2 packet. Therefore, W_{add} is as given by Equation (36) for $j = 2$ and is equal to zero for $j = 1$.

5 | Results and discussion

To validate the analytical model, extensive simulations were carried out in MATLAB [35]. A node is modelled to generate the Poisson traffic of class 1 and class 2. The Wi-Fi data rate is assumed to be 2 Mbps. A packet length of 2.5 MB is assumed. Wi-Fi connectivity is assumed to be available three-fourths of

the time ($AR = 0.75$). We consider the case in which high-class priority has a short deadline in terms of seconds, and low-class priority has a deadline in terms of minutes. Note that *MDO* corresponds to a standard variant of delayed mobile data offloading, and *MDOB* corresponds to the proposed scheme, mobile data offloading with balking.

The transmission delay for class 1 and class 2 for different λ_2 is plotted in Figures 4a,b, respectively. Figure 4a shows that class 1 transmission delay is not affected by the class 2 arrival rate for both schemes because class 1 has preemptive priority over class 2. Although class 1 has a delay tolerance of 30 s, the transmission delay is much lower in prioritising class 1 traffic, as seen in Figure 4a. However, class 2 transmission delay increases with an increase in the class 2 arrival rate for both schemes, as seen in Figure 4b. Even the high-priority class 2 packets have a delay tolerance of 5 min, and the transmission delay is less than the delay tolerance in the range of λ_2, as shown in Figure 4b.

The transmission delay for MDOB is lower than transmission delay curve for MDO for both classes and for all values of λ_2. Hence, the transmission delay of both the classes can be reduced upon including balking. Our results are supported by the conclusion drawn in Cheung and Huang [16]. The optimal solution derived in that work [16] showed that when there are many packets waiting to be transmitted or when the deadline is close, the user should start transmitting via a cellular network immediately to minimise the overall cost.

Figure 5 shows the offloading efficiency of class 1 and class 2 packets. Although class 1 is given preemptive priority over class 2, the offloading efficiency of class 1 is not high. This is because class 1 has a shorter deadline of 30 s. Upon increasing λ_2, the offloading efficiency of class 1 is not affected because it has preemptive priority over class 2. However, the offloading efficiency of class 2 decreases. The offloading efficiency curve for MDOB is slightly lower than that of the MDO scheme for class 1. However, the offloading efficiency curve for MDOB is slightly higher than the corresponding curve of MDO for class 2. This shows that for balking appropriate packets, although there is a slight decrease in offloading efficiency for class 1, there is an increase in offloading efficiency for class 2 as well.

The reneging and balking probability of class 1 and class 2 is plotted in Figure 6. The reneging probability of class 1 is unaffected by λ_2, and the reneging probability of class 2 increases upon increasing λ_2 for both MDO and MDOB. The reneging probability is lower for MDOB compared with MDO for both classes.

5.1 | Comparison with related works

Although a number of works are reported in the literature, as Table 1 shows, the references differ from one another in (1) system model and system parameters considered, (2) performance metrics analysed, (3) the analytical framework adopted, and (4) overall goal of the paper. Therefore, we confine comparison of our work with two closely related works [23,26]. Nonetheless, some differences in the system model and parameters probed exist [23,26]. As a result, we can make only qualitative comparisons of our work against those authors [23,26] and draw general inferences. The average delay (10.8 s) of high-priority data in our model is less compared with the delay of high- and low-priority data (25 s) of Xu et al. [26], and also with an average delay (23 to 42 s) of single-class data of Mehmeti and Spyropoulos [23]. This is achieved through prioritisation; as a result, the delay of low-priority data of our model is larger. Our model achieves a higher offloading efficiency η (44 and 36) compared with that of Xu et al. [26] is 18. Although the availability ratio of that work [26] is only 0.1 (which reduces η), the deadline of 30 to 90 min of Xu et al. [26] favours higher η. Most important, the reneging rate of our model (0.23 for class 1 and 0.4 for class 2) is much lower compared with that of 0.867 of Mehmeti and Spyropoulos [23].

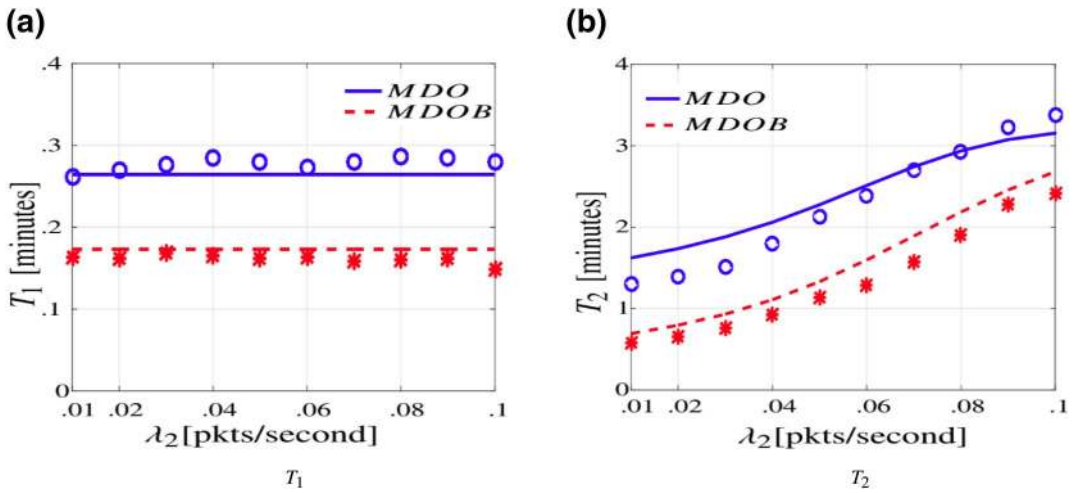

FIGURE 4 Transmission delay for class 1 and class 2 for different λ_2 (Parameters: $\lambda_1 = 0.09$ packets/s, $\varkappa_1 = \kappa_2 = 2.5$ MB, $R_W = 2$ Mbps, $d_1 = 30$ s, $d_2 = 5$ min, $AR = 0.75$). Lines (analysis) and corresponding symbols are simulation results. (a)T_1 and (b)T_2

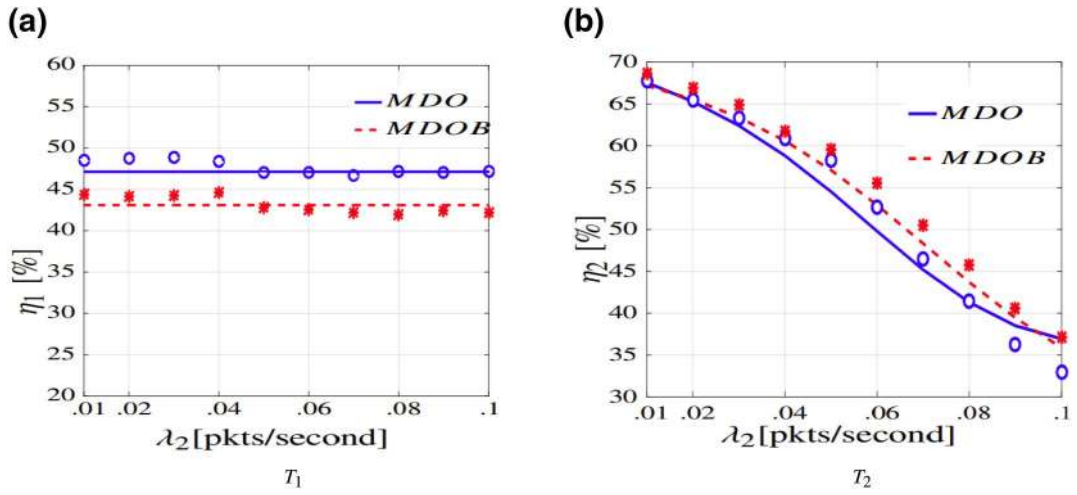

FIGURE 5 Offloading efficiency for class 1 and class 2 for different λ_2. Lines (analysis) and corresponding symbols are simulation results. (a)T_1 and (b)T_2

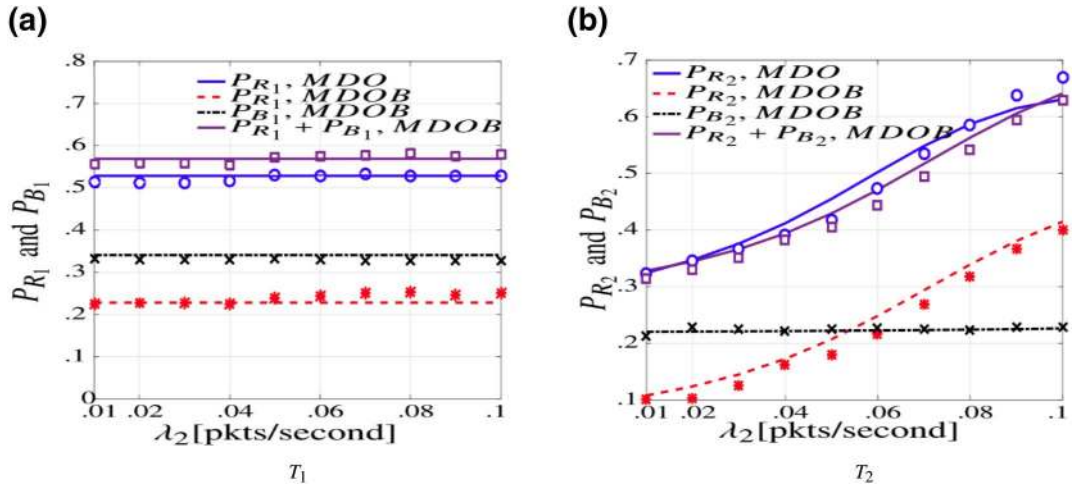

FIGURE 6 Reneging and balking probability for class 1 and class 2 for different λ_2. Lines (analysis) and corresponding symbols are simulation results. (a)T_1 and (b)T_2

5.2 | Effect of finite cellular delay

We now investigate the effect of considering a finite cellular delay. Assume that files transmitted over the cellular network incur a fixed delay $D_{cellular}$ ($D_{cellular}$ is the ratio between packet size κ and $R_{Cellular}$), capturing queueing delays over the cellular network. Let $T_{j,MDO}, j \in \{1, \text{two}\}$ denote the transmission delay of class j packet in the Wi-Fi queue without considering cellular delay. Then, in standard delayed offloading, the transmission delay is given by $T_{j,MDO} + P_{R_j} D_{cellular}$. In the proposed mobile data offloading with balking, the transmission delay is given by $T_{j,MDOB} + (P_{R_j} + P_{B_j}) D_{cellular}$. We show in the simulation as well as theoretically that $T_{j,MDOB}$ is smaller than $T_{j,MDO}$ in Figure 4. Furthermore, we have shown that P_{R_j} for standard delayed offloading is almost equal to $P_{R_j} + P_{B_j}$ for our scheme in Figure 6; thus, we are not seeing a reduction in offloading efficiency even after incorporating balking. Therefore, $T_{j,MDOB} + (P_{R_j} + P_{B_j}) D_{cellular}$ is smaller than $T_{j,MDO} + P_{R_j} D_{cellular}$. Thus, even if we consider a finite cellular delay, our proposed scheme has a lower transmission delay.

6 | Conclusion

We presented a delayed mobile data offloading scheme to support QoS traffic. We considered a set of mobile nodes under the coverage of cellular and Wi-Fi technology that carry packets with two priority levels. We have modelled the mobile nodes as an M/G/1 queue with an ON-OFF server along with reneging and balking. We validate the need for balking for QoS enhancement in delayed offloading by deriving a transmission delay of offloaded and reneged packets separately. A balking probability was derived. Finally, delayed offloading with balking along with a multiclass prioritisation scheme was analysed. We demonstrated that the average transmission delay of the high-priority packets can be substantially reduced without sacrificing

offloading efficiency. Our investigation has the potential to improve QoS in future mobile data offloading standards.

ACKNOWLEDGEMENT

The authors sincerely acknowledge the help offered by Mr. Ankit Kumar Gupta, Department of Electrical Engineering, Indian Institute of Technology, Madras.

ORCID

Tiruchirai Gopalakrishnan Venkatesh https://orcid.org/0000-0003-0676-512X

REFERENCES

1. Cisco Visual Networking Index: Global mobile data traffic forecast update (2016–2021) White paper, Cisco (2017) www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.pdf
2. Balasubramanian, N., Balasubramanian, A., Venkataramani, A.: Energy consumption in mobile phones: a measurement study and implications for network applications. In: Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement, Chicago, Illinois, USA, pp. 280–293 (2009)
3. Rebecchi, F., et al.: Data offloading techniques in cellular networks: a survey. Commun. Surveys Tuts. 17, 580–603 (2015)
4. Yu, H., et al.: Mobile data offloading for green wireless networks. IEEE Wirel. Commun. 24(4), 31–37 (2017)
5. Integration of Cellular and Wi-Fi Networks, White paper, 4G Americas (2015). http://www.5gamericas.org/files/3114/0622/2546/Intgration_of_Cellular_and_WiFi_Networks_White_Paper-_9.25.13.pdf
6. Understanding 3GPP Release 12 (2015). https://www.3gpp.org/ftp/Information/WORK_PLAN/Description_Releases/
7. Executive Summary: Inside 3GPP Release 13. White paper, 4G Americas (2015) http://www.5gamericas.org/files/4314/7700/6698/Inside_3GPP_Release_13_Understanding_the_Standards_for_LTE_Advanced_Enhancements_Final.pdf
8. LTE AggregationUnlicensed Spectrum. White paper, 4G Americas (2015)
9. Lee, K., et al.: Mobile data offloading: how much can Wi-Fi Deliver? IEEE/ACM Trans. Netw. 21, 536–550 (2013)
10. Sok-Ian, S.: Mobile data offloading with policy and charging control in 3GPP core network. IEEE Trans. Veh. Technol. 62(7), 3481–3486 (2013)
11. Chen, Q., et al.: Rethinking mobile data offloading for LTE in unlicensed spectrum. IEEE Trans. Wirel. Commun. 15(7), 4987–5000 (2016)
12. Zhou, H., et al.: A survey on mobile data offloading technologies. IEEE Access. 6, 5101–5111 (2018)
13. Song, X., et al.: Incentive framework for mobile data offloading market under QoE-aware users. IET Commun. 14(13), 2151–2161 (2020)
14. Li, M., Tony, Q.S., Courcoubetis, C.: Mobile data offloading with uniform pricing and overlaps. IEEE Trans. Mobile Comput. 18(2), 348–361 (2018)
15. Yang, F., et al.: Stackelberg-game-based mechanism for opportunistic data offloading using moving vehicles. IEEE Access. 7, 166435–166450 (2019)
16. Cheung, M.H., Huang, J.: DAWN: delay-aware WiFi offloading and network selection. IEEE J. Sel. Areas Commun. 33(6), 1214–1223 (2015)
17. Wu, Y., et al.: Optimal resource allocations for mobile data offloading via dual-connectivity. IEEE Trans. Mobile Comput. 17(10), 2349–2365 (2018)
18. Sun, W., Wang, R., Zhang, Y.: Reducing offloading latency for digital twin edge networks in 6G. IEEE Trans. Veh. Technol. 69(10), 12240–12251 (2020)
19. Yang, C., Radu, S.: CEO: cost-aware energy efficient mobile data offloading via opportunistic communication. In: 2020 International Conference on computing, Networking and Communications (ICNC). IEEE, Big Island, Hawaii, USA (2020)
20. Fragkos, G., et al.: Artificial intelligence empowered UAVs data offloading in mobile edge computing. In: ICC 2020-2020 IEEE International Conference on Communications (ICC), Virtual Conference Communications Enabling Shared Understanding. pp. 1–7. IEEE (2020)
21. Ke, H., et al.: Joint optimization of data offloading and resource Allocation with renewable energy aware for IoT devices: a deep reinforcement learning approach. IEEE Access. 7, 179349–179363 (2019)
22. Neuts, M.F.: Matrix Geometric Solutions in Stochastic Models: An Algorithmic Approach. The John Hopkins University Press, Baltimore (1981)
23. Mehmeti, F., Spyropoulos, T.: Performance modelling, analysis, and optimization of delayed mobile data offloading for mobile users. IEEE/ACM Trans. Netw. 25(1), 550–564 (2017)
24. Ajith, A., Venkatesh, T.G.: QoS enhanced mobile data offloading with balking. IEEE Commun. Lett. 21(5), 1143–1146 (2017)
25. Hu, Z., et al.: Stochastic-Geometry-based performance analysis of delayed mobile data offloading with mobility prediction in dense IEEE 802.11 networks. IEEE Access. 5, 23060–23068 (2017)
26. Xu, H., et al.: Performance analysis of delayed mobile data offloading with multi-level priority. In: IEEE 27th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), pp. 1–6.Vale (2016)
27. Sou, S., Peng, Y.: Performance modelling for multipath mobile data offloading in cellular/Wi-Fi networks. IEEE Trans. Commun. 65(9), 3863–3875 (2017)
28. Ross, S.: Introduction to probability models, 10th ed. Elsevier (2010)
29. Brill, P.H.: Level crossings methods in stochastic models. Springer, Berlin (2008)
30. Bae, J., Kim, S., Lee, E.Y.: The virtual waiting time of the M/G/1 queue with impatient customers. Queueing Syst. 38(4), 485–494.(2001)
31. Medhi, J.: Stochastic models in queueing theory, 2nd ed. Academic press (2002)
32. Polyanin, A.D., Manzhirov, A.V.: Handbook of integral equations. CRC press (2008)
33. Iravani, F., Balcıoglu, B.: On priority queues with impatient customers. Queueing Syst. 58(4), 239–260 (2008)
34. Stanford, R.E.: Reneging phenomena in single channel queues. Math. Oper. Res. 4(2), 162–178 (1979)
35. Matlab 9.0.0.341360, The MathWorks, Inc., Natick (R2016a)
36. Little, J.D.C.: A proof of the queueing formula L = λW. Oper. Res. 9(3), 383–387(1961)

How to cite this article: Ajith A, Venkatesh TG. Delayed mobile data offloading scheme for quality of service traffic: Design and analysis. *IET Netw.* 2021;1–13. https://doi.org/10.1049/ntw2.12012