Cacheable and Non-Cacheable Traffic Interplay in a Relay-Assisted Wireless Network

Ioannis Avgouleas, Nikolaos Pappas, and Vangelis Angelakis

Department of Science and Technology, Linköping University

Norrköping, 60174, Sweden

E-mails: {ioannis.avgouleas, nikolaos.pappas, vangelis.angelakis}@liu.se

Abstract—We study a discrete-time wireless network that serves both cacheable and non-cacheable traffic with assistance of a relay node with storage capabilities for both types of traffic. We investigate how allocating the storage capacity to cacheable and non-cacheable traffic affects the network throughput. Our numerical results provide useful insights by varying not only the allocation of cacheable to non-cacheable storage but also the rate by which non-cacheable content is transmitted, the rate by which cacheable content is requested, as well as different popularity distributions of the cached files.

I. INTRODUCTION

Wireless data traffic grew immensely over the past 10 years and is expected to comprise 20 percent of total IP traffic by 2022 with 1.5 mobile-connected device per capita. Almost three-fifths of traffic will be offloaded from cellular networks to Wi-Fi by the same year [1]. Mobile devices represented a large part of the total wireless traffic with wireless video being one of the main sources of wireless data traffic. Moreover, the introduction of high quality video formats such as 4K, 8K, 360° etc. will further contribute to degraded user experience due to increased delay and congestion. Video streaming services are interested in mitigating such performance issues by offloading content closer to the users [2].

Additionally, the advent of the Internet of Things (IoT) will necessitate serving a massive amount of devices with limited resources of energy, memory and computation. Thus, the attention of the research community is moving towards effectively supporting IoT communications. For instance, Named Data Networking (NDN) is an information-centric Internet architecture which has been recently considered as an enabling technology for IoT, due to its innovative features like named-based routing and in-network caching [3]. NDN allows for caching at intermediate nodes which proves to be effective in mitigating the network delay and traffic as well as the load on content producers [4].

Caching has been quite successful recently in reducing cellular traffic and delay as well as increasing network throughput and reliability. Cache-enabled 5G wireless systems and future network architectures will benefit from caching in terms of reduced costs for the network infrastructure and the quality of service available to the end users. A cache can typically store a small subset of the files library because of its limited capacity. Therefore, caching policies are necessary to decide which files are placed into the cache as well as which files

to evict from the cache when the cache is full and a new file should be cached. Many content placement strategies have been proposed in the literature e.g., caching the least frequently used content [5], caching the most popular content everywhere [5], probabilistic caching [6], cooperative caching [7], and geographical caching [8]. The role of caching for future communication systems is analyzed in [9].

Caching policies usually assume a model for file requests which, in practice, is a priori unknown and time-varying [10]. Additionally, there are caching architectures whose caches receive a low number of requests and, thus, realize request processes with high non-stationary popularity [11]. However, such models are challenging to fit and depend on strong assumptions about the popularity distribution. The authors in [12] develop a class of policies that make no assumptions on the file request distribution and, hence, is robust to popularity deviations by adjusting their caching decisions when the popularity model changes.

Typical performance criteria for a caching policy include the cache hit ratio (or probability), the density of successful receptions [6], energy efficiency [13] and the traffic load of the wireless links [14], among others. Additionally, a considerable amount of contemporary works consider throughput and/or delay with caching helpers [15]–[17]. To reduce delay, many works mitigate the backhaul or transmission delay under the assumption that traffic or requests are saturated. Works that do not make this assumption, but assume stochastic arrivals have also appeared e.g., [18].

The rise of wireless networks serving a massive amount of devices, such as 5G or IoT networks, will give birth to new traffic patterns. For example, traffic generated from Machine to Machine (M2M) devices is generally different compared to traditional smartphone traffic [19], [20]. These findings along with the aforementioned proliferation of caching techniques at the network edge suggest that understanding traffic is key to designing and optimizing the performance of future networking architectures.

A. Our work

In this paper, we study a wireless system that serves both cacheable and non-cacheable traffic. A relay node partially assists the non-cacheable transmissions by queueing the non-cacheable packets that failed to be transmitted to the destination. The queued packets are intended to be transmitted from

Fig. 1: An example configuration of our system model. User U_1 sends non-cacheable traffic to the destination user D with the assistance of relay node R. The storage capabilities of R can be split between cacheable and non-cacheable traffic. In case U_1's transmissions to D fails, the failed packet is stored at the relay's queue Q given that there is a successful transmission from U_1 to R. User U_2 has a cache and requests cached files from external resources with some probability. The relay node can serve U_2 given that it has the requested file and it is not serving U_1. Otherwise, the requested cached file can be fetched from the data center DC through the base station BS. The data center hosts the entire library of files and is available to serve the requests cached files with some probability.

the relay to the destination in a subsequent time slot. Moreover, when the relay node is not assisting the non-cacheable pair, it is available to serve cached files to another wireless user within in its coverage. The wireless user that requests cached content can also be served by a data center in case the relay misses the requested file or is not available for caching. The data center is assumed to contain the file library and is connected directly to a wireless base station through a backhaul link. Files from the data center are fetched to the cached user through the base station if the data center is available to serve cached files.

We analyze the network throughput considering the rate by which non-cacheable traffic is transmitted to the destination as well as the rate by which the relay attempts transmissions for non-cacheable traffic. In our numerical results, we vary the storage capacity dedicated to non-cacheable traffic and, thus, to cacheable traffic, to gain useful insights into the throughput of such systems and introduce the first step for the understanding of larger network topologies.

II. SYSTEM MODEL

A. Network Model

We consider the following network system: a user device U_1 serving non-cacheable traffic to a destination node D with the assistance of a wireless relay node R, and another user, U_2, requesting cached files in case of a local cache miss. The requested cached file can be served by external resources i.e., from the relay node R or the data center DC. We assume that packets and files are equally sized, so, thereafter, we use the terms packet and files interchangeably. The topology of the studied system can be found in Fig. 1.

We assume slotted time and that a packet transmission takes exactly one time-slot. Nodes have random access to the wireless medium with no coordination between them

regarding transmissions' scheduling. Thus, nodes attempt transmissions to the channel with some probability. An acknowledgment mechanism is assumed such that instantaneous and error-free acknowledgment/ negative-acknowledgment (ACK/NACK) packets are sent by the receiver over a separate channel. As a result, when D successfully receives a packet from U_1, the latter removes it from its buffer and is ready to attempt transmission of the next packet (in the next time slot) and the relay R discards it from its queue (if it has successfully received it). When R successfully receives a packet that did not reach D, node U_1 discards it from its buffer and is ready to attempt transmission of the next packet. The evolution of the relay's queue is analysed in Section III-A.

Additionally, the relay node R does not generate packets on its own and is equipped with a FD transceiver i.e., it can receive and transmit a packet within the same time slot. Its purpose is two-fold: (i) forward non-cacheable packets to the destination node, and (ii) serve cached files to user U_2. In each time slot, the relay is available to serve either non-cacheable or cacheable traffic. For that purpose, it hosts F files that can be used for serving both types of traffic. Non-cacheable incoming packets to R are stored in its queue with size for B packets. The rest of the storage capacity is devoted to cached files for user U_2. Consequently, the cache at the relay can hold $F - B$ files to help user U_2's file requests from external resources. The operation of the relay's cache is described in Section II-D.

The data center can be accessed through a wireless base station (BS) and stores the library of files i.e., all files that U_2 might request. We model DC's availability with probability α to model the fact that it can be out of service due to serving other users, failure, maintenance etc. If the DC is always available to U_2, then $\alpha = 1$. On the other hand, if the DC is not available for U, then $\alpha = 0$.

B. Transmission Model

In each time slot, U_1 attempts to transmit non-cacheable traffic to D with probability q_1 and the relay R can serve non-cacheable traffic to D with probability q_R. Thus, it is available to serve cacheable traffic to U_2 with probability $1 - q_R$. Moreover, in each time slot, U_2 requests a cached file from external resources with probability q_U. Therefore, if the relay node serves D with non-cacheable traffic, then R interferes

TABLE I: Notation table

Notation	Description
q_1	probability of U_1 attempting non-cacheable transmission.
q_R	probability of R being available to serve D.
q_U	probability of U_2 requesting a cached file from external resources (R or BS).
p_h	probability of cache hit at R.
α	probability of DC being available to serve U_2 requests.
$P_{i \to j}$	success probability of link $i \to j$, when node i is the only transmitter.
$P_{i \to j/T}$	success probability of link $i \to j$, when i and nodes in T are transmitting.

with the transmission from U_1 to D and any transmission from BS to U_2 which happens when U_2 has requested a file which will be served by the data center DC. On the other hand, if the relay node serves U_2 with cacheable traffic, then it only interferes with the transmission from U_1 to D. We summarize the aforementioned events and notation in Table I.

C. Physical Layer Model

We assume Rayleigh fading for the wireless channel and that a packet transmission from node i to node j is successful if and only if the link Signal-to-Interference-plus-Noise power ratio (SINR) between node i and j exceeds a minimal threshold θ. The received power at node j when i transmits is $P_{rx}(i,j) = A(i,j)h(i,j)$, where $A(i,j)$ is a unit-mean exponentially distributed random variable and the received power factor is: $h(i,j) = P_{tx}(i)/r(i,j)^p$, where $P_{tx}(i)$ is the power measured at $1\ m$ away from the transmitting antenna of node i, $r(i,j) \geq 1\ m$ is the distance in m between i and j, and p is the path-loss exponent.
The success probability of link $i \to j$, with $\mathcal{T}$ denoting the set of transmitting nodes, is given by [?]:

$$P_{i \to j/\mathcal{T}} = exp\left(-\theta \frac{n_j}{h(i,j)}\right) \prod_{k \in \mathcal{T}\setminus\{i,j\}} \left(1 + \theta \frac{h(k,j)}{h(i,j)}\right)^{-1},$$

where n_j is the noise power at receiver j.

D. Caches' Operation

We assume the content placement is given and that the cached nodes i.e., U_2 and R, follow the Collaborative Most Popular Content (CMPC) policy. According to the latter, user U_2 stores the first M_U most popular files in its cache, the relay node caches the next $F - B$ most popular files, and the data center hosts all files that U_2 might request. When the user node requests for a file that is not stored in its most popular files, it first probes R for it. The relay node serves U_2 if it is available for caching i.e., when not serving U_1, and has the requested file. Otherwise, user U_2 requests the file from the data center. If the latter is available for U_2 (which happens with probability α), then the file is fetched by the data center. We assume that the information exchange required for the operation of the CMPC policy e.g., the cache size of each device and the content placement in each device, is negligible.

Moreover, we consider a finite content library of N files with f_i denoting the i-th most popular file. For the sake of simplicity, we assume that all files have equal size and that access to cached files happens instantaneously. The request probability of the f_i is given by: $p_i = \Omega/i^\delta$, where $\Omega = (\sum_{j=1}^N j^{-\delta})^{-1}$ is the normalization factor and δ is the shape parameter of the Zip law which determines the correlation of user requests. As a result, the probability that user U_2 requests a file that is not located in its cache is: $q_U = 1 - \sum_{i=1}^{M_U} p_i$, and the cache hit probability at the relay node R is given by: $p_h = \sum_{i=M_U+1}^{M_U+F-B} p_i$, where F and B are the storage capacity and the queue size at the relay node, respectively.

III. ANALYSIS

In this section, we present the analysis for the states of the relays' queue and the throughput of the system in Fig. 1.

A. Relay's Queue Analysis

Let B denote the buffer size of the relay's queue Q. The latter follows the first-come-first-serve (FCFS) discipline. Moreover, we assume that when the queue is full i.e., holds B packets, and a new packet arrives at the relay, the relay rejects its arrival and acknowledgments user U_1 to attempt re-transmitting that packet in a subsequent time slot. We consider that this acknowledgment is instantaneous and error-free. The evolution of the Discrete Time Markov Chain (DTMC) of Q is shown in Fig. 2.

Fig. 2: Markov Chain of the relay's queue.

The transition matrix that models the DTMC above is given by the following stochastic column matrix:

$$\mathbf{P} = \begin{pmatrix} \bar{a}_1 & b_0 & & & \\ a_1 & b_1 & \ddots & & \\ & b_2 & \ddots & b_0 & \\ & & \ddots & b_1 & b_0 \\ & & & b_2 & \bar{b}_0 \end{pmatrix},$$

where $\bar{q} = 1 - q$. The entries of P are given by:

$$a_1 = \mathbb{P}(\text{``Q increases by 1 packet when Q=0''}) =$$
$$= q_1 \bar{q}_U (1 - P_{1 \to D}) P_{1 \to R}$$
$$+ q_1 q_U p_h (1 - P_{1 \to D/R}) P_{1 \to R}$$
$$+ q_1 q_U \bar{p}_h \alpha (1 - P_{1 \to D/BS}) P_{1 \to R/BS}$$
$$+ q_1 q_U \bar{p}_h \bar{\alpha} (1 - P_{1 \to D}) P_{1 \to R},$$

$$b_0 = \mathbb{P}(\text{``Q decreases by 1 packet when Q>0''}) =$$
$$= q_R \bar{q}_1 \left(\bar{q}_U P_{R \to D} + q_U \bar{p}_h (\alpha P_{R \to D/BS} + \bar{\alpha} P_{R \to D})\right)$$
$$+ q_R q_1 \bar{q}_U \times$$
$$\left[P_{R \to D/1}\left(P_{1 \to D/R} + (1 - P_{1 \to D/R})(1 - P_{1 \to R})\right)\right]$$
$$+ q_R q_1 q_U \bar{p}_h \left(\alpha \left[P_{R \to D/1,BS}\left(P_{1 \to D/R,BS} + \right.\right.\right.$$
$$\left.\left.(1 - P_{1 \to D/R,BS})(1 - P_{1 \to R/BS})\right)\right] +$$
$$\left.\bar{\alpha}\left[P_{R \to D/1}\left(P_{1 \to D/R} + (1 - P_{1 \to D/R})(1 - P_{1 \to R})\right)\right]\right),$$

$$b_1 = 1 - b_0 - b_2 = \mathbb{P}(\text{``Q does not change''}),$$

$$b_2 = \mathbb{P}(\text{``Q increases by 1 packet when Q>0''}) =$$
$$= q_1 q_R \bar{q}_U (1 - P_{R \to D/1})(1 - P_{1 \to D/R}) P_{1 \to R}$$
$$+ q_1 q_R q_U \bar{p}_h \alpha (1 - P_{R \to D/1,BS}) \times$$
$$(1 - P_{1 \to D/R,BS}) P_{1 \to R/BS}$$

$$+ \; q_1 q_R q_U \bar{p}_h \bar{\alpha}(1 - P_{R \to D/1})(1 - P_{1 \to D/R})P_{1 \to R}$$
$$+ \; q_1 \bar{q}_R [\bar{q}_U (1 - P_{1 \to D})P_{1 \to R} +$$
$$q_U p_h (1 - P_{1 \to D/R})P_{1 \to R}]$$
$$+ \; q_1 \bar{q}_R q_U \bar{p}_h \alpha(1 - P_{1 \to D/BS})P_{1 \to R/BS}$$
$$+ \; q_1 \bar{q}_R q_U \bar{p}_h \bar{\alpha}(1 - P_{1 \to D})P_{1 \to R},$$

To derive the steady state distribution $\pi = [\pi_0, \cdots, \pi_B]^T$, we need to solve the balance equations: $\mathbf{P}\pi = \pi$ which produce the following relation for calculating the probability of being in state i:

$$\pi_i = \rho^{i-1} t_0 \pi_0, \; \forall \, 1 \le i \le B, \text{ and } \pi_0 = \left[1 + t_0 \left(\frac{1 - \rho^B}{1 - \rho}\right)\right]^{-1},$$

where: $\rho = b_2/b_0$ and $t_0 = a_1/b_0$.

B. Throughput Analysis

First, we derive the direct throughput from user U_1 to the destination node D and the relayed throughput from the relay node R to D with the intention of deriving the throughput of D. Then, we formulate the cacheable throughput seen by user U_2. Recall that user U_1 and the relay attempt transmissions with probabilities q_1 and q_R, respectively. Moreover, U_2 requests a file from external resources with probability q_U, and p_h is the probability of a cache hit at the relay's cache. Also, α denotes the probability with which the data center is available to serve file requests of U_2 and $P(Q > 0)$ denotes the probability that Q i.e., the queue at the relay, is not empty (please see Table I for the description of our notation). The *direct throughput from user U_1 to D* is:

$$T_{1 \to D} = q_1 q_R P(Q > 0) \big[q_U \bar{p}_h \alpha P_{1 \to D/R,BS}$$
$$+ \; q_U \bar{p}_h \bar{\alpha} P_{1 \to D/R} + \bar{q}_U P_{1 \to D/R} \big]$$
$$+ \; q_1 [1 - q_R P(Q > 0)] \big[\bar{q}_U P_{1 \to D} + q_U p_h P_{1 \to D/R}$$
$$+ \; q_U \bar{p}_h \alpha P_{1 \to D/BS} + q_U \bar{p}_h \bar{\alpha} P_{1 \to D} \big].$$

The *relayed throughput from R to D* is given by:

$$T_R = q_1 P(Q = 0) \big[\bar{q}_U (1 - P_{1 \to D})P_{1 \to R}$$
$$+ q_U p_h (1 - P_{1 \to D/R})P_{1 \to R}$$
$$+ q_U \bar{p}_h \alpha(1 - P_{1 \to D/BS})P_{1 \to R/BS}$$
$$+ q_U \bar{p}_h \bar{\alpha}(1 - P_{1 \to D})P_{1 \to R} \big]$$
$$+ q_1 P(0 < Q < B) q_R \big[\bar{q}_U (1 - P_{1 \to D/R})P_{1 \to R}$$
$$+ q_U \alpha(1 - P_{1 \to D/R,BS})P_{1 \to R/BS}$$
$$+ q_U \bar{\alpha}(1 - P_{1 \to D/R})P_{1 \to R} \big]$$
$$+ q_1 P(0 < Q < B)\bar{q}_R \times$$
$$\big[\bar{q}_U (1 - P_{1 \to D})P_{1 \to R} +$$
$$q_U p_h (1 - P_{1 \to D/R})P_{1 \to R} + q_U \bar{p}_h \times$$
$$\big(\alpha(1 - P_{1 \to D/BS})P_{1 \to R/BS} + \bar{\alpha}(1 - P_{1 \to D})P_{1 \to R} \big) \big]$$
$$+ q_1 P(Q = B) q_R \big[\bar{q}_U P_{R \to D/1}(1 - P_{1 \to D/R})P_{1 \to R}$$

$$+ q_U \alpha P_{R \to D/1,BS}(1 - P_{1 \to D/R,BS})P_{1 \to R/BS}$$
$$+ q_U \bar{\alpha} P_{R \to D/1}(1 - P_{1 \to D/R})P_{1 \to R} \big].$$

The *non-cacheable throughput seen by D* is given by:

$$T_D = T_{1 \to D} + T_R.$$

The *cacheable throughput seen by U_2* is given by:

$$T_2 = q_U q_R P(Q > 0)\bar{p}_h \alpha [q_1 P_{BS \to 2/R,1} + \bar{q}_1 P_{BS \to 2/R}]$$
$$+ \; q_U [1 - q_R P(Q > 0)] q_1 [p_h P_{R \to 2/1} + \bar{p}_h \alpha P_{BS \to 2/1}]$$
$$+ \; q_U [1 - q_R P(Q > 0)] \bar{q}_1 [p_h P_{R \to 2} + \bar{p}_h \alpha P_{BS \to 2}].$$

IV. NUMERICAL RESULTS

In this section, we present numerical evaluations of the analysis in Section III. The transmission power of each device and the distances between nodes are set as per Table II. Please notice that we use the same SINR θ for every wireless link in our system and the same noise power n for every receiver.

TABLE II: Wireless links parameters for our numerical results.

Parameter	Value
$r(1, 2)$	$100 \; m$
$r(1, BS)$	$100\sqrt{2} \; m$
$r(1, D)$	$100 \; m$
$r(1, R)$	$50\sqrt{2} \; m$
$r(BS, 2)$	$100 \; m$
$r(D, BS)$	$100 \; m$
$r(R, 2)$	$50\sqrt{2} \; m$
$r(R, BS)$	$50\sqrt{2} \; m$
$r(R, D)$	$50\sqrt{2} \; m$

Parameter	Value
P_1	$1 \; mW$
P_R	$2 \; mW$
P_{DC}	$10 \; mW$
n	$10^{-11} \; W$
p	4
θ	$0 \text{ or } 5 \text{ dB}$

Parameter	Value	Parameter	Value
θ	$0 \; dB$	θ	$5 \; dB$
$P_{1 \to D}$	0.368	$P_{1 \to D}$	0.042
$P_{1 \to D/R}$	0.041	$P_{1 \to D/R}$	0.002
$P_{1 \to D/BS}$	0.033	$P_{1 \to D/BS}$	0.001
$P_{1 \to D/R,BS}$	0.004	$P_{1 \to D/R,BS}$	0
$P_{1 \to R}$	0.779	$P_{1 \to R}$	0.454
$P_{1 \to R/BS}$	0.071	$P_{1 \to R/BS}$	0.014
$P_{R \to D}$	0.883	$P_{R \to D}$	0.674
$P_{R \to D/1}$	0.784	$P_{R \to D/1}$	0.483
$P_{R \to D/BS}$	0.392	$P_{R \to D/BS}$	0.136
$P_{R \to D/1,BS}$	0.349	$P_{R \to D/1,BS}$	0.098
$P_{BS \to 2}$	0.905	$P_{BS \to 2}$	0.729
$P_{BS \to 2/1}$	0.823	$P_{BS \to 2/1}$	0.554
$P_{BS \to 2/R}$	0.503	$P_{BS \to 2/R}$	0.207
$P_{BS \to 2/1,R}$	0.457	$P_{BS \to 2/1,R}$	0.157
$P_{R \to 2}$	0.883	$P_{R \to 2}$	0.674
$P_{R \to 2/1}$	0.784	$P_{R \to 2/1}$	0.483

Regarding caching, we assume that the cache of node U_2 hosts M_U files, and that the relay node stores $F = 10$ files for both types of traffic. Its queue has finite size B for cacheable traffic and, hence, R holds $F - B$ files the non-cacheable traffic. We also assume that the whole library (at the data center) holds $N = 10000$ files. The caches follow the Collaborative Most Popular Content (CMPC) policy (which we describe in Section II-D). The random availability of the data center for U_2 was set to $\alpha = 0.7$. We summarize the cache parameters in Table III. In the following results, we vary $B \in [0, F]$ to gain insight into its effect on the throughput at the destination nodes (U_2 and D) and the distribution of the packets at the queue of the relay R.

A. Throughput T_D, T_2, and T vs. Queue Size B

In this section, we study how B i.e., the storage at the relay R for non-cacheable traffic affects T_D, T_2, and T i.e., the non-cacheable throughput at destination node D, the cacheable throughput at user node U_2, and the network throughput given by: $T = T_D + T_S$, respectively. We present two cases in each plot: (i) $\delta = 0.5$ and $M_U = 5$, and (ii) $\delta = 1.2$ and $M_U = 0$. It should be noted that q_U i.e., the probability of requesting content from external resources decreases with δ for fixed M_U, and $q_U = 1$ for $M_U = 0$ (see Section II-D). In Fig. 3(a) and (b), we plot the aforementioned throughputs versus B for $q_1 = 0.4$ and $q_1 = 0.8$, respectively. We observe that T_D is increased with q_1 since increasing q_1 results in U_1 attempting transmissions more frequently. As a result, more interference to U_2 is realized from U_1 and R, and, hence, T_2, is decreased with q_1. Additionally, we observe that the network throughput T is decreased with q_1.

Furthermore, when $q_1 = 0.4$, we observe that increasing B above a minimum value is not beneficial for the three throughputs when $\delta = 0.5$ and $M_U = 5$. However, this is not the case when $\delta = 1.2$ and user U_2 has no cache i.e., $q_U = 1$. We observe that T_2 obtains its maximum value when $B = 0$ i.e., when R is not set to hold any non-cacheable packets, and, thus, is only available for serving cacheable traffic. There is a considerable drop in T_2, when R is set to hold non-cacheable packets ($B > 0$) as well, since q_R i.e., the probability of R being available to serve non-cacheable traffic, was set to 0.8 in our results. The network throughput T behaves similarly to T_2.

When raising q_1 to 0.8, we observe that T_D increases with B no matter the values for δ and M_U. Regarding T_2, we observe a similar behavior to the case in which $q_1 = 0.4$, but with lower values. The slight decrease of T_2 with B that is observed for $B < 4$ can be attributed to the fact that when the queue is increased, the cache size at the relay is decreased and, hence, user U_2 requests more frequently content from DC instead of R. Similarly to $q_1 = 0.4$, T_2 starts to drop with increased B up to 3 and, then, increases with B.

B. Distribution of relay's queue states

In this section we study $\pi(i)$ i.e., the probability of the queue Q at the relay being in a specific state i or, equivalently, holding i non-cacheable packets, versus i for the cases of Section IV-A: (i) $\delta = 0.5$ and $M_U = 5$, and (ii) $\delta = 1.2$ and $M_U = 0$. We used three different values for B i.e., the

TABLE III: Cache parameters and attempt probabilities for the transmitters in our numerical results.

Parameter	Description	Value
M_U	cache size at user U_2	0 or 5
F	number of files at the relay R	10
N	number of files at the data center DC	10000
δ	shape parameter for the correlation of the files	0.5 or 1.2
q_U	probability of U_2 requesting a cached file	0.984 or 1
q_R	probability of R being available to serve D. from external resources (R or BS)	0.8
α	probability of DC being available to serve U_2	0.7
θ	links SINR minimum value	5 dB

(a) $q_1 = 0.4$.

(b) $q_1 = 0.8$.

Fig. 3: Throughput T_D, T_2, and T vs. relay queue size B (non-cacheable packets at the relay) when the relay node R holds $F = 10$ files for both types of traffic, the links SINR $\theta = 5$ dB and either (i) $\delta = 0.5$ and $M_U = 5$, or (ii) $\delta = 1.2$ and $M_U = 0$.

size of the queue at the relay, to gain insight into its effect on the distribution of the queue states.

For $B = 1$, it is more probable for the relay to store no packets at all when $\delta = 0.5$ and $M_U = 5$ and more probable to store one packet when $\delta = 1.2$ and $M_U = 0$ (see Fig. 4). Furthermore, when $B \in \{5, 10\}$, the queue is more probable to hold on average more packets when $\delta = 1.2$ and $M_U = 0$ than when $\delta = 0.5$ and $M_U = 5$ no matter the value of q_1. We observe that, when $B = 5$ i.e., the storage at the relay is equally split among cacheable and non-cacheable purposes, or when $B = 10$ i.e., the storage at the relay is dedicated to non-cacheable traffic, then the probability of the queue holding more than 4 packets is almost zero for $q_1 = 0.4$ (see Fig. 4(a)). This is anticipated since, in our results, $q_R = 0.8$ i.e., the probability by which the relay is available to serve non-cacheable traffic is double the rate by which non-cacheable traffic is transmitted to the network by U_1.

However, this is not the case if q_1 is increased to 0.8. In general, increasing q_1 yields higher values of $\pi(i)$ for higher states since increasing the rate by which U_1 attempts transmissions results in more failed transmissions to D and, hence, more attempts to queue the failed packets at the relay. Consequently, the queue is more probable to store more packets for higher

(a) $q_1 = 0.4$.

(b) $q_1 = 0.8$.

Fig. 4: Probability of the relay's queue Q holding i packets vs. i when the relay node R holds $F = 10$ files for both types of traffic, the links SINR $\theta = 5$ dB, and either (i) $\delta = 0.5$ and $M_U = 5$, or (ii) $\delta = 1.2$ and $M_U = 0$.

values of q_1 compared to lower ones. Moreover, for fixed q_1, when $\delta = 1.2$ and $M_U = 0$, increasing B over five has a decreasing effect on $\pi(i)$ for higher states i.e., increasing B produces a queue that has a higher probability of holding less packets. This is not the case when $\delta = 0.5$ and $M_U = 5$ where the distributions of the queue states for the first six states are very close for $B \in \{5, 10\}$.

V. CONCLUSION

In this work, we studied the effect of a relay node with storage capabilities in a wireless system that serves two types of traffic: cacheable and non-cacheable traffic. The relay's storage can be split to accommodate the needs of both types of traffic. We derived the network throughput taking into consideration the number of files for cacheable and non-cacheable traffic at the relay, the wireless links' parameters, the availability of the data center, and the rate by which cacheable and non-cacheable content is requested and transmitted, respectively.

Our numerical results provide insight into the network throughput and distribution of the relay's files in terms of the aforementioned parameters. It is shown how the network throughput is affected by allocation of the relay storage to cacheable and non-cacheable files as well as how the distribution of the cached files affects the network's performance.

REFERENCES

[1] "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 20172022 White Paper," 2019. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-738429.html

[2] S. Dernbach, N. Taft, J. Kurose, U. Weinsberg, C. Diot, and A. Ashkan, "Cache content-selection policies for streaming video services," *IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications*, pp. 1–9, April 2016.

[3] Cheng Yi and Alexander Afanasyev and Ilya Moiseenko and Lan Wang and Beichuan Zhang and Lixia Zhang, "A case for stateful forwarding plane," *Computer Communications*, vol. 37, no. 7, pp. 779 – 791, 2013.

[4] M. A. Hail, M. Amadeo, A. Molinaro, and S. Fischer, "Caching in Named Data Networking for the wireless Internet of Things," *International Conference on Recent Advances in Internet of Things (RIoT)*, pp. 1–6, April 2015.

[5] G. Paschos, E. Bastug, I. Land, G. Caire, and M. Debbah, "Wireless caching: technical misconceptions and business barriers," *IEEE Communications Magazine*, vol. 54, no. 8, pp. 16–22, Aug. 2016.

[6] Z. Chen, N. Pappas, and M. Kountouris, "Probabilistic Caching in Wireless D2D Networks: Cache Hit Optimal versus Throughput Optimal," *IEEE Communications Letters*, vol. 21, no. 3, pp. 584–587, Mar. 2017.

[7] J. Ma, J. Wang, and P. Fan, "A Cooperation-Based Caching Scheme for Heterogeneous Networks," *IEEE Access*, vol. 5, pp. 15 013–15 020, 2017.

[8] Z. Chen and M. Kountouris, "D2D caching vs. small cell caching: Where to cache content in a wireless network?" *IEEE 17th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pp. 1–6, July 2016.

[9] L. Li, G. Zhao, and R. S. Blum, "A Survey of Caching Techniques in Cellular Networks: Research Issues and Challenges in Content Placement and Delivery Strategies," *IEEE Communications Surveys Tutorials*, vol. 20, no. 3, pp. 1710–1732, thirdquarter 2018.

[10] N. Carlsson and D. Eager, "Ephemeral Content Popularity at the Edge and Implications for On-Demand Caching," *IEEE Transactions on Parallel and Distributed Systems*, vol. 28, no. 6, pp. 1621–1634, June 2017.

[11] Elayoubi, Salah-Eddine and Roberts, James, "Performance and Cost Effectiveness of Caching in Mobile Access Networks," *Proceedings of the 2nd ACM Conference on Information-Centric Networking (ACM-ICN)*, pp. 79–88, 2015.

[12] G. Paschos, A. Destounis, L. Vigneri, and G. Iosifidis, "Learning to Cache With No Regrets," *IEEE Conference on Computer Communications (INFOCOM)*, pp. 235–243, April 2019.

[13] D. Liu and C. Yang, "Energy Efficiency of Downlink Networks With Caching at Base Stations," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 4, pp. 907–922, Apr. 2016.

[14] J. Rao, H. Feng, C. Yang, Z. Chen, and B. Xia, "Optimal caching placement for D2D assisted wireless caching networks," *IEEE International Conference on Communications (ICC)*, pp. 1–6, May 2016.

[15] N. Pappas, Z. Chen, and I. Dimitriou, "Throughput and Delay Analysis of Wireless Caching Helper Systems With Random Availability," *IEEE Access*, vol. 6, pp. 9667–9678, Feb. 2018.

[16] I. Avgouleas, N. Pappas, and V. Angelakis, "Performance Evaluation of Wireless Caching Helper Systems," *15th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, Oct. 2019.

[17] G. Smpokos, N. Pappas, Z. Chen, and P. Mohapatra, "Wireless Caching Helper System with Heterogeneous Traffic and Secrecy Constraints," *IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pp. 1–5, July 2019.

[18] F. Rezaei and B. H. Khalaj, "Stability, Rate, and Delay Analysis of Single Bottleneck Caching Networks," *IEEE Transactions on Communications*, vol. 64, no. 1, pp. 300–313, Jan. 2016.

[19] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, and J. Wang, "Large-Scale Measurement and Characterization of Cellular Machine-to-Machine Traffic," *IEEE/ACM Transactions on Networking*, vol. 21, no. 6, pp. 1960–1973, Dec. 2013.

[20] S. Di Domenico, M. De Sanctis, E. Cianca, L. Silvestri, V. Curcur, and A. Betti, "Classification of Heterogenous M2M/IoT Traffic Based on C-plane and U-plane Data," *IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, pp. 1–5, Sep. 2018.