CHALMERS
UNIVERSITY OF TECHNOLOGY

Energy-and Bandwidth-Efficient, QoS-Aware Edge Caching in Fog-Enhanced Radio Access Networks

Citation for the original published paper (version of record):
Bhar, C., Agrell, E. (2021). Energy-and Bandwidth-Efficient, QoS-Aware Edge Caching in Fog-Enhanced Radio Access Networks. IEEE Journal on Selected Areas in Communications, 39(9): 2762-2771. http://dx.doi.org/10.1109/JSAC.2021.3064659

(article starts on next page)

Energy- and Bandwidth-Efficient, QoS-Aware Edge Caching in Fog-Enhanced Radio Access Networks

Chayan Bhar and Erik Agrell, *Fellow, IEEE*

***Abstract*—The emerging video services are associated with stringent quality-of-service (QoS) requirements and place high bandwidth demands on the core networks. Edge caching can facilitate the stringent QoS demands while easing the bandwidth requirement from core networks. However, such schemes require on-field caching equipment, in which energy consumption is a function of cache utilization. Designing opportunistic caching strategies for energy efficiency is therefore essential in such schemes. This paper studies the possibilities for achieving high energy efficiency, QoS, and low bandwidth consumption from the core network, in an optically fronthauled fog-enhanced radio access network that implements edge caching. An analytical model for such a network has been derived to measure latency, bandwidth consumption, and cache utilization. It is deduced from the results that low latency (high QoS) and bandwidth consumption can be ensured in such schemes while reducing the energy consumption by up to 93%. The derived model allows to design caching strategies for addressing the trade-off between energy efficiency, QoS, and bandwidth efficiency.**

***Index Terms*—Energy-efficient edge caching, QoS-aware content delivery, optically fronthauled wireless networks.**

I. Introduction

There has been a rapid increase in the mobile video traffic, which is predicted to account for 79% of the overall mobile traffic in 2020 [1]. End-user mobility adversely affects the network performance, due to the involved network dynamics [2], [3]. However, delivering bandwidth-intensive multimedia videos with stringent quality-of-service (QoS) constraints in 5G communications [4], [5], to mobile users from content servers located in the core network causes high latency (poor QoS) and significant bandwidth consumption in the core network and other network segments. The deployment of fog-enhanced radio access networks (Fe-RANs) is envisioned to be a potential solution to the above network issues [6]. Fe-RANs involve caching of bandwidth-intensive videos in edge caches (ECs) that are co-located with on-field base stations called enhanced remote radio heads (eRRHs) [4]. Moreover, optical fronthauling of Fe-RANs has been proposed to satisfy the stringent QoS demands of low latency services [7], provision high bandwidth [8], and allow simple deployment [9] of eRRHs. This allows the content providers to deliver QoS-aware bandwidth-intensive videos from network locations, that are closer to end-users. Optimal caching strategies in ECs are illustrated to decrease the video download latency [10].

The optimal caching time in ECs increases when the caching energy consumption is low [11], thereby increasing the involved complexity. However, the on-field base stations are proposed to have low complexity radio processing functionalities in 5G communications to facilitate easy deployment, energy efficiency, and centralized processing. Since installing extensive caching and processing facilities closer to the end users may increase the caching energy requirement, limited caching capabilities are typically installed at the eRRHs to limit the energy consumption [12]. The energy and complexity constraints of eRRHs enforce limitations on the cache size and energy consumption of ECs.

These challenges necessitate deployment of hierarchical caching, in which complex functionalities are implemented in centralized cloud units (CUs), while the on-field eRRHs are lightweight [13]. Implementation of hierarchical caching over Fe-RANs [14] allows centralization of complex processing elements, and decreases the network load in the fronthaul segment by up to 35% [15]. It was shown that caching schemes with more than two hierarchical levels experience high overall bandwidth consumption [16], [17]. In such schemes, videos from content providers are hosted in CUs that are located in the core network and are owned by CU providers [18]. The videos can be temporarily cached in the ECs and delivered to end-users from either the ECs or the CUs, depending on their demand. Therefore, video delivery to mobile users involves ECs at eRRHs that can be owned by different mobile network operators (MNOs), CUs, a transport network (TN) connecting the CUs and ECs (consisting of the core network, metro network, and fronthaul networks), the access network (AN), and all intermediate network equipment. Business partners, called tenants, can own the involved network segments and equipment, and lease resources to the content providers [18]. A network orchestrator manages end-to-end resource allocation in these segments for facilitating delivery of videos to end-users from ECs and CUs, through a suitable eRRH.

In a multi-tenant Fe-RAN, designing an optimal caching strategy is essential to minimize the resource consumption from different tenants and support QoS-aware applications. The latency of end-user services increases with network load [7]. On the other hand, stringent QoS specifications for decreasing latency can increase the optimal caching time [11]. Instead, hierarchical caching limits the content download delay by facilitating high end-user data-rates [4], [19]. Latency minimization through heterogeneous cache allocation to entire videos in an Fe-RAN employing hierarchical caching assum-

This work was funded by Vinnova under grant 2017-05228 and the Knut and Alice Wallenberg Foundation under grant 2013.0021.

C. Bhar was with the Department of Electrical Engineering, Chalmers University of Technology, SE-41296 Gothenburg, Sweden. He is now with the Department of Electronics and Communication Engineering, National Institute of Technology Warangal, Telangana, India. (e-mail: cbhar@nitw.ac.in).

E. Agrell is with the Department of Electrical Engineering, Chalmers University of Technology, SE-41296 Gothenburg, Sweden.

ing different capacities for the CU–user and EC–user links has been studied in [20], [21]. The latency of streaming videos having different file sizes can be minimized by optimizing video delivery from ECs and the CU [4]. Moreover, caching fractions of a video (video segments (VSs)) maximizes their utility compared to caching the entire video when the request process is memoryless [20]. This facilitates low network and EC utilization in scenarios where users may consume only a part of a video [22]. Different caching strategies like the least-recently-used, least-frequently-used, etc., can be implemented in the ECs. However, such schemes involve complicated analysis [23]. On the other hand, timer-based strategies that involve setting a timer when a video is stored in the cache and deleting it when the timer expires have simple analytical frameworks [24]. The timer duration is called the time-to-live (TTL) of the video. TTL-based strategies can accurately replicate the performance of the least-recently-used policy when the content requests are Poisson distributed [24].

It is illustrated that proactively caching the requested VSs in all nearby locations, where a mobile user can possibly migrate (given its current location) ensures low latency and seamless mobility [25]. Content download latency increases with end-user mobility [26], while the network throughput decreases when the number of mobile users approaches the threshold value that can be supported [27]. Moreover, costs of migration (network resource consumption) and latency increase if end users are highly mobile and have the opportunity to migrate to different neighbourhoods.

Although end-user mobility is a critically important aspect in Fe-RANs, its effect on energy, bandwidth, and QoS in a realistic environment has not been explicitly studied in the literature. For example, the following research illustrates QoS as a function of: EC size and fronthaul network capacity limitations for a simple scenario with only two ECs and two users [21], [20]; eRRH placement and video popularity [19]; the number of end-users [27]; the network load [7]; the effect of centralized and de-centralized decisions for caching and video size [4]; and EC capacity [26]. On the other hand, [25] studies the trade-off between QoS and cost-of-caching, [28] discusses the effect of the number of video requests on energy efficiency, and [14] studies the effect of cache size on its utilization. However, the design and the effect of energy-efficient caching strategies on QoS, bandwidth consumption in the core network and ANs, and energy consumption in the EC, AN, and TN, and the trade-off between these parameters are not studied in existing literature. Similarly, energy consumption and latency are illustrated to increase when more than two hierarchical caching levels are employed [16], [17]. Yet, existing studies do not illustrate optimization of bandwidth and energy consumption, and latency in networks employing hierarchical caches in the presence of mobile users. Finally, caching of VSs with TTL timers, called soft-TTL [22] are proposed. However, its effect on the improvement in QoS, transportation and caching energy consumption, and bandwidth consumption at different network segments, in a scenario with mobile users is absent in existing literature.

The main contributions of this paper are as follows:

1) We formulate analytical models for VS delivery to mobile users in an Fe-RAN from ECs that employ fractional caching; soft-TTL [22]; and proactive caching [25]; and the CU. We utilize realistic models for user mobility [26], [3] and VS request arrival [22], [3], [29].
2) We analyze the effect of user mobility on QoS and energy and bandwidth efficiency. Using the analytical models, we design caching strategies for ensuring QoS to mobile users, minimize bandwidth consumption in the TN and AN, and energy consumption in ECs, TN, and AN.
3) We illustrate that there is a trade-off between above parameters. Using results from the analytical framework, we discuss strategies available to the network orchestrator and content providers for leasing network slices from other tenants that address this trade-off and minimize the amount of resources to be leased.

The rest of this paper is arranged as follows. In the next section we formulate analytical models for the assumed network scenario. This is followed by results on the performance of such schemes in Section III and a discussion and conclusions in Section IV.

II. System Description and Modeling

In this section, we first describe the considered system in Fig. 1. Thereafter, the assumptions made for modeling the Fe-RAN of Fig. 1 are stated. Finally, an analytical model is formulated for the proposed system.

A. Assumptions

We consider the Fe-RAN illustrated in Fig. 1 that implements two levels of caches, at the ECs and CUs, respectively, (for minimizing energy and bandwidth consumption [16], [17]) similar to [7], [21], [14], [26], [20], and [30]. The network configuration and the assumptions made for modeling the system are stated below using Fig. 1. The symbols used in this paper are described in Table I. K ECs are assumed to be deployed. The formulated model and the discussion below is for a particular VS which may be present in an eRRH E_i, where $i \in \{1, \ldots, K\}$. The eRRH E_i consists of an EC and provides services to the users of its associated cell i. The CUs are located in the core network as in [7], [21], [26]. An on-field eRRH E_i provides connection and services to the users of its respective cell i. The Fe-RAN employs an optical fronthaul network similar to [7], [27], [31]. The fronthaul network consists of an optical line terminal (OLT) at the central office and K on-field optical network units (ONUs), one in each cell. The OLT forms an interface between the AN and TN and is co-located with the baseband unit (BBU) that performs complex baseband functionalities. The ONUs are co-located with the respective eRRHs $E_1, \ldots, E_K$. The AN, eRRHs, and BBU are owned by MNOs, the CUs by cloud unit providers, the videos by content providers, and the intermediate TN consisting of the core network, metro network, and fronthaul network by network providers. In each of the above cases, there can be a single or multiple tenants at the same horizontal level. In order to facilitate end-user content delivery, the content providers lease slices consisting of AN bandwidth and EC space from the MNO, TN bandwidth

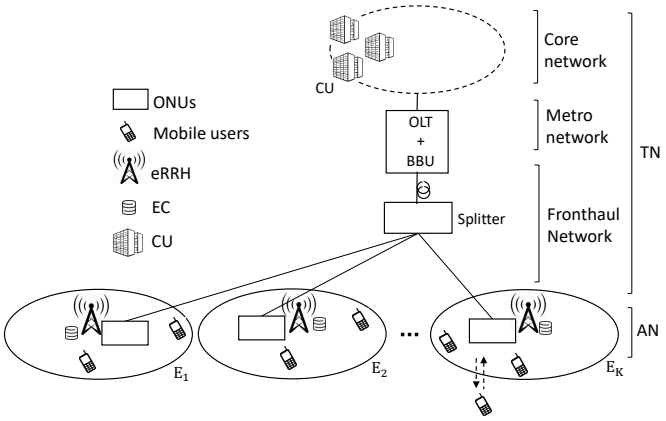

Fig. 1: Optically fronthauled Fe-RAN with edge caching at eRRHs.

from the network providers, and CU space from the cloud unit provider, while the network orchestrator designs pricing strategies that facilitate QoS, energy efficiency, and bandwidth efficiency.

The Fe-RAN implements fractional caching and soft-TTL [22] in the ECs implying that VSs are cached and deleted from the ECs as functions of their requests and the caching strategy implemented. VSs are streamed at higher download rates from ECs with rate r_{AN}, compared to when they are streamed from the CU with rate r_{TN}, i.e., $r_{\text{AN}} > r_{\text{TN}}$, since the ECs are located closer to the users [26], [30], while network congestion and statistical multiplexing are involved in the metro network and core network segments. The VS download times are assumed to be exponentially distributed for simplicity of analysis.

In this paper the analysis is performed with respect to a particular VS of size κ in a particular eRRH E_i. Users arrive to cell i associated with E_i following a Poisson process with a mean inter-arrival time $\bar{\upsilon}_i$, and depart from cell i after an exponentially distributed interval τ_i with mean $\bar{\tau}_i$ similar to [26], [3] assuming a small handover margin (< 0 dB) and small τ_i ($\leq$ 100 s) [32]. A small $\bar{\tau}_i$ corresponds to fast end-user mobility. Since the end-user mobility follows a Poisson process [3], [26], the average time after starting the VS download, when the user moves away from the cell, can be derived from splitting of Poisson process [33, Chapter 2.2] as $\bar{\tau}_i - 1/\lambda_i$, where $1/\lambda_i$ is the average inter-request interval. Our model does not capture the behaviour of users after they download the VS. A maximum of N_i users can simultaneously request and download the same VS from E_i, while requests for the same VS from other users are dropped. The users first arrive to the cell and thereafter may request for the VSs. The request process for VSs follows a Poisson distribution with exponentially distributed inter-request intervals, similar to [22], [3], [29], [34]. The popularity of a VS is assumed to affect its per-user request arrival rate λ_i (overall request arrival rate normalized by the number of users considered to be present in the cell). Streaming of VSs to the requesting users is performed according to the flowchart of actions in Fig. 2, as described below. The possible events are marked as e_1 – e_6 in Fig. 2. Users' request (Request) for VSs are served by the eRRHs either from the EC, or CU, in the order of availability. Moreover, a VS is cached at ECs from the CU when the number of requests ε_i from a particular eRRH E_i exceeds a threshold ζ_i for that particular VS in E_i. λ_i is assumed to be equal for all users and can be calculated using the methods outlined in [14]. Since any caching strategy affects the TTL of a VS we select the VS TTL in E_i σ_i, as the design parameter for caching. The TTL results in deletion (Deleted during download in Fig. 2) of the VS from E_i after an exponentially distributed interval with mean σ_i of caching it (as a function of the caching strategy). This causes on-going VS downloads from E_i to be re-streamed from the CU as in [35]. Content providers employ proactive caching [25], in which the VS can be immediately downloaded (without requesting) from all cells to which a user can possibly move, thereby ensuring seamless connectivity to mobile users and low delay. A fraction p_i of users arriving to cell i are assumed to be downloading the VS at the time of migration. p_i is independent of N_i. Therefore, VS requests from users that require proactive caching are independent of the requests from other users connected to E_i. Therefore, the overall user arrival rate $[N_i - (\omega_i + \varepsilon_i + \alpha_i)] \cdot 1/\bar{\upsilon}_i$ can be divided into $[N_i - (\omega_i + \varepsilon_i + \alpha_i)] \cdot 1/\bar{\upsilon}_i(1 - p_i)$ that can request the VS with Poisson intensity λ_i and $[N_i - (\omega_i + \varepsilon_i + \alpha_i)]p_i/\bar{\upsilon}_i$ that immediately start downloading the VS using Poisson splitting. Moreover, if proactive caching is absent, then $p_i = 0$.

We analyze the effect of user mobility $\bar{\upsilon}_i$, $\bar{\tau}_i$; VS parameters λ_i, κ; network parameters r_{TN}, N_i; and caching policy ζ_i, σ_i on VS download time δ_i; energy consumption in AN χ_{AN}, TN χ_{TN}, and EC χ_{EC}; and the bandwidth consumption in the AN γ_{AN} and TN γ_{TN}, by formulating queuing models for the above scenario. These parameters are grouped into parameters that are controllable by tenants σ_i, ζ_i, r_{TN}, κ, and user dependent parameters N_i, λ_i, $\bar{\upsilon}_i$, and $\bar{\tau}_i$. *Using the analytical model, we derive values of σ_i, r_{TN}, κ, and ζ_i that can simultaneously minimize the expectations of δ_i, χ_{AN}, χ_{TN}, χ_{EC}, γ_{AN}, and γ_{TN} in Section II-D.*

The queuing models are formulated from the perspective of a single VS. This implies that:

(i) the request arrival process of each VS is independent [24]. The performance for a particular video is obtained by assuming that respective VSs have independent request arrival processes [22]. Although the VSs influence each other due to the size restriction of ECs, we have neglected this aspect for simplicity of analysis.
(ii) the download rate of a VS is independent of other VSs being downloaded.
(iii) ECs allow caching of a VS independently of other VSs.

B. Markov model for ECs

A continuous-time Markov model with discouraged arrivals and finite calling population [33, Chapter 3.8] is formulated for E_i. The model specifies the states of N_i users that are connected to E_i and can request or have requested the VS. The state variables for analyzing the system performance with respect to a particular VS in E_i can be the number of downloads in progress from E_i ε_i, the number of downloads

TABLE I: List of symbols used in the text

Symbol	Description
E_i	i^{th} eRRH, consists of an EC and provides service to users of cell i
K	Number of ECs deployed
N_i, N	The maximum number of users that are yet to download the VS from E_i. N_i consists of the users that are downloading, that have arrived but are yet to request, and that are yet to arrive, $N = \sum_{i=1}^{K} N_i$
$\bar{\upsilon}_i$	Mean of the inter-arrival time of users to cell i
κ	VS size
λ_i	Average per-user VS request arrival rate in cell i
τ_i, $\bar{\tau}_i$	Exponentially distributed time interval after entering cell i when users move away and its mean
σ_i	Mean caching time of a particular VS in E_i
r_{AN}, r_{TN}	Average VS download rates from the AN and TN
β_i, ω_i, ε_i, α_i	Components of the state variable: Indicator variable for the availability of the VS in E_i, the number of users that have arrived to the cell but are yet to request the VS, downloading the VS from the EC, and downloading from the CU
$\bar{\beta}_i$, $\bar{\omega}_i$, $\bar{\varepsilon}_i$, $\bar{\alpha}_i$	Expectations of the state variables β_i, ω_i, ε_i, and α_i, respectively
α, $\bar{\alpha}$	State variable for the number of VS downloads in progress from the CU and its expectation
ζ_i	Maximum number of VS downloads from the CU beyond which the VS is stored in EC
R_i	average data-rate observed by a user while downloading the VS in E_i
δ_i	Average delay while downloading the VS in cell i
χ_{EC}	Energy consumption for caching the VS in EC
χ_{AN}, χ_{TN}	Energy consumption for VS transport at the AN and TN segments due to VS download in E_i
p_i	Fraction of arriving users that require proactive caching
γ_{AN}, γ_{TN}	Bandwidth consumption in the AN and TN

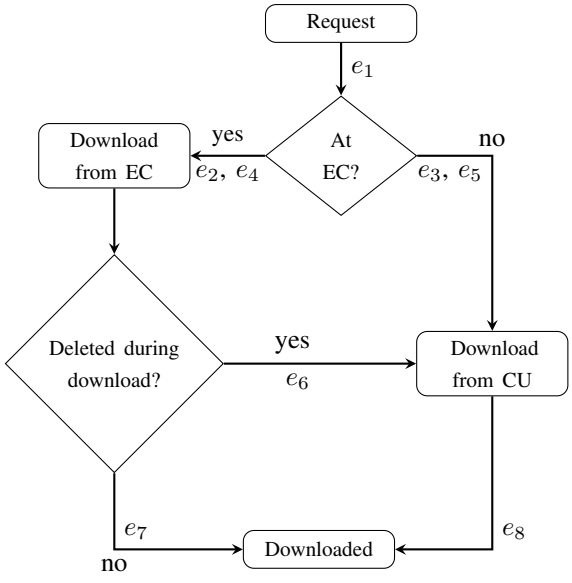

Fig. 2: Flowchart of actions for streaming VS to a user.

in progress from the CU in E_i α_i, indicator for availability of the VS in E_i β_i, the number of users that can request the VS in cell i, and the number of users that can arrive to cell i. β_i can be either 0 or 1, while other state variables lie in the range $\{0, \ldots, N_i\}$. Such a system description will encompass $\sum_{i=1}^{K} N_i$ users and the state space will be of the order $\sum_{i=0}^{K} N_i^{4i} \times 2^i$. Solving for the steady state probabilities of such a large state space is complicated. Instead, we analyze the system by reducing the state space size and making separate models for eRRHs and the CU.

In the model for a VS in E_i we concentrate on the users that (can) potentially affect the network performance, i.e., the users downloading the VS from E_i ε_i, or the CU α_i, and the users that have arrived to the cell but have not requested for the VS ω_i. We do not consider other users present in the cell. We assume that $\varepsilon_i + \alpha_i + \omega_i$ is upper limited by N_i. Therefore, $N_i - (\varepsilon_i + \alpha_i + \omega_i)$ users can arrive to cell i. Although this places a limitation on the arrival of users, thereby introducing a memory between the involved processes, we illustrate through network simulations that if $1/\lambda_i$ is in the order of seconds, the analytical results are sufficiently accurate. We define the state vector $\boldsymbol{s}_1 = (\beta_i, \omega_i, \varepsilon_i, \alpha_i)$ and $\mathcal{S}_1$ as the set of possible values for $\boldsymbol{s}_1$. The events e_1 – e_8 in Fig. 2 result in the following state transitions of the Markov chain, mobility of end-users cause e_9 – e_{11}, while VS transfer from CU to EC causes e_{12}. The state transitions caused by the events along with the necessary conditions for the transitions are also described in Table II, in which the initial state is assumed to be $\boldsymbol{s}_1 = (\beta_i, \omega_i, \varepsilon_i, \alpha_i)$ before the respective events.

(i) e_1: Arrival of a non-requesting user, i.e., the user is yet to request the VS and proactive caching is not done for it, to cell i if $\omega_i + \varepsilon_i + \alpha_i < N_i$, with rate $[N_i - (\omega_i + \varepsilon_i + \alpha_i)](1 - p_i)/\bar{\upsilon}_i$ s^{-1}.

(ii) Arrival of a VS request with rate $\omega_i \lambda_i$ s^{-1} to
- e_2: EC if $\beta_i = 1$,
- e_3: CU if $\beta_i = 0$.

(iii) Arrival of VS request with rate $[N_i - (\omega_i + \varepsilon_i + \alpha_i)]p_i/\bar{\upsilon}_i$ s^{-1} due to proactive caching to
- e_4: EC if $\beta_i = 1$,
- e_5: CU if $\beta_i = 0$.

(iv) e_6: VS deletion from the EC after σ_i s, along with transfer of the ε_i VS downloads from the EC to CU.

(v) Completion of VS download from either the
- e_7: EC at rate r_{AN}/κ VSs s^{-1},
- e_8: or CU with rate r_{TN}/κ VSs s^{-1}.

(vi) Mobility of a user that
- e_9: has not requested VS $\bar{\tau}_i$ s after entering the cell
- e_{10}: is downloading from the EC $1/(\bar{\tau}_i - 1/\lambda_i)$ s after starting VS download.
- e_{11}: is downloading from CU $1/(\bar{\tau}_i - 1/\lambda_i)$ s after starting VS download.

(vii) e_{12}: VS transfer to the EC from CU at rate r_{TN} if $\alpha_i \geq \zeta_i$.

The state variables follow the conditions $\omega_i + \varepsilon_i + \alpha_i \leq N_i$ (finite calling population) and $\beta_i \in \{0, 1\}$ with the additional constraints $\varepsilon_i = 0$ if $\beta_i = 0$, resulting in $|\mathcal{S}_1| = \sum_{n=0}^{N} \binom{n+2}{2} + \binom{n+1}{1}$ states. The above description of the Markov chain

TABLE II: Description and consequences of events at E_i from a state $\boldsymbol{s}_1 = (\beta_i, \omega_i, \varepsilon_i, \alpha_i)$

Event	Rate (s^{-1})	Condition	New state
e_1	$[N_i - (\omega_i + \varepsilon_i + \alpha_i)](1 - p_i)/\bar{v}_i$	$\omega_i + \varepsilon_i + \alpha_i < N_i$	$\beta_i, \omega_i + 1, \varepsilon_i, \alpha_i$
e_2	$\omega_i \lambda_i$	$\beta_i = 1$, $\omega_i > 0$	$1, \omega_i - 1, \varepsilon_i + 1, \alpha_i$
e_3	$\omega_i \lambda_i$	$\beta_i = 0$, $\omega_i > 0$	$0, \omega_i - 1, \varepsilon_i, \alpha_i + 1$
e_4	$[N_i - (\omega_i + \varepsilon_i + \alpha_i)]p_i/\bar{v}_i$	$\beta_i = 1$	$1, \omega_i, \varepsilon_i + 1, \alpha_i$
e_5	$[N_i - (\omega_i + \varepsilon_i + \alpha_i)]p_i/\bar{v}_i$	$\beta_i = 0$	$0, \omega_i, \varepsilon_i, \alpha_i + 1$
e_6	$1/\sigma_i$	$\beta_i = 1$	$0, \omega_i, 0, \varepsilon_i + \alpha_i$
e_7	r_{AN}/κ	$\varepsilon_i > 0$, $\beta_i = 1$	$1, \omega_i, \varepsilon_i - 1, \alpha_i$
e_8	r_{TN}/κ	$\alpha_i > 0$	$\beta_i, \omega_i, \varepsilon_i, \alpha_i - 1$
e_9	$1/\bar{\tau}_i$	$\omega_i > 0$	$\beta_i, \omega_i - 1, \varepsilon_i, \alpha_i$
e_{10}	$1/(\bar{\tau}_i - 1/\lambda_i)$	$\beta_i = 1$	$1, \omega_i, \varepsilon_i - 1, \alpha_i$
e_{11}	$1/(\bar{\tau}_i - 1/\lambda_i)$	$\alpha_i > 0$	$\beta_i, \omega_i, \varepsilon_i, \alpha_i - 1$
e_{12}	r_{TN}/κ	$\alpha_i \geq \zeta_i,\ \beta_i = 0$	$1, \omega_i, \varepsilon_i, \alpha_i$

allows us to formulate the $|\mathcal{S}_1| \times |\mathcal{S}_1|$ state transition rate matrix $\boldsymbol{Q}$ for the transition from one state to another.

The Markov chain so formed has the following properties: (i) is time-homogeneous as the state transitions are independent of time, (ii) has a finite number of states, and (iii) is connected. The steady-state probability for a state $\boldsymbol{s}_1$ is denoted by $\pi_{\boldsymbol{s}_1}$, while the steady-state probability vector is $\boldsymbol{\pi}_1 = \{\pi_{\boldsymbol{s}_1}\}_{\boldsymbol{s}_1 \in \mathcal{S}_1}$. These properties of the Markov chain allow us to solve for the steady state probabilities using $\boldsymbol{\pi}_1 \boldsymbol{Q} = \boldsymbol{0}$ together with $\sum_{\boldsymbol{s}_1 \in \mathcal{S}_1} \pi_{\boldsymbol{s}_1} = 1$. We have solved the steady state probabilities in MATLAB. Furthermore, we define

$$(\bar{\beta}_i, \bar{\omega}_i, \bar{\varepsilon}_i, \bar{\alpha}_i) = \sum_{\boldsymbol{s}_1 \in \mathcal{S}_1} \pi_{\boldsymbol{s}_1}(i)\ \boldsymbol{s}_1, \tag{1}$$

where $\bar{\varepsilon}_i$ and $\bar{\alpha}_i$ are the average number of VSs in download from the EC in E_i and the CU at any given instant, respectively. The occupancy of the EC in E_i is $\bar{\beta}_i$.

C. Markov model for the CU

The model for CU is an abstract model and utilizes parameters from the EC model for each cell that summarize all processes at the respective cells. We assume that $N = \sum_{i=1}^{K} N_i$ users can simultaneously download the VS from the CU and concentrate only on those users that request and download the VS from the CU. Other users and the streaming of VS from CU to ECs are not considered in the model for CU. The state variable consists of a single non-negative integer α, representing the number of downloads in progress from the CU at a given time. The possible values for α are $\mathcal{S}_2 = [0, 1, \ldots, N]$. A user can request the VS from cell i if $\beta_i = 0$, whereas j users downloading the VS from cell i can request the VS from the CU if it is deleted from the EC. The state transitions caused by the events along with the necessary conditions for the transitions are described in Table III, in which the initial state is assumed to be α before the respective events. The case of VS deletion during download is represented here as $J = \max_i N_i$ separate events $c_{2,1}, \ldots, c_{2,J}$, in which each have different arrival rates.

(i) c_1: For a user requesting from cell i, the VS is streamed from the CU only if $\beta_i = 0$ with rate $\lambda_i \omega_i + \{N_i - (\omega_i + \varepsilon_i + \alpha_i)\} p_i/\bar{v}_i$ from Section II-B. Therefore, requests for the VS from cell i is with the average rate $\sum_{\boldsymbol{s}_1 \in \mathcal{S}_1, \beta_i = 0} \pi_{\boldsymbol{s}_1}(i)\ [\lambda_i \omega_i + \{N_i - (\omega_i + \varepsilon_i + \alpha_i)\} p_i/\bar{v}_i]$. The overall request rate is the sum of request rates from individual cells. The per-user request rate from the CU is obtained by dividing the overall request rate by the maximum number of downloads possible for a VS N. Assuming that α downloads are in progress, $N - \alpha$ more users can request the VS from K cells. Therefore, the per-user request rate is $(N - \alpha) \sum_{i=1}^{K} \sum_{\boldsymbol{s}_1 \in \mathcal{S}_1, \beta_i = 0} \pi_{\boldsymbol{s}_1}(i)\ \{\lambda_i \omega_i + [N_i - (\omega_i + \varepsilon_i + \alpha_i)] p_i/\bar{v}_i\}/N$ due to non-availability of the VS in ECs, i.e., e_3 and e_5 in Section II-B.

(ii) $c_{2,j}$: VS deletion in E_i causes the ongoing j downloads to be re-streamed, i.e, requested from the CU, as the download process is assumed to be memoryless. Since the probability for multiple events at the same instant is almost zero, only one EC can delete the considered VS at any time. Therefore, bulk arrival [33, Chapter 4.1] of $j = 1, \ldots, J$ VS requests from cell i to the CU can occur with the rate $\sum_{\boldsymbol{s}_1 \in \mathcal{S}_1, \varepsilon_i = j} \pi_{\boldsymbol{s}_1}(i)\ 1/\sigma_i$ and probability $1/\sigma_i / \sum_{i=1}^{K} 1/\sigma_i$. This results in the overall arrival rate $(\sum_{i=1}^{K} \sum_{\boldsymbol{s}_1 \in \mathcal{S}_1, \varepsilon_i = j} \pi_{\boldsymbol{s}_1}(i)\ 1/{\sigma_i}^2) / \sum_{i=1}^{K} 1/\sigma_i$ due to VS deletion from ECs e_6 in Section II-B.

(iii) c_3: Completion of a VS download.

(iv) c_4: Mobility of of an user downloading from CU.

Since $\alpha \leq N$ (finite calling population), there are $N + 1$ states. Table III allows us to formulate the $(N+1) \times (N+1)$ state transition rate matrix $\boldsymbol{Q}$. The Markov chain so formed is time-homogeneous as the state transitions are independent of time, has a finite number of states N, and is connected, as in Section II-B. Assuming the steady state probability for a state α is denoted by π_α, the steady state probability vector is $\boldsymbol{\pi}_2 = \{\pi_\alpha\}_{\alpha=0}^{N}$. These properties of the Markov chain allow us to solve for the steady state probabilities using $\boldsymbol{\pi}_2 \boldsymbol{Q} = \boldsymbol{0}$ together with $\sum_{\alpha=0}^{N} \pi_\alpha = 1$. We have solved the steady state probabilities in MATLAB. Furthermore, we define

$$\bar{\alpha} = \sum_{\alpha=0}^{N} \pi_\alpha \alpha, \tag{2}$$

where $\bar{\alpha}$ is the average number of VS downloads from the CU at any given instant.

TABLE III: Description and consequences of events at CU from a state α

Event	Rate (s^{-1})	Condition	New state
c_1	$(N-\alpha)\sum_{i=1}^{K}\sum_{\boldsymbol{s}_1\in\mathcal{S}_1,\beta_i=0}\pi_{\boldsymbol{s}_1}(i)\ \{\lambda_i\omega_i+(N_i-(\omega_i+\varepsilon_i+\alpha_i))p_i/\bar{\upsilon}_i\}/N$	$\alpha<N$	$\alpha+1$
$c_{2,j},\ j=1,\ldots,J$	$(\sum_{i=1}^{K}\sum_{\boldsymbol{s}_1\in\mathcal{S}_1,\varepsilon_i=j}\pi_{\boldsymbol{s}_1}(i)\ 1/{\sigma_i}^2)/\sum_{i=1}^{K}1/\sigma_i$	$\alpha+j<N$	$\alpha+j$
c_3	$\alpha\ r_{\text{TN}}/\kappa$	$\alpha>0$	$\alpha-1$
c_4	$1/\bar{\tau}_i$	$\alpha>0$	$\alpha-1$

TABLE IV: Energy consumption of network components [12]

Network device	Energy consumption (nJ/bit)	Symbolic representation
Core router	17	ρ_{CR}
Edge router	26.3	ρ_{ER}
Aggr. switch	8.21	ρ_{AS}
OLT	19.2	ρ_{OLT}
eRRH (eNodeB)	$2\cdot10^3$	ρ_{eRRH}
Solid state drive	$1.97\cdot10^5$	ρ_{SSD} [a]

[a] Assuming operation over a one year period, $\rho_{\text{SSD}} = 6.25\cdot10^{-12}$ W/b $\times$ $31536\cdot10^3$ s

D. Performance parameters from the Markov model

The average data-rate R_i observed by a user in cell i while downloading a VS is the sum of r_{AN} and r_{TN} multiplied with the average number of downloads $\bar{\varepsilon}_i$ and $\bar{\alpha}_i$, respectively, divided by the average number of VS downloads from EC and CU in cell i. Thus,

$$R_i = \frac{\bar{\varepsilon}_i\ r_{\text{AN}} + \bar{\alpha}_i\ r_{\text{TN}}}{\bar{\varepsilon}_i + \bar{\alpha}_i}. \tag{3}$$

The delay is calculated from (3) as $\delta_i = \kappa/R_i$. Since all processes are assumed to be memoryless, the analytical model considers that upon deletion of the VS from an EC, it is streamed again from CU to the users that were downloading from the EC. However, (3) does not take this repeated VS streaming into consideration. Therefore, (3) is accurate only when r_{AN} and r_{TN} are high or κ is small.

The energy consumed by the transportation and caching of VSs at different network levels are modeled by:

- TN: core and edge routers, OLT and aggregation switch, $\rho_{\text{TN}} = \rho_{\text{CR}} + \rho_{\text{ER}} + \rho_{\text{OLT}} + \rho_{\text{AS}}$,
- AN: eRRH, $\rho_{\text{AN}} = \rho_{\text{eRRH}}$, and
- EC: Solid state drives, ρ_{SSD},

where the energy consumed by the respective network components are given in Table IV. The baseline energy consumption in solid state drives is neglected, as other VSs may be also cached. Therefore, the baseline energy consumption is not directly influenced by a single VS. $\bar{\beta}_i$ depends on the probability that the VS is present at E_i. Since a cache needs to be operational for caching the VS, the energy consumption of EC is $\chi_{\text{EC}} = \bar{\beta}_i\rho_{\text{SSD}}$. The transport of VSs from eRRH or CU due to the switches located in these network segments, is assumed to cause energy consumption at the AN

$$\chi_{\text{AN}} = (\bar{\varepsilon}_i + \bar{\alpha}_i)\rho_{\text{AN}}, \tag{4}$$

and at the TN

$$\chi_{\text{TN}} = \bar{\alpha}\rho_{\text{TN}}, \tag{5}$$

because the CU utilizes the TN and AN, while the EC utilizes only the AN to stream a VS.

The average bandwidth consumed in the AN $\varepsilon_i r_{\text{AN}}$ is normalized with the bandwidth available from the AN to N_i users $N_i r_{\text{AN}}$ to derive the normalized bandwidth consumption in the AN,

$$\gamma_{\text{AN}} = \frac{\bar{\varepsilon}_i}{N_i}. \tag{6}$$

Similarly, the average bandwidth consumed in the TN due to VS download from the CU $\bar{\alpha}\ r_{\text{TN}}$ is normalized with the bandwidth available from the TN to N users, $N r_{\text{TN}}$, to derive the normalized bandwidth consumption in the TN,

$$\gamma_{\text{TN}} = \frac{\bar{\alpha}}{N}. \tag{7}$$

III. Results

In this section, we perform a case study using the model derived in Section II for the scenario depicted in Table V along with an additional round-trip delay of 300 ms for VS download from the CU [36]. The average time spent by a user in a cell $\bar{\tau}_i$ is calculated by assuming that cells cover an area of 1 km^2 and vehicle speeds 180 km/h or 50 km/h [27]. Thus, low $\bar{\tau}_i$ corresponds to high mobility. Although, the analytical and simulation frameworks have been implemented for $N_i = \{25,\ 30\}$ users that are yet to download the VS from E_i, these models can also be scaled up to analyze scenarios with higher N_i. The values of parameters that are common in all simulations are grouped under the scenario $\boldsymbol{C}$ and listed in Table V. The performance results for $\boldsymbol{C}$ are a benchmark in the calculations for energy and bandwidth efficiency. The scenarios $\boldsymbol{C}_1$–$\boldsymbol{C}_3$ consist of the parameters in $\boldsymbol{C}$ with variation in the parameters controllable by tenants, namely the maximum number of VS downloads from the CU beyond which it is stored in E_i ζ_i, average download rate from TN r_{TN}, and VS size κ. Furthermore, the scenarios $\boldsymbol{C}_4$–$\boldsymbol{C}_7$ consist of user-dependent parameters, namely N_i, per-user VS request arrival rate in cell i λ_i, mean inter-arrival time of users to cell i $\bar{\upsilon}_i$, and $\bar{\tau}_i$ as mentioned in Table V. The delay and bandwidth consumption obtained from the analytical models proposed in Section II are validated using network simulations in OMNET++.

A. Analysis of VS download time

There is a finite probability of VS deletion from E_i before download completion when κ is high, resulting in high delay

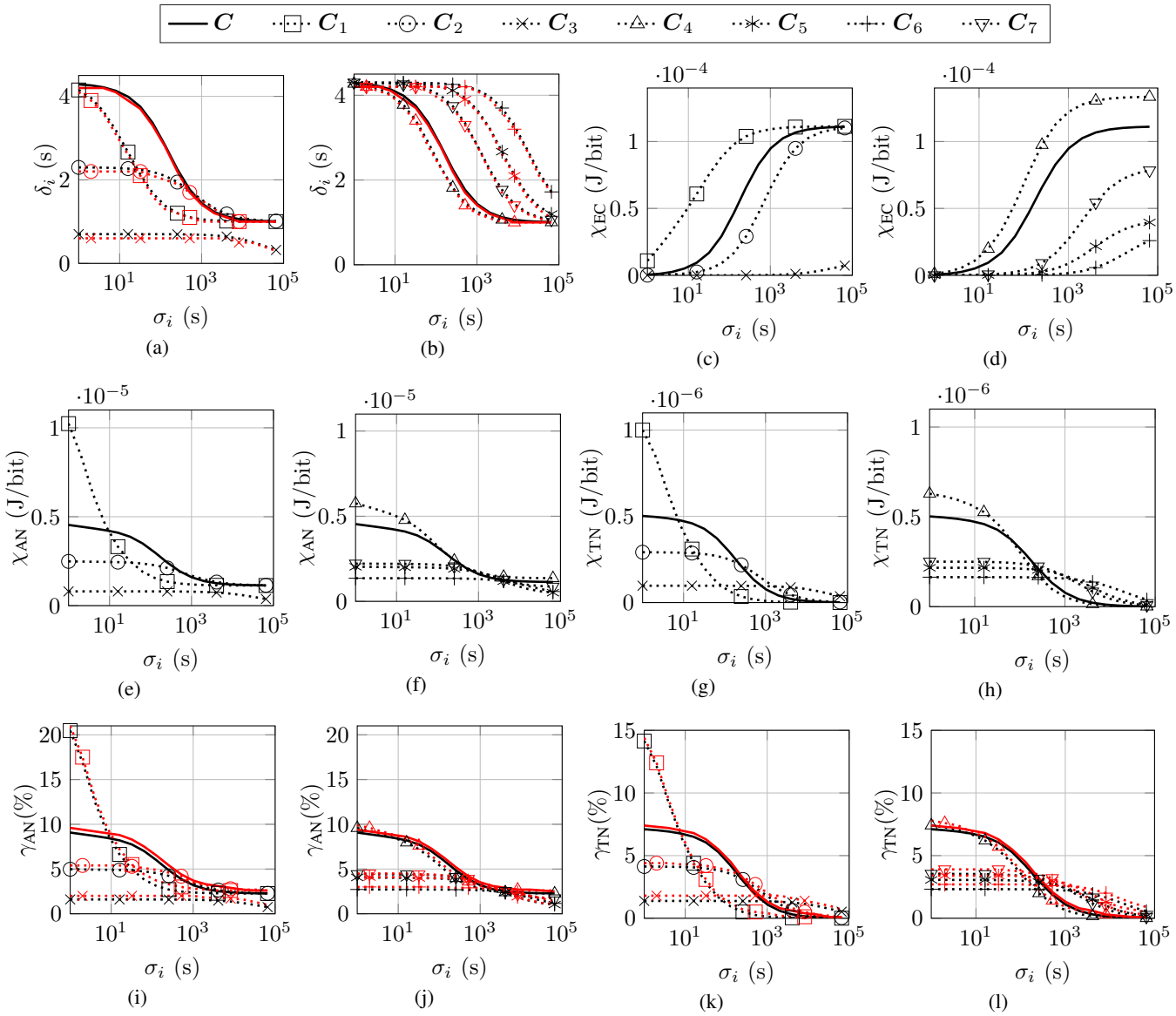

Fig. 3: δ_i, χ_{EC}, χ_{AN}, χ_{TN}, γ_{AN}, γ_{TN}, as a function of σ_i. Analytical model - black and network simulations - red.

TABLE V: Values assumed for different scenarios (Sc.). Arrows indicate changes compared with the benchmark $\boldsymbol{C}$.

Sc.	Parameter		Value	Sc.	Parameter		Value
$\boldsymbol{C}$	κ		10^9 bits	$\boldsymbol{C}$	ζ_i		5
	$\bar{v}_i$		20 s		N_i		25
	λ_i		0.05 s^{-1} [29]		r_{TN}		250 Mbps
	$\bar{\tau}_i$		72 s [27]		r_{AN}		1 Gbps [37]
	K		4		N_i		25
	p_i		0.1				
$\boldsymbol{C}_1$	ζ_i	$\searrow$	2	$\boldsymbol{C}_2$	r_{TN}	$\nearrow$	500 Mbps
$\boldsymbol{C}_3$	κ	$\searrow$	10^8 bits	$\boldsymbol{C}_4$	N_i	$\nearrow$	30
$\boldsymbol{C}_5$	λ_i	$\searrow$	0.01 s^{-1} [29]	$\boldsymbol{C}_6$	$\bar{v}_i$	$\nearrow$	100 s
$\boldsymbol{C}_7$	$\bar{\tau}_i$	$\searrow$	25 s [27]				

δ_i, as $r_{\text{AN}} > r_{\text{TN}}$. This highlights the importance of fractional caching [22], i.e., storing VSs with small κ at ECs. Decreasing ζ_i, forces E_i to cache the VS even for a small ε_i. This increases $\bar{\beta}_i$. Similarly, high σ_i causes high $\bar{\beta}_i$ and $\bar{\varepsilon}_i$. Therefore, the VS is streamed to the demanding users at $r_{\text{AN}} > r_{\text{TN}}$ and consequently low δ_i. The overall VS request rate in E_i decreases with λ_i and on increasing $\bar{v}_i$, resulting in low $\bar{\beta}_i$. Under such circumstances the VS is streamed from the CU, resulting in a high δ_i. High user mobility, i.e., low $\bar{\tau}_i$ implies that more users move out of the cell faster. Hence, $\bar{\varepsilon}_i$ and $\bar{\beta}_i$ decrease. Therefore, δ_i decreases with decreasing κ, ζ_i, and $\bar{v}_i$; and with increasing λ_i, $\bar{\tau}_i$, and σ_i, as illustrated in Figs. 3(a) and 3(b).

B. Energy-efficient VS caching

The caching energy consumption χ_{EC} is a function of $\bar{\beta}_i$ and $\bar{\varepsilon}_i$ (Section II-D) which are affected by different parameters

as discussed below. Frequent requests due to high N_i, and λ_i, low $\bar{\upsilon}_i$, or low ζ_i, cause ε_i to reach ζ_i faster. Therefore, cells serving VSs with high popularity will experience high χ_{EC} as discussed in [27]. On the other hand, $\bar{\beta}_i$ increases with σ_i. These factors also cause high $\bar{\beta}_i$ and correspondingly high χ_{EC}. A high $\bar{\tau}_i$, i.e., low mobility, allows more users to complete the VS download before moving out from the cell, resulting in high $\bar{\varepsilon}_i$ and χ_{EC}. In summary, χ_{EC} increases with increasing N_i, λ_i, σ_i, and $\bar{\tau}_i$; and with decreasing $\bar{\upsilon}_i$ and $\bar{\zeta}_i$, as illustrated in Figs. 3(c) and 3(d).

C. Energy-efficient VS transport

The analytical model assumes memoryless processes. Therefore, on deletion of a VS from E_i, it is assumed to be re-streamed entirely from the CU. VS transportation energy consumption in the AN χ_{AN} and TN χ_{TN} as a function of σ_i is discussed below.

1) VS transport in the AN: Increasing σ_i and decreasing κ increases the probability that a VS is completely streamed from E_i. For low σ_i and ζ_i, $\bar{\alpha}_i$ and consequently χ_{AN} are high. Users arrive faster to the cell for low $\bar{\upsilon}_i$ which increases the overall VS request rate. Finally, increasing $\bar{\tau}_i$ allows more users to complete VS download, before moving away from cell i. Therefore, χ_{AN} increases with increasing κ, and $\bar{\tau}_i$; and with decreasing σ_i, ζ_i, r_{TN}, $\bar{\upsilon}_i$.

2) VS transport in the TN: A low overall VS request rate at high $\bar{\upsilon}_i$ and incomplete downloads at low $\bar{\tau}_i$ cause low χ_{TN}. When σ_i and ζ_i are simultaneously low, $\bar{\beta}_i$ is also low and the VS is rapidly deleted from ECs, resulting in high χ_{TN}. However, $\bar{\beta}_i$ increases rapidly at high σ_i and low ζ_i thereby causing χ_{TN} to decrease. $\bar{\varepsilon}_i$ increases with σ_i (Sections III-A, III-B) which causes $\bar{\alpha}_i$ and χ_{TN} to decrease. In summary, χ_{TN} decreases with decreasing $\bar{\tau}_i$ and with increasing σ_i, $\bar{\upsilon}_i$, r_{TN}, and ζ_i.

χ_{AN} is observed to be more significant than χ_{TN} from Figs. 3(e)–3(h). This is because VSs from both the CU and the EC are streamed through the AN. Finally, the energy consumed for downloading a video can be calculated by multiplying χ_{EC}, χ_{AN}, and χ_{TN} with the number of VSs.

D. Analysis of bandwidth consumption

The effect of σ_i on bandwidth consumption in AN γ_{AN} and TN γ_{TN} is discussed below.

1) Bandwidth consumption in AN: R_i increases with σ_i and r_{TN} (discussed in Sections III-A and III-C), whereas, increasing $\bar{\upsilon}_i$ decreases the request rate (Section III-C) resulting in low γ_{AN}. $\bar{\alpha}_i$ decreases with ζ_i which causes γ_{AN} to simultaneously decrease. An increase in $\bar{\tau}_i$ allows more users to complete VS downloads, before moving out of the considered cell, resulting in high γ_{AN}. In summary, γ_{AN} decreases with increasing σ_i, r_{TN}, and $\bar{\upsilon}_i$; and with decreasing ζ_i and $\bar{\tau}_i$, as observed from Figs. 3(i) and 3(j).

2) Bandwidth consumption in TN: γ_{TN} decreases on increasing σ_i, ζ_i, r_{TN}, and $\bar{\upsilon}_i$, and on decreasing $\bar{\tau}_i$ as observed from Figs. 3(k) and 3(l). The reasons for these observations are similar to that for γ_{AN} (Section III-D1). Fig. 4 illustrates the effect of different $\bar{\upsilon}_i$ and λ_i on γ_{TN}. Heterogeneous user

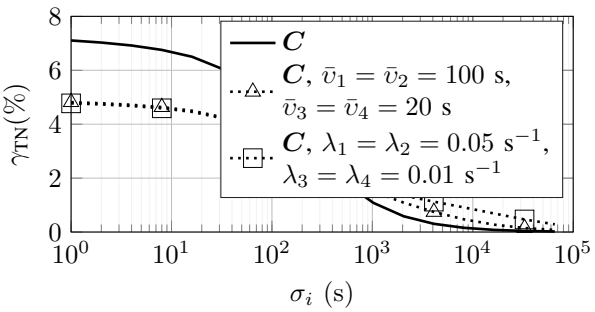

Fig. 4: γ_{TN} as a function of σ_i.

mobility patterns can cause different $\bar{\upsilon}_i$ in the respective cells, whereas different VSs in the same or different cells can have different λ_i. Moreover, a particular VS can also be associated with different λ_i due to the spatio-temporal variation of VS popularity [15]. It is observed from Fig. 4 that γ_{TN} has a similar trend as γ_{TN} in Fig. 3(l). However, $\bar{\upsilon}_i \in \{20 \text{ s}, 100 \text{ s}\}$ in Fig. 3(l), whereas $\bar{\upsilon}_1 = \bar{\upsilon}_2 = 100$ s and $\bar{\upsilon}_3 = \bar{\upsilon}_4 = 20$ s in Fig. 4. Therefore, γ_{TN} is intermediate between the plots of $\bar{\upsilon}_i = 20$ s and $\bar{\upsilon}_i = 100$ s of Fig. 3(l). A similar observation is made by comparing the plots for $\lambda_1 = \lambda_2 = 0.05$ s^{-1} and $\lambda_3 = \lambda_4 = 0.01$ s^{-1} in Fig. 4, with $\lambda_i = 0.05$ s^{-1} and $\lambda_i = 0.01$ s^{-1} in Fig. 3(l).

E. Design of caching strategies

In this sub-section we discuss the strategies that can be adopted by the content providers, MNOs, and network provider, owning the VSs, AN, and the TN, respectively. The discussion below is with respect to σ_i, ζ_i, r_{TN}, and κ that are controllable by tenants. We assume that MNOs aim for caching energy efficiency, i.e., low χ_{EC}, while network providers target bandwidth efficiency in the TN, i.e., low γ_{TN}. The content providers aim to achieve delay efficiency, i.e., low δ_i, by leasing required network slices from MNOs and network providers depending on N_i, $\bar{\tau}_i$, λ_i, and $\bar{\upsilon}_i$. As discussed in Section I, ensuring low γ_{TN} (7) is the prime motivation behind deploying an Fe-RAN. However, there is a trade-off between χ_{EC}, γ_{TN}, and δ_i. This results in conflicting goals, thereby requiring slice management by the network orchestrator. Since the on-field ECs are more numerous compared to a centrally located CU and $\chi_{\text{EC}} > \chi_{\text{TN}}$, the minimization of caching energy efficiency prioritizes low χ_{EC}. On the other hand, γ_{TN} is prioritized for transport energy efficiency, as TN resources are more expensive [18].

In a scenario with highly mobile users (low $\bar{\tau}_i$) or low request arrival rate (low λ_i and $\bar{\upsilon}_i$), the content provider can minimize δ_i if $\sigma_i \geq 30$ s. Although δ_i can also be minimized using high r_{TN}, it is concluded from Figs. 3(c) and 3(d) that content providers can support services with stringent QoS bounds using the required σ_i, instead of leasing bandwidth slices with high r_{TN} from network providers that are more expensive. This highlights the importance of edge caching [7]. Therefore, millisecond order δ_i (≈ 700 ms) can be achieved if κ is limited to 10^8 bits.

The energy efficiency of ECs can improve by more than $1 - 4.5 \cdot 10^{-5}/1.1 \cdot 10^{-4} = 59\%$ if $\sigma_i \leq 100$ s. Furthermore, if $\sigma_i \leq 100$ s and $\kappa = 10^8$ bits simultaneously, the energy efficiency can improve by $1 - 7.2 \cdot 10^{-6}/1.1 \cdot 10^{-4} = 93\%$. Such a choice of σ_i is essential for the MNOs to ensure low χ_{EC} in scenarios with high $\bar{\tau}_i$, N_i, and λ_i, or low $\bar{v}_i$. On the other hand MNOs will tend to select $\kappa = 10^8$ bits and high σ_i, i.e., $\sigma_i \geq 30$ s to reduce bandwidth consumption. MNOs select small κ for delay, bandwidth, and energy efficiency which highlights the importance of fractional caching and soft-TTL [22]. The size of ECs is assumed to be 1 TB in the network simulations. A change in the EC size will not affect χ_{EC} for respective VSs under the independence assumption of Section II. However, it will allow ECs to cache more VSs. The overall energy consumed by an EC can be calculated by summing χ_{EC} of all VSs.

For energy efficiency in the TN, it is concluded from Figs. 3(g) and 3(h) that if $\sigma_i \geq 30$ s, then χ_{TN} is limited to $4.2 \cdot 10^{-7}/1 \cdot 10^{-6} = 42\%$ of the worst case energy consumption $1 \cdot 10^{-6}$ J/bit. Instead, decreasing κ to 10^8 bits limits χ_{TN} to $9.7 \cdot 10^{-8}/1 \cdot 10^{-6} = 9.7\%$, while if $\sigma_i \geq 100$ s, χ_{TN} can be $\leq 2.9 \cdot 10^{-7}/1 \cdot 10^{-6} = 29\%$. Since the TN is involved in delivering multiple services, network providers may be interested in limiting χ_{TN}. On the other hand, χ_{AN} is limited to $3.8 \cdot 10^{-6}/4.53 \cdot 10^{-6} = 83\%$–$2.9 \cdot 10^{-6}/4.53 \cdot 10^{-6} = 64\%$ for 30 s $\leq \sigma_i \leq 100$ s, and to $8 \cdot 10^{-7}/4.53 \cdot 10^{-6} = 17.7\%$ if $\kappa = 10^8$ bits, resulting in an energy efficiency of 17%–36% (Figs. 3(e) and 3(f)).

It is observed from Figs. 3(c)–3(f) that low κ will enhance energy efficiency when users do not download an entire VS [22]. Such conditions arise in high user mobility scenarios (low $\bar{\tau}_i$), where some users may move to a new cell before download completion. Due to proactive caching, the VS is streamed afresh from the new cell. Thus, the same VS is downloaded over a longer period, resulting in low χ_{TN} in one cell, but high overall energy consumption. Without proactive caching, the p_i fraction of concerned users will experience high latency [25]. However, our model does not capture the effect of proactive caching on overall latency experienced by these mobile users and energy consumption in the cells through which the users move. The proposed model also does not consider the scenario in which proactive caching is implemented only when the ECs are vacant [25].

It is also concluded that the network orchestrator can encourage network providers (through pricing strategies) to increase statistical multiplexing on the TN by installing more ANs with low r_{TN}. This will result in economic benefits for the network providers. In such scenarios, degradation of QoS and energy efficiency in the AN can be prevented if the MNOs can be incentivized to maintain high σ_i. In scenarios with high user mobility, more caching resources (high σ_i) are required to ensure low δ_i. Although χ_{EC} decreases when $\bar{\tau}_i$ increases, thereby profiting MNOs, limited cache availability may allow MNOs to exploit content providers through overcharging. MNOs and network providers may also overcharge content providers for VSs with high λ_i which causes an increase in χ_{EC} and γ_{TN}. The network orchestrator can prevent such situations by providing incentives to MNOs and network providers for leasing more resources. This ensures fair pricing due to market competition among network providers and MNOs, while allowing MNOs and network providers to achieve energy and bandwidth efficiency respectively; ensuring QoS to mobile users; and sufficient profits to content providers. It is also observed from Figs. 3(c)–3(h), that ECs form the most energy-expensive segment, thereby highlighting the importance of this study.

It is concluded from Figs. 3(i)–3(l) that network providers can limit γ_{AN} and γ_{TN} to 7.7% and 6.1% respectively, if $\sigma_i \geq 30$ s. Furthermore, γ_{TN} is limited to $\leq 1.3\%$, if $\kappa = 10^8$ bits. Since the TN bandwidth slice is a more expensive resource than the AN slice [18], it is recommended that content providers lease sufficient resources (high σ_i) from the MNOs so that $\sigma_i \geq 30$ s can be maintained.

The above benefits are observable only if there is no constraint on the total cache size deployed in the entire network. Instead, if there is a fixed budget on the cache size deployed in the network, then cache deployment at intermediate levels is essential to maximize resource utilization [34].

IV. Discussion and Conclusions

Real-time networks are often faced with a trade-off between delay, energy efficiency, and bandwidth efficiency. In the network scenario considered in Section III, reasonable delay, transport energy, and bandwidth efficiency are achieved if the mean caching time $\sigma_i \geq 30$ s, while caching energy efficiency is achieved for low σ_i, i.e., $\sigma_i \leq 100$ s. Therefore, delay, energy efficiency, and bandwidth efficiency are simultaneously achieved if σ_i is maintained between 30 s and 100 s for the considered scenario. This allows the network orchestrator to address the trade-off between bandwidth efficiency, energy efficiency, and delay in an Fe-RAN, and facilitates simple EC design for 5G to MNOs. Moreover, caching energy efficiency improves marginally on decreasing σ_i (say from 10 s to 1 s) which highlights the importance of the derived model. Therefore, arbitrarily choosing high σ_i to enhance QoS or low σ_i for energy efficiency will incrementally improve the respective figures at the cost of a significant decrease in other performance metrics.

Moreover, it is concluded that in a scenario with a low delay δ_i requirement, energy efficiency can be achieved by selecting low VS size κ and moderate σ_i (Figs. 3(a), 3(c)). Otherwise, the content provider will have to pay a high amount to the MNO in order to maintain a high σ_i. Furthermore, a QoS-aware energy- and bandwidth-efficient caching strategy should also aim to appropriately select the maximum number of VS downloads from the CU beyond which it is transferred to E_i. Finally, low user mobility and high caching time adversely affects caching energy efficiency as discussed in Section III.

Properly designed pricing strategies by the network orchestrator can enable MNOs to maintain σ_i in the desirable range [38]. This will facilitate bandwidth efficiency to the network providers by limiting the bandwidth consumed in the TN γ_{TN} to less than 6.1% (Section III-E), and allow content providers to satisfy stringent QoS requirements by limiting δ_i to ≈ 1 s. It will also allow MNOs to decrease VS

caching and transportation energy consumption by up to 93% and 90.3%, respectively, under the assumed network scenario (Section III-E). This corresponds to relatively better bandwidth efficiency and QoS compared to [15], in which a fronthaul load reduction of $\approx 35\% - 50\%$ was reported.

The performance results derived from the model illustrate the importance of setting the VS deletion rates at ECs to achieve delay, energy, and bandwidth efficiency. Moreover, the ability to set the desired delay (QoS, Fig. 3(a)) can allow the network orchestrator to design caching schemes for heterogeneous services. The analytical framework also provides insights into a possible game-theoretic setup, in which multiple tenants own the ECs and TN, and independently control σ_i and ζ_i to achieve the desired network performance with high energy efficiency.

REFERENCES

[1] Cisco VNI, "Cisco visual networking index: global mobile data traffic forecast update, 2017-2022," Tech. Rep., 2019.

[2] A. Aissioui, A. Ksentini, A. M. Gueroui, and T. Taleb, "On enabling 5G automotive systems using follow me edge-cloud concept," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 6, pp. 5302–5316, 2018.

[3] T. Taleb, A. Ksentini, and P. A. Frangoudis, "Follow-me cloud: When cloud services follow mobile users," *IEEE Transactions on Cloud Computing*, vol. 7, no. 2, pp. 369–382, 2019.

[4] S. He, Y. Huang, A. Nallanathan, C. Qi, and Q. Hou, "Two-level transmission scheme for cache-enabled fog radio access networks," *IEEE Transactions on Communications*, vol. 67, no. 1, pp. 445–456, 2018.

[5] C.-F. Liu, M. Bennis, and H. V. Poor, "Latency and reliability-aware task offloading and resource allocation for mobile edge computing," in *IEEE GLOBECOM Workshops*, 2017.

[6] R. Ferrús, O. Sallent, J. Pérez-Romero, and R. Agustí, "On 5G radio access network slicing: radio interface protocol features and configuration," *IEEE Communications Magazine*, vol. 56, no. 5, pp. 184–192, 2018.

[7] J. Li, X. Shen, L. Chen, J. Ou, L. Wosinska, and J. Chen, "Delay-aware bandwidth slicing for service migration in mobile backhaul networks," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 11, no. 4, pp. B1–B9, 2019.

[8] H. Chen, Y. Li, S. Bose, W. Shao, L. Xiang, Y. Ma, and G. Shen, "Cost-minimized design for TWDM-PON-based 5G mobile backhaul networks," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 8, no. 11, pp. B1–B11, 2016.

[9] X. Liu and F. Effenberger, "Emerging optical access network technologies for 5G wireless," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 8, no. 12, pp. B70–B79, 2016.

[10] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, "FemtoCaching: Wireless content delivery through distributed caching helpers," *IEEE Transactions on Information Theory*, vol. 59, no. 12, pp. 8402–8413, 2013.

[11] H. Hsu and K. C. Chen, "Optimal caching time for epidemic content dissemination in mobile social networks," in *IEEE International Conference on Communications*, Kuala Lumpur, Malaysia, 2016.

[12] O. Ayoub, F. Musumeci, M. Tornatore, and A. Pattavina, "Energy-efficient video-on-demand content caching and distribution in metro area networks," *IEEE Transactions on Green Communications and Networking*, vol. 3, no. 1, pp. 159–169, 2019.

[13] Z. Zhao, M. Peng, Z. Ding, W. Wang, and H. V. Poor, "Cluster content caching: an energy-efficient approach to improve quality of service in cloud radio access networks," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 5, pp. 1207–1221, 2016.

[14] Y. Jiang, M. Ma, M. Bennis, F. C. Zheng, and X. You, "User preference learning based edge caching for fog radio access network," *IEEE Transactions on Communications*, vol. 67, no. 2, pp. 1268–1283, 2019.

[15] J. Kwak, Y. Kim, L. B. Le, and S. Chong, "Hybrid content caching in 5G wireless networks: cloud versus edge caching," *IEEE Transactions on Wireless Communications*, vol. 17, no. 5, pp. 3030–3045, 2018.

[16] Y. T. Hou, J. Pan, B. Li, X. Tang, and S. Panwar, "Modeling and analysis of an expiration-based hierarchical caching system," in *IEEE Global Telecommunications Conference*, 2002, pp. 2468–2472.

[17] Y. T. Hou, J. Pan, C. Wang, and B. Li, "On prefetching in hierarchical caching systems," in *IEEE International Conference on Communications*, Anchorage, AK, USA, 2003, pp. 814–818.

[18] I. Afolabi, T. Taleb, K. Samdanis, A. Ksentini, and H. Flinck, "Network slicing and softwarization: A survey on principles, enabling technologies, and solutions," *IEEE Communications Surveys and Tutorials*, vol. 20, no. 3, pp. 2429–2453, 2018.

[19] S. Zhang, P. He, K. Suto, P. Yang, L. Zhao, and X. Shen, "Cooperative edge caching in user-centric clustered mobile networks," *IEEE Transactions on Mobile Computing*, vol. 17, no. 8, pp. 1791–1805, 2018.

[20] J. Goseling, O. Simeone, and P. Popovski, "Delivery latency trade-offs of heterogeneous contents in fog radio access networks," in *IEEE Global Communications Conference*, Singapore, 2017.

[21] R. Tandon and O. Simeone, "Cloud-aided wireless networks with edge caching: Fundamental latency trade-offs in fog Radio Access Networks," in *IEEE International Symposium on Information Theory*, Barcelona, Spain, 2016, pp. 2029–2033.

[22] J. Goseling and O. Simeone, "Soft-TTL: Time-Varying Fractional Caching," *IEEE Networking Letters*, vol. 1, no. 1, pp. 18–21, 2018.

[23] V. Martina, M. Garetto, and E. Leonardi, "A unified approach to the performance analysis of caching systems," in *IEEE INFOCOM*, Toronto, Canada, 2014, pp. 2040–2048.

[24] H. Che, Y. Tung, and Z. Wang, "Hierarchical web caching systems: Modeling, design and experimental results," *IEEE Journal on Selected Areas in Communications*, vol. 20, no. 7, pp. 1305–1314, 2002.

[25] X. Vasilakos, V. A. Siris, G. C. Polyzos, and M. Pomonis, "Proactive selective neighbor caching for enhancing mobility support in information-centric networks," in *ACM Proceedings of the Information-Centric Networking Workshop*, Helsinki, Finland, 2012, pp. 61–66.

[26] K. Zhang, S. Leng, Y. He, S. Maharjan, and Y. Zhang, "Cooperative content caching in 5G networks with mobile edge computing," *IEEE Wireless Communications*, vol. 25, no. 3, pp. 80–87, 2018.

[27] Y. Lu, M. Zhang, C. Song, L. Guan, D. Wang, S. Li, and Y. Zhan, "A multi-migration seamless handover scheme for vehicular networks in fog-based 5G optical fronthaul," in *24th OptoElectronics and Communications Conference/International Conference Photonics in Switching and Computing 2019*, Fukuoka, 2019.

[28] M. Savi, O. Ayoub, F. Musumeci, Z. Li, G. Verticale, and M. Tornatore, "Energy-efficient caching for video-on-demand in fixed-mobile convergent networks," in *IEEE Online Conference on Green Communications*, 2015, pp. 17–22.

[29] E. Cohen, E. Halperin, and H. Kaplan, "Performance aspects of distributed caches using TTL-based consistency," *Theoretical Computer Science*, vol. 331, no. 1, pp. 73–96, 2005.

[30] Y. Sun, M. Peng, S. Member, S. Mao, and S. Yan, "Hierarchical radio resource allocation for network slicing in fog radio access networks," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 4, pp. 3866–3881, 2019.

[31] W. Wang, Y. Zhao, M. Tornatore, H. Li, J. Zhang, and B. Mukherjee, "Coordinating multi-access edge computing with mobile fronthaul for optimizing 5G end-to-end latency," in *Optical Fiber Communication Conference*, San Diego, USA, 2018.

[32] S. Kourtis and R. Tafazolli, "Evaluation of handover related statistics and the applicability of mobility modelling in their prediction," in *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, London, UK, 2000.

[33] L. E. Clarke, D. Gross, and C. M. Harris, *Fundamentals of queueing theory*, 4th ed. Hoboken, New Jersey: John Wiley & Sons, 2008.

[34] O. Ayoub, F. Musumeci, D. Andreoletti, M. Mussini, M. Tornatore, and A. Pattavina, "Optimal cache deployment for video-on-demand delivery in optical metro-area networks," in *IEEE Global Communications Conference*, Abu Dhabi, Dubai, 2018.

[35] N. Wang, G. Shen, S. K. Bose, and W. Shao, "Zone-based cooperative content caching and delivery for radio access network with mobile edge computing," *IEEE Access*, vol. 7, pp. 4031–4044, 2019.

[36] B. P. Rimal, D. Pham Van, and M. Maier, "Mobile-edge computing versus centralized cloud computing over a converged FiWi access network," *IEEE Transactions on Network and Service Management*, vol. 14, no. 3, pp. 498–513, 2017.

[37] X. Ma, S. Zhang, W. Li, P. Zhang, C. Lin, and X. Shen, "Cost-efficient resource provisioning in cloud assisted mobile edge computing," in *IEEE Global Communications Conference*, Singapore, 2017.

[38] W. Huang, W. Chen, and H. V. Poor, "Request delay-based pricing for proactive caching: A Stackelberg game approach," *IEEE Transactions on Wireless Communications*, vol. 18, no. 6, pp. 2903–2918, 2019.