Resource-Efficient QoS-Aware Video Streaming Using UAV-Assisted Networks

Citation for the original published paper (version of record):
Bhar, C., Ghosh, D., Agrell, E. (2024). Resource-Efficient QoS-Aware Video Streaming Using UAV-Assisted Networks. IEEE Transactions on Cognitive Communications and Networking, 10(2): 649-659. http://dx.doi.org/10.1109/TCCN.2023.3336908

Resource-efficient QoS-Aware Video Streaming using UAV-Assisted Networks

Chayan Bhar, *Member, IEEE*, Debayani Ghosh, Erik Agrell, *Fellow, IEEE*

***Abstract*—Emerging video services are associated with stringent quality-of-service (QoS) and high data-rate requirements. Moreover, the presence of data-rate-hungry mobile users in future networks necessitate sophisticated design strategies. The deployment of unmanned aerial vehicle access point (UAP)-assisted networks (UANs) has been proposed to ensure high data-rates to mobile users. Moreover, UAPs can be equipped with energy-efficient caches to facilitate video delivery with stringent QoS. However, the mobility of users and UAPs may cause temporal variations in the QoS experienced by users. This paper conducts an extensive performance evaluation of a UAN, by studying the effect of user behavior, mobility of users and UAPs, and a temporal variation of video popularity on the QoS. The QoS is measured in terms of the delay experienced by the users. To that end, a time-dependent queueing model and its associated fluid approximation models are derived, which are illustrated to be reasonably accurate in an appropriate asymptotic regime. A detailed analysis of these models reveals that low delay, i.e., high QoS, can be ensured in UANs. Finally, a reinforcement-learning (RL) approach based on these models is utilized to minimize the number of deployed UAPs and the playout buffer size while guaranteeing a certain QoS.**

***Index Terms*—UAV-assisted networks, QoS-aware video delivery, time-dependent queueing model.**

I. Introduction

There has been a rapid increase in data traffic generated by mobile video streaming services in recent years. Internet video is expected to contribute up to 82% of all Internet traffic in 2022 [1]. Video traffic is also expected to increase with the increasing use of high resolution video formats. However, such traffic can congest existing networks in the near future due to their bandwidth-intensive nature. Network congestion will in turn affect the quality-of-service (QoS) [2]. However, over-provisioning the network infrastructure to accommodate the growing video traffic can result in high energy consumption and an expensive transport network. Furthermore, existing and future networks comprise highly mobile users that can cause a spatio-temporal variation in the user density and video request patterns. This necessitates migration to new networks, like Unmanned aerial vehicle access point-assisted networks (UANs) that can offer stringent QoS [1] with efficient usage of the access points and the playout buffers, which in this paper are called "resource efficiency" for short.

UANs employing mobile unmanned aerial vehicle access points (UAPs) have been proposed to provide consistent communication links to mobile end users in next generation networks [3]. The UAPs can be co-deployed with the existing base stations, whenever there is a sudden increase in the number of mobile users. Moreover, UAPs can be strategically deployed to follow mobile users on a need basis in geographic regions with high user density or data requirements [4]–[6]. This makes UAPs more effective and cost-efficient compared to deploying static ultra-dense small cells [7]. Therefore, UAPs provide the flexibility of managing a time-varying density of users [3], [8].

On the other hand, edge caching is proposed as a potential solution to the bandwidth congestion problem in core and metro networks [9]. Edge caching can be implemented in UANs by equipping UAPs with light and energy-efficient small fog caches that can cache popular bandwidth-intensive on-demand videos [10]. Such UAPs can stream popular videos to mobile users with extremely high data rates [11], [12]. Therefore, such a scheme can ensure stringent QoS and network bandwidth efficiency, while simultaneously decreasing the bandwidth consumption of the transport network, as videos are streamed from network locations near the end users.

A. Literature Survey

We divide the related literature in three categories: (i) UAP placement and trajectory optimization for its power efficiency, (ii) user-dependent UAP deployment strategies, and (iii) cache placement and video caching strategies in UANs. Trajectory optimization of a unmanned aerial vehicle (UAV) using various algorithms is illustrated to maximize the throughput and power efficiency of an individual UAP in [5], [6], [13], [14]. Power-efficient allocation of a UAP from a set of deployed UAPs that have fixed locations for a particular coverage area using optimal transport theory and user behavior prediction is discussed in [8] and [12], respectively. Semi-definite relaxation and coordinate descent methods are used to design a bandwidth- and power-efficient scheme in which UAPs equipped with caches serve users having heterogeneous data-rate requirements [10]. Placement of a UAP to maximize its coverage in a scenario with QoS bounds using the exhaustive search method is discussed in [15].

The deployment and placement of a UAP as a function of the users' parameters like (i) user density and mobility [3]–[5], [11], [16], (ii) end user throughput [6], (iii) video

This work was funded in parts by SERB under grant SRG/2020/001364 and by the Swedish Research Council (VR) under grant 2021-03709.

C. Bhar is with the Department of Electronics and Communication Engineering, National Institute of Technology Warangal, Telangana, India. (e-mail: cbhar@nitw.ac.in).

D. Ghosh is with the Department of Electronics and Communication Engineering, Thapar Institute of Engineering and Technology, Patiala, Punjab, India.

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

requests [12], and (iv) QoS [15], [17]–[19] requirements have also been studied in the literature. It is illustrated that mobile users experience poor QoS, i.e., high delay if the time to deploy a UAP is too high [5]. High data-rates resulting in good QoS can be achieved by increasing the number of deployed caches in small-cell stationary access points using a greedy method [20]. Optimal cache placement, power allocation, and trajectory optimization subject to the mobility constraints of a UAP is illustrated to maximize user throughput using the block-coordinate descent and successive convex approximation methods [21]. UAP deployment as a function of the time-averaged request arrival and download rates is illustrated to minimize the video streaming delay in [18], [22]. User mobility and energy-efficiency-aware UAP movement are illustrated to limit delay [17] and the average transmit power [19]. Finally, caching of contents in a high-altitude platform using federated learning is implemented for estimating content popularity in [23]. Optimal cache placement using the entropy weight method is illustrated to minimize the content access delay in a scenario with multiple access points (APs) and cache units [24]. The delay experienced by users at concurrent time intervals is used for designing the transmission power of UAVs in [12]. Delay can be experienced by a user due to (i) content retrieval from the caches [21]–[24], (ii) request queueing at UAPs, and (iii) video streaming from a UAP [5], [12], [20], [22].

Reinforcement-learning (RL)-based strategies for energy minimization in UAV-assisted wireless powered sensor networks and for learning the optimal control policy to minimize delay in queueing networks have been proposed in [25], [26], respectively. However, such methods have not been implemented for designing UAN deployment as a function of QoS and user mobility parameters.

B. Motivation

The deployment of UANs has been proposed for edge caching and last mile content delivery using UAVs thereby reducing backhaul traffic and improving QoS [12], [27], [28]. Even a modest cache size of 100 Mbits can reduce the backhaul traffic by more than 50% in favorable cases [29, Fig. 4]. Mobile UAPs can be deployed in or traverse through different geographical locations in which they can experience a spatio-temporal variation in the video popularity [12], [30]. Similarly, the simultaneous mobility of UAPs and users can cause loss of connectivity and a variation in the download times observed by users [16]. Inconsistent download rates while streaming videos from the UAPs can result in poor QoS. Moreover, UAPs suffer from energy, space, and size limitations, which limit the hardware capability and equipment complexity of UAPs [3], [5], [8], [9], [13]. Therefore, the maximum user density and frequency of video requests that can be served by a UAP is limited and affected by its energy [15]. For example, if a UAP is low on charge, it can stream to a small number of users [8] or stream videos with low popularity [31], [7]. The number of users that can be served by UAPs in a UAN depends on the deployment of UAPs and is time-dependent. Thus, the deployment of UAPs must be managed in real time as a function of the videos' popularity and size, user mobility pattern [7], and the constraints related to UAP deployment [12]. Otherwise, the efficiency and effectiveness of UANs and the real-time QoS experienced by users can be affected. Thus, QoS is an important parameter to decide the optimal number of UAPs required.

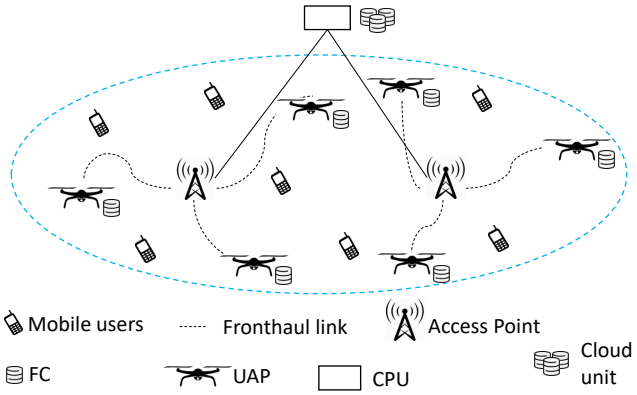

Fig. 1: Architecture for the proposed UAN.

UAPs can be deployed to serve users assuming that the QoS bounds allow some waiting time to the network provider before starting to stream a video. However, the user parameters required for UAP deployment in a geographical region are not available to individual UAPs. Furthermore, the energy limitation in UAPs discourages on-board implementation of computationally complex deployment strategies. However, existing literature does not discuss QoS-aware deployment of UAPs as functions of the time varying request pattern, user arrival, and channel conditions. Furthermore, minimizing the resources, i.e., number of UAPs deployed and the playout buffer size is also absent in existing literature.

C. Our contributions

Whereas previous literature on cache-enabled UANs has considered system parameters such as video popularity, user mobility, and channel conditions to be constant, this paper takes one step further and investigates how the experienced QoS (queueing delay) is affected under the more realistic scenario of time-varying system parameters. Using an analytic model and extensive simulations, it is shown that the temporal variations of these parameters, not only their averages, significantly affect the experienced QoS. This allows real-time UAP deployment for a resource-efficient UAN design.

The rest of the paper is arranged as follows. In the next section we describe the considered UAN and provide a detailed description of the system model. This is followed by an analysis of the derived model, results on QoS performance of the proposed UAN, and design of UAP deployment strategies using RL in Section III. Finally the paper concludes in Section IV with a discussion and conclusion.

II. System Model

The considered UAN architecture is illustrated in Fig. 1. It is similar to the architecture in [9], [24] but with coordinated

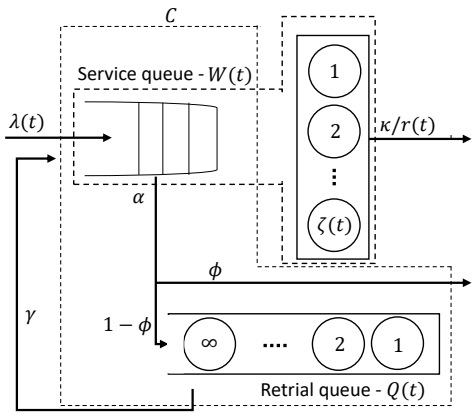

Fig. 2: Network of queues for the UAN.

UAPs. It is also similar to the architecture in [3], [31], which we extend by considering a time-varying deployment of UAPs. In this section, we formulate an analytical model for the system of Fig. 1. The symbols used for developing the queueing model are described in Table I while those used for developing the user mobility, channel, and video popularity models are described in Table II.

A. Functions of the central processing unit (CPU) and UAPs

The proposed UAN simultaneously employs traditional base stations and UAPs to provide consistent network throughput to all users. Each UAP connects to the CPU C via base stations using fronthaul links. The CPU C is centrally located and performs processing of the payload data from UAPs and base stations, and video popularity estimation. This allows UAPs and base stations to carry minimal processing elements for energy efficiency. Thus, in our proposed scheme, the UAPs are used only for content delivery. All baseband functionalities are executed in C, which also connects UAPs and base stations to the core network. For this purpose, at any given time, each UAP is connected to C through a base station using wireless fronthaul links. UAPs forward the channel state information and video requests from users to the CPU over fronthaul links using the scheme discussed in [32]. Thus, at any given point of time, the CPU has knowledge of the channel state information and video request arrival from all UAPs. The CPU informs UAPs about video streaming requirements. UAPs position themselves in the two-dimensional space with a fixed height to serve the users mandated by C while conforming to their energy limitations. The UAPs prevent high interference to users in their respective coverage area by using the reinforcement learning-based mobility strategy of [33]. UAPs are shown to cause minimum interference to users when they hover below 10 m [34]. Thus, in most of our simulated scenarios we fix the UAP height to 10 m. Moreover, the UAPs utilize the interference-minimization techniques described in [35] to maximize end-user data rates. The CPU C utilizes an estimated streaming rate for all users throughout the UAN. It also utilizes UAPs to facilitate fast streaming of bandwidth-intensive videos, thereby providing high QoS at spatio-temporal locations that have a high intensity of video requests [31]. The UAPs operate as cache helpers in a femto-caching environment [20]. For this purpose, each UAP D_i, $i \in \{1, \ldots, I\}$ is equipped with a small, energy-, weight-, and space-efficient femto cache (FC) F_i [5], [31], while the CPU C is equipped with a cloud unit. The FCs are significantly light-weight compared to the cloud unit. The decision to stream a video v from the FC of a UAP to a requesting user is taken by C on a per-request basis.

TABLE I: List of symbols used for the queueing model. All time-varying functions are defined for time $t > -T$, unless otherwise stated.

Symbol	Description
I	Number of UAPs deployed
D_i	i^{th} UAP consisting of FC F_i
C	CPU
v	Considered video
κ	Size of v
$\zeta(t)$	The maximum number of simultaneous v streams supported by C, i.e., UAPs deployed at t, upper and lower limits ζ_{h}, ζ_{l}, period of deployment T_ζ
$\lambda(t)$	Ensemble mean of v request inter-arrival time to C at t
$r(t)$	Ensemble mean of the rate of download at t
ϕ	Probability of users to disconnect from C without downloading v
γ	Mean time after which requests from $Q(t)$ retry v download, desired γ_{M}
α	Mean time after which requests are abandoned while waiting for v in $W(t)$
$\tau,\ s$	Running variables of t. The "test" request arrives at $t = \tau$
$W(t)$	Number of requests enqueued in the service queue at C for requests of v, scaled queue $W^\eta(t)$, fluid limit $W^{\text{f}}(t)$, diffusion limit $W^d(t)$
$\hat{W}^{\text{f}}(t)$, $\hat{W}^{\text{d}}(t)$	Fluid and diffusion limits of $W(t)$ for $t > \tau$
$Q(t)$	Number of requests enqueued in the retrial queue for users that left $W(t)$ and can request v from C, scaled queue $Q^\eta(t)$, fluid limit $Q^{\text{f}}(t)$, diffusion limit $Q^d(t)$
$\mathbf{\Omega}(t)$	The system of queues consisting of $W(t)$ and $Q(t)$ considered for analysis, scaled system $\mathbf{\Omega}^\eta(t)$, fluid limit $\mathbf{\Omega}^{\text{f}}(0)$, diffusion limit $\mathbf{\Omega}^{\text{d}}(t)$
$A(t)$	Request arrival process for v at C, scaled process $A^\eta(t)$, fluid limit $A^{\text{f}}(t)$, diffusion limit $A^{\text{d}}(t)$
$\Delta(t)$	Departure process due to completion of v streaming from C, scaled process $\Delta^\eta(t)$, fluid limit $\Delta^{\text{f}}(t)$, diffusion limit $\Delta^{\text{d}}(t)$
$E^{\text{f}}(t)$	Fluid limit of service initiation process at UAPs with derivative $e^{\text{f}}(t)$, diffusion limit $E^{\text{d}}(t)$, and scaled process $E^\eta(t)$
$P^{\text{f}}(t)$	Fluid limit of attainment process, scaled process $P^\eta(t)$
$\hat{}$	Operator for processes observed at $t \geq \tau$
$-T$	Lower limit of time for observing all processes at C to calculate delay
$\delta^{\text{f}}(t)$	Fluid limit of delay in downloading v, scaled delay $\delta^\eta(t)$, diffusion limit $\delta^{\text{d}}(t)$, desired δ_{M}
$X_i(t)$	Intensity for a time-inhomogeneous Poisson process $\Pi_i()$

TABLE II: List of symbols used for the user mobility, channel, and video popularity models and the resource optimization problem.

Symbol	Description
$\langle \Delta r \rangle$	User displacement from initial position
α_1, α_2	Parameters of distribution for Δt, Δr
$S(t)$	number of distinct locations visited by a user till t
ρ, ν	parameter of user mobility, predicting a user's tendency to explore a new location
α_3	Path loss exponent
$f_{\rm c}$, c, B	Carrier frequency, speed of light, and bandwidth available to a UAP-user link
h, $d(t)$	height of UAP, distance of user from UAP on ground plane
δ_1, δ_2	Excessive path loss exponent for line-of-sight and non-line-of-sight links between UAP and user
$p_{\rm LoS}$, $p_{\rm NLoS}$	probability of line-of-sight and non-line-of-sight links
σ^2	AWGN noise power
ψ, ψ_1	Environment-dependent constant for modeling the wireless channel
$p(t)$, $P(t)$	UAP-to-user transmit power, path loss between UAP and user
$P_i(t)$	Path loss experienced by interefering signals from UAP i to a particular user
α_4, α_5	Parameters for direct and word-of-mouth recommendation of v
q	Intrinsic popularity of the video
$N(t)$	Number of users in a geographical region, lower limit $N_{\rm l}$, upper limit $N_{\rm h}$, period of variation T_N
β_1, $a_1, \cdots, a_5$	Constants of the optimization problem in Section III-B
Ξ	Total interference experienced by a user from all UAPs

B. Network model

The model formulated in this paper is with respect to a particular video v, CPU C, and UAPs D_i, $i \in \{1, \ldots, I\}$. It is assumed that physical resources of UAPs like storage capacity, battery, energy efficiency, cache placement, etc., are managed according to [5]. The interplay between different videos due to the size limitation of FCs is not considered for simplicity of analysis. Since video popularity varies spatio-temporally, a CPU can experience a temporal variation in its observed video popularity [30]. The popularity is modeled by the inter-arrival time between requests for a video v. Empirical evidence suggests that video popularity is modeled by a Zipf distribution. However, for a single video on a short timescale, the request arrival process can be modelled by a pseudo-stationary Poisson point process [36]. Thus, the request inter-arrival time, a measure of video popularity, is exponentially distributed with ensemble mean $\lambda(t)$ at time t [30], [36], [37]. When a user requests v, the request is transferred to C, which employs UAPs to deliver v to requesting users. A network operator owns the terrestrial network and C and pays the mobile network operators owning D_i for streaming videos. The business model for the UAN can be similar to [38], [39]. For deciding the payment for streaming videos, the network operator considers the ensemble mean of inter-arrival time $\lambda(t)$ and the QoS experienced by end users. Poor channel conditions, low battery charge, or low $\lambda(t)$ can require a high payment.

A UAP D_i caches v depending on its popularity [30], [36] and the payment available from C [40]. Therefore, a measure of video popularity is required both at the CPU and by mobile network operators owning D_i. The CPU shares the estimated $\lambda(t)$ with the mobile network operators to decide video caching in D_i. When a user requests a video v, C coordinates its delivery from the FC of a UAP connected to C and having v [40]. The v streaming rate $r(t)$ depends on the propagation channel between the UAP and user and the battery state of the concerned UAP [16], [41]. The download time for v with size κ is assumed to be exponentially distributed with mean $\kappa/r(t)$ for all users. As UAPs can be mobile, the users downloading v can experience a temporal variation in the streaming rates and hence in the download data-rates. The ensemble mean of the download time is assumed to be equal for all users for simplicity of analysis. The total number of users that can simultaneously stream v from UAPs associated with C is limited to $\zeta(t)$. The CPU controls $\zeta(t)$ and I by varying the payments provided to UAPs.

At any point in time, if $\zeta(t)$ UAPs simultaneously stream v to an equal number of users, then further requests for v wait in a buffer located at C for completion of the ongoing streams. A user with a queued request may lose interest in v, resulting in deletion of the queued request. Moreover, the link quality between a user with a queued request and the CPU may momentarily deteriorate due to network outage resulting from high interference, airflow disturbances, beam steering errors, etc. [42]. Such events result in deletion of the queued requests after an exponentially distributed interval with mean α [43]. Therefore, high α can arise from frequent channel disruptions or high user mobility. If the link with a user deteriorates, then the request is dropped with probability ϕ and reenqueued otherwise. If it is reenqueued, it happens after an exponentially distributed interval with mean γ [16], [44]–[46]. If v corresponds to a video segment in a fractional caching scheme [47], [48], γ can be proportional to the time for emptying the playout buffer [11]. Therefore, if users are equipped with small playout buffers, then γ must be small.

C. Queueing model

Below we formulate a queueing model for the CPU C using the assumptions of Section II-A and the queue illustration of Fig. 2. The model consists of a service queue and a retrial queue. The service queue $W(t)$ comprises the users downloading v from UAPs and the users for which the corresponding requests are queued at C. The retrial queue $Q(t)$ consists of users that requested v but their requests were removed from $W(t)$ due to channel deterioration. Thus, $W(t)$ and $Q(t)$ are positive integers that represent lengths of the service and retrial queues, respectively. Requests from $Q(t)$ are again placed in $W(t)$ after γ s. We assume that $W(t)$ is a first-in-first-out type of queue. If there is an exogenous request for v at time t and the number of users streaming v at t is less than $\zeta(t)$, then $W(t) < \zeta(t)$. In this case, v is streamed immediately by a UAP to the requesting user. In contrast, if $\zeta(t)$ users are simultaneously streaming v at time t then $W(t) = \zeta(t)$ and

the exogenous request is placed in $W(t)$. γ and $\zeta(t)$ are upper limited to γ_{M} and ζ_{h}, respectively. Deletion of queued requests from $W(t)$ is with rate α due to the reasons mentioned in Section II-A. The network resource optimization performed in this paper is with respect to γ and $\zeta(t)$ as described in Section III-B using (34) and (35).

In the context of the proposed UAN, the arrival and service processes in the queueing system of Fig. 2 are time-dependent Markovian processes. Moreover, the number of simultaneous v streams from C is also time-dependent. Therefore, the queueing model for C is an $M(t)/M(t)/\zeta(t)$ queue, in which $W(t)$ allows abandonment, while $Q(t)$ facilitates retrial [49]. In general, such a queueing model is analytically intractable [49]. However, to make this model amenable to analysis, a fluid approximation is often adopted in the queuing literature. In the asymptotic regime wherein the arrival rate and the number of simultaneous streams from the service queue are large, the fluid approximation has been shown to be accurate. Motivated by this, the proposed system in this work is modeled along similar lines as in [49].

We denote with $\Pi_i(X_i(t))$ independent, inhomogeneous Poisson point processes with time-varying intensity functions $X_i(t)$ [50, Sec. 5.1]. The length of the service queue at time t increases with the number of v requests (i) present initially, $W(0)$, (ii) arriving from the retrial queue, $(1/\gamma)\Pi_1(Q(t))$, and (iii) arriving exogenously, $\Pi_2(1/\lambda(t))$. On the other hand, the service queue decreases with the number of v requests (i) leaving the service queue, $(1/\alpha)\Pi_3((W(t)-\zeta(t))^+)$ where $x^+ = \max(x,0)$ and (ii) completing v streaming $(1/\kappa)\Pi_4(\min(W(t),\zeta(t)))r(t)$.

The length of the retrial queue at time t increases with the number of v requests present initially at $Q(0)$ and the requests routed to $Q(t)$ from $W(t)$. It decreases with the enqueueing of $(1/\gamma)\int_0^t \Pi_1(Q(s))\ \mathrm{d}s$ requests, from $Q(t)$ to $W(t)$. Therefore, the sample paths [51, Sec. 5.2] for $W(t)$ and $Q(t)$ are the unique solutions of the equations

$$\begin{aligned} &W(t) \\ &= W(0) + \frac{1}{\gamma}\int_0^t \Pi_1\left(Q(s)\right)\ \mathrm{d}s + \int_0^t \Pi_2\left(\frac{1}{\lambda(s)}\right)\ \mathrm{d}s \\ &\quad - \frac{1}{\alpha}\int_0^t \Pi_3\left(\left(W(s)-\zeta(s)\right)^+\right)\mathrm{d}s \\ &\quad - \frac{1}{\kappa}\int_0^t \Pi_4\left(\min\left(W(s),\zeta(s)\right)\right)r(s)\ \mathrm{d}s, \end{aligned} \tag{1}$$

$$\begin{aligned} Q(t) &= Q(0) + \frac{(1-\phi)}{\alpha}\int_0^t \Pi_3\left(\left(W(s)-\zeta(s)\right)^+\right)\mathrm{d}s \\ &\quad - \frac{1}{\gamma}\int_0^t \Pi_1\left(Q(s)\right)\ \mathrm{d}s \end{aligned} \tag{2}$$

where $1/\lambda(t)$, $r(t)$, and $\zeta(t)$ are locally integrable. We define $\mathbf{\Omega}(t) = (W(t), Q(t))$. Using the theory of strong approximations for Poisson processes [49], a random sample path construction of C is performed to do an asymptotic sample path analysis and obtain the fluid limit theorems. In the asymptotic regime, $\zeta(t)$ is scaled up in response to a similar scaling up of the arrival rate by customers. The asymptotic regime in C denoted by $^\eta$ corresponds to $\mathbf{\Omega}^\eta(t) = (W^\eta(t), Q^\eta(t))$, where $\mathbf{\Omega}^\eta(t)$ is the fluid approximation of $\mathbf{\Omega}(t)$. The scaled parameters are the initial conditions $\mathbf{\Omega}^\eta(0) = \lceil \eta\mathbf{\Omega}^{\mathrm{f}}(0) + \sqrt{\eta}\mathbf{\Omega}^{\mathrm{d}}(0)\rceil + \mathrm{o}(\sqrt{\eta})$ for constants $\mathbf{\Omega}^{\mathrm{f}}(0) = (W^{\mathrm{f}}(0),\ Q^{\mathrm{f}}(0))$ and $\mathbf{\Omega}^{\mathrm{d}}(0) = (W^{\mathrm{d}}(0),\ Q^{\mathrm{d}}(0))$, $\eta/\lambda(t)$, and $\eta\zeta(t)$ [49]. The scaled processes are

$$\begin{aligned} &W^\eta(t) \\ &= W^\eta(0) + \frac{1}{\gamma}\int_0^t \Pi_1\left(Q^\eta(s)\right)\ \mathrm{d}s + \int_0^t \eta\Pi_2\left(\frac{1}{\lambda(s)}\right)\ \mathrm{d}s \\ &\quad - \frac{1}{\alpha}\int_0^t \Pi_3\left(\left(W^\eta(s)-\eta\zeta(s)\right)^+\right)\mathrm{d}s \\ &\quad - \frac{1}{\kappa}\int_0^t \Pi_4\left(\min\left(W^\eta(s),\eta\zeta(s)\right)\right)r(s)\ \mathrm{d}s, \end{aligned} \tag{3}$$

$$\begin{aligned} &Q^\eta(t) \\ &= Q^\eta(0) + \frac{(1-\phi)}{\alpha}\int_0^t \Pi_3\left(\left(W^\eta(s)-\eta\zeta(s)\right)^+\right)\mathrm{d}s \\ &\quad - \frac{1}{\gamma}\int_0^t \Pi_1\left(Q^\eta(s)\right)\ \mathrm{d}s. \end{aligned} \tag{4}$$

To obtain the fluid approximation, an asymptotic regime is considered wherein the arrival rate $1/\lambda(t)$ and $\zeta(t)$ are scaled up by a factor $\eta > 0$. Hence, in this new system $(W^\eta(t), Q^\eta(t))$, the arrival and service rates are $\eta/\lambda(t)$ and $\eta r(t)/$, respectively [52, Chapter 6]. Using the strong law of large numbers [51, App. A.2], the fluid limits are obtained for all $t > 0$ as

$$\lim_{\eta\to\infty}\frac{1}{\eta}W^\eta(t) = W^{\mathrm{f}}(t),\ \ \lim_{\eta\to\infty}\frac{1}{\eta}Q^\eta(t) = Q^{\mathrm{f}}(t) \tag{5}$$

with the initial conditions

$$\lim_{\eta\to\infty}\frac{1}{\eta}W^\eta(0) = W^{\mathrm{f}}(0),\ \ \lim_{\eta\to\infty}\frac{1}{\eta}Q^\eta(0) = Q^{\mathrm{f}}(0). \tag{6}$$

The convergences of (5) are uniform on any compact subset of $t \geq 0$ [49], while the convergences at time $t = 0$ follow from the strong law of large numbers [51, App. A.2] [53]. The validity of this assumption is illustrated in Section III.

The asymptotic assumptions made in (5) and (6) allow us to derive $W^{\mathrm{f}}(t)$ and $Q^{\mathrm{f}}(t)$ by dividing (3) and (4) by η on both sides and taking the limits as $\eta \to \infty$

$$\begin{aligned} &W^{\mathrm{f}}(t) \\ &= W^{\mathrm{f}}(0) + \frac{1}{\gamma}\int_0^t \Pi_1\left(Q^{\mathrm{f}}(s)\right)\ \mathrm{d}s + \int_0^t \Pi_2\left(\frac{1}{\lambda(s)}\right)\ \mathrm{d}s \\ &\quad - \frac{1}{\alpha}\int_0^t \Pi_3\left(\left(W^{\mathrm{f}}(s)-\zeta(s)\right)^+\right)\ \mathrm{d}s \\ &\quad - \frac{1}{\kappa}\int_0^t \Pi_4\left(\int_0^t \min\left(W^{\mathrm{f}}(s),\zeta(s)\right)\right)r(s)\ \mathrm{d}s, \end{aligned} \tag{7}$$

$$\begin{aligned} &Q^{\mathrm{f}}(t) \\ &= Q^{\mathrm{f}}(0) + \frac{(1-\phi)}{\alpha}\int_0^t \Pi_3\left(\left(W^{\mathrm{f}}(s)-\zeta(s)\right)^+\right)\mathrm{d}s \\ &\quad - \frac{1}{\gamma}\int_0^t \Pi_1\left(Q^{\mathrm{f}}(s)\right)\ \mathrm{d}s. \end{aligned} \tag{8}$$

Finally, differentiating (7) and (8) with respect to t, we obtain the rate of change in the two fluid limits $W^{\mathrm{f}}(t)$ and $Q^{\mathrm{f}}(t)$, which are given by the solutions of

$$\frac{\mathrm{d}}{\mathrm{d}t}W^{\mathrm{f}}(t) = \frac{1}{\lambda(t)} + \frac{Q^{\mathrm{f}}(t)}{\gamma} - \frac{r(t)\min\left(W^{\mathrm{f}}(t), \zeta(t)\right)}{\kappa} - \frac{\left(W^{\mathrm{f}}(t) - \zeta(t)\right)^{+}}{\alpha}, \tag{9}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}Q^{\mathrm{f}}(t) = \frac{(1-\phi)\left(W^{\mathrm{f}}(t) - \zeta(t)\right)^{+}}{\alpha} - \frac{Q^{\mathrm{f}}(t)}{\gamma}. \tag{10}$$

The queue length $W^{\mathrm{f}}(t)$ increases due to the arrival of exogeneous and retrial requests for v with rates $1/\lambda(t)$ and $Q^{\mathrm{f}}(t)/\gamma$, respectively. On the other hand, $W^{\mathrm{f}}(t)$ decreases with the completion of v streams and with requests leaving $W^{\mathrm{f}}(t)$ to enter $Q^{\mathrm{f}}(t)$ or leave the system permanently with rates $r(t)/\kappa$ and $1/\alpha$, respectively. The queue length $Q^{\mathrm{f}}(t)$ increases with retrial requests coming from $W^{\mathrm{f}}(t)$ and decreases when requests leave $Q^{\mathrm{f}}(t)$ to enter $W^{\mathrm{f}}(t)$ with rates $(1-\phi)/\alpha$ and $1/\gamma$, respectively. The queue lengths $W^{\mathrm{f}}(t)$ and $Q^{\mathrm{f}}(t)$ are used for deriving the delay as discussed in Section II-G.

D. User mobility model

The user mobility model discussed below is adopted from [54]. It is assumed that at $t_1 = 0$, the user is at a given location. After waiting for a random time Δt, the user moves to a new location with probability $\rho t^{-\alpha_1 \nu/(1+\nu)}$ and otherwise remains in the same location. Thereafter, the process repeats with another random time Δt. The time interval Δt is chosen from a heavy-tailed distribution $\propto |\Delta t|^{-1-\alpha_1}$. When the user moves, the distance Δr is chosen from another heavy-tailed distribution $\propto |\Delta r|^{-1-\alpha_2}$ and the direction is uniform in $[\,0, 2\pi\,)$. The mean-square user displacement with respect to their initial position is

$$E[\Delta r^2] \sim \left[\log\left(\frac{1 - S(t_1)^{1-\nu}}{\nu - 1}\right)\right]^{2/\alpha_2} + \text{constant}. \tag{11}$$

Moreover, the number of users in a particular geographical region is assumed to have a periodic variation [3]. Thus we assume

$$N(t) = \frac{N_{\mathrm{h}} - N_{\mathrm{l}}}{2}\sin\left(\frac{2\pi t}{T_N}\right) + \frac{N_{\mathrm{h}} + N_{\mathrm{l}}}{2}. \tag{12}$$

E. Video popularity model

The request inter-arrival time of a video is measured in real time by measuring the time elapsed from arrival of the previous request from all users at the CPU [36]. Video popularity is modeled by $\lambda(t)$ assuming direct and word of mouth recommendations α_4 and α_5, respectively, videos' intrinsic popularity q, and total number of users at time t $N(t)$, resulting in [55]

$$\lambda(t) = \frac{qN(t)\frac{\alpha_4}{\alpha_5 qN(t)}e^{(\alpha_4 + \alpha_5 qN(t))t} - \frac{\alpha_4}{\alpha_5}}{\frac{\alpha_4}{\alpha_5 qN(t)}e^{(\alpha_4 + \alpha_5 qN(t))t} + 1}. \tag{13}$$

Equations (4.1)–(4.7) of [56] can be used to derive $\lambda(t)$ at the CPU from real-time video request inter-arrival time data.

F. Channel model

The channel between the UAP and a user consists of line-of-sight and non-line-of-sight links. Assuming elevation of a UAP h, carrier frequency f, speed of light c, path loss exponent α_3, excessive path loss co-efficient for line-of-sight links δ_1 and non-line-of-sight links δ_2, horizontal distance of the user from the UAP $d(t)$, bandwidth B, environment-dependent constants ψ_1 and ψ_2, additive white Gaussian noise power σ^2, and UAP-to-user transmit power $p(t)$, the probabilities of the line-of-sight and non-line-of-sight links are

$$p_{\mathrm{LoS}}(t) = \frac{1}{1 + \psi\exp\left(-\psi_1\left[\frac{180}{\pi}\arctan(\frac{h}{d(t)}) - \psi\right]\right)} \tag{14}$$

and $p_{\mathrm{NLoS}}(t) = 1 - p_{\mathrm{LoS}}(t)$, respectively [57]–[60]. The path loss between a UAP and user is given by

$$P(t) = \left[\frac{4\pi f h}{c}\right]^{\alpha_3}\left(\delta_1 p_{\mathrm{LoS}}(t) + \delta_2 p_{\mathrm{NLoS}}(t)\right). \tag{15}$$

The CPU C uses the worst-case interference Ξ to estimate the download rate $r(t)$. For this purpose, C assumes that a user receives interference from all UAPs except its serving UAP. Moreover, other users are assumed to be present within Δr of a user while the UAPs are positioned to serve these users. Assuming signal power used by UAPs $p(t)$ and path loss between an interfering UAP i and the considered user $P_i(t)$, Ξ is derived similar to equation (5) of [35] as

$$\Xi = \sum_{i=1}^{\zeta(t)-1}\frac{p(t)}{P_i(t)}. \tag{16}$$

The path losses $P_i(t)$ in (16) are derived similarly to $P(t)$ in (15) for real-time positions of UAPs and users. For simplicity of analysis, C is assumed to estimate the same interference for all users. On the other hand, each UAP maximizes the signal-to-interference-plus-noise ratio in real time using the interference-management equations (6)–(28) of [35]. This results in higher streaming rates compared to that estimated by C. The Doppler shift effect due to mobility of the UAPs and users is assumed to be perfectly compensated. As an optimistic estimate of the time-varying download rate, we use

$$r(t) = B\log_2\left(1 + \frac{p(t)}{P(t)(\sigma^2 + \Xi)}\right), \tag{17}$$

at C which equals the channel capacity if $p(t)$ and $P(t)$ vary slowly enough.

G. Performance parameter: Delay $\delta^{\mathrm{f}}(t)$

Delay is experienced by users when the number of requests for v arriving from users is greater than the number of requests being currently streamed by UAPs. In this paper, the delay is defined in terms of the virtual delay, which is the waiting time experienced by a "test" user that requests v exogenously at any instant t, assuming that the channel conditions do not deteriorate before a UAP starts to stream v to it [51, Sec. 3.3].

In this section, we derive an expression for the delay. We first define the diffusion processes corresponding to the length of the service and retrial queues and thereafter assume that $\mathbf{\Omega}(t)$ is observed in the absence of exogenous arrivals. Delay is defined as an attainment process, i.e., the first instant at which a waiting request begins to stream v. For this purpose we make the following assumptions.

1) $\zeta(t)$ is continuously differentiable at all $t \geq 0$.
2) $r(t)$ is continuous at all $t \geq 0$.
3) No new exogenous requests arrive after the arrival of a request from a "test" user at time $\tau \geq 0$.

Furthermore, we assume that the limit $\lim_{\eta\to\infty} \sqrt{\eta}(\mathbf{\Omega}^{\eta}(0)/\eta - \mathbf{\Omega}^{\mathrm{f}}(0))$ exists and converges in distribution to $\mathbf{\Omega}^{\mathrm{d}}(0) = (W^{\mathrm{d}}(0),\ Q^{\mathrm{d}}(0))$ for $\eta \to \infty$. The fluid approximation $\mathbf{\Omega}^{\mathrm{f}}(t)$ can be refined using the functional central limit theorem [53]. Thus, for $t > 0$ and the diffusion process $\mathbf{\Omega}^{\mathrm{d}}(t)$ [61, Sec. 3.1], [49]

$$\lim_{\eta\to\infty} \sqrt{\eta}\left(\frac{1}{\eta}\mathbf{\Omega}^{\eta}(t) - \mathbf{\Omega}^{\mathrm{f}}(t)\right) \xrightarrow{d} \mathbf{\Omega}^{\mathrm{d}}(t), \tag{18}$$

is a convergence in distribution $\xrightarrow{d}$ of the stochastic processes in an appropriate functional space [53]. Moreover, if the set of time points $\{t \geq 0 \mid W^{\mathrm{f}}(t) = \zeta(t)\}$ has measure zero for the given queueing system, then $\mathbf{\Omega}^{\mathrm{d}}(t)$ is a Gaussian process [49] for $t \geq 0$.

The arrival process of v requests to C and the departure process associated with the streaming of v from $W(t)$ are denoted by $A(t)$ and $\Delta(t)$ with scaled processes $A^{\eta}(t)$ and $\Delta^{\eta}(t)$, respectively, for $t \geq 0$. In order to calculate the delay, $\mathbf{\Omega}(t)$ is observed after time $t = \tau$ for which the queues are denoted by the operator $\widehat{\ }$. Since there are no exogenous future arrivals, $A^{\eta}(0) = \widehat{W}^{\eta}(0)$, $\Delta^{\eta}(0) = 0$, $A^{\eta}(t) - \Delta^{\eta}(t) = \widehat{W}^{\eta}(t)$ for $t \geq 0$ [62, Chapter 2], and $A^{\mathrm{f}}(t) = A^{\mathrm{f}}(\tau)$ for $t \geq \tau$. The fluid limit results of (5) and (6) are written for $t \geq \tau$ as

$$\lim_{\eta\to\infty} \frac{1}{\eta}\left(\widehat{\mathbf{\Omega}}^{\eta}(t), A^{\eta}(t), \Delta^{\eta}(t)\right) \xrightarrow{d} \left(\widehat{\mathbf{\Omega}}^{\mathrm{f}}(t), A^{\mathrm{f}}(t), \Delta^{\mathrm{f}}(t)\right). \tag{19}$$

$\Delta^{\mathrm{f}}(t)$ is a continuously differentiable nondecreasing function of t for $t \geq 0$. For $0 \leq t \leq \tau$, $\widehat{W}^{\mathrm{f}}(t)$ satisfies (9) while for $t \geq \tau$, it is assumed that C does not have any new arrivals (exogenous or retrial) after time τ. Removing the terms corresponding to the arrivals after time τ from (9) results in

$$\frac{\mathrm{d}}{\mathrm{d}t}\widehat{W}^{\mathrm{f}}(t) = -\frac{\min\left(\widehat{W}^{\mathrm{f}}(t), \zeta(t)\right) r(t)}{\kappa} - \frac{\left(\widehat{W}^{\mathrm{f}}(t) - \zeta(t)\right)^{+}}{\alpha}. \tag{20}$$

The diffusion limits, which are obtained from the convergence in distribution similarly to (18), [61, Sec. 3.1], are

$$\lim_{\eta\to\infty} \sqrt{\eta}\Big(\frac{1}{\eta}\widehat{\mathbf{\Omega}}^{\eta}(t) - \widehat{\mathbf{\Omega}}^{\mathrm{f}}(t), \frac{1}{\eta}A^{\eta}(t) - A^{\mathrm{f}}(t), \frac{1}{\eta}\Delta^{\eta}(t) - \Delta^{\mathrm{f}}(t)\Big) \xrightarrow{d} \left(\widehat{\mathbf{\Omega}}^{\mathrm{d}}(t), A^{\mathrm{d}}(t), \Delta^{\mathrm{d}}(t)\right). \tag{21}$$

From (18), (21), and [62, Chapter 2], the diffusion limit $\widehat{W}^{\mathrm{d}}(t)$ can be written as

$$\widehat{W}^{\mathrm{d}}(t) = A^{\mathrm{d}}(t) - \Delta^{\mathrm{d}}(t). \tag{22}$$

We define the η-scaled process $E^{\eta}(t)$ as the sum of the processes corresponding to the completion of streaming and ongoing v streams as

$$E^{\eta}(t) = \Delta^{\eta}(t) + \eta\zeta(t), \qquad t \geq 0. \tag{23}$$

It follows that $(1/\eta)E^{\eta}(t)$ converges to $E^{\mathrm{f}}(t)$ almost surely for $\eta \to \infty$ where the convergence is uniform on compact sets of t and $E^{\mathrm{f}}(t) = \Delta^{\mathrm{f}}(t) + \zeta(t)$, $t \geq 0$. Since $\zeta(t)$ and $\Delta^{\mathrm{f}}(t)$ are continuously differentiable by earlier assumptions, $E^{\mathrm{f}}(t)$ is also continuously differentiable with derivative $e^{\mathrm{f}}(t)$. It is assumed that $e^{\mathrm{f}}(t)$ is strictly positive and $\lim_{t\to\infty} E^{\mathrm{f}}(t) > A^{\mathrm{f}}(\tau)$. According to previous definitions, $A^{\eta}(t)$ and $A^{\mathrm{f}}(t)$ are constant for $t \geq \tau$ assuming that the processes considered above are defined for $t \geq -T$ with

$$T = \frac{\zeta(0)}{e^{\mathrm{f}}(0)}. \tag{24}$$

This extension is made by assuming that nothing is happening in $-T \leq t < 0$ (no arrivals or departures) except that the number of users simultaneously streaming v increases linearly from 0 to $\eta\zeta(0)$. Therefore, (19) and (21) are retained without modification with all the functions being defined for $t \geq -T$.

The processes $A^{\mathrm{f}}(t)$, $E^{\mathrm{f}}(t)$, $A^{\mathrm{d}}(t)$, and $E^{\mathrm{d}}(t)$ are continuous and $E^{\mathrm{f}}(-T) = E^{\mathrm{d}}(-T) = 0$, where $\Delta^{\mathrm{d}}(t) = E^{\mathrm{d}}(t)$. Therefore, the processes discussed next are finite with probability 1 for all sufficiently large η. The first attainment processes are defined by the first instant s at which the number of times v is streamed exceeds the number of times it is requested after the arrival of request from the "test" user, i.e., the first instant when this request can start to stream v and are defined for all $t \geq -T$ as [49]

$$P^{\eta}(t) = \min\left\{s \geq -T : E^{\eta}(s) > A^{\eta}(t)\right\}, \tag{25}$$

$$P^{\mathrm{f}}(t) = \min\left\{s \geq -T : E^{\mathrm{f}}(s) > A^{\mathrm{f}}(t)\right\}. \tag{26}$$

The first attainment processes correspond to the first hitting time in a normal queuing process without a virtual queue [63, Sec. 6.2]. Thus, $P^{\mathrm{f}}(t)$, is the smallest instant $\tau \geq t$ such that $\widehat{W}^{\mathrm{f}}(\tau) = \zeta(\tau)$, i.e., the test user starts streaming v at time τ [49]. The delay experienced by the test user is given by the attainment waiting time processes as [51, Sec. 3.3]

$$\delta^{\eta}(t) = P^{\eta}(t) - t, \tag{27}$$

$$\delta^{\mathrm{f}}(t) = P^{\mathrm{f}}(t) - t. \tag{28}$$

Moreover, $\widehat{\delta}^{\eta}(\tau)$ denotes the virtual waiting time at τ, i.e., the delay experienced by the test user arriving to the service node at time τ, until it starts streaming v, assuming that this user does not quit or enter $Q(t)$ while waiting. Then the relation between the virtual waiting time and the attainment waiting time is $\widehat{\delta}^{\eta}(t) = \delta^{\eta}(t)^{+}$ for all $t \geq 0$. Furthermore, the following convergences in distribution follow from (19), (21), [61, Sec. 3.1], [49]

$$\lim_{\eta\to\infty} \left(\frac{1}{\eta}\widehat{\mathbf{\Omega}}^{\eta}(t), \frac{1}{\eta}A^{\eta}(t), \frac{1}{\eta}E^{\eta}(t), \frac{1}{\eta}\delta^{\eta}(t)\right) \xrightarrow{d} \left(\widehat{\mathbf{\Omega}}^{\mathrm{f}}(t), A^{\mathrm{f}}(t), E^{\mathrm{f}}(t), \delta^{\mathrm{f}}(t)\right), \tag{29}$$

$$\lim_{\eta\to\infty}\sqrt{\eta}\left(\frac{1}{\eta}\widehat{\mathbf{\Omega}}^{\eta}(t)-\widehat{\mathbf{\Omega}}^{\mathrm{f}}(t),\frac{1}{\eta}A^{\eta}(t)-A^{\mathrm{f}}(t),\right.$$
$$\left.\frac{1}{\eta}E^{\eta}(t)-E^{\mathrm{f}}(t),\delta^{\eta}(t)-\delta^{\mathrm{f}}(t)\right)$$
$$\xrightarrow{d}\left(\widehat{\mathbf{\Omega}}^{\mathrm{d}}(t),A^{\mathrm{d}}(t),E^{\mathrm{d}}(t),\delta^{\mathrm{d}}(t)\right), \tag{30}$$

where

$$\delta^{\mathrm{d}}(t)=\frac{A^{\mathrm{d}}(t)-E^{\mathrm{d}}\left(P^{\mathrm{f}}(t)\right)}{e^{\mathrm{f}}\left(P^{\mathrm{f}}(t)\right)} \tag{31}$$

is derived in [49]. Since the processes $A^{\mathrm{d}}(t)$, $E^{\mathrm{d}}(t)$, $\widehat{Q}^{\mathrm{d}}(t)$, and $\delta^{\mathrm{d}}(t)$ are continuous, the finite-dimensional distributions converge. In particular, considering the nontrivial case $P^{\mathrm{f}}(\tau)\geq\tau$ (i.e., $\widehat{W}^{\mathrm{f}}(\tau)\geq\zeta(\tau)$), if $0\leq t\leq\tau$, then the set of points $\{t\mid\widehat{W}^{\mathrm{f}}(t)=\zeta(t)\}$ has measure zero, and $\delta^{\eta}(t)$ converges to $\delta^{\mathrm{f}}(t)$ almost surely for $\eta\to\infty$. Thus, from (22), (31), the definition of $\Delta^{\mathrm{d}}(t)$, and assumption 3 above that results in $A^{1}(t)=A^{1}(P^{\mathrm{f}}(t))$ for all $t>-T$,

$$\delta^{\mathrm{d}}(t)\xrightarrow{d}\frac{\widehat{W}^{\mathrm{d}}\left(P^{\mathrm{f}}(t)\right)}{e^{\mathrm{f}}\left(P^{\mathrm{f}}(t)\right)}. \tag{32}$$

Also, $e^{\mathrm{f}}\left(P^{\mathrm{f}}(t)\right)=\zeta(P^{\mathrm{f}}(t))r(P^{\mathrm{f}}(t))/\kappa$ when $P^{\mathrm{f}}(t)\geq t$.

Since a delay is experienced by users when the number of v requests arriving from the request arrival process and waiting to stream v, $\widehat{W}^{\mathrm{f}}(\tau)$ is larger than the requests being streamed $\zeta(\tau)$. $P^{\mathrm{f}}(t)$ from (26) can be written as the first instant when the above occurs, i.e.,

$$P^{\mathrm{f}}(t)=\min\left\{\tau\geq t\mid\widehat{W}^{\mathrm{f}}(\tau)=\zeta(\tau)\right\}. \tag{33}$$

To compute $P^{\mathrm{f}}(t)$, (9) is used for $t<\tau$ and (20) for $t\geq\tau$.

A delay is experienced by the users if the UAPs simultaneously stream to $\zeta(t)$ users only, i.e., $\widehat{W}^{\mathrm{f}}(t)\geq\zeta(t)$. In such a scenario, the fluid limit of the delay $\delta^{\mathrm{f}}(t)$ is the difference between the instants when the request from the test user arrived and when a UAP started to stream it, $P^{\mathrm{f}}(t)$. Therefore, the delay is given by $\delta^{\mathrm{f}}(t)=P^{\mathrm{f}}(t)-t$ in analogy with (27).

III. Results

In this section, we conduct simulations using the queueing model derived in Section II for the scenarios depicted in Table III, and compare them with simulations carried out in OMNET++. Our simulations illustrate that the analytical models are fairly accurate in the asymptotic regime considered in this work. Further, using an RL approach, we also provide guidelines for resource-efficient deployment strategies of UAPs and CPU design.

A maximum of $I=144$ UAPs are assumed to serve up to 50000 users in [16]. The period of simulation is limited to 50 s. The variation in the number of video requests arriving as a function of time is illustrated in [65]. The values of parameters for user mobility, channel model, and video popularity model are taken from [64], [57]–[60], and [55], respectively. The UAP-to-user transmit power $p(t)$ is assumed to be a constant 10 mW [58]. The values of the parameters that are common in all simulations are grouped under the scenario $\boldsymbol{S}$ and listed in Table III. The performance results for $\boldsymbol{S}$ are a benchmark in the calculations of the delay. In scenario $\boldsymbol{S}_1$, κ is increased to observe the effect if a more bandwidth-intensive video is requested. The height h of UAPs is increased in $\boldsymbol{S}_2$ to observe the effect of interference while UAPs implement interference management. The effect of a different deployment strategy is illustrated in $\boldsymbol{S}_3$. In $\boldsymbol{S}_4$, the effect of increased link bandwidth B is explored. The user arrival pattern is changed by increasing the period of variation, decreasing the maximum number of users in a geographical area, and decreasing the users' tendency to explore a new region in $\boldsymbol{S}_5$–$\boldsymbol{S}_7$, respectively. The maximum number of UAPs deployable is increased in $\boldsymbol{S}_8$ while the effect of a video with low intrinsic popularity is explored in $\boldsymbol{S}_9$.

TABLE III: Values assumed for the parameters [55], [57]–[60], [64]

Sc.	Parameter	Value	Sc.	Parameter	Value
$\boldsymbol{S}$	κ	10^6 bits	$\boldsymbol{S}$	T_N	50
	N_{h}	50000		N_{l}	1
	ϕ	0.5		c	$3{\cdot}10^8$m/s
	ρ	0.6		h	10 m
	α_1, α_2	0.55, 0.8		ν	1.21
	α_3	4		q	0.05
	α_4, α_5	0.1, 10^{-7}		ψ	4.88
	ψ_1	0.43		δ_1	0.01
	δ_2	21		f_{c}	2 GHz
	B	1 MHz		σ^2	10^{-13}
	a_3, a_4, a_5	0.5, 0.45, 0.05		δ_{M}	20 s
				q	0.5
$\boldsymbol{S}_1$	$\kappa\nearrow$	10^9 bits	$\boldsymbol{S}_2$	h$\nearrow$	100 m
$\boldsymbol{S}_3$	$a_3\nearrow$ $a_4\searrow$ a_5	0.9, 0.05, 0.05	$\boldsymbol{S}_4$	$B\nearrow$	50 MHz
$\boldsymbol{S}_5$	$T_N\nearrow$	100	$\boldsymbol{S}_6$	$N_{\mathrm{h}}\searrow$	1000
$\boldsymbol{S}_7$	$\nu\nearrow$	2	$\boldsymbol{S}_8$	$I\nearrow$	200
$\boldsymbol{S}_9$	$q\searrow$	$5{\cdot}10^{-5}$			

The delay experienced in $\boldsymbol{S}$–$\boldsymbol{S}_9$ is reported in Figs. 4(a)–4(j). The analytical model is first solved numerically using MATLAB assuming a periodic deployment of UAPs of the form $\zeta(t)=\frac{\zeta_{\mathrm{h}}-\zeta_{\mathrm{l}}}{2}\sin(\frac{2\pi t}{T_\zeta})+\frac{\zeta_{\mathrm{h}}+\zeta_{\mathrm{l}}}{2}$ with $T_\zeta=50$s. The delay obtained from the analytical model proposed in Section II-C is validated using network simulations in OMNET++, assuming that the user base of C is limited to at most 50000 users. In the network simulations, the delay is measured as the difference in the time instants when a user requests v and some UAP starts to stream it. We observe from Fig. 4 that the qualitative nature of the delay obtained from network simulations and the analytical model match well. This validates the fluid approximation models derived in Section II. Thus, our proposed analytical framework is fairly accurate in the asymptotic regime considered in this work, and can aid analysis of the real-time QoS in UANs. However, a UAN should employ resource-efficient strategies to achieve a desired QoS, which requires involvement of the underlying model dynamics. We implement an RL agent that utilizes our fluid approximation model to achieve the above objectives.

The following subsections discuss the effect of varying the parameters listed in Table III on the delay $\delta^{\mathrm{f}}(t)$ experienced by a user while downloading v from the UAN associated

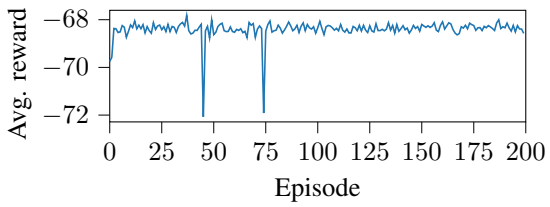

Fig. 3: Convergence of rewards in $\mathcal{S}$.

with C and the UAP deployment strategies implemented. The discussion proceeds by comparing $\delta^{\mathrm{f}}(t)$ observed in $\boldsymbol{S}_1$–$\boldsymbol{S}_9$ with that of $\boldsymbol{S}$.

A. Delay

The requests wait in the service or retrial queue at C before being streamed when $\widehat{W}^{\mathrm{f}}(t) > \zeta(t)$, resulting in a delay in $\boldsymbol{S}$–$\boldsymbol{S}_9$. The delay increases with $N(t)$, k, h, N_h, ν, and q as observed from the delay plots of $\boldsymbol{S}$–$\boldsymbol{S}_2$, $\boldsymbol{S}_6$, $\boldsymbol{S}_7$, and $\boldsymbol{S}_9$, respectively, while it decreases on increasing B and T_N in $\boldsymbol{S}_4$ and $\boldsymbol{S}_5$, respectively, as observed from Fig. 4. In $\boldsymbol{S}$–$\boldsymbol{S}_9$, $\lambda(t)\kappa/\zeta(t)r(t) > 1$ for certain time intervals, resulting in a very high delay in Fig. 4. Therefore, the request queuing delay is high in many scenarios. Furthermore, a variation in $\zeta(t)$ affects the delay more than $N(t)$. Thus, an efficient UAN design involves jointly optimizing the available resources and the real-time QoS experienced by the users. However, $\delta^{\mathrm{f}}(t)$ is a non-convex function. An optimization problem involving $\delta^{\mathrm{f}}(t)$, γ, and $\zeta(t)$ is difficult to solve. In the next subsection, we employ an RL agent in the CPU to solve the optimization problem and decide UAP deployment that can yield a high QoS.

B. Design of UAP deployment strategies

The CPU can change UAP deployment as a function of user mobility and the popularity of v for improving the QoS experienced by users while maximizing resource efficiency. Varying $\zeta(t)$ and γ, resulting in changes in the UAP deployment and CPU design, respectively, affects resource utilization. The goal is to maximize resource efficiency by minimizing $\zeta(t)$ and γ while minimizing delay which is also upper-bounded. Thus, the optimization problem is formulated as

$$\begin{aligned} \mathbf{P}_0: \quad & \min a_1\zeta(t) + a_2\gamma \\ \text{s.t.} \quad & 0 \geq \delta^{\mathrm{f}}(t) < \delta_{\mathrm{M}}, \\ & \zeta_{\mathrm{l}} \leq \zeta(t) \leq \zeta_{\mathrm{h}} \\ & 0 \leq \gamma \leq \gamma_{\mathrm{M}} \end{aligned} \tag{34}$$

for positive constants a_1, a_2. The problem $\mathbf{P}_0$ is a non-convex problem. We propose a deep Q network (DQN)-based solution method for (34). The state spaces $\mathcal{S}(t)$, action spaces $\mathcal{A}(t)$, and reward of the proposed DQN $\mathcal{R}(t)$ are designed by [25] for positive constant β_1, a_3, a_4, a_5, $0 \leq \gamma$, and $\zeta(t) \geq \zeta_{\mathrm{l}}$ such that

$$\begin{aligned} &\mathcal{S}(t) = \delta^{\mathrm{f}}(t) \\ &\mathcal{A}(t) = \gamma, \zeta(t) \\ &\mathcal{R}(t) = \begin{cases} \beta_1(1 - \zeta(t)/\zeta_{\mathrm{h}}), \text{if } \delta^{\mathrm{f}}(t) < \delta_{\mathrm{M}} \\ -\beta_1(a_3\zeta(t)/\zeta_{\mathrm{h}} - a_4\delta^{\mathrm{f}}(t)/\delta_{\mathrm{M}} - a_5\gamma/\gamma_{\mathrm{M}}), \\ \quad \text{otherwise.} \end{cases} \end{aligned} \tag{35}$$

The constants a_1 and a_2 refer to the weights given to the number of UAPs deployed and the mean retrial time, respectively, in the optimization problem of (34). The problem $\mathbf{P}_0$ is translated to derive the reward function $\mathcal{R}(t)$ of an RL algorithm in (35). In $\mathcal{R}(t)$, the constants a_3, a_4, and a_5 can be used to minimize the number of UAPs deployed, the maximum delay, and the mean retrial time, respectively, as done in $\boldsymbol{S}_3$. The RL simulations were performed in TensorFlow using Python 3.9. The RL agent is made to explore for 30000 episodes followed by the exploitation phase. The RL agent runs after each episode of duration 0.05 s. It is observed from Fig. 4 and Table IV that the RL agent can jointly minimize the delay and the number of UAPs required in different scenarios. Convergence of the proposed algorithm is tested numerically by running the RL agent for 200 consecutive epochs. Each epoch is assumed to be of 200 s. The average reward obtained by the RL agent in each epoch is plotted in Fig. 3. The average reward obtained in an epoch is observed to increase initially followed by saturation around -62. Thus, we conclude that the proposed algorithm converges in $\boldsymbol{S}$. The RL agent is also observed to converge in other scenarios for which the average reward is not plotted due to space restrictions. High k and h in $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$, respectively, result in the requirement for high $\zeta(t)$. In such scenarios, the RL agent prevents a high delay by adjusting $\zeta(t)$ and γ, as observed from Figs. 4(b), 4(c), and Table IV, respectively. A stringent policy for reducing the UAPs deployed $\zeta(t)$, causes a marginally higher delay in $\boldsymbol{S}_3$ and Fig. 4(d) compared to $\boldsymbol{S}$. A high bandwidth availability allows the RL agent to ensure lower resource consumption and delay in $\boldsymbol{S}_4$ compared to $\boldsymbol{S}$. A high period of variation in user arrival T_N, i.e., a less bursty nature of user arrival and less number of users N_h, results in lower delay and low $\zeta(t)$ and γ requirements in $\boldsymbol{S}_5$ and $\boldsymbol{S}_6$ compared to $\boldsymbol{S}$. This is illustrated in Figs. 4(f), 4(g) and in table IV. A high tendency of users to come back to their original locations ν results in marginally higher delay in Fig. 4(h) for $\boldsymbol{S}_7$ with higher $\zeta(t)$ requirement compared to $\boldsymbol{S}$. Availability of a higher number of deployable UAPs in $\boldsymbol{S}_8$ compared to $\boldsymbol{S}$ does not cause any changes in delay, γ or $\zeta(t)$, since $\zeta(t) < I$ in $\boldsymbol{S}$. A low video popularity also causes low delay, γ and $\zeta(t)$ in $\boldsymbol{S}_9$. Fig. 4 illustrates the importance of employing the RL agent in maintaining a good QoS in different network scenarios. Fig. 4 also compares the delay achieved with that in [18], [66]. Delay has been derived as a closed-form expression of the time-averaged parameters in [18] and is observed to be higher compared to that achieved by our RL agent from Fig. 4. In fact, the delay achieved by this scheme in scenarios $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ are in the order of 100s of seconds and could not fit into the plots of Figs. 4(b) and 4(c). On the other hand, the scenario of [66] is a system

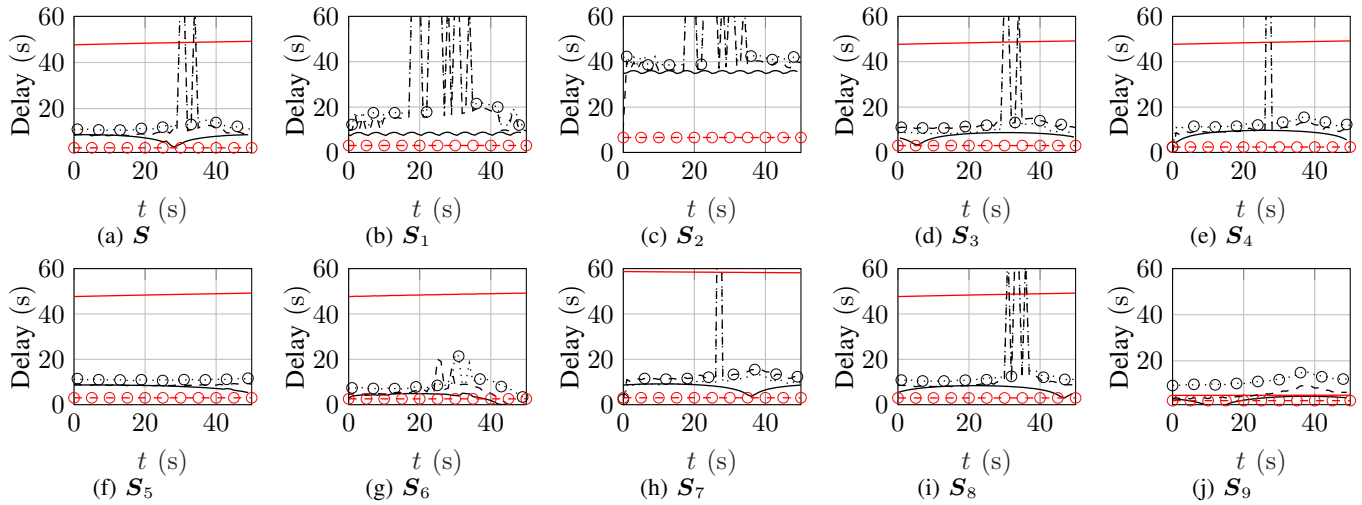

Fig. 4: Evolution of delay. Dashed black line: $\delta^{\mathrm{f}}(t)$ from MATLAB. Dotted black line with marker: Delay from network simulations. Black line: Delay from RL. Dashed red line with marker: Queueing delay of [66]. Red line: Queueing delay of [18].

TABLE IV: Values obtained from RL for γ and $\zeta(t)$

	$\boldsymbol{S}$	$\boldsymbol{S}_1$	$\boldsymbol{S}_2$	$\boldsymbol{S}_3$	$\boldsymbol{S}_4$	$\boldsymbol{S}_5$	$\boldsymbol{S}_6$	$\boldsymbol{S}_7$	$\boldsymbol{S}_8$	$\boldsymbol{S}_9$
$\zeta(t)$	93	94	127	98	59	97	78	112	92	62
γ (s)	10	10	10	2.5	5	5	2.5	1.42	5	5

with time-averaged Markovian arrival and service processes with finite servers and calling population. Thus, according to the analysis of [66], the queueing delay experienced by users before starting to stream the video is significantly lower and has an almost constant value. However, the time-averaged delay obtained by [18], [66] does not provide a true picture of the real-time QoS experienced by users, as observed from the delay obtained from MATLAB and network simulations in Fig. 4. This affects the UAP deployment. In contrast, our work captures the real-time queuing delay, i.e., QoS, and provides deployment strategies to ensure a desired QoS with resource efficiency.

IV. Conclusions

Present-day networks often face a challenge to deliver data-intensive multimedia video to highly mobile users with high QoS. Although UANs can offer high QoS, they can result in high deployment cost and impose logistic challenges due to the limitations imposed by UAPs. The correct UAN parameters must be varied to enable high QoS and cost-efficient deployment of UAPs simultaneously. This highlights the importance of the analytical model derived in this paper.

References

[1] Cisco, "Cisco Annual Report," Tech. Rep., 2020. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.pdf, visited on 10/11/2021

[2] C. Zhan, H. Hu, X. Sui, Z. Liu, J. Wang, and H. Wang, "Joint resource allocation and 3D aerial trajectory design for video streaming in UAV communication systems," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 8, pp. 3227–3241, 2021.

[3] J. Lu, S. Wan, X. Chen, Z. Chen, P. Fan, and K. B. Letaief, "Beyond empirical models: Pattern formation driven placement of UAV base stations," *IEEE Transactions on Wireless Communications*, vol. 17, no. 6, pp. 3641–3655, 2018.

[4] M. Banagar and H. S. Dhillon, "Performance characterization of canonical mobility models in drone cellular networks," *IEEE Transactions on Wireless Communications*, vol. 19, no. 7, pp. 4994–5009, 2019.

[5] J. Ji, K. Zhu, D. Niyato, and Ran Wang, "Joint cache placement, flight trajectory, and transmission power optimization for multi-UAV assisted wireless networks," *IEEE Transactions on Wireless Communications*, vol. 19, no. 8, pp. 5389–5403, 2020.

[6] Z. Becvar, P. Mach, J. Plachy, and M. F. P. De Tudela, "Positioning of flying base stations to optimize throughput and energy consumption of mobile devices," in *IEEE Vehicular Technology Conference*, 2019.

[7] Z. Becvar, M. Vondra, P. Mach, J. Plachy, and D. Gesbert, "Performance of mobile networks with UAVs," *European Wireless Conference*, pp. 261–267, 2017.

[8] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Optimal transport theory for power-efficient deployment of unmanned aerial vehicles," *IEEE International Conference on Communications*, 2016.

[9] N. Zhao, R. Yu, L. Fan, Y. Chen, J. Tang, A. Nallanathan, and V. C. M. Leung, "Caching unmanned aerial vehicle-enabled small-cell networks," *IEEE Vehichular Technology Magazine*, vol. 14, pp. 71–79, 2019.

[10] E. Kalantari, H. Yanikomeroglu, and A. Yongacoglu, "Wireless networks with cache-enabled and backhaul-limited aerial base stations," vol. 19, no. 11, pp. 7363–7376, 2020.

[11] L. Vigneri, T. Spyropoulos, and C. Barakat, "Low cost video streaming through mobile edge caching: Modelling and optimization," *IEEE Transactions on Mobile Computing*, vol. 18, no. 6, pp. 1302–1315, 2019.

[12] M. Chen, M. Mozaffari, W. Saad, C. Yin, M. Debbah, and C. S. Hong, "Caching in the sky: Proactive deployment of cache-enabled unmanned aerial vehicles for optimized quality-of-experience," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 5, pp. 1046–1061, 2017.

[13] C. Di Franco and G. Buttazzo, "Energy-aware coverage path planning of UAVs," *IEEE International Conference on Autonomous Robot Systems and Competitions*, pp. 111–117, 2015.

[14] J. Ji, K. Zhu, C. Yi, and D. Niyato, "Energy consumption minimization in UAV-assisted mobile-edge computing systems: Joint resource allocation and trajectory design," *IEEE Internet of Things Journal*, vol. 8, no. 10, pp. 8570–8584, 2021.

[15] M. Alzenad, A. El-Keyi, and H. Yanikomeroglu, "3-D placement of an unmanned aerial vehicle base station for maximum coverage of users with different QoS requirements," *IEEE Wireless Communications Letters*, vol. 7, no. 1, pp. 38–41, 2018.

[16] A. Fotouhi, M. Ding, and M. Hassan, "Flying drone base stations for macro hotspots," *IEEE Access*, vol. 6, pp. 19 530–19 539, 2018.

[17] S. H. Cheng, Y. T. Shih, and K. C. Chang, "Proactive deployment of cache-enabled aerial base stations for optimized energy-delay cost," in *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, (PIMRC)*, 2022, pp. 493–498.

[18] S. Kumar, S. Suman, and S. De, "Backhaul and delay-aware placement of UAV-enabled base station," *IEEE Conference on Computer Communications Workshops*, pp. 634–639, 2018.

[19] S. Anokye, D. Ayepah-Mensah, A. M. Seid, G. O. Boateng, and G. Sun, "Deep reinforcement learning-based mobility-aware UAV content caching and placement in mobile edge networks," *IEEE Systems Journal*, vol. 16, no. 1, pp. 275–286, 2022.

[20] 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.

[21] D. H. Tran, S. Chatzinotas, and B. Ottersten, "Satellite- and cache-assisted UAV: A joint cache placement, resource allocation, and trajectory optimization for 6G aerial networks," *IEEE Open Journal of Vehicular Technology*, vol. 3, pp. 40–54, 2022.

[22] J. Luo, J. Song, F. C. Zheng, L. Gao, and T. Wang, "User-centric UAV deployment and content placement in cache-enabled multi-UAV networks," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 5, pp. 5656–5660, 2022.

[23] A. Masood, T. V. Nguyen, T. P. Truong, and S. Cho, "Content caching in HAP-assisted multi-UAV networks using hierarchical federated learning," *International Conference on ICT Convergence*, pp. 1160–1162, 2021.

[24] L. Zhao, W. Sun, Y. Shi, and J. Liu, "Optimal placement of cloudlets for access delay minimization in SDN-based internet of things networks," *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 1334–1344, 2018.

[25] L. Liu, K. Xiong, Y. Lu, P. Fan, and K. B. Letaief, "Age-constrained energy minimization in UAV-assisted wireless powered sensor networks: A DQN-based approach," *IEEE Conference on Computer Communications Workshops*.

[26] B. Liu, Q. Xie, and E. Modiano, "RL-QN: A reinforcement learning framework for optimal control of queueing systems," in *Annual Allerton Conference on Communication, Control, and Computing*, Allerton, 2019.

[27] T. Zhang, Z. Wang, Y. Liu, W. Xu, and A. Nallanathan, "Joint resource, deployment, and caching optimization for AR applications in dynamic UAV NOMA networks," *IEEE Transactions on Wireless Communications*, vol. 21, no. 5, pp. 3409–3422, 2022.

[28] B. Zeng, C. Zhan, Y. Yang, and J. Liao, "Access delay minimization for scalable videos in cache-enabled multi-UAV networks," in *Proceedings of the IEEE Global Communications Conference, GLOBECOM 2022*. IEEE, 2022, pp. 389–394.

[29] T. Zhang, Y. Wang, Y. Liu, W. Xu, and A. Nallanathan, "Cache-enabling UAV communications: Network deployment and resource allocation," *IEEE Transactions on Wireless Communications*, vol. 19, no. 11, pp. 7470–7483, 2020.

[30] J. Wu, Y. Zhou, D. M. Chiu, and Z. Zhu, "Modeling dynamics of online video popularity," *IEEE Transactions on Multimedia*, vol. 18, no. 9, pp. 1882–1895, 2016.

[31] E. Kalantari, H. Yanikomeroglu, and A. Yongacoglu, "Wireless networks with cache-enabled and backhaul-limited aerial base stations," *IEEE Transactions on Wireless Communications*, vol. 19, no. 11, pp. 7363–7376, 2020.

[32] E. Björnson and L. Sanguinetti, "Scalable cell-free massive MIMO systems," *IEEE Transactions on Communications*, vol. 68, no. 7, pp. 4247–4261, 2020.

[33] H. V. Abeywickrama, Y. He, E. Dutkiewicz, B. A. Jayawickrama, and M. Mueck, "A reinforcement learning approach for fair user coverage using UAV mounted base stations under energy constraints," *IEEE Open Journal of Vehicular Technology*, vol. 1, no. January, pp. 67–81, 2020.

[34] M. Ding and D. L. Perez, "Please lower small cell antenna heights in 5G," *2016 IEEE Global Communications Conference, GLOBECOM 2016 - Proceedings*, pp. 2–7, 2016.

[35] U. Challita, W. Saad, and C. Bettstetter, "Interference management for cellular-connected UAVs: A deep reinforcement learning approach," *IEEE Transactions on Wireless Communications*, vol. 18, no. 4, pp. 2125–2140, 2019.

[36] T. Wang, C. Jayasundara, M. Zukerman, A. Nirmalathas, E. Wong, C. Ranaweera, C. Xing, and B. Moran, "Estimating video popularity from past request arrival times in a VoD system," *IEEE Access*, vol. 8, pp. 19 934–19 947, 2020.

[37] 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.

[38] R. I. Ansari, N. Ashraf, and C. Politis, "An energy-aware distributed open market model for UAV-assisted communications," *IEEE Vehicular Technology Conference*, vol. 2020-May, 2020.

[39] Y. K. Tun, Y. M. Park, T. H. T. Le, Z. Han, and C. S. Hong, "A business model for resource sharing in cell-free UAVs-assisted wireless networks," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 8, pp. 8839–8852, 2022.

[40] 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.

[41] S. Kourtis and R. Tafazolli, "Evaluation of handover related statistics and the applicability of mobility modelling in their prediction," *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, (PIMRC)*, pp. 665–670, 2000.

[42] B. Lin, W. Wang, J. Guo, and Z. Fei, "Outage Performance for UAV Communications under Imperfect Beam Alignment: A Stochastic Geometry Approach," *International Conference on Communication Technology Proceedings, ICCT*, pp. 632–637, 2021.

[43] F. A. Cruz-Pérez, A. Seguín-Jiménez, and L. Ortigoza-Guerrero, "Effects of handoff margins and shadowing on the residence time in cellular systems with link adaptation," *IEEE Vehicular Technology Conference*, 2004.

[44] P. K. Sharma and D. I. Kim, "Coverage probability of 3D mobile UAV networks," *IEEE Wireless Communications Letters*, vol. 8, no. 1, pp. 2018–2021, 2018.

[45] L. Libman and A. Orda, "Optimal retrial and timeout strategies for accessing network resources," *IEEE/ACM Transactions on Networking*, vol. 10, no. 4, pp. 551–564, 2002.

[46] C. M. Chen, C. W. Lin, and Y. C. Chen, "Cross-layer packet retry limit adaptation for video transport over wireless LANs," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 11, pp. 1448–1461, 2010.

[47] J. Goseling and O. Simeone, "Soft-TTL: Time-varying fractional caching," *IEEE Networking Letters*, vol. 1, no. 1, pp. 18–21, 2019.

[48] M. A. Maddah-Ali and U. Niesen, "Decentralized coded caching attains order-optimal memory-rate tradeoff," *IEEE/ACM Transactions on Networking*, vol. 23, no. 4, pp. 1029–1040, 2015.

[49] A. Mandelbaum, W. A. Massey, M. I. Reiman, A. Stolyar, and B. Rider, "Queue lengths and waiting times for multiserver queues with abandonment and retrials," *Telecommunication Systems*, vol. 21:2-4, pp. 149–171, 2002.

[50] L. C. Drazek, "Intensity estimation for Poisson processes," *Dissertation, Master of Science in Statistics, School of Mathematics, The University of Leeds*, 2013. [Online]. Available: http://129.11.36.6/ voss/projects/2012-Poisson/Drazek.pdf, visited on 20/10/2022

[51] M. El-Taha and S. Stidham, *Sample-Path Analysis of Queueing Systems*, 1st ed. Springer US, 1999.

[52] J. G. Dai and J. M. Harrison, *Processing Networks Fluid Models and Stability*. Cambridge University Press, 2020.

[53] A. Mandelbaum, W. A. Massey, and M. I. Reiman, "Strong approximations for Markovian service networks," *Queueing Systems*, vol. 30, pp. 149–201, 1998.

[54] C. Song, T. Koren, P. Wang, and A. L. Barabási, "Modelling the scaling properties of human mobility," *Nature Physics*, vol. 6, no. 10, pp. 818–823, 2010.

[55] J. Wu, Y. Zhou, D. M. Chiu, and Z. Zhu, "Modeling dynamics of online video popularity," *IEEE Transactions on Multimedia*, vol. 18, no. 9, pp. 1882–1895, 2016.

[56] J. Bigot, S. Gadat, T. Klein, and C. Marteau, "Intensity estimation of non-homogeneous Poisson processes from shifted trajectories," *Electronic Journal of Statistics*, vol. 7, no. 1, pp. 881–931, 2013.

[57] S. Ahmed, M. Z. Chowdhury, and Y. M. Jang, "Energy-efficient UAV-to-user scheduling to maximize throughput in wireless networks," *IEEE Access*, vol. 8, pp. 21 215–21 225, 2020.

[58] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Mobile unmanned aerial vehicles (UAVs) for energy-efficient internet of things communications," *IEEE Transactions on Wireless Communications*, vol. 16, no. 11, pp. 7574–7589, 2017.

[59] A. Bhowmick, S. D. Roy, and S. Kundu, "Throughput maximization of a UAV assisted CR network with NOMA-based communication and energy-harvesting," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 1, pp. 362–374, 2022.

[60] K. M. Liao, G. Y. Chen, Y. J. Chen, and Y. F. Chen, "End-to-end delay analysis in mmWave UAV-assisted wireless caching networks," in *IEEE Wireless Communications and Networking Conference Workshops*, 2020.
[61] G. Weiss, "Lecture 3: Queueing Networks and their Fluid and Diffusion." [Online]. Available: https://web.stanford.edu/class/msande324/handouts/Lecture3.pdf, visited on 10/11/2021
[62] L. Kleinrock, *Queueing Systems*. John Wiley & Sons, 1975.
[63] R. G. Gallager, *Stochastic Processing: Theory For Applications*. Cambridge University Press, 2013.
[64] C. Song, Z. Qu, N. Blumm, and A. L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.
[65] J. Liu, H. Yan, Y. Li, D. Karamshuk, N. Sastry, D. Wu, and D. Jin, "Discovering and understanding geographical video viewing patterns in urban neighborhoods," *IEEE Transactions on Big Data*, vol. 7, no. 5, pp. 873–884, 2021.
[66] Q. Liu, T. Xia, L. Cheng, M. Van Eijk, T. Ozcelebi, and Y. Mao, "Deep reinforcement learning for load-balancing aware network control in IoT edge systems," *IEEE Transactions on Parallel and Distributed Systems*, vol. 33, no. 6, pp. 1491–1502, 2022.