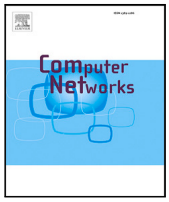
Queuing models of links carrying streaming and elastic services

Computer Networks 244 (2024) 110306

Contents lists available at ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet



Queuing models of links carrying streaming and elastic services

Andrea Marin [a], Marco Ajmone Marsan [b,c,*], Michela Meo [c], Matteo Sereno [d]

[a] Università Ca' Foscari Venezia, Italy
[b] IMDEA Networks Institute, Spain
[c] Politecnico di Torino, Italy
[d] Università di Torino, Italy

ARTICLE INFO

Keywords:
Streaming and elastic services
Queuing model
Product form and insensitivity

ABSTRACT

We consider an access link carrying data generated by streaming and elastic services requested by fixed or mobile end users, and subjected to an admission control (AC) algorithm. For the performance analysis of such link we develop a new queuing model and we show that, with the considered AC, the queuing model admits a product form expression for the joint limiting probability distribution of the numbers of active services of the different types. In addition, we prove that, when mobility can be neglected, i.e., in the case of either fixed access or slow mobility, the queuing model is insensitive to the distribution of the amount of data to be transferred for the fulfillment of the different service requests. Numerical results show unexpected oscillating behaviors for several performance metrics, and provide interesting insight into the link performance.

1. Introduction

1.1. Motivation

Since the early days of data networks, the offering of different types of constant bit rate services – that we call *streaming* services in this paper – has been a playground for network design and planning. The advent of packet networks has further complicated the scenario, with the introduction of *elastic* services that can adapt their data rate to better exploit the resources not used by streaming services.

While this mix of services appeared first in fixed networks, it then moved also to mobile networks. Now, with the roll-out of 5G, the mobile network operators (MNOs) goal is to allow their radio access networks (RANs) to carry data generated by an extremely wide range of different services, from video (like TV broadcasting) to audio (like voice or music), to messages, to real-time interactions supporting gaming, automated driving, factory automation, and all aspects of the tactile Internet. The gamut of services offered by operators can only be expected to increase with the arrival of 6G and the subsequent generations of networks. Designers are considering the applications related to the metaverse, including holographic virtual presence, and cooperative populations of IoT devices [1] This means that the need of designing and managing networks that can support many different services, each with its own traffic patterns and quality requirements, meeting for each service some specific set of key performance indicators (KPIs), is today very pressing for both fixed and mobile networks.

The prediction of the performance of such complex networks by simulation is extremely costly, since the network internal behaviors are complex, and the set of parameters that can be varied is extremely large. The availability of analytical models that can be solved with limited complexity is a critical asset that can be of enormous value in this case, since analytical models allow network managers to explore the RAN performance as a function of many system parameters, to understand the main trade-offs, to possibly exploit models to drive admission control (AC) and scheduling in real time, or to restrict choices to a manageable number of options that can then be explored with more detailed simulation models.

Unfortunately, not many tools are available for the performance analysis of networks offering different types of services. Traditional analytical models can be applied to either streaming or elastic services, not to their mixes. Only few papers have tried to analytically model the case of links loaded by a mixture of the two types of services, mostly considering scheduling and AC, rarely considering the effect of user mobility, and never before accounting for streaming services with bit rate that can adapt to resource availability (as it normally happens today with video). We briefly report on some of the relevant literature in the related work section.

The natural approach for the development of an analytical model of the type of system we consider lies in resorting to queuing theory. Indeed, queuing theory is the standard instrument for the investigation

* Corresponding author at: Politecnico di Torino, Italy.
E-mail addresses: marin@unive.it (A. Marin), marco.ajmone@imdea.org (M. Ajmone Marsan), michela.meo@polito.it (M. Meo), matteo.sereno@unito.it (M. Sereno).

https://doi.org/10.1016/j.comnet.2024.110306
Received 21 November 2023; Received in revised form 7 February 2024; Accepted 4 March 2024
Available online 12 March 2024

of systems where a finite set of resources (the access link data rate, in our case) is requested by a population of users (services, in our case). If resources are available, service can start. If resources are not immediately available, service requests could in general be forced to wait, or be rejected (as it happens in our case, giving rise to blocking of services).

1.2. Problem statement

The main objective of this paper is to propose a queuing model that can bypass the limitations of previous approaches in terms of complexity and scalability for the analysis of a portion of a RAN with mixes of services and an AC scheme. In particular, in this paper, extending the discussion in our conference paper [2], we present a queuing model that considers an arbitrary number of streaming service classes (e.g., real-time audio and video with different – and possibly adaptable – data rates) and elastic services, corresponding to variable data transfer rates (e.g., email, social network apps, non real-time audio/video services or web traffic). The contention for the finite link resources that arises in this scenario is solved by applying AC, and using the resources not allocated to streaming for elastic services. Moreover, mobility of users (hence of services) is taken into account to capture the dynamics of a RAN. Given the different characteristics of streaming and elastic services, the AC scheme acts differently on these two classes of traffic. Streaming service requests are accepted if they do not exceed a predefined maximum accepted number of requests and there is enough available bandwidth to satisfy the new request without compromising a minimum bandwidth dedicated to elastic services. For elastic services, instead, the AC aims at trading off two quality metrics: starting time of the service and throughput. The AC adjusts the acceptance of elastic services to the data rate not used by streaming services so that the load of elastic services is kept constant. This avoids that, by serving at too low bitrate, the quality of elastic services deteriorates. Elastic service in excess with respect to the target constant load are delayed.

When exponential assumptions are introduced, and with AC, the Markov process underlying the queuing system is proved to admit a product form solution, whose numerical tractability is shown to be high for practical scenarios. The product form among streaming service classes relies on the theory of Erlang-loss networks, whereas the product form between the streaming service classes and the elastic service classes was considered unlikely in the literature, and is unveiled in our work for the first time, extending here the results we presented in [2].

Thanks to the queuing model, we can efficiently compute a number of relevant interesting performance indices for each service: e.g., the distribution of the number of active services of the different types, the rate of finished services due to either completion or mobility, and the service request blocking probability.

When mobility is not considered, either because the model refers to a fixed network or because its effects are negligible with respect to service completion, we prove for the first time that performance indices depend only on the first moment of the amount of data transferred by services. This means that the crucial insensitivity property of processor sharing and Erlang-B queues is maintained also in this more complex scenario.

1.3. Contributions

The main contributions of the paper are the following.

- We define a novel queuing model of an access link supporting several classes of streaming and elastic services, also accounting for user mobility and access control.
- We study the queuing model under Markovian assumptions.
- We prove the existence of a product form solution for the joint limiting probability distribution of the number of streaming and elastic services when the proposed AC for elastic services is adopted. This product form solution was considered too difficult to prove in the previous literature.
- When mobility can be neglected, we prove the insensitivity of the joint limiting probabilities to the distribution of the amount of data to be transferred for the provision of the different services. This insensitivity allows studying services with non-exponential durations, which were considered too complex to analyze in the previous literature.
- We show how our queuing model can account for streaming services whose bit rate can adapt to resource availability. The case of video services with adaptive data rate was considered out of the modeling reach in the previous literature.
- We derive numerical results that show unexpected oscillating behaviors for blocking probabilities and other metrics. Non-monotonic behaviors of loss probabilities were only reported in one previous work [3], for a simpler case only including streaming services, and of much lower amplitude.
- We validate the exponential assumptions introduced in the model against results of simulations with non-exponential distributions for service requirements and service request processes.
- We provide interesting insight from the discussion of numerical results.

In particular, the proofs concerning product form and insensitivity are an original contributions to both queuing theory and network modeling.

1.4. Paper structure

The rest of this paper is organized as follows. Section 2 describes the characteristics of the access link we consider, and Section 3 introduces the corresponding queuing model. Section 4 presents the solution of the queuing model in the exponential case, and Section 5 proves the existence of a product form solution as well as the insensitivity of the model to the distribution of the service requirements in the case of negligible user mobility. Section 6 discusses the use of the analytical model to study the case of video streaming with adaptable data rate. Section 7 presents numerical results. Section 8 shows that the assumption of independence between streaming and elastic services often produces incorrect results. Section 9 presents simulation results that validate the exponential assumptions introduced in the model. Section 10 discusses some relevant previous work. Eventually, Section 11 concludes the paper.

2. The access link

In this section, we describe the access link we consider in this paper. Relevant parameters and corresponding notation are reported in Table 1.

We consider an access link supporting both elastic and streaming services. Streaming services are grouped in S different classes, each requiring a predefined data rate for a random time interval corresponding to the service duration. The data rate of streaming services can be either constant for the whole service duration or slowly varying to adapt to periods of high (or low) service demand. We will initially consider the constant data case, and show how to account for adaptivity later in the paper. We denote by $R_i^{(s)}$ and $\tau_i^{(s)}$, respectively, the required data rate and the random service duration of streaming services of class i, for $i \in \{1, 2, \dots, S\}$. Thus, a streaming service of class i requires the transfer of $R_i^{(s)}\tau_i^{(s)} = \varphi_i^{(s)}$ bits. Durations of class i services are independent and identically distributed (i.i.d.). The different classes of service can for example represent voice calls, multiparty voice/video conferences, real-time audio or video distribution of sport events, etc. Elastic services are

Table 1
Notation used for the system's parameters.

System's parameter	Notation
Link data rate	C
Dwell time (time in the cell), r.v.	δ
Number of streaming service classes Number of elastic service classes	S E
Maximum number of streaming services of class i Maximum number of elastic services of class i	$N_i^{(s)}$ $N_i^{(e)}$
Number of active streaming services of class i at time t Number of active elastic services of class i at time t	$n_i^{(s)}(t)$ $n_i^{(e)}(t)$
Minimum data rate for elastic services Required data date for a streaming service of type i	η $R_i^{(s)}$
Duration of a streaming service of type i, r.v. Data to be transferred in a streaming service of class i, r.v.	$\tau_i^{(s)}$ $\varphi_i^{(s)}$
Data to be transferred in an elastic service of class i, r.v.	$\varphi_i^{(e)}$

grouped in E different classes, each requiring the transfer of a random amount of data at the maximum possible data rate. We denote with the i.i.d. random variables $\varphi_j^{(e)}$, with $j \in \{1, 2, \dots, E\}$, the amount of data transferred by elastic services of class j. The different classes of elastic service can for example represent text or images or short videos transferred through messaging applications, web browsing, download of songs or videos (or video chunks) from repositories, etc.

The considered link is capable of providing a maximum user-plane data rate C, that can be allocated to end user services. The link can accommodate a maximum number $N_i^{(s)}$ of simultaneous streaming services of class i, reserving a minimum data rate η to elastic services[1], so that:

$$C \geq \eta + \sum_{i=1}^{S} R_i^{(s)} n_i^{(s)}(t) \tag{1}$$

where $n_i^{(s)}(t) \leq N_i^{(s)}$, $\forall i \in \{1, 2, \dots, S\}$ are the numbers of streaming services of each class active at time t, since each service instance of class i must be allocated the fixed data rate $R_i^{(s)}$ for the whole service duration. Clearly, it must hold that $N_i^{(s)} \leq \lfloor (C-\eta)/R_i^{(s)} \rfloor$ for $i = 1, \dots, S$.

Elastic services accept a variable service rate, and equally share the capacity that at any time instant is not allocated to active instances of streaming services. The maximum number of simultaneous elastic services of class j that can be activated on the link is denoted by $N_j^{(e)}$.

The state of the link at time t is described by the vector $\mathcal{N}(t) = [\mathcal{N}^{(s)}(t), \mathcal{N}^{(e)}(t)]$, with:

$$\mathcal{N}^{(s)}(t) = [n_1^{(s)}(t), \dots, n_S^{(s)}(t)] \tag{2}$$

$$\mathcal{N}^{(e)}(t) = [n_1^{(e)}(t), \dots, n_E^{(e)}(t)] \tag{3}$$

where $0 \leq n_j^{(e)}(t) \leq N_j^{(e)}$, $\forall j \in \{1, 2, \dots, E\}$, are the numbers of elastic services of each class at time t.

In the case of wireless access, base stations (BSs) interact with one another because of handovers of service instances generated by the movement of end users. The time spent by users within the cell defined by the BS, i.e., the end user *dwell time* in the cell, is described by the random variable δ. In the case of fixed access, users do not move, and the dwell time can be considered infinite.

With respect to a BS access link, a service request can either correspond to a new request for service that is started by a user located within the cell, or to an incoming handover due to a user with an active service entering the cell. A service completion in a BS access link can either be due to the end of a service while a user is within the cell, or to an outgoing handover because of the movement of the user out of the cell.

The processes of new service requests and incoming handovers can be considered to be either dependent or independent of the BS state. The approach often adopted in the literature is to consider those processes as related to the number of end users within the cell defined by the BS, but otherwise independent of the BS state, unless a state-dependent AC algorithm is applied.

In RANs, management procedures normally try to balance the load in nearby cells by governing the end user associations to BSs, as well as the selection of the cells for handovers, implementing an AC algorithm with the objective of avoiding congestion of BSs. Similarly, in wired access networks, new service requests can be rejected or delayed if they jeopardize the quality of active services. We can thus assume that some level of dependency of the processes of new service requests (and incoming handovers) on the link state exists. It can be reasonable to assume that dependency is higher for elastic services, whose quality is more sensitive to the BS load, since they can only use the data rate that remains after the allocation of the required data rate to all active streaming services.

For these reasons, in this paper we assume that both the processes of streaming and elastic service requests are subjected to an AC policy. For streaming service requests, we simply assume that access of a request of class i is permitted until the admission of a request violates the constraint represented by η, the minimum data rate reserved for elastic services, or the maximum number of admissible class i services has been reached. Requests violating the constraint are rejected, and must be resubmitted by the end user after a delay. In the case of elastic services, we consider a quasi-optimal (as we will show) AC policy, which is easy to implement, and leads to a very low complexity analysis that can guide in real time the AC algorithm. The AC algorithm has the objective of modulating the rates of accepted service requests according to the data rate not used by streaming services so that the load of elastic services on the bandwidth not used by streaming services is kept constant (as we will see later, this is the key for product form.) Requests for elastic services (both newly arriving and waiting) are admitted when the load is low. Conversely, when the load is high, requests are delayed to a period in which the link utilization is low again or, in the case of a portion of a RAN comprising several neighboring cells, requests are moved toward cells with more available resources. In addition, class j requests are rejected when the maximum number of admissible class j services has been reached. Note that the implementation of the AC algorithms for streaming and elastic services only requires information about the data rate collectively used by streaming services and the number of active streaming and elastic services of the different classes. These values can be easily monitored by the AC algorithm implementation software.

3. Queuing model of the access link

We now describe the queuing model we propose to study the considered system. Notation is reported in Table 2.

The operation of a wired or wireless link naturally maps onto a queuing model where the service speed corresponds to the link data rate C, and service instances are represented by customers. We will primarily consider the wireless case, including thus the user dwell time in the cell. In the wired case, the analysis is identical with an infinite value for the dwell time. The customers corresponding to streaming services (termed *streaming customers*) are handled according to a multiclass multiserver paradigm with losses, while the customers corresponding to elastic services (*elastic customers*) are handled according to a multiclass processor sharing paradigm with losses, with an overall service speed corresponding to the data rate not used by streaming customers.

Arrivals of streaming customers are assumed to follow Poisson processes, with losses induced by the AC algorithm (in the case of a BS we must account for both new service requests and incoming

[1] In most of this paper, we assume $\eta = 0$, but results can be straightforwardly generalized to the more general case.

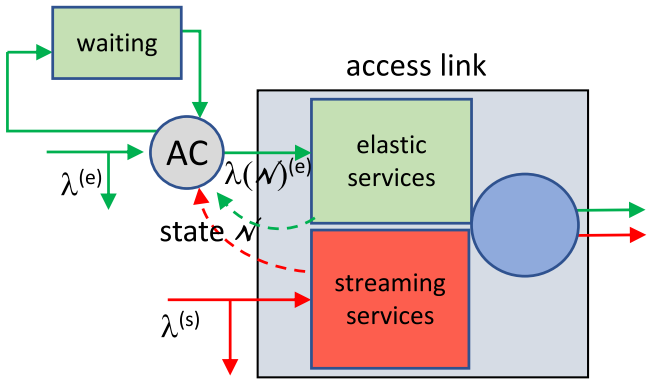


Fig. 1. Sketch of the queuing model of the access link.

Table 2
Notation used for the model.

Model parameters and assumptions	Notation
State of the Markov chain model, vector	$\mathcal{N}(t)$
Sub-vector of the state with active streaming services	$\mathcal{N}^{(s)}(t)$
Sub-vector of the state with active elastic services	$\mathcal{N}^{(e)}(t)$
Steady-state probability vector	$\pi_{\mathcal{N}}$
Poisson arrival rate of class i streaming services	$\lambda_i^{(s)}$
Poisson arrival rate of class i elastic services	$\lambda_i^{(e)}$
Rate of exponential dwell time δ	μ_H
Rate of exponential duration of a streaming service of class i	$\mu_i^{(s)}$
Rate of exponential duration of an elastic service of class i	$\mu_i^{(e)}(\mathcal{N})$
Rate of exponential amount of data in an elastic service of class i	$\alpha_i^{(e)}$
Data rate used by an elastic service of class i	$r_i^{(e)}(\mathcal{N})$

handovers; in the case of a wired link, handovers are not present). The arrival rate before losses of streaming customers of class i is denoted by $\lambda_i^{(s)}$.

Arrivals of elastic customers are also assumed to follow Poisson processes with rates $\lambda_j^{(e)}$ before AC, but become *state-dependent* Markovian processes, after the AC algorithm is applied. The AC algorithm modulates the rate of requests admitted in the queue according to the data rate available to elastic services. In a real system this means that during periods of scarce data rate for elastic services, the AC can either offload a request to a neighboring cell or postpone a request to a period of more abundance. Conversely, during periods of abundant data rate for elastic services, the AC can either accept request offloads from a neighboring cell or retrieve postponed requests.

The arrival rate of elastic customers of class j in state $\mathcal{N}$ due to new service requests (and incoming handovers in the wireless case) after AC is denoted by $\lambda_j^{(e)}(\mathcal{N})$. Note that the state dependency does not alter the average arrival rate, i.e. $E[\lambda_j^{(e)}(\mathcal{N})] = \lambda_j^{(e)}$.

Blocking of streaming customers of class i can be due to reaching the maximum number of permitted simultaneous services $N_i^{(s)}$ or to reaching the maximum usable bandwidth at the BS. Blocking of elastic customers of class j is due to reaching the maximum number of permitted simultaneous services $N_j^{(e)}$ (we allow the data rate of elastic services to temporarily drop to zero).

Service times of streaming customers of class i are i.i.d. random variables $\tau_i^{(s)}$, with general probability density function (pdf) and average $E[\tau_i^{(s)}] = 1/\mu_i^{(s)}$. The service times of elastic customers of class j are computed from the i.i.d. random variables $\varphi_j^{(e)}$, with general pdf.

Customers, both streaming and elastic, are *impatient*: they remain at the queue only for a random time δ (that models the dwell time in the case of wireless access and can instead be taken to be infinite in the case of fixed access), which is assumed to be distributed according to a negative exponential pdf with rate μ_H (in case of fixed access $\mu_H = 0$).

The queuing model of the access link is sketched in Fig. 1. Solid lines represent customer flows. Dashed lines indicate the state information given to the AC. The *waiting* box represents the offloading/postponement component of the AC.

4. Analysis of the queue in the exponential case

In order to study the queue with a continuous-time Markov chain (CTMC), it is necessary to introduce exponential assumptions for customer service times. The duration of streaming customer services of class i is thus assumed to be an exponentially distributed random variable with rate $\mu_i^{(s)}$. The duration of elastic customer services of class j depends on the BS data rate not allocated to streaming services. We assume that the amount of data transferred by elastic services of class j, $\varphi_j^{(e)}$, is an exponentially distributed random variable with rate $\alpha_j^{(e)}$ bits. The average duration of the file transfer is comprised between a minimum $1/(\alpha_j^{(e)}C)$ seconds (if the elastic service can use the whole server capacity C) and a maximum

$$\frac{\sum_{j=1}^{E} N_j^{(e)}}{\alpha_j^{(e)}\left[C - C_{\max}\right]} \tag{4}$$

where

$$C_{\max} = \max\Big(\sum_{i=1}^{S} R_i^{(s)} n_i^{(s)} s.t.\, \forall i\ n_i^{(s)} \le N_i^{(s)} \wedge \sum_{i=1}^{S} R_i^{(s)} n_i^{(s)} \le C - \eta\Big).$$

Expression (4) corresponds to a service rate equal to the ratio between the minimum data rate left by streaming services to elastic services and the maximum number of simultaneous elastic services. It should be noted that when the BS data rate for elastic services is large and mobility is slow, elastic customers observe a system behaving like a multiclass M/M/1-PS queue (i.e., a queue with Markovian arrivals, exponential service times and one server that equally divides its capacity among all customers – possibly belonging to multiple classes – at the queue in a processor sharing fashion) where service times are driven by the elastic file sizes. On the contrary, when the BS data rate for elastic services is small and mobility is fast, elastic customers observe a system behaving like a multiclass M/M/m/m queue (i.e., a queue with Markovian arrivals, exponential service times, no waiting line, and m servers with equal capacity, each serving one customer) where service times are driven by dwell times.

The rate of service of an elastic customer of class j in state $\mathcal{N}$ is:

$$\mu_j^{(e)}(\mathcal{N}) = \frac{\alpha_j^{(e)}\left[C - \sum_{i=1}^{S} R_i^{(s)} n_i^{(s)}\right]}{\sum_{\ell=1}^{E} n_\ell^{(e)}} \tag{5}$$

The individual departure rate of a streaming customer of class i in any state $\mathcal{N}$ with $n_i^{(s)} > 0$ is equal to:

$$\Delta_i^{(s)}(\mathcal{N}) = \mu_H + \mu_i^{(s)} \tag{6}$$

The individual departure rate of an elastic customer of class j in any state $\mathcal{N}$ with $n_j^{(e)} > 0$ is equal to:

$$\Delta_j^{(e)}(\mathcal{N}) = \mu_H + \mu_j^{(e)}(\mathcal{N}) \tag{7}$$

The expressions derived above allow the construction of a finite CTMC with a state space S comprising states such that $0 \le n_i^{(s)} \le N_i^{(s)}$, $0 \le n_j^{(e)} \le N_j^{(e)}$, for all $i = 1, 2, \ldots, S$ and $j = 1, 2, \ldots, E$ and $\sum_{i=1}^{S} n_i^{(s)} R_i^{(s)} \le C - \eta$.

The CTMC is finite and irreducible, hence it admits a steady-state distribution, which can be computed numerically, obtaining the limiting state probabilities $\pi_{\mathcal{N}} = \lim_{t\to\infty} P\{\mathcal{N}(t)\}$, where $P\{\mathcal{N}(t)\}$ is the probability that the CTMC state at time t is $\mathcal{N}(t)$.

The average number of active streaming services of class i is computed as

$$E[n_i^{(s)}] = \sum_{\mathcal{N}} n_i^{(s)}\, \pi_{\mathcal{N}} \tag{8}$$

and the average number of active elastic services of class j is computed as

$$E[n_j^{(e)}] = \sum_{\mathcal{N}} n_j^{(e)}\ \pi_{\mathcal{N}} \tag{9}$$

We denote by $E[n^{(s)}]$, and by $E[n^{(e)}]$ the average total (i.e., adding over all classes) number of active streaming and elastic services, respectively. We have:

$$E[n^{(s)}] = \sum_{i=1}^{S} E[n_i^{(s)}] \tag{10}$$

and:

$$E[n^{(e)}] = \sum_{j=1}^{E} E[n_j^{(e)}]. \tag{11}$$

Finally, the total average number of active services is $E[n] = E[n^{(s)}] + E[n^{(e)}]$.

The average utilization of the access link is

$$\rho = \sum_{\substack{\mathcal{N} \\ \mathcal{N}^{(e)}=\mathbf{0}}} \frac{n_i^{(s)} R_i^{(s)}}{C}\ \pi_{\mathcal{N}} + \sum_{\substack{\mathcal{N} \\ \mathcal{N}^{(e)}\neq\mathbf{0}}} \pi_{\mathcal{N}} \tag{12}$$

The residual capacity left by streaming services to elastic ones is given by

$$C^{(e)}(\mathcal{N}) = C - \sum_{i=1}^{S} R_i^{(s)} \tag{13}$$

The data rate used by an elastic service of class j in state $\mathcal{N}$ with $n_j^{(e)} \geq 1$ is

$$r_j^{(e)}(\mathcal{N}) = \frac{\left[C - \sum_{i=1}^{S} R_i^{(s)} n_i^{(s)}\right]}{\sum_{\ell=1}^{E} n_\ell^{(e)}} \tag{14}$$

and the corresponding probability is

$$P\{r_j^{(e)}(\mathcal{N})\} = \frac{\pi_{\mathcal{N}}}{1 - \sum_{\mathcal{N}^{(e)}=0} \pi_{\mathcal{N}}} \tag{15}$$

Hence, the average data rate used by an elastic service of class j is computed as

$$E[r_j^{(e)}] = \sum_{\substack{\mathcal{N} \\ \mathcal{N}^{(e)}>0}} r_j^{(e)}(\mathcal{N})\ P\{r_j^{(e)}(\mathcal{N})\} \tag{16}$$

The blocking (or loss) probabilities for streaming services of class i are

$$P\{\text{loss}_i^{(s)}\} = \sum_{\substack{\mathcal{N} \\ n_i^{(s)}=N_i^{(s)} \vee R_i^{(s)} + \sum_{k=1}^{S} n_k^{(s)} R_k^{(s)} > C-\eta}} \pi_{\mathcal{N}} \tag{17}$$

and the blocking (or loss) probability for elastic services of class j is

$$P\{\text{loss}_j^{(e)}\} = \frac{\sum_{\substack{\mathcal{N} \\ n_j^{(e)}=N_j^{(e)}}} \lambda_j^{(e)}(\mathcal{N})}{\sum_{\substack{\mathcal{N} \\ n_j^{(e)}\leq N_j^{(e)}}} \lambda_j^{(e)}(\mathcal{N})} \tag{18}$$

5. Product form and insensitivity

The CTMC defined in Section 4 can be seen as a Markov modulated process, where the active streaming calls are the modulating environment and the elastic request queue is the modulated system. Hence, the marginal steady-state probabilities for the occupancy of streaming requests can be computed independently of the state of the elastic services queue.

It is well-known that the model of streaming services admits a product form solution based on the theory of loss networks [4]. In this section, we discuss the existence of a product form solution for the limiting pdf of the number of streaming and elastic customers, and the insensitivity of the limiting pdf to the service time distribution. For the sake of a simple notation, we consider the case of just one class for both customer types, but the extension to multiple classes is straightforward. In addition, for the sake of simplicity, we let $\eta = 0$. The state of the queue in this case is simply defined as $\mathcal{N} = [n^{(s)}, n^{(e)}]$, with $n^{(s)} \in \{0, 1, 2, \ldots, N^{(s)}\}$ and $n^{(e)} \in \{0, 1, 2, \ldots, N^{(e)}\}$.

5.1. Product form

The marginal limiting pdf for streaming customers $\pi_{n^{(s)}} = \lim_{t\to\infty} P\{n^{(s)}(t)\}$, is expressed as:

$$\pi_{n^{(s)}} = \frac{\frac{1}{n^{(s)}!}\left(\frac{\lambda^{(s)}}{\mu_H+\mu^{(s)}}\right)^{n^{(s)}}}{\sum_{k=0}^{N^{(s)}} \frac{1}{k!}\left(\frac{\lambda^{(s)}}{\mu_H+\mu^{(s)}}\right)^{k}} \tag{19}$$

since the behavior of streaming customers is not influenced by the number of elastic customers, and corresponds to that in an $M/M/m/m$ queue.

The number of streaming customers at the queue modulates the service rate of elastic customers. Indeed, the total service rate of elastic customers in state $\mathcal{N}$ is

$$\mu_T^{(e)}(\mathcal{N}) = \mu_T^{(e)}(n^{(s)}) = \alpha^{(e)}\left[C - R^{(s)} n^{(s)}\right] \tag{20}$$

and thus depends on the number of streaming customers at the queue $n^{(s)}$.

According to the results in [5,6], the product form between the modulating and modulated processes exists if the steady-state distribution of the latter conditioned on the state of the former remains the same for all the possible states. We emphasize that, in our case, the transition rates of the process associated with elastic services vary according to the state of the streaming queue, at least for what concerns the service rate induced by the residual capacity.

Intuitively, we are required to define an adaptation of the arrival process intensity such that this is higher when the available bandwidth is high, and lower otherwise. Intriguingly, this is exactly what an AC policy is expected to do, and we will observe that this leads to the definition of a quasi-optimal AC policy.

More formally, under the condition

$$\frac{\lambda^{(e)}(\mathcal{N})}{\mu_T^{(e)}(\mathcal{N}) + n^{(e)}\mu_H} = \rho^{(e)}(n^{(e)}) \tag{21}$$

(i.e., $\rho^{(e)}(n^{(e)})$ does not depend on the number of streaming customers $n^{(s)}$) the marginal limiting pdf for elastic customers $\pi_{n^{(e)}} = \lim_{t\to\infty} P\{n^{(e)}(t)\}$, is simply expressed as:

$$\pi_{n^{(e)}} = \frac{\prod_{\ell=1}^{n^{(e)}} (\rho^{(e)}(\ell))}{\sum_{k=0}^{N^{(e)}} \prod_{\ell=1}^{k} (\rho^{(e)}(\ell))} \tag{22}$$

and the joint limiting pdf for the number of streaming and elastic customers can be expressed in product form as:

$$\pi_{\mathcal{N}} = \pi_{n^{(s)},n^{(e)}} = \pi_{n^{(s)}}\pi_{n^{(e)}} \tag{23}$$

Note that this implies that the elastic customer arrival rate $\lambda^{(e)}(\mathcal{N})$ is modulated by the value of $n^{(s)}$ so as to obtain (21).

Although the product form holds for any AC policy that satisfies condition (21), we define a policy with the following properties: (i) a job arriving at a saturated queue is simply discarded and (ii) on average the arrival intensity served when the queue has room is $\lambda^{(e)} > 0$.

Let $C_R(\mathcal{N})$ and $\overline{C}_R$ be the residual capacity not used by the streaming service in state $\mathcal{N}$ and its expectation, respectively. Then, the modulation performed by the following rule:

$$\lambda^{(e)}(\mathcal{N}) = \begin{cases} \lambda^{(e)} \frac{\alpha^{(e)} C_R(S) + (n^{(e)}+1)\mu_H}{\alpha^{(e)} \overline{C}_R + (n^{(e)}+1)\mu_H} & n^{(e)} < N^{(e)} \\ \lambda^{(e)} & n^{(e)} = N^{(e)} \end{cases} \tag{24}$$

satisfies condition (21) and is nearly optimal, as we will show with numerical experiments.

From a practical point of view this condition can be satisfied by an AC algorithm that either offloads or postpones elastic service requests in periods of high load, and accepts offloads or reactivates postponed requests in periods of low load. Note that this is done without altering the average elastic request arrival rate $\lambda^{(e)}$.

We remark that previous works considered product form impossible, and focused on deriving approximations or bounds. We instead proved that product form holds under the mild condition expressed by (21), which can be interpreted in terms of AC. Note however that for the queue without AC, where streaming and elastic services coexist and compete with each other, the product form does not hold; hence, the product form derivation under the mild condition expressed by (21) can be also seen as an approximation of the system without AC, making up for the lack of computationally efficient methods for this type of models.

5.2. Insensitivity

Insensitivity is an important property of some stochastic models in equilibrium that states that its stationary probabilities depend only on the first moment of a distribution that describes its behavior [7]. For example, certain queuing systems, such as the M/G/1/PS, are insensitive to the pdf of the service time.

The insensitivity to the service time distribution of the M/M/m/m queue is also well known even in the case of multiple classes with different resource demands [8], and is sufficient to prove the insensitivity of $\pi_{\mathcal{N}}$ to the pdf of the service time of streaming customers.

In this subsection, we thus focus on the insensitivity of $\pi_{\mathcal{N}}$ to the pdf of the amount of service requested by elastic customers, in the special case in which the mobility of elastic customers can be neglected, either because of a wired access link or due to the fact that the elastic customer service time is much shorter than the user dwell time in the cell (which is typically true for all those services in which the amount of data to transfer is small, hence for most elastic services).

Assume that elastic customers have a K-phase Coxian distributed service requirement, to be fulfilled by a server whose speed is modulated by the queue state $\mathcal{N}$ (i.e., by the number of streaming and elastic customers in service). Recall that the set of Coxian distributions is dense in the field of all positive-valued distributions, thus they can arbitrarily well approximate any service requirement pdf.

Each phase of the Coxian distribution is exponentially distributed with rate $\xi_k \mu_T^{(e)}(\mathcal{N})$, where $k \in \{1, 2, \dots, K\}$ denotes the Coxian phase and $\mu_T^{(e)}(\mathcal{N})$, defined in (20), is the server speed when the system is in state $\mathcal{N}$. In this context, we interpret $1/\alpha^{(e)}$ as the average service demand of the distribution. Let p_k denote the probability of moving from phase k to phase $k+1$ (so that $1-p_k$ is the probability of service completion at the end of phase k), with $p_K = 0$. Note that p_k does not depend on $\mathcal{N}$ if we neglect customer impatience, but this would not be the case if we included impatience, i.e., user mobility.

Then, the average individual elastic customer service rate, given $\mathcal{N}$, is:

$$\mu^{(e)}(\mathcal{N}) = \mu_T^{(e)}(\mathcal{N}) \left(\sum_{k=1}^{K} \left(\sum_{t=1}^{k} \frac{1}{\xi_t} \right) (1-p_k) \prod_{j=1}^{k-1} p_j \right)^{-1}$$

The state of the PS queue with Coxian service requirement pdf is $\mathbf{m} = (m_1, \dots, m_K)$, where $0 \le m_i \le N^{(e)}$ denotes the number of elastic customers in the system that reached phase of service i. Now assume, like in the previous subsection, that the arrival process of elastic customers is a state-dependent Poisson process following Condition (21) with $\mu_H = 0$, and let $\lambda^{(e)}(\mathcal{N})$ be its intensity.

Condition (21) is sufficient for product form even in the case of Coxian distributions. From a generic state $\mathbf{m}$ the outgoing transitions conditioned to $\mathcal{N}$ are shown in Fig. 2. If we divide all the rates

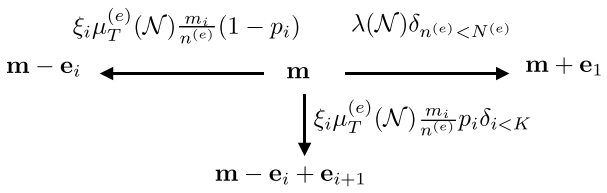


Fig. 2. Transition diagram conditioned to $\mathcal{N}$ of the Coxian queue. $\mathbf{e}_i$ is the vector with all 0s and a 1 in position i, δ_P is 1 when P is true and 0 otherwise.

by $\mu_T^{(e)}(\mathcal{N})$ under Condition (21), we observe that the Markov chains obtained by conditioning on $\mathcal{N}$ have the same steady-state distributions and hence the product form result holds by [5,6]:

$$\pi_{n^{(s)},\mathbf{m}} = \pi_{n^{(s)}} \pi^{(e)}(\mathbf{m})$$

where $\pi^{(e)}(\mathbf{m})$ is the stationary distribution of the PS queue with Coxian service times. Since the joint state space of the system is the Cartesian product of the state space of the PS queue and that of the M/M/m/m queue, then there is no need to re-normalize probabilities.

Now, let us aggregate the state space so that macro-state $(n^{(s)}, n^{(e)})$ is the set of states such that:

$$(n^{(s)}, n^{(e)}) = \left\{ (n^{(s)}, \mathbf{m}) : \sum_{k=1}^{K} m_k = n^{(e)} \right\}$$

Then, we have:

$$\pi^*(n^{(s)}, n^{(e)}) = \sum_{\substack{\mathbf{m}: \\ \sum_{k=1}^{K} m_k = n^{(e)}}} \pi_{n^{(s)}} \pi^{(e)}(\mathbf{m}) = \pi_{n^{(s)}} \sum_{\mathbf{m}: \sum_{k=1}^{K} m_k = n^{(e)}} \pi^{(e)}(\mathbf{m}) = \pi_{n^{(s)}} \pi_{n^{(e)}}$$

with $\pi_{n^{(e)}}$ given by (22), where the last equality follows by the insensitivity property of the $M/G/1/PS$ queuing system (possibly with state dependent arrival rates as in [9]). Since $\pi^*(n^{(s)}, n^{(e)})$ depends only on the first moment of the Coxian distribution, we conclude that the PS system in product form maintains the insensitivity property of the well-know PS system with constant arrival and service rates.

We remark that this is a very powerful result, that greatly generalizes previous works, showing that in the case of general service time distributions, which was considered intractable in the literature, we not only have product form, but also insensitivity, under the condition of no (or limited) mobility.

6. Video streaming at adaptable data rate

Video streams can be delivered to users with different qualities that typically correspond to quite different data rates. Standards for video streaming specify data rates that vary over several orders of magnitude, from few Mb/s for 640 by 480 pixels, 30 frames per second and aspect ratio 4:3 (480p), up to several Gb/s for Ultra HD 8 K (uncompressed) video.

Video streams are normally started at one of the available data rates, but their data rate can dynamically change to another possible value, according to the resources available along the path from transmitter to receiver.

The model we presented in the previous sections is capable of accommodating adaptable bit rate video with a finite number of video qualities, hence of data rates.

For the sake of simplicity, we consider the case of just two data rates for video: a lower data rate R_{vl} and a higher data rate R_{vh}, with $R_{vl} < R_{vh}$, but similar arguments can be applied to a larger number of data rates. In addition, we assume that, while the video data rate can be scaled according to data rate availability, the data rate of audio streams remains constant.

Table 3
Parameters used in the numerical evaluation. Basic scenario.

Parameter	Value
Total data rate	300 Mb/s
Data rate video	6 Mb/s
Video average duration	1800 s
Data rate audio	0.1 Mb/s
Audio average duration	600 s
Elastic file average size	0.5 Mb
Maximum number of elastic services	50
Average time in the cell	600 s
Fraction of elastic arrivals	0.8
Fraction of audio/video arrivals	0.1

Video service requests are accepted if the available data rate is at least R_{vl}, possibly after reducing the data rate of some ongoing video services from R_{vh} to R_{vl}, and blocked otherwise. During video streaming, the video data rate can be increased from R_{vl} to R_{vh} if resources allow, as well as reduced from R_{vh} to R_{vl} in order to make room for an incoming streaming service request. When several video streams are simultaneously active, as many as possible use the higher data rate R_{vh}, while the others only use R_{vl}. Blocking of a new video streaming service request occurs when the data rate available on the link is less than R_{vl}.

The state space of the resulting queuing model is the same as the one obtained by just considering videos at the lower rate R_{vl}. Indeed, the fact that a higher data rate for video is possible in some states just induces a modification of the residual data rate for elastic services.

If the number of active streaming services at time t is represented by the pair $(n_a(t), n_v(t))$, where $n_a(t)$ is the number of active audio streams and $n_v(t)$ is the number of active video streams, the number of video streams that use the high data rate R_{vh}, denoted by $n_{vh}(t)$, with $0 \le n_{vh}(t) \le n_v(t)$, can be computed as:

$$n_{vh}(t) = \left\lfloor \frac{C - \eta - n_a(t)\ R_a - n_v(t)\ R_{vl}}{R_{vh} - R_{vl}} \right\rfloor \tag{25}$$

Hence, the number of video streams that use the low data rate R_{vl}, denoted by $n_{vl}(t)$, with $0 \le n_{vl}(t) \le n_v(t)$, can be expressed as:

$$n_{vl}(t) = n_v(t) - n_{vh}(t) \tag{26}$$

The residual data rate available to elastic services, including the reserved data rate η, is

$$C - n_a(t)\ R_a - n_{vl}(t)\ R_{vl} - n_{vh}(t)\ R_{vh} \tag{27}$$

It is interesting to observe that the blocking probability for video service requests in the case of multiple data rates is the same as for the case in which videos can use just the lowest data rate.

7. Numerical results

We present numerical results for a wireless access link corresponding to a BS loaded by video and audio streaming traffic, and by elastic traffic subjected to the AC algorithm described before. We set $\eta = 0$ and use the parameter values listed in Table 3, which define our basic scenario.

7.1. Oscillations of KPIs

We start the analysis by considering a basic scenario and we investigate an interesting oscillating behavior for the KPIs.

The BS can simultaneously accommodate at most 50 video streaming services, and up to 3000 audio services. We assume a maximum of 50 simultaneous elastic services. We plot results as a function of the total service request arrival rate, before losses and before AC. For all values of the arrival rate, 10% of the requests refer to video services, 10% refer to audio services, and 80% to elastic services. Note that this implies that streaming video produces most of the BS load, like in real systems, due to its data rate requirement and duration.

Fig. 3 reports blocking probabilities for audio and elastic services (a) and for video services (b); blocking probabilities are computed as in (17) and (18) for streaming and elastic services, respectively. The figure reports also the residual capacity in Mb/s not used by streaming services that is therefore used by elastic services (c); this KPI is computed as in (13). It is extremely interesting to observe the oscillations induced by the mixture of streaming and elastic services with very different data rate requirements: oscillations are due to the step reduction in the number of active video services for increasing arrival rate, as we will explain later on. In the literature, only one study observed the possible non-monotonicity of blocking probabilities with respect to traffic intensity [3]. No previous work however unveiled the possibility of oscillating behaviors for both blocking probability of, and bandwidth available to, elastic services, or showed oscillations of the amplitude observed in Fig. 3. The oscillations observed for audio services are likely to intermittently violate the service level agreement constraints (typically an average blocking probability limit of the order of 1%), even for moderate load. The oscillations in blocking probability of elastic services are due to the non-monotonicity of the data rate not utilized by streaming services (dot-dashed blue line in the figure).

In order to explain the root causes of the behavior observed in Fig. 3, Fig. 4 shows the steady-state probabilities to have a large number of active video calls (values between 46 and 50). The probability of 50 active video services is always extremely low, since the presence of one audio service is sufficient to prevent access to the 50th video. Other probabilities show peaks due to the interplay between audio and video arrivals. The distance between peaks corresponds to an increase of the average number of active audio services equal to 60 (remember that 60 audio services consume the same data rate of 1 video). Indeed, with an audio service average duration in the cell equal to 300 s (also accounting for mobility), an increase of 60 audio services corresponds to an increase in arrival rate equal to 0.2, which is close to the distance between consecutive peaks that we observe in the plots.

The behavior we just observed is shown also in the curves of blocking probabilities reported in Fig. 5, for a larger range of values and in semilogarithmic scale. While the video service blocking probability exhibits a familiar monotonic behavior, the audio service blocking probability (which is much lower because of the lower data rate) oscillates, but the amplitude of oscillations reduces as video blocking probability grows.

Fig. 6 shows the average number of active audio and video services, to be read in the y-axis on the left and right, respectively. While the average number of active audio services grows, again with some oscillations due to the phenomena described before, the average number of active video services reaches a maximum for a video service request arrival rate close to 0.2 s^{-1}, and then starts declining due to the increasing amount of the BS resources occupied by audio services, that increase the blocking probability for video.

The lesson learnt from these results is that the interplay among classes of services with quite different requirements, even when exhibiting the same kind of resource usage (i.e., constant bit rate), is complex, and might lead to undesired behaviors, such as the non-monotonic growth of blocking probability with load.

7.2. Effect of AC

Let us now consider what happens to elastic services. Fig. 7 reports, on the y-axis on the left, the blocking probability of elastic services, and, on the y-axis on the right, the average bandwidth that is available to elastic services, i.e., the residual capacity not used by streaming services. The blocking probability is reported for three cases: (i) elastic services with the considered AC, (ii) with an optimal AC which relies on full information (arrival times and amount of data to be transferred of all service requests are known) and (iii) with no AC, labeled as

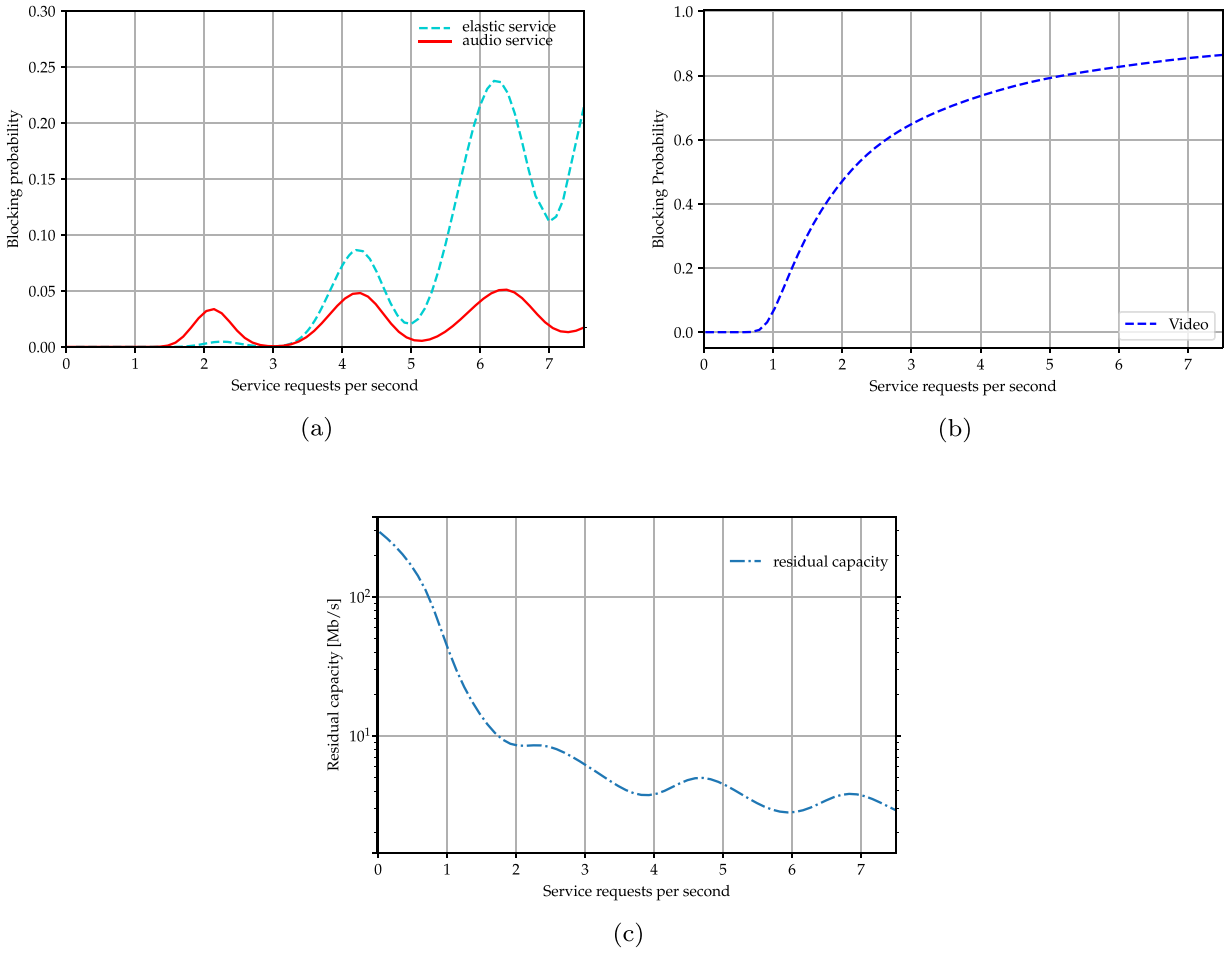


Fig. 3. Blocking probability for audio, elastic (a) services, video (b), and expected residual capacity for elastic services (c), versus total service request rate. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

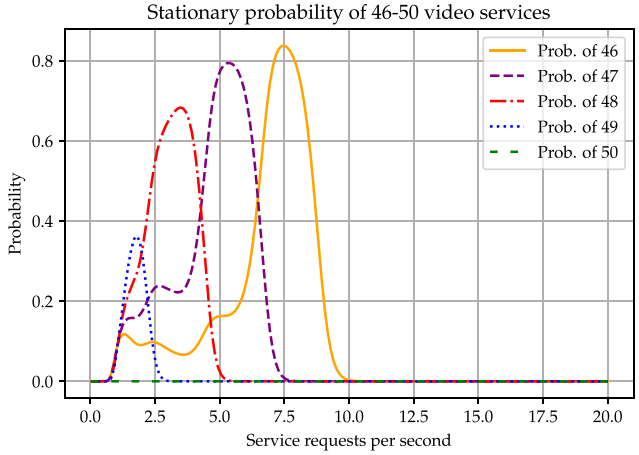


Fig. 4. Stationary probabilities of numbers of active video services between 46 and 50; basic scenario.

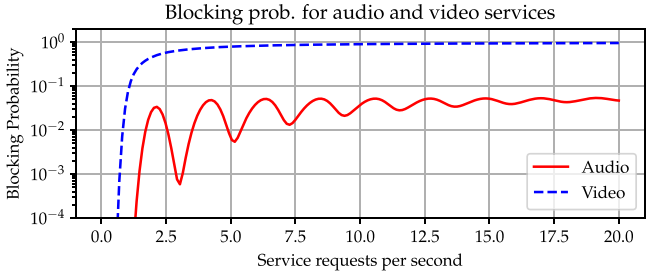


Fig. 5. Blocking probability for audio and video services in the basic scenario.

"constant rate" in the figure. Note that only the results with our proposed AC are obtained with the product form solution derived before. Results for the other two cases required the solution of Markovian models with extremely large state spaces, exploiting the capabilities of sophisticated solvers [10], which have very high cost in terms of both memory and computation. We see that our proposed AC exhibits almost identical performance to the optimal AC (the two curves overlap), while it outperforms the case of no AC. Oscillations in residual capacity occur with a decreasing pattern due to the increasing arrival rate of service requests of all types. An interesting effect of the interplay between audio and video streaming services is that, even under very heavy load conditions, streaming services leave some resources for elastic services. This effect is caused by the inability of streaming services to exploit the entire available bandwidth, due to the very different data rate requirements of the two classes of streaming services. This has the very desirable effect of preventing starvation of elastic services even at very high streaming traffic intensity.

The average number of elastic services in progress computed as in (11) is shown in Fig. 8, again in the cases with or without the AC algorithm and for the optimal control. With the considered AC, we observe a phase transition for an overall service request arrival

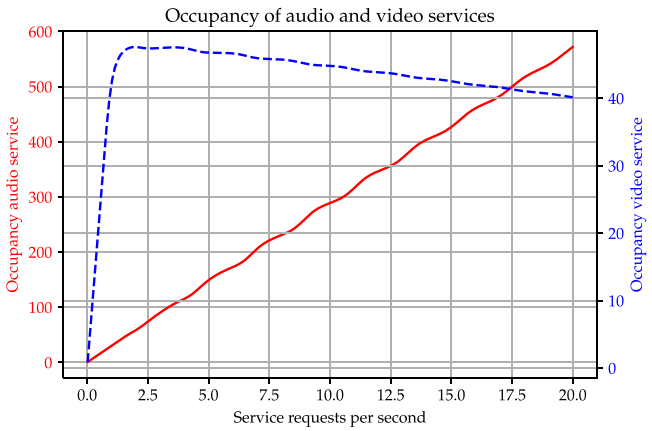


Fig. 6. Average number of active audio and video services in the basic scenario.

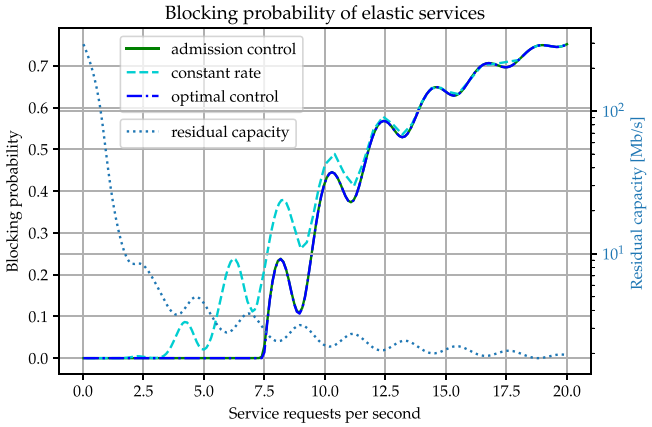


Fig. 7. Blocking probabilities and residual capacity for elastic services in the basic scenario with and without admission control.

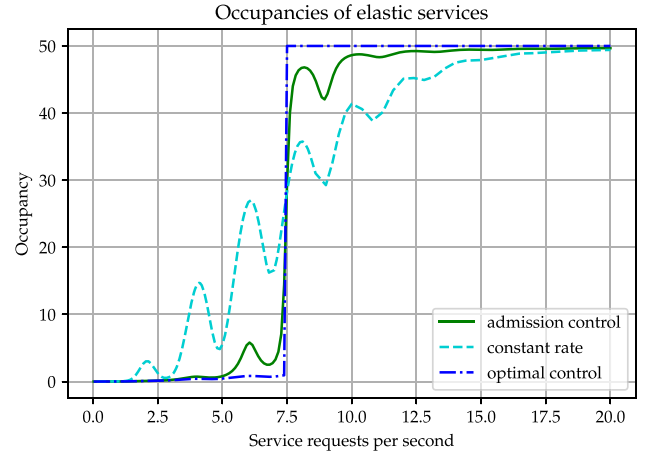


Fig. 8. Average number of active elastic services in the basic scenario with and without admission control.

rate equal to 7.5 requests per second, which corresponds to the value for which the load of elastic services equals 1. For lower values, the residual bandwidth is enough to satisfy elastic traffic, and the number of services in progress is low. For higher values, the residual bandwidth is low and the system is overloaded; hence, the number of elastic services in progress quickly approaches 50, the maximum permitted value. In the case of no AC, the average number of elastic services in progress oscillates, with peaks corresponding to minima of the residual bandwidth. Before saturation, the average number of elastic services in progress is larger than in the case of AC, because of the expected beneficial effect of the algorithm. After saturation, the average number of elastic services in progress is lower than in the case of AC, because of the higher blocking probability, as can be seen in Fig. 7.

Figs. 7 and 8 show the quasi optimality of the AC algorithm for elastic services. When the bandwidth available to elastic services is such that their load is less than 1, no loss occurs, and the average number of elastic services in progress is close to the case of the ideal control algorithm. When the bandwidth available to elastic services is such that load exceeds 1, losses occur and the number of elastic services in progress is close to 50 (exactly 50 for the full information algorithm).

7.3. Constant video request rate

As a second scenario we consider a variation of the basic scenario that shows that the phenomena described above occur also when the growth of the service request rates of the three classes of service is not the same. In particular, in the modified scenario the service request rate for video is fixed at 0.2 requests per second, and the request rates for audio and elastic services grow, keeping the same ratio between elastic and audio request rate as in the basic scenario.

Fig. 9 shows the audio and video blocking probabilities versus the total service request arrival rate. The audio blocking probability oscillates, with a frequency similar to the case of the basic scenario. The video blocking probability now has a smoothed staircase behavior (in spite of the constant load generated by videos) because of the progressive erosion of the available resources due to audio services.

The explanation of these behaviors can again be found in the plot of the steady-state probabilities that the number of active video calls is equal to values between 46 and 50, which is reported in Fig. 10.

Finally, Fig. 11 reports the average bandwidth in Mb/s that is not occupied by video and voice services, and that can be exploited by elastic services. Once more, the oscillations are clearly visible.

7.4. Video at adaptable data rate

We now consider the case of video streams that adapt their data rate to the amount of resources that are available on the access link. In particular, we assume that two data rates are available for video streaming: 6 and 10 Mb/s (on the contrary, the data rate for audio services remains fixed at 0.1 Mb/s, for simplicity). If enough resources are available to allow all or some of the active video streams to operate at the higher data rate, then they use 10 Mb/s. The remaining video streams operate at the lower data rate. For example, if audio services collectively consume 9 Mb/s (i.e., 90 audio services are in progress) and 30 video streams are active, the 291 Mb/s not used by audio are used to serve 27 video streams at 10 Mb/s and 3 video streams at 6 Mb/s, for a total of 288 Mb/s, so that 3 Mb/s remain to serve elastic traffic.

In Fig. 12 we plot the blocking probability for audio and elastic services, as well as the expected residual capacity for elastic services, versus the total service request rate in the case of the basic scenario with variable data rate video services. These results must be compared with those in Fig. 3 that refer to constant video data rate. We can see that the effect of the adaptable data rate of video is a faster decrease of the residual capacity and larger oscillations of the blocking probability of elastic services. Note that the curve of the blocking probability of audio services is the same in the two figures. Also the curve of the blocking probability of video services (not shown in Fig. 12) is the same in the two cases.

In Fig. 13 we plot the average number of active elastic services in the basic scenario with and without admission control in the case of adaptable data rate video services. These results must be compared with

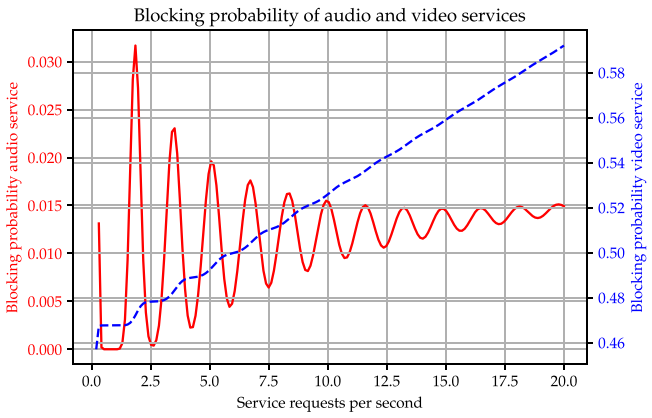


Fig. 9. Blocking probability for audio and video services in the modified scenario.

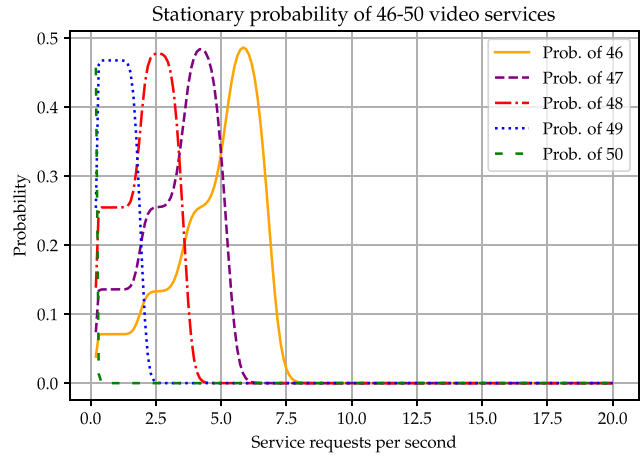


Fig. 10. Stationary probability of occupancy between 46 and 50 for video services in the modified scenario.

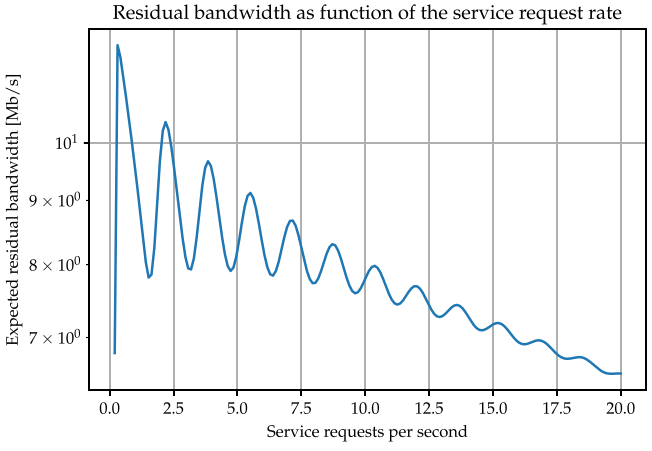


Fig. 11. Residual bandwidth for elastic services in the modified scenario.

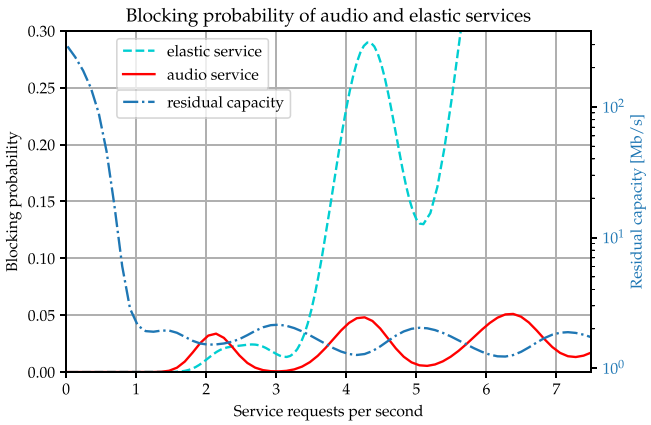


Fig. 12. Blocking probability for audio and elastic services, and expected residual capacity for elastic services, versus total service request rate with adaptable data rate video services. Basic scenario.

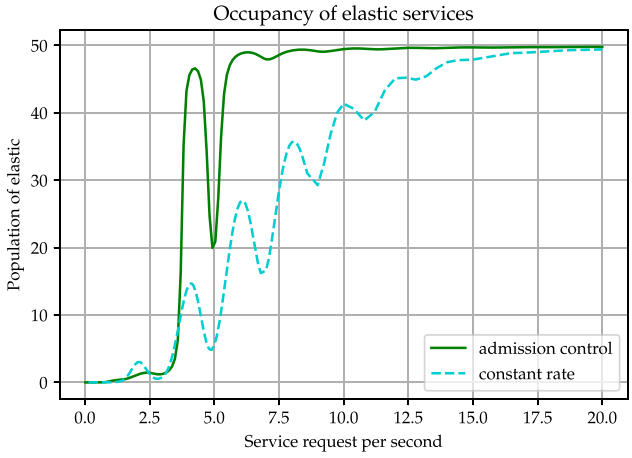


Fig. 13. Expected occupancy of elastic services with adaptable data rate for the video services.

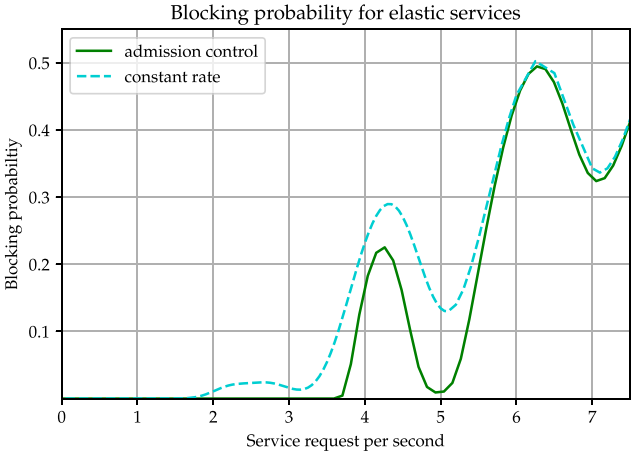


Fig. 14. Blocking probability for elastic services with adaptable data rate video.

those in Fig. 8, that refer to the case of constant data rate video. We can see that the variability of the video data rate makes the number of active elastic services grow much earlier than with constant data rate video. This is coherent with the faster decrease of the residual data rate, which makes elastic services last longer.

In Fig. 14 we plot the blocking probability for elastic services in the case of adaptable data rate video services. These results must be compared with those in Fig. 7, that refer to the case of constant data rate video. Also this metric shows that elastic services suffer due to the variability of the video data rate and the subsequent reduction in the residual capacity.

8. Assuming independence

The complexity of scenarios with heterogeneous service types derives by the interplay between the allocation of resources and their use

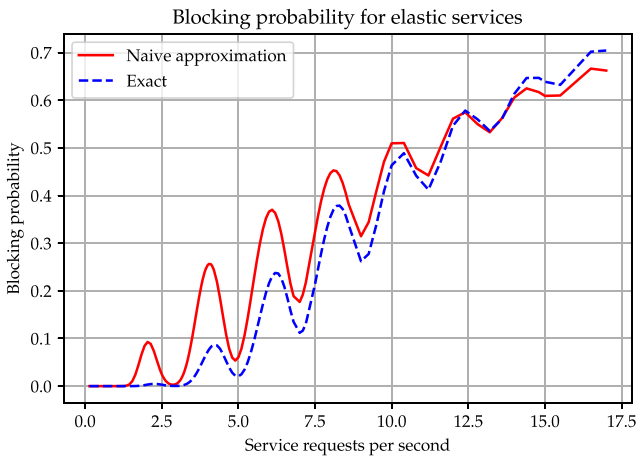


Fig. 15. Comparison between the exact model and the one that assumes independence between streaming and elastic traffic: blocking probability for elastic services.

by requests of different types. Thanks to the existence of a product form solution, our model is effective and scalable. However, in this section, we aim to further highlight the importance of a detailed model that takes into account the complex interplay among services of different types. To do so, we compare our model against a simpler approach in which the interplay between services is taken into account in a simplified, but approximate, way.

In particular, we consider results obtained by assuming that the two traffic types are independent. Since streaming traffic is not influenced by the presence of elastic traffic, the steady-state number of active streaming services can be simply derived as in (19).

By assuming independence, the steady-state probability to have $n^{(e)}$ active elastic services can be derived (by the theorem of total probability) as the weighted sum of the conditional steady-state probabilities of $n^{(e)}$ active elastic services given that $n^{(s)}$ streaming services are active multiplied by the steady-state probability to have $n^{(s)}$ active services. In formulas:

$$\pi_{n^{(e)}} = \sum_{n^{(s)} \in \mathcal{N}^{(s)}} P\{n^{(e)}|n^{(s)}\}\pi_{n^{(s)}} \tag{28}$$

where $P\{n^{(e)}|n^{(s)}\}$ is obtained from the solution of a processor sharing M/M/1 queue in which the server works at the rate corresponding to the residual capacity.

Fig. 15 reports the blocking probability of elastic services $E[n^{(e)}]$ computed with (28) and with the exact model, when no AC is assumed and the basic scenario is considered.

We can see that (28) reproduces the general oscillating behavior observed for the loss probability, as expected because this directly derives from the oscillation of available bandwidth for elastic traffic. However, the value of the loss probability is overestimated for low values of load and underestimated for high load.

The reason for this discrepancy lies in the fact that the assumption of independence and expression (28) consider both processes of streaming and elastic services to be in steady state. However, this is not the case in practice, since a variation in active streaming services induces a transient on the behavior of elastic services that requires time to approach the steady state, and is likely disrupted by another change in the number of active streaming services. This means that elastic services cannot be considered in steady state for a given configuration of streaming services, and the system behavior can only be captured with a model like ours, that accounts for the interplay of the two service types.

Table 4
Average number of active audio and video services and blocking probability for audio and video, with variable coefficient of variation for the duration of audio and video services. Basic scenario.

Coefficient of variation	Average active audio	Average active video	$P\{loss\}$ audio	$P\{loss\}$ video
1	147.6	46.8	0.01	0.79
1.2	144.4	46.8	0.01	0.79
1.4	138.6	46.9	0.02	0.79
1.6	134.9	46.8	0.03	0.78
1.8	130.7	47.0	0.03	0.78
2.0	124.7	47.2	0.05	0.77
2.2	121.0	47.3	0.05	0.77
2.4	118.2	47.4	0.05	0.77
2.6	114.6	47.5	0.05	0.76
2.8	113.1	47.6	0.04	0.76
3.0	109.8	47.6	0.03	0.75

9. Validation of exponential assumptions

In order to validate our analytical results, that in the case of a RAN cell serving mobile users require exponential assumptions, we compare analytical results against simulation estimates obtained in the non-exponential case. In particular, in Table 4 we show what happens to the number of active video and audio services and to the blocking probabilities when we increase the coefficient of variation of the service duration in the basic scenario, defined by the parameters reported in Table 3. Considering that the duration of streaming services have been observed to have distributions with variances higher than those of an exponential, we vary the coefficient of variation from 1 (corresponding to an exponential distribution) to 3 (corresponding to a rather large variance, equal to 9 times the square of the mean). This is obtained by using a 2-stage hyperexponential distribution where one stage (having probability 0.95) describes the core of the distribution and the other (having probability 0.05) models the tail of the distribution.

Results in Table 4 show that the average number of active video services is only marginally impacted by the changes of the coefficient of variation: an increase from 1 to 3 corresponds to an increase from 46.8 to 47.6 (i.e., an increase of 1.7% only). More relevant is the variation in the average number of active audio services, that decrease from 147.6 to 109.8 (i.e., by 25%). This is due to the much lower amount of bandwidth consumed by audio with respect to video.

If we turn our attention to blocking probabilities, we see that for video the increase of the coefficient of variation implies a limited reduction of the blocking probability from 0.79 to 0.75. Quite interestingly, we instead observe the blocking probability for audio services to first grow with the coefficient of variation and then decrease (always remaining in the range between 1 and 5%), again showing an unexpected non-monotonic behavior.

10. Related work

The coexistence of streaming and elastic traffic flows, and their interaction, is studied in the literature for both wired and wireless networks. The paper [11] studies an integrated AC scheme for a wired network loaded by both streaming and elastic flows, using a fluid model that provides a good approximation under rather general and realistic traffic conditions. In [12,13] the authors study a UMTS cell loaded with both streaming and elastic traffic submitted to AC with no mobility, and develop approximate analysis approaches; the authors state that exact analysis is non-tractable in general. Channel-aware scheduling algorithms for streaming and elastic services in a base station are discussed in [14] and evaluated using an approximate approach which is shown to be very accurate against simulations, again disregarding user mobility.

The papers [15,16] delve into the analysis of a multi-server queuing system characterized by two categories of service requests, namely

Table 5
Comparison between this paper and the most relevant previous works modeling mixes of streaming and elastic services (AC stands for access control; MC stands for Markov chain; PF stands for product form; QN stands for queuing network).

Reference	Main content of reference	Improvement of this paper
[11]	AC scheme; fluid model	Mobility; exact analysis; PF; insensitivity
[12,13]	AC scheme; approximate model	Mobility; exact analysis; PF; insensitivity
[14]	Scheduler; approximate model	Mobility; exact analysis; PF; insensitivity
[15]	MC model	Mobility, PF; insensitivity
[16]	MC model; elastic mobility	PF; insensitivity
[17,18]	Resource management; AC; QN model	Exact analysis; PF; insensitivity
[19]	MC model	Exact analysis; insensitivity
[20]	AC; approximate MC; mobility	PF; insensitivity
[21]	MC model	Mobility; PF; insensitivity
[22]	Approximate MC	Mobility; exact analysis, PF; insensitivity

inelastic and elastic. In the context of this queuing system, the computation of performance metrics is carried on through the utilization of the stationary distribution derived from the multidimensional Markov chain governing the system's dynamics.

In [17,18], and [19] the authors use queuing models to study a resource management scheme for an integrated cellular/WLAN network to support streaming and elastic service classes. Resource sharing is based on virtual partitioning, and a dynamic load balancing policy is proposed to distribute the traffic load.

The work in [20] studies an AC policy for cellular networks supporting both streaming and elastic traffic, also accounting for user mobility. The AC is based on virtual partitioning. Resources are quantized, and an approximate Markovian analysis is used to derive performance metrics. When the total traffic demand of elastic flows exceeds the available capacity some flows might be aborted due to impatience. A similar approach is used in [21] to derive the performance of streaming and elastic services in a multiservice network with a dynamic channel allocation scheme where different classes of traffic have different quality of service requirements. The authors of [22] develop an approximate Markovian model of a wired or wireless system loaded with streaming and elastic traffic. The model is based on a two-dimensional Markov process, which approximates the real process in the system. To validate and verify the model, an original and purpose-made simulator was used. The results of the simulation confirm the model accuracy.

The main characteristics of the above references and the differences with respect to our work are very concisely summarized in Table 5.

It is worth underlining that the AC strategy adopted in our paper has different characteristics compared to those proposed in the above cited references. First of all, our AC strategy, by using a re-shaping of the flow of incoming requests, reduces the waste of bandwidth due to session premature terminations (see for instance [20,23], and [24]). Moreover, our AC strategy, allowing a re-shaping of the input that can be both spatial (using adjacent cells to compensate for over/under-load) and temporal (adding a delay), can be applied to both wired and wireless environments.

In [25], the authors propose a new approximate technique for the analysis of Markov modulated models called MARC, where the response time of the modulated process can be approximated up to an additive constant. The method differs from that considered here since it requires to compute a function that is the solution of a Poisson equation that allows one to understand the trend of the response time as function of the intensity of the workload. Notice that the additive constant may be relevant in low and moderate load conditions and since we aim to estimate the blocking probabilities of elastic services this may cause inaccurate estimates in practical scenarios.

Queuing models combining elastic and inelastic service requests find extensive application in modeling communication networks, distributed systems, and computing systems. For example, in [26] the authors consider data center workloads consisting in inelastic and elastic jobs, and they develop analytical models to study properties of scheduling algorithms.

The papers [27,28] offer a compelling exploration of intriguing research problems associated with these models, and of their applications to the performance evaluation of data centers and cloud systems.

11. Conclusions

In this paper, we have developed queuing models of access links of wired and wireless multiservice access networks, showing unexpected oscillatory phenomena in the system performance that were mostly neglected in the literature.

We have defined a quasi-optimal AC algorithm for elastic services that leads to a product form expression of the joint probability of the numbers of active services of the different classes, and to an insensitivity of the system performance to the distribution of the amount of data to be transferred, in the case of no mobility.

The phenomena that we observed, and in particular the oscillations of several performance metrics, are mainly due to the large difference in the data rate requirements of different classes of streaming services. It is important to state that we have not exaggerated the data rate differences in our analysis. We used 6 Mb/s for video and 100 kb/s for audio, but today a HD streaming of a sport event easily requires over 10 Mb/s, and a one-on-one voice-only call on any of the multiparty voice conference platforms that have become so popular in times of pandemic normally consumes about 10 kb/s.

The use of AC for elastic services is very important for the provision of quality of service. In our setting, we observed that with a service request rate around 6 requests per second, i.e., with an elastic service load of the order of 80% of the bandwidth not used by streaming services, Poisson arrivals generate a blocking probability around 20%, while both our proposed scheduler and the full information scheduler produce hardly any blocking.

The extension of the work presented in this paper will consider portions of a radio access network, including several base stations that define macro and small cells over which end users roam while accessing the network.

CRediT authorship contribution statement

Andrea Marin: Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Marco Ajmone Marsan:** Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Michela Meo:** Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Matteo Sereno:** Writing – review & editing, Writing – original draft, Methodology, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Marco Ajmone Marsan has been for several years on the editorial board of Computer Networks The other authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

This work was supported in part by the European Union under the Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU, PRIN 2022 project "Deploying Artificial Intelligence in 6G Network Management using Digital Twins (6GTWINS) and by the partnership on "Telecommunications of the Future" PE00000001 - program "RESTART", project ITA NTN.

References

[1] BRAIN-IoT, Model-based framework for dependable sensing and actuation in intelligent decentralized IoT systems, 2020, URL https://www.brain-iot.eu/about/concept/.
[2] A. Marin, M. Meo, M. Sereno, M. Ajmone Marsan, Modeling Service Mixes in Access Links: Product Form and Oscillations, in: 2022 IEEE 23rd International Symposium on a World of Wireless, Mobile and Multimedia Networks, WoWMoM, 2022, pp. 312–318.
[3] J. Roberts, U. Mocci, J. Virtamo (Eds.), Multi-rate models, in: Broadband Network Traffic: Performance Evaluation and Design of Broadband Multiservice Networks, Springer Berlin Heidelberg, Berlin, Heidelberg, 1996, pp. 513–545.
[4] F.P. Kelly, Loss Networks, Ann. Appl. Probab. 1 (3) (1991) 319–378.
[5] A. Economou, Generalized Product-Form Stationary Distributions for Markov Chains in Random Environments with Queueing Applications, Adv. Appl. Probab. 37 (1) (2005) 185–211.
[6] S. Balsamo, A. Marin, Separable solutions for Markov processes in random environments, European J. Oper. Res. 229 (2) (2013) 391–403.
[7] P.G. Taylor, Insensitivity in Stochastic Models, in: R.J. Boucherie, N.M. van Dijk (Eds.), Queueing Networks: A Fundamental Approach, Springer US, Boston, MA, 2011, pp. 121–140.
[8] P. Whittle, Systems in Stochastic Equilibrium, John Wiley, New York, 1986, (Wiley Series in Probability & Mathematical Statistics).
[9] F. Baskett, K.M. Chandy, R.R. Muntz, F.G. Palacios, Open, Closed, and Mixed Networks of Queues with Different Classes of Customers, J. ACM 22 (2) (1975) 248–260.
[10] G. Ciardo, R.L. Jones, A.S. Miner, R. Siminiceanu, Logical and Stochastic Modeling with Smart, in: P. Kemper, W.H. Sanders (Eds.), Computer Performance Evaluation. Modelling Techniques and Tools, Springer Berlin Heidelberg, Berlin, Heidelberg, 2003, pp. 78–97.
[11] N. Benameur, S. Ben Fredj, F. Delcoigne, S. Oueslati-Boulahia, J.W. Roberts, Integrated Admission Control for Streaming and Elastic Traffic, in: M.I. Smirnov, J. Crowcroft, J. Roberts, F. Boavida (Eds.), Quality of Future Internet Services, Springer Berlin Heidelberg, 2001, pp. 69–81.
[12] O. Boxma, A. Gabor, R. Nunez Queija, H. Tan, Integration of streaming and elastic traffic in a single UMTS cell: modeling and performance analysis, Stoch. Process. Appl. (2006).
[13] O. Boxma, A. Gabor, R. Nunez Queija, H.-P. Tan, Performance analysis of admission control for integrated services with minimum rate guarantees, in: 2006 2nd Conference on Next Generation Internet Design and Engineering, 2006. NGI '06, 2006.
[14] S. Borst, N. Hegde, Integration of Streaming and Elastic Traffic in Wireless Networks, in: IEEE INFOCOM 2007 - 26th IEEE International Conference on Computer Communications, 2007, pp. 1884–1892.
[15] R. Nunez Queija, J. van den Berg, M. Mandjes, Performance Evaluation of Strategies for Integration of Elastic and Stream Traffic, in: Proceedings of International Teletraffic Congress, ITC16, 1999, pp. 1039–1050.
[16] A. Dudin, S. Dudin, O. Dudina, Analysis of a Queueing System with Mixed Service Discipline, Methodology and Computing in Applied Probability 25 (2) (2023).
[17] W. Song, W. Zhuang, Multi-Class Resource Management in a Cellular/WLAN Integrated Network, in: 2007 IEEE Wireless Communications and Networking Conference, 2007, pp. 3070–3075.
[18] W. Song, W. Zhuang, Resource Allocation for Conversational, Streaming, and Interactive Services in Cellular/WLAN Interworking, in: IEEE GLOBECOM 2007 - IEEE Global Telecommunications Conference, 2007, pp. 4785–4789.
[19] I.-A. Chousainov, I.D. Moscholios, P.G. Sarigiannidis, Congestion Probabilities in a Multi-Cluster C-RAN Servicing a Mixture of Traffic Sources, Electronics 9 (12) (2020).
[20] E. Bernal-Mor, V. Pla, J. Martinez-Bauset, Robust admission control for streaming and elastic services in cellular networks, in: The IEEE Symposium on Computers and Communications, 2010, pp. 372–374.
[21] G. Basharin, T. Aterekova, Analytical model of streaming and elastic traffic with dynamic channel allocation scheme, in: International Congress on Ultra Modern Telecommunications and Control Systems, 2010, pp. 1086–1090.
[22] S. Hanczewski, M. Stasiak, J. Weissenberg, A Model of a System With Stream and Elastic Traffic, IEEE Access 9 (2021) 7789–7796.
[23] R. Ramjee, D. Towsley, R. Nagarajan, On optimal call admission control in cellular networks, Wirel. Netw. 3 (1997) 29–41.
[24] D. García, J. Martínez, V. Pla, Admission Control Policies in Multiservice Cellular Networks: Optimum Configuration and Sensitivity, in: G. Kotsis, O. Spaniol (Eds.), Wireless Systems and Mobility in Next Generation Internet, Springer Berlin Heidelberg, 2005, pp. 121–135.
[25] I. Grosof, Y. Hong, M. Harchol-Balter, A. Scheller-Wolf, The RESET and MARC techniques, with application to multiserver-job analysis, Perform. Eval. 162 (2023) 102378.
[26] B. Berg, M. Harchol-Balter, B. Moseley, W. Wang, J. Whitehouse, Optimal Resource Allocation for Elastic and Inelastic Jobs, in: Proceedings of SPAA 2020, 2020, pp. 75–87.
[27] M. Harchol-Balter, Open Problems in Queueing Theory Inspired by Datacenter Computing, Queueing Syst. Theory Appl. 97 (1–2) (2021) 3–37.
[28] W. Wang, Q. Xie, M. Harchol-Balter, Zero Queueing for Multi-Server Jobs, Proc. ACM Meas. Anal. Comput. Syst. 5 (1) (2021).

Andrea Marin received his Ph.D. in Computer Science from the University of Venice in 2009. He is Associate Professor of Computer Science at the same university. He is the (co-)author of over 100 technical papers in refereed international journals and conference proceedings. His current research focuses on the performance and reliability evaluation of computer systems using stochastic modeling techniques.

Marco Ajmone Marsan is a part-time research professor at the IMDEA Networks Institute in Spain and an Emeritus Professor of Politecnico di Torino. From 1974 to 2021 he was at the Politecnico di Torino, in the different roles of an academic career, with an interruption from 1987 to 1990, when he was a full professor at the Computer Science Department of the University of Milan. He obtained degrees in EE from the Politecnico di Torino and the University of California, Los Angeles (UCLA).

He served in the editorial board of several international journals, and chaired the steering committee of the ACM/IEEE Transactions on Networking. He was the General Co-chair of Infocom 2013, and of ICC 2023. He is a Fellow of the IEEE, and a member of the Academia Europaea and of the Academy of Sciences of Torino. He is qualified as "ISI Highly Cited researcher" in computer science. He received a honorary degree in Telecommunication Networks from the Budapest University of Technology and Economics. He was named Commander of the Order of Merit of the Republic of Italy. He was the Vice-Rector for Research, Innovation and Technology Transfer at the Politecnico di Torino, and the Director of IEIIT-CNR.

Michela Meo is a professor at Politecnico di Torino, Italy, in Telecommunication Engineering. Her research interests include green networking, energy-efficient mobile networks and data centers, machine-learning for Internet traffic classification and characterization. She co-authored more than 200 papers, edited a book with Wiley on Green Communications and several special issues of international journals. She was chairing the International Advisory Council of the International Teletraffic Conference from 2015 to 2021. She is senior editor of IEEE Transactions on Green Communications and she was associate editor of ACM/IEEE Transactions of Networking, Green Series of the IEEE Journal on Selected Areas of Communications Networking and IEEE Communication Surveys and Tutorials,. In the role of general or technical chair, she has led the organization of several conferences, including ACM e-Energy, ITC, Infocom Miniconference, ICC symposia, ISCC. She was Deputy Rector of Politecnico di Torino from March 2017 to March 2018.

Matteo Sereno was born in Nocera Inferiore, Italy. In 1987, he earned a Laurea degree in Computer Science from the University of Salerno, followed by a Ph.D. in Computer Science from the University of Torino in 1992.

Presently, he holds the position of Full Professor at the Computer Science Department of the University of Torino.

His current research focuses on the performance evaluation of computer systems, communication networks, distributed systems, queueing networks, stochastic Petri net models, and their application in analyzing computer and telecommunication systems.