Asymptotic properties of sequential streaming leveraging users' cooperation

Delia Ciullo, Valentina Martina, Michele Garetto, Emilio Leonardi, Giovanni Luca Torrisi

***Abstract*—We consider a communication system in which a given digital content has to be delivered sequentially at constant rate to a set of users who asynchronously request it according to a Poisson process. Users can retrieve data: i) from one or more sources that statically store the entire content; ii) from users who have previously requested the content, and contribute (for limited time) a random amount of upload bandwidth to the system. We propose a stochastic fluid framework that allows characterizing the aggregate streaming rate necessary at the sources to satisfy all active requests. In particular, we establish the conditions under which the system becomes asymptotically scalable as the number of users grows. Our theoretical results apply to increasingly popular Video-on-Demand systems exploiting users' cooperation.**

***Index Terms*—Stochastic models, cooperative networking, Video-On-Demand.**

I. Introduction

Online streaming over the Internet is becoming the dominant way of distributing multimedia contents to large populations of users. According to recent forecasts [2], video traffic is expected to exceed 90% of all global consumer Internet traffic by 2016, posing a tremendous challenge to both content providers and network operators.

The current approach to handle the increasing demand of bandwidth-hungry contents in the Internet is based on Content Delivery Networks (CDNs): thanks to the proliferation of proxy servers, contents are "moved" close to the users, significantly reducing the Internet core traffic and improving the perceived quality of service (*e.g.*, the latency). However, any solution based on CDNs has severe limitations in terms of scalability: indeed, the aggregate resources required at data centers (bandwidth/storage/processing), and the corresponding costs incurred by content providers, inevitably scale linearly with users' demand and data volume. Content providers are thus forced to continuously upgrade their CDN infrastructure, or acquire additional resources from cloud services.

The only scalable solution proposed so far to distribute multimedia contents at massive scale is to exploit users' cooperation: while they retrieve and watch contents, users contribute their resources (bandwidth/storage/processing) to the system, thus offloading the servers [3], [4].

On the other hand, streaming architectures which primarily rely on users' cooperation can hardly guarantee the strict quality-of-service requirements of multimedia contents (e.g., in the case of online video a steady download rate no smaller than the playback rate is necessary for a smooth watching experience). For these reasons, user-assisted architectures should be supported by properly dimensioned CDNs (or cloud services) that intervene whenever the resources provided by users are not enough to satisfy the current demand.

In our work, we are specifically interested in characterizing the additional bandwidth that servers must supply (in addition to the bandwidth contributed by the users) to guarantee ideal service to all users (*i.e.*, requests are immediately satisfied and contents can be enjoyed without interruption till the end).

In this paper, we focus on a single content (*e.g.*, a video), and we theoretically analyse the communication system that allows this content to be sequentially delivered to an arbitrarily large number of users, who contribute a random amount of upload bandwidth while they stay in the system. Our main contribution is a stochastic analytical framework that allows us to derive general upper and lower bounds to the additional bandwidth requested from the servers to guarantee constant download rate to all users. In particular, our bounds permit tightly characterizing the asymptotic system behavior as the number of users increases.

We observe that the mathematical formulation of the system considered in this paper has been already proposed in the literature [3], [5]. However in previous work authors have resorted to Monte-Carlo approaches to evaluate it. To the best of our knowledge, we are the first to provide an analytical characterization of the solution and rigorously prove its asymptotic properties in the large users limit.

One of our main finding is that the conditions under which the system becomes scalable (i.e., the bandwidth requested from the servers does not increase with the number of users) depend critically on the underlying assumptions about the video request process. In particular, we will consider both the case in which the video request rate is constant, and the case of a newly introduced video whose popularity changes over time, determining a non-homogeneous video request rate. We find that completely different conditions on the system parameters must be satisfied in the above two cases to guarantee the scalability of the video distribution as the system size increases.

As another contribution, we propose a methodology to derive exact estimate of the bandwidth requested from the

D. Ciullo is with EURECOM Sophia Antipolis, France (delia.ciullo@eurecom.fr).
V. Martina and E. Leonardi are with Dipartimento di Elettronica e delle Telecomunicazioni, Politecnico di Torino, Italy ({valentina.martina, emilio.leonardi}@polito.it).
M. Garetto is with Dipartimento di Informatica, Università di Torino, Italy (michele.garetto@unito.it).
G.L. Torrisi is with Istituto per le Applicazioni del Calcolo "Mauro Picone", CNR, Italy (torrisi@iac.rm.cnr.it).
M. Garetto is supported by project AMALFI (Università di Torino/Compagnia di San Paolo).
A preliminary version of this paper appeared at IEEE INFOCOM 2012 [1].

servers under the assumption that the user upload bandwidth is exponentially (or more generally phase-type) distributed.

Although we restrict our attention to a single video, our theoretical analysis provides the building block to assess the performance of general user-assisted Video-On-Demand (VoD) systems, in which users browse an online catalogue of available videos and asynchronously issue requests to watch a given content. We remark that VoD systems should not be confused with websites offering live streaming service, where users join the distribution of a given TV channel at random points in time, but users connected to the same channel watch the content almost synchronously.

We emphasize that our analysis is orthogonal to two ongoing streams of theoretical research: the one targeting optimal replication strategies and push/pull schemes for content (re)placement [6], [7], [8], and the one dealing with practical issues related to chunk scheduling [9] and peer selection [10]. This will become clear later on in the paper.

The paper is organized as follows. In Sect. II we describe the system model; in Sect. III we present our stochastic fluid framework to characterize the instantaneous bandwidth requested from the servers, as function of the average number of users present in the system at a given time. The obtained results are then exploited in Sect. IV to analyse the asymptotic behavior of the system as the number of users increases, under both constant and time-varying video popularity. In Sect. V we give an overview of the related literature. In Sect.VI we briefly discuss the impact of the main modeling assumptions. We conclude the paper in Sect. VII.

II. Model

A. System assumptions

In our system users run applications that allow them to browse an online catalogue of videos. When a user selects a video, we assume that the request is immediately satisfied and the selected video can be watched uninterruptedly till the end, *i.e.*, the system is able to steadily provide to the user a data flow greater than or equal to the video playback rate. Users contribute their upload bandwidth to the video distribution, thus they can retrieve part of the video (or even the entire video) from other peers[1], saving servers' resources.

We focus on a given video of size L bytes. We assume that the video is downloaded by each user at constant rate d bytes/s, greater than or equal to the playback rate. Let $\tau_d = L/d$ be the time needed to download the whole video. In general, the download rate of a peer could be adapted over time, and made dependent on certain peer's characteristics (such as its upload bandwidth). By assuming a constant download rate d at each user we greatly simplify the analysis, while obtaining a conservative prediction with respect to the case in which the download rate can be adapted over time maintaining an average value equal to d.

The amount of upload bandwidth with which users contribute to the redistribution of the video, instead, may or may not be under the control of the system. In our analysis, we assume that the upload bandwidth available at a peer is a random variable with a given distribution. This way, we encompass both the realistic case of users with heterogeneous Internet connections (*i.e.*, ADSL, fiber, LAN) and cross-traffic fluctuations, and the case in which the peer upload bandwidth allocated to the given video is tuned by the system (such as in universal streaming architectures). More specifically, the amount of upload bandwidth with which users contribute at a given time to the redistribution of the considered video is modelled as a random variable U with cumulative distribution function $F_U()$ and mean $\overline{U}$. The random variables denoting the instantaneous upload bandwidths of the users are assumed to be i.i.d. (independent and identically distributed).

[1]In this paper we use the terms peer and user interchangeably.

B. Users dynamics

We need to incorporate in our analysis a model describing how peers join the distribution of the considered video, and when and how they leave the system, stopping to contribute their upload bandwidth. To this aim, we adopt a flexible model that allows to consider a non-stationary video request process. In particular, we assume that the arrival process of requests for the considered video follows a time-varying Poisson process of intensity $\lambda(t)$. Assuming that at a given time the arrival process is Poisson is reasonable, since users behave independently of each other, and their requests are immediately satisfied. On the other hand, a video (*e.g.*, a typical movie) can be long enough that the rate at which it is requested varies significantly during the playing time, due to daily traffic fluctuations, or rapidly-changing video popularity. By considering a non-homogeneous Poisson process of video requests with time-varying intensity $\lambda(t)$, we account for quite general non-stationary conditions.

As soon as users issue their request to watch the considered video, they start downloading it and assisting other peers. We define as *activity period* the duration of the interval during which a peer contributes its upload bandwidth to the distribution of the video, starting from the instant at which the video has been requested. Activity periods of the users are highly heterogeneous, as observed in several measurement studies [4]: some users stop watching the video after a very short time since the beginning, because they realize they are no longer interested in it; most users who decide to watch the video shut down the computer/Internet-TV towards the end of it; some of them keep the application running for prolonged time after the end of the video; those running set-top-boxes can be considered to be always active and serving other peers (as long as the set-top-box keeps a copy of the considered video, contributing to its distribution).

We account for general user behavior assuming that the activity period of a user is described by an arbitrary random variable T with finite mean $\overline{T}$ and complementary cumulative distribution function $G_T()$. The activity periods of the users are assumed to be i.i.d.

It follows from our assumptions that the number of active users $N(t)$ at time t is distributed as the number of customers in an M/G/∞ queue with time-varying arrival rate, hence it follows a Poisson distribution with time-varying mean $\overline{N}(t)$

given by

$$\overline{N}(t) = \int_{-\infty}^{t} \lambda(z) G_T(t-z)\,\mathrm{d}z \tag{1}$$

In our analysis we need to distinguish two classes of active users: those who are still *downloading* the video, and those who have completed the download (referred to as *seeds* in the following). The number of *downloading* peers at a given time instant t, denoted by $N_d(t)$, follows a Poisson distribution of mean $\overline{N}_d(t)$ given by

$$\overline{N}_d(t) = \int_{t-\tau_d}^{t} \lambda(z) G_T(t-z)\,\mathrm{d}z \tag{2}$$

Then, standard properties of Poisson processes allow to say that the number of *seeds* at time t, $N_{\text{seed}}(t)$, follows a Poisson distribution of mean $\overline{N}_{\text{seed}}(t) = \overline{N}(t) - \overline{N}_d(t)$.

We define as instantaneous system load $\gamma(t)$ the quantity

$$\gamma(t) = \frac{d \cdot \overline{N}_d(t)}{\overline{U} \cdot \overline{N}(t)} \tag{3}$$

which is the ratio between the average amount of data rate requested at time t by downloading peers and the average upload rate provided by all active users at time t. Borrowing the terminology adopted in previous work [3], [5] we say that at time t the system operates in *deficit* mode if $\gamma(t) > 1$, in *balanced* mode if $\gamma(t) = 1$, and in *surplus* mode if $\gamma(t) < 1$.

We denote by $\overline{T}_d = \int_0^{\tau_d} G_T(z)\,\mathrm{d}z$ the average time spent downloading the video by a user. Note that, in general, $\overline{T}_d$ can be shorter than τ_d due to peer churn (premature abandons). We define as per-user system load γ_p the quantity

$$\gamma_p = \frac{d \cdot \overline{T}_d}{\overline{U} \cdot \overline{T}} \tag{4}$$

which is the ratio between the average amount of data that is downloaded by a peer, and the average amount of data that a peer is able to offer to other peers. Note that by construction γ_p is equal to the (constant) instantaneous system load $\gamma(t)$ in the case of a time-invariant user arrival process.

C. Performance metrics

A fundamental goal of a VoD system is to minimize the bandwidth requested from the servers. To save server bandwidth, the system tries first to exploit the upload capacity of the peers as much as possible, so as to meet the strict constraints of video distribution (*i.e.*, maximum delay, minimum rate). In this paper we consider an ideal bandwidth allocation mechanism that is able, at any given time, to fully exploit the upload bandwidth that users can supply to the system (subject to physical constraints). In particular, we first take into account the cooperation among downloading users, allocating their upload bandwidth according to an (optimal) sequential delivery scheme. Then we allocate the bandwidth provided by the seeds. If this is not enough (i.e., some additional bandwidth is needed to support all active downloads), the system finally resorts to servers' bandwidth.

Let $S(t)$ be the random variable denoting the additional bandwidth that the servers must supply at time t to satisfy all active downloads of the considered video. Let $F_S(w,t)$ be the cumulative distribution of $S(t)$. At last, we denote by $\overline{S}(t)$ the mean of $S(t)$. Since in practice there are multiple videos to be served concurrently by the system, statistical multiplexing arguments suggest that a good design goal is to minimize the mean value $\overline{S}(t)$ of the servers bandwidth required by a single video. Therefore, this will be the main metric that we will look at in our performance analysis. Table I summarizes the notation of our model.

TABLE I
NOTATION

Symbol	Definition
d	user download rate
$\overline{U}$	average user upload bandwidth
$\overline{T}_d$	average time spent downloading the video (s)
$\overline{T}$	average user activity period (s)
$\lambda(t)$	arrival rate of new requests for the video at time t
$\overline{N}(t)$	average number of users at time t
$\overline{N}_d(t)$	average number of *downloading* users at time t
$\overline{N}_{\text{seed}}(t)$	average number of *seeds* at time t
$\overline{S}_d(t)$	average bandwidth requested by downloading users at time t
$\overline{S}_{\text{seed}}(t)$	average bandwidth offered by seeds at time t
$\overline{S}(t)$	average bandwidth requested from the servers at time t
$\gamma(t)$	instantaneous system load at time t
$f \sim g$	$f \in \Theta(g)$, *i.e.*, f is bounded both above and below by g asymptotically

III. Analysis

A. Preliminaries

In our analysis we take a snapshot of the system at an arbitrary instant t, and seek to characterize the random variable $S(t)$ denoting the instantaneous bandwidth requested from the servers at time t. This quantity depends on the instantaneous number of downloading users $N_d(t)$ and on the instantaneous number of seeds $N_{\text{seed}}(t)$. Let $S_d(t)$ be the bandwidth requested by the downloading users from both servers and seeds at time t (i.e., $S_d(t)$ is the bandwidth that servers and seeds must supply to downloading users at time t). Observe that, since downloading users can assist the download of other users with their own bandwidth, $S_d(t)$ significantly differs from the total download rate $d \cdot N_d(t)$. The mathematical definition of $S_d(t)$ will be specified in the following. We define $S_{\text{seed}}(t) = \sum_{i=1}^{N_{\text{seed}}(t)} U_i$ as the aggregate upload bandwidth offered by the seeds at time t. The bandwidth requested from the servers at time t is given by the difference between the bandwidth requested by the downloading users and the bandwidth offered by seeds, provided that such difference is positive:

$$S(t) = \max\{0, S_d(t) - S_{\text{seed}}(t)\}. \tag{5}$$

Under our system assumptions, both $N_d(t)$ and $N_{\text{seed}}(t)$ have a Poisson distribution, which is completely characterized by its mean. It follows that $S(t)$ is essentially a function of the mean value $\overline{N}_d(t)$ of downloading users at time t and of the mean value $\overline{N}_{\text{seed}}(t)$ of seeds at time t:

$$S(t) = f\left(\overline{N}_d(t), \overline{N}_{\text{seed}}(t)\right).$$

In this section, to simplify the notation, we will drop the dependency of all variables on time, focusing our attention on

the random variable

$$S = f(\overline{N}_d, \overline{N}_{\text{seed}}) \triangleq \max\{0, S_d - S_{\text{seed}}\} \tag{6}$$

representing the bandwidth requested from the servers in the presence of a generic average number $\overline{N}_d$ of downloaders and a generic average number $\overline{N}_{\text{seed}}$ of seeds (both Poisson distributed). Notice that S completely characterizes the system performance under constant video request rate, since in this case the average number of downloaders/seeds does not vary over time. The extension of the analysis to time-varying video request rate will be done in Section IV.

To analyse variable S in (6), we first observe that $S_{\text{seed}} = \sum_{i=1}^{N_{\text{seed}}} U_i$ is simply characterized as a compound Poisson random variable whose moment generating function is

$$\mathbb{E}[e^{zS_{\text{seed}}}] = e^{\overline{N}_{\text{seed}}(\phi_U(z)-1)} \tag{7}$$

where $\phi_U(z) \triangleq \mathbb{E}[e^{zU}]$ is the moment generating function of the peer upload bandwidth U, which is supposed to be known. In particular, the average of S_{seed} is $\overline{S}_{\text{seed}} = \overline{N}_{\text{seed}}\overline{U}$. Notice also that S_{seed} is independent from S_d. We will denote by $F_{S_{\text{seed}}}()$ its cumulative distribution function.

The main challenge of our analysis is thus the characterization of the bandwidth S_d requested by the downloading users.

B. Universal lower bound

Focusing on S_d, we first condition it on the number of downloading users, defining

$$S_d(k) \triangleq (S_d \mid N_d = k) \tag{8}$$

An easy lower bound to $S_d(k)$ can be obtained assuming that the upload bandwidth of each downloading user can always be fully utilized by the system, irrespective of the arrival time of the user into the system. We obtain

$$S_d(k) \geq \max\{d, k\ d - \sum_{i=1}^{k} U_i\} \geq k\ d - \sum_{i=1}^{k} U_i$$

and thus $\mathbb{E}[S_d(k)] \geq k(d - \overline{U})$. Deconditioning with respect to k we obtain $\mathbb{E}[S_d] \geq \overline{N}_d(d - \overline{U})$.

At last, we can obtain a lower bound to the average server bandwidth $\overline{S}$ requested from the servers, since by construction

$$\begin{array}{rcl} \overline{S} & = & \mathbb{E}[\max\{0, S_d - S_{\text{seed}}\}] \\ & \geq & \max\{0, \mathbb{E}[S_d - S_{\text{seed}}]\} \\ & \geq & \max\{0, \overline{N}_d(d - \overline{U}) - \overline{N}_{\text{seed}}\overline{U}\} \\ & = & \max\{0, d\,\overline{N}_d - \overline{U}\,\overline{N}\} \end{array} \tag{9}$$

which provides a universal lower bound to $\overline{S}$ for any chunk distribution scheme. The above lower bound is trivially zero for $\gamma<1$, whereas it is equal to $d\,\overline{N}_d - \overline{U}\,\overline{N}$ for $\gamma \geq 1$.

C. Upper bounds

An upper bound to the bandwidth requested from the servers can be obtained assuming that all peers download the video chunks sequentially. We observe that many implemented applications inspired by BitTorrent allow also non-sequential chunk dissemination in a swarm-like fashion, although this is typically only enabled within a limited portion of the video to meet the hard delay constraints of individual chunks. Actually, an almost in-order download is the only choice when the download rate d is close to the video playback rate (and the start-up delay is small).

Besides being analytically tractable (as we will see), the sequential download is also simple to implement in a peer-assisted VoD system, as it does not require the complex chunk/peer selection mechanisms which are necessary in BitTorrent-like swarms. In any case, the main point is that the server bandwidth required under sequential download is an upper bound to the bandwidth required by a more general (non-sequential) download scheme.

Below we show how to obtain analytical upper bounds to $\overline{S}$, the average of S, in the case of sequential download, obtaining upper bounds valid for any distribution scheme.

We start looking at quantity $S_d(k)$ defined in (8). We observe that, if all peers download the video sequentially at common rate d, a peer can only redistribute video pieces to peers arrived later on in time.

When there is only one downloading user, we trivially have $S_d(1) = d$. If there are two downloaders, the first arrived makes its entire upload bandwidth available to the second, and we have

$$S_d(2) = d + \max\{0, d - U_1\} = \max\{d, 2d - U_1\}$$

where d represents the external bandwidth necessary to sustain the download of the first arrived peer and $\max\{0, d - U_1\}$ represents the bandwidth needed to sustain the download of the second arrived peer.

When there are three downloaders, the last arrived can exploit the upload bandwidth of the second plus the residual upload bandwidth of the first, *i.e.*, a total upload rate of $U_2 + \max\{U_1 - d, 0\}$. Summing up the download rates needed by the three peers, we obtain

$$\begin{array}{rcl} S_d(3) & = & d + \max\{0, d - U_1\} \\ & & + \max\{0, d - U_2 - \max\{U_1 - d, 0\}\} \\ & = & d + \max\{0, d - U_1\} \\ & & + \max\{0, \min\{2d - U_1 - U_2, d - U_2\}\} \\ & = & \max\{S_d(2), 3d - U_1 - U_2\} \end{array} \tag{10}$$

The last equation follows from the fact that if $d - U_2 < 2d - U_1 - U_2$ then $d - U_1 > 0$ and thus $\max\{0, d - U_1\} + \max\{0, \min\{2d - U_1 - U_2, d - U_2\}\} = \max\{0, d - U_1, 2d - U_1 - U_2\}$.

In general the k-th downloader (assuming downloaders to be numbered in order of arrival) can receive the content from every other downloader preceding it. However, if the preceding peers are not able to fully support the download of the k-th downloader (*i.e.*, if $\sum_{i=1}^{k-1} U_i - kd < 0$), the missing bandwidth must be provided either by servers or by seeds. We obtain the following recursive equation for $S_d(k)$:

$$S_d(k) = \begin{cases} d & k = 1 \\ \max\{S_d(k-1), kd - \sum_{j=1}^{k-1} U_j\} & k > 1 \end{cases} \tag{11}$$

If we iterate back up to $S_d(1)$ we can obtain an explicit

expression for $S_d(k)$ in terms of the upload bandwidths of peers U_i, for $i < k$, and of the download rate d, as:

$$\begin{aligned} S_d(k) &= d + \max\Big\{0, d-U_1, 2d-(U_1+U_2), \\ &\quad 3d-(U_1+U_2+U_3), \ldots, (k-1)d - \textstyle\sum_{i=1}^{k-1} U_i\Big\} \\ &= d + \max_{1\le j\le k-1}\Big\{0, \sum_{i=1}^{j}(d-U_i)\Big\} \end{aligned} \tag{12}$$

We emphasize that (12) has already been obtained in [3], [5]. However in previous work authors have resorted to Monte-Carlo approaches to evaluate it.

To proceed in the analysis, we define the auxiliary variable $Z_d(k)$:

$$Z_d(k) \triangleq \max_{1\le j\le k}\Big\{\sum_{i=1}^{j}(d-U_i)\Big\} \tag{13}$$

where $Z_d(k) = 0$ if $k = 0$. Then $S_d(k)$ can be expressed in terms of $Z_d(k-1)$ according to

$$S_d(k) = d + \max\{0, Z_d(k-1)\}. \tag{14}$$

Now, $Z_d(k)$ can be regarded as the maximum value (up to time k) reached by a unidimensional random walk with increments $X_i = d - U_i$. Thus we can exploit the existing literature on random walks, and especially their application to risk theory, to characterize the distribution of $Z_d(k)$.

For our purposes, we need the following classic result, known as the Lundberg's inequality (see for example [11]).

Lemma 1: (**Lundberg inequality**) Consider a sequence of i.i.d. variables $(X_i)_{i\ge1}$, satisfying the following three properties: i) $\mathbb{E}[X_1] < 0$; ii) $\mathbb{P}(X_1 > 0) > 0$; iii) $\mathbb{E}[e^{tX_1}]$ is finite in a neighborhood of the origin. Define the r.v. $Q(k) \triangleq \sum_{i=1}^{k} X_i$, $k \ge 1$, $Q(0) \triangleq 0$. Then, denoting θ^* the strictly positive solution of $\mathbb{E}[\mathrm{e}^{\theta^* X_1}] = 1$, which exists unique under i), ii), and iii), we have, for all $n \ge 1$:

$$\mathbb{P}(\max_{1\le k\le n} Q(k) > w) \le \mathrm{e}^{-\theta^* w} \ , \ \forall w \ge 0. \tag{15}$$

Remark: condition iii) requires X_1 to be light-tailed (*i.e.*, to have a tail that decays at least exponentially fast).
For completeness, in Appendix A we report a proof of Lemma 1 based on a Martingale approach.

Lundberg inequality can be generalized and adapted to our context, to obtain an upper bound to $\mathbb{P}(S_d > w)$:

Theorem 1: Assume the following properties hold for U: i) $\overline{U} > 0$, ii) $\mathbb{E}[e^{tU}]$ is finite in a neighborhood of the origin, iii) $F_U(w) > 0$ for every $w > 0$. For $\epsilon \in [(\overline{U}-d)^+, \overline{U})$, define $A \triangleq d - \overline{U} + \epsilon$ (note that $\max\{0, d-\overline{U}\} \le A < d$). Let θ^* be the unique strictly positive solution of the equation $\mathbb{E}[e^{\theta(d-U-A)}] = 1$. For any $w \ge 0$, it holds

$$\mathbb{P}(S_d > w) \le \begin{cases} \min\{C_2 e^{-\theta^*(w-d)}, C_3\} & w \ge d \\ C_1 & 0 \le w < d, \end{cases}$$

where $C_1 \triangleq 1 - e^{-\overline{N}_d}$, $C_2 \triangleq e^{-\theta^* A} e^{-\overline{N}_d}(e^{\overline{N}_d e^{\theta^* A}} - \overline{N}_d e^{\theta^* A} - 1)$ and $C_3 \triangleq 1 - e^{-\overline{N}_d} - \overline{N}_d e^{-\overline{N}_d}$.

A detailed proof of Theorem 1 is reported in Appendix B. Observe that, when $d < \overline{U}$, we can obtain an upper bound on $\mathbb{P}(S_d > w)$ applying the Lundberg inequality to $\mathbb{P}(S_d(k) > w)$ for any k. Instead, when $d > \overline{U}$, since $\mathbb{E}[d - U_i] > 0$, we cannot apply Lundberg inequality directly to $\mathbb{P}(S_d(k) > w)$. Therefore we need to define an auxiliary sequence of random variables, tightly related to $S_d(k)$, on which Lundberg bound can be applied. Then we can derive a bound on $\mathbb{P}(S_d(k) > w)$. The approach of the auxiliary sequence of variables is generalized also to the case $d < \overline{U}$, to obtain a possibly tighter upper bound.

Exploiting the result in Theorem 1, we derive an upper bound to the average bandwidth $\overline{S}_d$ requested by the downloading peers:

Corollary 1: The average bandwidth requested by downloading peers satisfies:

$$\overline{S}_d \le \begin{cases} C_1 d + C_3(w^* - d) + C_3/\theta^* & \text{if } C_2 > C_3 \\ C_1 d + C_2/\theta^* & \text{o.w.} \end{cases} \tag{16}$$

where $w^* \triangleq \frac{1}{\theta^*}\log\left(\frac{C_2}{C_3}\right) + d$.

The proof of Corollary 1 can be found in Appendix C.

Remark: note that (1) and (16) hold under an arbitrary choice of $\epsilon \in [(\overline{U}-d)^+, \overline{U})$. The tightest bound is obtained by minimizing the expressions (1) and (16) with respect to ϵ.

From Corollary 1 it immediately follows that the average bandwidth requested by downloaders is finite even when the average number of downloaders diverges (*i.e.*, $\overline{N}_d \to \infty$), provided that $d < \overline{U}$. Indeed, by selecting $\epsilon = \overline{U} - d$ (and thus $A = 0$), we obtain $\overline{S}_d \le d + \frac{1}{\theta^*}$, being θ^* the unique positive solution to $\mathbb{E}[e^{\theta(d-U)}] = 1$.

By taking into account also the impact of the seeds, we obtain an upper bound to the average bandwidth requested from the servers, according to (6):

Theorem 2: The following bound holds for the average bandwidth requested from the server:

$$\begin{aligned} \overline{S} \ \le \min\Big\{ & C_1 F_{S_{\text{seed}}}(d) d + C_2 \tfrac{e^{\theta^* d}}{\theta^*}\mathbb{E}[e^{-\theta^* S_{\text{seed}}}], C_1 F_{S_{\text{seed}}}(d) d \\ & + C_3\left(\tfrac{2}{\theta^*} + w^* - \overline{S}_{\text{seed}} + \mathbb{E}[e^{\theta^* S_{\text{seed}}}] e^{-\theta^* w^*}/\theta^*\right)\Big\} \end{aligned} \tag{17}$$

where $w^* = \frac{1}{\theta^*}\log\left(\frac{C_2}{C_3}\right) + d$.

The proof of Theorem 2 is reported in Appendix D.

D. Exact solutions

In this section we describe a methodology to obtain an exact solution of (12) when the upload bandwidth distribution is exponentially or phase-type distributed.

The first step consists in deriving an integral equation satisfied by the cumulative distribution function of the quantity $Z_d(k)$ defined in (13). Observe that $Z_d(k)$ can be written as:

$$Z_d(k) = \max\Big\{d - U_1, \max_{2\le j\le k}\big\{\sum_{i=2}^{j}(d-U_i)\big\} + d - U_1\Big\}.$$

Now since the U_i are i.i.d., we can permute the indices i of

$(U_i)_{i\geq 1}$, obtaining a new random variable $\hat{Z}_d(k)$ defined as

$$\hat{Z}_d(k) \triangleq \max\Big\{d-U_k, \max_{1\leq j\leq k-1}\big\{\sum_{i=2}^{j}(d-U_i)\big\}+d-U_k\Big\}$$

which has the same distribution of $Z_d(k)$. Note that $\hat{Z}_d(k)$ can be written as:

$$\hat{Z}_d(k) = \max\Big\{d-U_k, \hat{Z}_d(k-1)+d-U_k\Big\}. \tag{18}$$

Denoting by $F_Z(w \mid k)$ the cumulative distribution function of $Z_d(k)$ (and thus of $\hat{Z}_d(k)$), we have:

$$\begin{aligned} F_Z(w \mid k) &= \mathbb{P}(\hat{Z}_d(k)\leq w) \\ &= \mathbb{P}\Big(\max\Big\{d-U_k, \hat{Z}_d(k-1)+d-U_k\Big\}\leq w\Big) \end{aligned} \tag{19}$$

We now condition on the value assumed by $X_k \triangleq d-U_k$:

$$\begin{aligned} F_Z(w \mid k) &= \int_{-\infty}^{\infty}\mathbb{P}\Big(\max\Big\{X_k, \hat{Z}_d(k-1)+X_k\Big\}\leq w \\ &\quad \mid X_k=\alpha\Big)\,\mathrm{d}F_{X_k}(\alpha) \\ &= \int_{-\infty}^{w}\mathbb{P}\left(\hat{Z}_d(k-1)+\alpha\leq w\right)\,\mathrm{d}F_{X_k}(\alpha) \\ &= \int_{-\infty}^{w}F_Z(w-\alpha \mid k-1)\,\mathrm{d}F_{X_k}(\alpha). \end{aligned} \tag{20}$$

Observing that by construction $F_Z(\alpha \mid 1) = F_{X_k}(\alpha)$, from (20) we get:

$$F_Z(w \mid k) = \int_{-\infty}^{w}F_Z(w-\alpha \mid k-1)\,\mathrm{d}F_Z(\alpha \mid 1) \tag{21}$$

Explicit solutions of the above functional equation can be given when the peer upload bandwidth is phase-type distributed. In particular, in the case of peer upload bandwidth exponentially distributed, we obtain (see Appendix E):

$$\begin{aligned} &F_Z(w \mid k) = F_Z(d \mid k)e^{\frac{w-d}{\overline{U}}}\mathbb{I}_{w<0} + \mathbb{I}_{w>kd} \\ &+e^{\frac{w-d}{\overline{U}}}\sum_{i=0}^{k}(-1)^i F_Z(d \mid k-i)\frac{(w-id)^i}{\overline{U}^i i!}e^{-id/\overline{U}}\mathbb{I}_{id\leq w\leq kd}, \end{aligned} \tag{22}$$

where $\mathbb{I}$ is the indicator function. In (22) the constants $F_Z(d \mid k)$ can be obtained imposing the condition $F_Z(kd \mid k) = 1$ for all k, as shown in Appendix E.

From $F_Z(d \mid k)$ we immediately obtain $\mathbb{P}(S_d(k) > w)$:

$$\mathbb{P}(S_d(k)>w) = \begin{cases} 1-F_Z(w-d \mid k-1) & \text{if } w\geq d \\ 1 & \text{if } w<d \end{cases} \tag{23}$$

Finally, we can derive the average server bandwidth $\overline{S}_d$ requested by the downloading peers, and the average server bandwidth $\overline{S}$ requested from the servers.

In the case $d<\overline{U}$, since the sequence of increasing random variables $Z_d(k)$ converges w.p.1 to a finite random variable $Z_d(\infty)$ (as direct consequence of Lemma 1), we can find the distribution of $Z_d(\infty)$ from the stationary version of (21). We state this result in the following theorem.

Theorem 3: Under the condition $d<\overline{U}$, the cumulative function of $Z_d(\infty)$ satisfies the stationary version of (21), *i.e.*,

$$F_Z(w \mid \infty) = \int_{-\infty}^{w}F_Z(w-\alpha \mid \infty)\,\mathrm{d}F_X(\alpha) \tag{24}$$

When the U_k are exponentially distributed the solution of (24) can be obtained following the approach described in Appendix E, obtaining:

$$\begin{aligned} &F_Z(w \mid \infty) = F_Z(d \mid \infty)e^{\frac{w-d}{\overline{U}}}\mathbb{I}_{w<0} \\ &+F_Z(d \mid \infty)e^{\frac{w-d}{\overline{U}}}\sum_{i=0}^{\infty}(-1)^i\frac{(w-id)^i}{\overline{U}^i i!}e^{-id/\overline{U}}\mathbb{I}_{w>id} \end{aligned}$$

where

$$F_Z(d \mid \infty) = \lim_{k\to\infty}\frac{1}{\sum_{i=0}^{k}(-1)^i\frac{[(k-i)d]^i}{i!}e^{(k-i)d/\overline{U}}}$$

Notice that $Z_d(\infty)$ provides a tight bound to the distribution of the server bandwidth requested by a large number of downloading users, when the system operates in the surplus mode.

We emphasize that the approach described in Appendix E can be generalized in a rather straightforward way to obtain the exact solution of (21) and (24) under any phase-type distribution of peers upload bandwidth. In this regard, recall that any distribution whose moment generating function is finite in a neighborhood of the origin (*i.e.*, it is light-tailed) can be approximated by a phase-type distribution with an arbitrary degree of accuracy (see [11, Ch.4]). Thus the methodology presented in this section can be applied to derive analytical approximations of the bandwidth requested from the servers in the case in which the peer upload bandwidth is arbitrarily distributed.

As a concluding remark, we wish to emphasize that upper bounds obtained in Section III-C and exact solutions presented in this section are complementary tools for the analysis of peer-assisted VoD systems properties. Indeed, the upper bounds presented in Section III-C provide very general and easy-to-handle expressions from which we can derive qualitative/asymptotic properties of the system. The methodology described in this section provides more accurate estimates (which are exact for phase-type distributions) of the bandwidth requested from the servers, but are computationally more expensive, especially for large numbers of users.

E. Numerical Illustration

We provide a graphical illustration of our results considering a scenario in which the video request rate is constant, and the average activity period of the users is twice the time spent downloading the movie, representing users who tend to keep their application/devices active after watching the movie. Notice that in this case $\overline{N}_d = \overline{N}_{\text{seed}}$. We normalize the parameters $d=1$ and $\overline{T}_d=1$, and thus we set $\overline{T}=2$.

Figure 1 reports on a log-log scale the average server bandwidth $\overline{S}$ as function of the average number of users $\overline{N}$, for different values of the per-user system load γ_p. The peer upload bandwidth is exponentially distributed with mean $\overline{U}=1/(2\gamma_p)$. We compare the upper bound (17) (labeled UB) with the exact solution presented in Section III-D. Specifically, exact results are derived from (23), de-conditioning with respect to the number $n_d(t)$ of downloading users in the system. We also report for $\gamma_p>1$ the lower bound (9).

Comparing the upper bound with the exact solution, we observe that, although the bound can be pessimistic up to a

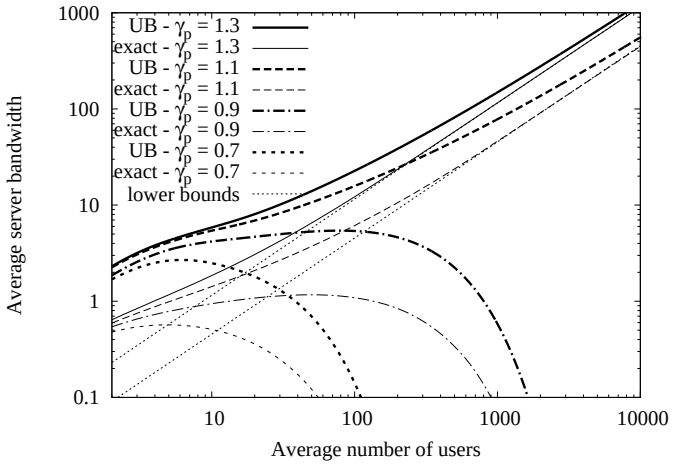

Fig. 1. Average server bandwidth $\overline{S}$ versus the average number of users $\overline{N}$, for different values of the per-user system load γ_p, in the case $d = 1$, $\overline{T}_d = 1$, $\overline{T} = 2$, and exponentially distributed upload bandwidth.

factor about 4, the bound captures well the qualitative behavior of the exact curve.

As expected, in the deficit mode ($\gamma_p > 1$) the server bandwidth diverges for increasing number of users. Moreover, the upper bound becomes asymptotically tight to the lower bound, which grows linearly with the number of users. This interesting property will be proved in the next section.

In the surplus mode, instead ($\gamma_p < 1$), the server bandwidth achieves a maximum for a given number of users, and then decreases to zero as $\overline{N} \to \infty$. This is another fundamental asymptotic property of our system, which will be proved in the next section.

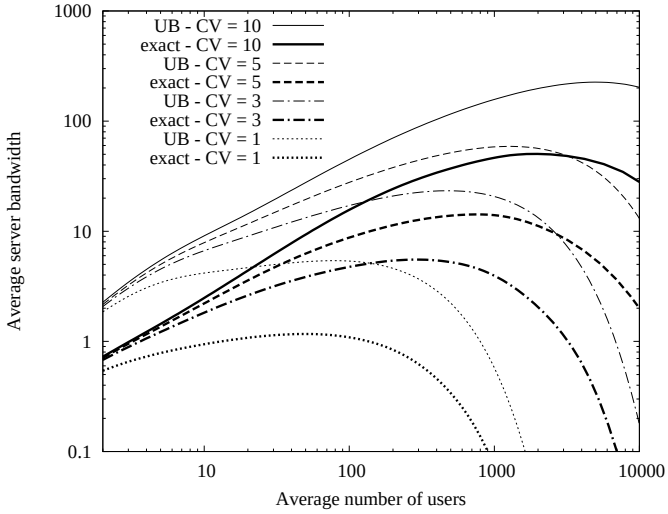

Fig. 2. Average server bandwidth $\overline{S}$ versus the average number of users $\overline{N}$, for different values of the coefficient of variation (CV) of user upload bandwidth, in the case $d = 1$, $\overline{T}_d = 1$, $\overline{T} = 2$, $\gamma_p = 0.9$.

To show the impact of peer bandwidth heterogeneity, we consider the same scenario as before, keeping the load fixed to $\gamma_p = 0.9$ (surplus mode), and varying the coefficient of variation (CV) of the upload bandwidth distribution of the users. In particular, we assume that the upload bandwidth is distributed according to a second-order hyper-exponential distribution with balanced means, which could well describe the situation in which we have many peers with low upload bandwidth (*e.g.*, behind ADSL lines) and few peers with large upload bandwidth (*e.g.*, connected with fiber or LAN). We observe the strong impact of the CV on the resulting server bandwidth. Since $\gamma_p = 0.9 < 1$ (surplus mode), we expect that $\overline{S}$ goes to zero as $\overline{N} \to \infty$ (as predicted by Theorem 4 in the next section). However, it is interesting to observe that the maximum value of $\overline{S}$ is achieved for quite large number of users (in the order of thousands) for large values of the CV. Again, the analytical upper bound follows well the qualitative behavior of the system, in all considered cases.

IV. Asymptotic analysis

The numerical results reported in the previous section suggest interesting asymptotic properties of our system as the number of users grows large. In this section, we will precisely characterize how the average bandwidth requested from the servers scales as we increase the number of users, i.e., when $\lambda \to \infty$.

We will first consider the simpler case in which the video request rate is constant, *i.e.*, $\lambda(t) = \lambda$, referred to as *time-invariant* video popularity. Then, we will analyse a scenario in which the video request rate varies over time, referred to as *time-varying* video popularity. We will see that, while in the deficit mode the above two cases scale in a similar way, in the surplus mode the asymptotic system behavior is radically different. In particular, different conditions on the system parameters must be satisfied in the two considered scenarios to achieve scalability as the users population size increases.

A. Time-invariant video popularity

The asymptotic system behavior in this case is characterized by the following fundamental result:

Theorem 4: Assume U not constant. Then, as $\lambda \to \infty$, the following asymptotic regimes hold for any chunk distribution scheme: For $\gamma_p < 1$ and, additionally $\overline{T} > \overline{T}_d$, the average bandwidth requested from the servers tends to zero, *i.e.*, $\lim_{\lambda\to\infty} \overline{S} = 0$. For $\gamma_p > 1$, the average bandwidth requested from the servers grows linearly with the number of users. In particular, $\lim_{\lambda\to\infty} \frac{\overline{S}}{(\overline{N}_d d - \overline{U}\ \overline{N})} = 1$.

The proof of Theorem 4 can be found in Appendix F. Notice that for $\gamma_p > 1$ the upper bound becomes asymptotically tight to the lower bound (9), as observed in the numerical example in Figure 1.

Theorem 4 suggests that, for very popular contents (large number of users concurrently watching the same video), a peer-assisted video distribution is scalable, provided that the system is in surplus mode (*i.e.*, $\gamma_p < 1$) and users stay in the system, on average, for a time larger (by an arbitrarily small constant) than the time needed to download the whole video ($\overline{T} > \overline{T}_d$). This holds for any chunk distribution scheme, including the simple sequential scheme.

In the deficit mode, *i.e.*, $\gamma_p > 1$, the system is obviously not scalable, since an additional bandwidth at least equal to the

bandwidth deficit ($\overline{N}_d d - \overline{U}\ \overline{N}$) must be provided by servers. However, in this case there is (asymptotically) no gain in adopting a non-sequential chunk delivery scheme with respect to the simple sequential download. Indeed, as the number of users grows large, the system in which users download the content sequentially performs as well as an ideal (infeasible for VoD applications) system in which the content can be downloaded in arbitrary order.

B. Time-varying video popularity

In this section we extend the asymptotic analysis to a scenario in which the video request rate varies over time, while still letting the average number of users downloading the video grow to infinity. In particular, we assume that the arrival process of requests for the considered video follows an non-homogeneous Poisson process with intensity $\lambda(t) = \Lambda q(t)$, where $q(t)$ is a shaping function modelling the popularity evolution of a video inserted into the catalogue at time $t = 0$.

Without loss of generality, we assume $q(t)$ to be an integrable function such that $\int_0^\infty q(t)\,\mathrm{d}t = 1$ and $q(t) = 0$ for all $t < 0$. By so doing, Λ becomes equivalent to the average overall number of times that the video is requested during its lifetime in the system. The asymptotic analysis is then performed by letting $\Lambda \to \infty$.

We characterize the system performance in terms of the average amount of data V that servers must supply to satisfy all requests for the considered video:

$$V = \int_0^\infty \overline{S}(t)\,\mathrm{d}t.$$

Our main results are summarized in the following:

Theorem 5: Consider a bounded, integrable popularity function $q(t)$ having finite mean, i.e., $\int_0^\infty tq(t)\,\mathrm{d}t < \infty$. If $d < \overline{U}$, we distinguish the following three cases:

- if $q(t)$ has an heavy tail, *i.e.*, $\exists\ T_0, K > 0$ and $\alpha > 1$ such that $q(t) < Kt^{-(\alpha+1)}$ for $t > T_0$, then the average amount of data requested from the servers is $V = O(\Lambda^{1/\alpha})$ as $\Lambda \to \infty$.
- if $q(t)$ has an exponential tail, *i.e.*, $\exists\ T_0, K > 0$ and $\alpha > 0$ such that $q(t) < Ke^{-\alpha t}$ for $t > T_0$, then $V = O(\log \Lambda)$ as $\Lambda \to \infty$.
- if $q(t)$ has finite support, *i.e.*, $\exists\ T_2 > 0$ such that $q(t) = 0$ for all $t > T_2$, then $V = O(1)$ as $\Lambda \to \infty$.

On the other hand, when $d > \overline{U}$, V grows linearly with Λ for any function $q(t)$, *i.e.*, $V = \Theta(\Lambda)$, $\Lambda \to \infty$.

A detailed proof of Theorem 5 is reported in Appendix G. In the following we just sketch the rationale of the proof.

First of all, from (2) we compute the following upper bound to the average number of downloading users at time t,

$$\overline{N}_d(t) \le \Lambda \int_{(t-\tau_d)^+}^{t} q(z)\,\mathrm{d}z \tag{25}$$

obtained considering the worst case in which users remain in the system at least for a period equal to the downloading time τ_d (*i.e.*, $G_T(z) = 1$ for all $z < \tau_d$). Given the bound on $\overline{S}(t)$ in (17), we compute V as the sum of two contributions: $V = \int_0^{T_1} \overline{S}(t)\,\mathrm{d}t + \int_{T_1}^\infty \overline{S}(t)\,\mathrm{d}t$, where T_1 is a threshold defined as $T_1 \triangleq \max\{t : \overline{N}_d(t) > 1\}$. Note that the average number of downloaders $\overline{N}_d(t)$ can be arbitrarily large for $t \in (0, T_1)$, while this number becomes negligible for $t > T_1$. When $d < \overline{U}$, we found that the first integral is bounded above by T_1. When the popularity distribution has an exponential decreasing tail, the most significant contribution is given by the first integral (*i.e.*, by T_1) which scales logarithmically with Λ. When the popularity distribution is heavy-tailed, the two integrals give the same contribution, asymptotically.

We emphasize that the results in Theorem 5 differ substantially from those stated in Theorem 4. Indeed, in the case of time-invariant video popularity the scalability of the system is governed by the system load γ_p. Notice that in this case the system can be in surplus mode ($\gamma_p < 1$) even when $d > \overline{U}$, since the system can also efficiently exploit the upload bandwidth offered by the seeds. In the case of time-varying video popularity, instead, the scalability of the system is governed by the relationship between $\overline{U}$ and d: if $d > \overline{U}$, the system cannot scale to large number of users, no matter how long users stay available in the system as seeds, after watching the movie. If $d < \overline{U}$, the scalability of the system depends on the tail behavior of the popularity shape function: the faster the tail decreases, the smaller the data volume requested from the servers. In particular we have shown that, at least in the case of a popularity function with finite support, the data volume requested from servers remains bounded as $\Lambda \to \infty$. More in general, the data volume V is sublinear in Λ.

V. Related Work

We restrict ourselves to mentioning theoretical performance studies of content distribution systems leveraging users' cooperation, which are closely related to our work.

Stochastic fluid models for classical peer-assisted file distribution systems, such as Bit-Torrent (in which the content can be enjoyed by users only after completing the download), have been proposed for both transient and steady-state regimes [12], [13], but they are not directly applicable to streaming systems. In [14], authors adapt the fluid model in [13] to VoD systems, investigating the impact of different piece selection policies (rarest-first and in-order) on download latency and startup delay, in the case of users with homogeneous bandwidth. In contrast to [14], we focus on the scalability of VoD systems with strict service guarantees and heterogeneous user upload bandwidths.

A stochastic fluid approach to analyse peer-assisted video distribution has been proposed in [15] in the context of live streaming, in which (heterogeneous) peers download and playback content synchronously. Here we apply the stochastic fluid approach to VoD streaming systems, whose dynamics are quite different from live streaming, since users can watch the video asynchronously.

The mathematical formulation (12) for the server bandwidth needed by a VoD system based on sequential delivery, appeared in [4], in which authors resort to a Monte-Carlo approach to get basic insights into the system behavior (like surplus and deficit modes).

The same formulation (12) has been considered in [5], where authors explore by simulation the effectiveness of different replication strategies to minimize the server load in the slightly surplus mode, as well as distributed replacement algorithms to achieve it. To the best of our knowledge, we are the first to analytically study the stochastic process (12), establishing its connection with random walks and risk theory.

An alternative analysis of equation (12), based on a second-order gaussian approximation, has been proposed in [16] to obtain an efficient methodology to evaluate the performance of both stationary and non-stationary systems.

In [17], a per-chunk capacity model is developed to show the tradeoff that exists between system throughput, sequentiality of downloaded data and robustness to heterogeneous network conditions. Optimal content placement strategies to maximize the upload capacity of (homogeneous) set-top-boxes (and thus minimize the servers workload) in VoD systems have been recently investigated in [8] under many-user asymptotic.

VI. Discussion on model assumptions

In this section we critically revisit the main assumptions of our model discussing their impact on the analysis of a VoD system.

The first strong assumption consists in assuming the video playback rate constant and equal to d. This assumption actually does not hold in practice since most video encoding schemes produce variable bitrate streams. However, rapid bitrate fluctuations are usually averaged out by the playout buffer of the decoder, so that assuming a download rate constant and equal to the average playback rate can be an acceptable assumption, while being also a reasonable design choice to simplify the system. Nevertheless, it would be possible to incorporate in the model a variable download rate, equal to the instantaneous playback rate of the video, under some specific assumptions: i) the bitrate distribution of the video is known; ii) bitrate fluctuations are sufficiently fast that any two users concurrently downloading the video have uncorrelated playback rates; iii) the bitrate distribution does not change throughout the video. Under the above assumptions, since different users in a VoD system are retrieving at time t different and independent segments of the video content, we could well assume instantaneous play-back rates of users to be described by i.i.d. random variables with known distribution. The effect of variable play-back rates could then be easily incorporated in our modeling framework without changing its mathematical structure. Observe, indeed, that the structure of $Z_d(k)$ in (13) remains unchanged when $d-U_i$ is replaced with d_i-U_i where d_i is a random variable.

The second important assumption of our work consists in modeling the users' upload bandwidth as i.i.d. random variables U_i with assigned distribution. We do not consider this assumption particularly restrictive, since it permits to represent pretty well bandwidth heterogeneity of users' access links, as well as random fluctuations in the available upload bandwidth due to cross traffic and other forms of bandwidth restriction. Notice that such fluctuations could be also correlated over time at each user, without affecting our analysis, which is essentially based on an instantaneous analysis of the system. The assumption that upload bandwidths are uncorrelated among users is also reasonable, since in a VoD system users are geographically spread, and thus they are likely to experience independent bandwidth fluctuations.

Furthermore, we assume that users download the video in a perfect sequential fashion, although this is not strictly required as long as a minimum play-out buffer level is maintained at the receiver. Indeed, notice that in many real systems videos are cut into segments, usually called chunks, and the reception of out-of-sequence chunks is allowed within a limited *sliding window* of data starting from the point currently played. Therefore, a perfect sequential delivery can be regarded as the limit case in which the size of the sliding window tends to zero (say it becomes much smaller than the total video size).

At last, in our work we have ignored implementation issues such as: i) the effects of protocol overheads and signalling bandwidth (necessary to reconfigure the cooperation among users); ii) possible constraints on the number of peers from which a user can simultaneously download data; iii) the effect of congestion inside the network. All these issues can potentially affect the performance of a realistic system, but we have not incorporated them for the sake of simplicity and analytical tractability.

VII. Conclusion

We have developed a stochastic fluid methodology to derive analytical upper bounds to the bandwidth requested from the servers in an ideal streaming system leveraging the upload bandwidth of the users, studying the performance achieved by the simple sequential distribution scheme. Our bounds hold under the only assumption that the upload bandwidth distribution of peers is light-tailed. We have also proposed an analytical methodology to exactly estimate the bandwidth requested from servers when user upload bandwidth is phase-type distributed. Besides being analytically tractable, the simple sequential delivery scheme is also an attractive solution in real systems, for two main reasons: i) it allows users to immediately start watching the requested movie; ii) it is simple to manage and control. Moreover, we have proved that the sequential delivery scheme leads to an asymptotically optimal exploitation of the peers' upload bandwidths as the number of users grows large. Indeed, our bounds tightly characterize the asymptotic performance of large-scale peer-assisted content distribution systems employing both sequential and non-sequential delivery schemes.

Appendix A
Proof of Lemma 1

Consider a sequence of i.i.d. variables $(X_i)_{i\geq 1}$, satisfying the three properties: i) $\mathbb{E}[X_1] < 0$; ii) $\mathbb{P}(X_1 > 0) > 0$; iii) $\mathbb{E}[e^{tX_1}]$ is finite in a neighborhood of the origin. Define $Q(k) = \sum_{i=1}^{k} X_i$, $k \geq 1$, $Q(0) := 0$. Define the filtration $\mathcal{F}_0 := \{\emptyset, \Omega\}$, $\mathcal{F}_k := \sigma\{X_1, \ldots, X_k\}$, $k \geq 1$ (*i.e.* the σ-algebra generated by $\{X_1 \ldots X_k\}$). Consider the r.v.

$$\tau_w := \inf\{k \geq 1 : \ Q(k) > w\}$$

where the infimum is equal to ∞ if $\{k \geq 1 :\ Q(k) > w\} = \emptyset$. Note that τ_w is the first time at which the process $Q(k)$ exceeds the quantity w.

Let θ^* be such that $\mathbb{E}[e^{\theta^* X_1}] = 1$. It can be proved that under the three conditions described above, there exists a unique $\theta^* > 0$. It can be easily checked that the process $\{\mathrm{e}^{\theta^* Q(k)}\}$ is an $\mathcal{F}_k$-martingale. Therefore, using the stopping theorem (note that τ_w is an $\mathcal{F}_k$-stopping time) the process $\{e^{\theta^* Q(k\wedge\tau_w)}\}$ is an $\mathcal{F}_k$-martingale, for each $w > 0$.

Consequently, we have

$$\begin{aligned} 1 &= \lim_{k\to\infty} \mathbb{E}[e^{\theta^* Q(k\wedge\tau_w)}] \\ &\geq \mathbb{E}[(\liminf_{k\to\infty} e^{\theta^* Q(k\wedge\tau_w)})\mathbf{1}\{\tau_w < \infty\}] \\ &= \mathbb{E}[\mathbf{1}\{\tau_w < \infty\} e^{\theta^* Q(\tau_w)}] \\ &\geq \mathbb{P}(\max_{1\leq k\leq n} Q(k) > w) e^{\theta^* w} \end{aligned}$$

where the first inequality follows by Fatou's lemma. We conclude that

$$\mathbb{P}(\max_{1\leq k\leq n} Q(k) > w) \leq e^{-\theta^* w}, \quad w \geq 0.$$

Appendix B
Proof of Theorem 1

Define $X_i \triangleq d - U_i - A$, for all $i \geq 1$ and $Q(k) \triangleq \sum_{i=1}^{k} X_i$. Since $\max\{0, d - \overline{U}\} \leq A < d$ we have $\mathbb{E}[X_i] < 0$ and $\mathbb{P}(X_i > 0) > 0$. Note that by (13), $Z_d(0) = 0$, while, for $k > 0$, it holds:

$$\begin{aligned} Z_d(k) &= \max_{1\leq j\leq k} \sum_{i=1}^{j} (d - U_i) \leq \max_{1\leq j\leq k} \Big\{ \sum_{i=1}^{j} (d - U_i) \\ &+ (k - j)A \Big\} = \left(\max_{1\leq j\leq k} Q(j) \right) + kA. \end{aligned}$$

For $k > 0$

$$\mathbb{P}(Z_d(k) > w + kA) \leq \mathbb{P}(\max_{1\leq j\leq k} Q(j) + kA > w + kA) \leq e^{-\theta^* w},$$

and then $\mathbb{P}(Z_d(k) > w) \leq e^{-\theta^* w} e^{\theta^* kA}$.

By (14), $S_d(k) = d + \max\{0, Z_d(k-1)\}$. It is easy to prove that the event $\{\max\{0, Z_d(k-1)\} > w'\}$ is equal to the event $\{Z_d(k-1) > w'\}$, for all $w' \geq 0$. Therefore, $\mathbb{P}(S_d(k) > w) = \mathbb{P}(Z_d(k-1) > w - d)$ for all $w \geq d$ and $\mathbb{P}(S_d(k) > w) = 1$ for $w < d$. Thus, for $w < d$ it holds

$$\begin{aligned} \mathbb{P}(S_d > w) &= \textstyle\sum_{k=1}^{\infty} \mathbb{P}(S_d(k) > w)\mathbb{P}(N_d = k) \\ &= \textstyle\sum_{k=1}^{\infty} \mathbb{P}(N_d = k) \\ &= 1 - \mathbb{P}(N_d = 0) = 1 - e^{-\overline{N}_d} = C_1. \end{aligned}$$

On the other hand, for $w \geq d$, we have

$$\begin{aligned} \mathbb{P}(S_d > w) &= \textstyle\sum_{k=1}^{\infty} \mathbb{P}(S_d(k) > w)\mathbb{P}(N_d = k) \\ &= \textstyle\sum_{k=1}^{\infty} \mathbb{P}(Z_d(k-1) > w - d) \frac{\overline{N}_d{}^k e^{-\overline{N}_d}}{k!} \\ &\leq \mathbb{P}(Z_d(0) > w - d)\, \overline{N}_d e^{-\overline{N}_d} \\ &\quad + e^{-\theta^*(w-d)} e^{-\overline{N}_d} e^{-\theta^* A} \textstyle\sum_{k=2}^{\infty} \frac{(e^{\theta^* A}\overline{N}_d)^k}{k!} \\ &= e^{-\theta^*(w-d)} e^{-\overline{N}_d} e^{-\theta^* A} \textstyle\sum_{k=2}^{\infty} \frac{(e^{\theta^* A}\overline{N}_d)^k}{k!} \\ &= C_2 e^{-\theta^*(w-d)}. \end{aligned}$$

Moreover, if $w \geq d$, we can get another simple bound as follows:

$$\begin{aligned} \mathbb{P}(S_d > w) &= \textstyle\sum_{k=1}^{\infty} \mathbb{P}(S_d(k) > w)\mathbb{P}(N_d = k) \\ &= \textstyle\sum_{k=1}^{\infty} \mathbb{P}(Z_d(k-1) > w - d) \frac{\overline{N}_d{}^k e^{-\overline{N}_d}}{k!} \\ &= \textstyle\sum_{k=2}^{\infty} \mathbb{P}(Z_d(k-1) > w - d) \frac{\overline{N}_d{}^k e^{-\overline{N}_d}}{k!} \\ &\leq \textstyle\sum_{k=2}^{\infty} \frac{\overline{N}_d{}^k e^{-\overline{N}_d}}{k!} \\ &= 1 - e^{-\overline{N}_d} - \overline{N}_d e^{-\overline{N}_d} = C_3. \end{aligned}$$

Therefore

$$\mathbb{P}(S_d > w) \leq \begin{cases} \min\left\{C_2 e^{-\theta^*(w-d)}, C_3\right\} & w \geq d \\ C_1 & 0 \leq w < d \end{cases}$$

Appendix C
Proof of Corollary 1

We compute the average bandwidth requested by downloading users:

$$\begin{aligned} \overline{S}_d &= \int_0^{\infty} \mathbb{P}(S_d > w)\, \mathrm{d}w \\ &\leq \int_0^{d} C_1\, \mathrm{d}w + \int_d^{\infty} \min\left\{C_2 e^{-\theta^*(w-d)}, C_3\right\}\, \mathrm{d}w, \end{aligned}$$

where the last inequality follows from Theorem 1. The quantity $\min\left\{C_2 e^{-\theta^*(w-d)}, C_3\right\}$ is equal to C_3 if $w < w^* = (1/\theta^*)\log\left(\frac{C_2}{C_3}\right) + d$. Thus, if $w^* > d$ we have:

$$\begin{aligned} \overline{S}_d &\leq \int_0^{d} C_1\, \mathrm{d}w + \int_d^{w^*} C_3\, \mathrm{d}w + \int_{w^*}^{\infty} C_2 e^{-\theta^*(w-d)}\, \mathrm{d}w \\ &= C_1 d + C_3(w^* - d) + C_2 e^{-\theta^*(w^*-d)}/\theta^* \\ &= C_1 d + C_3(w^* - d) + C_3/\theta^* \end{aligned}$$

where the last equality comes from the fact that $C_3 = C_2 e^{-\theta^*(w^*-d)}$ by the way we defined w^*.

On the other hand, if $w^* \leq d$, we have:

$$\overline{S}_d \leq \int_0^{d} C_1\, \mathrm{d}w + \int_d^{\infty} C_2 e^{-\theta^*(w-d)}\, \mathrm{d}w = C_1 d + C_2/\theta^*.$$

Note that $w^* > d$ if, and only if, $C_2 > C_3$. Thus, we have:

$$\overline{S}_d \leq \begin{cases} C_1 d + C_3(w^* - d) + C_3/\theta^* & \text{if } C_2 > C_3 \\ C_1 d + C_2/\theta^* & \text{o.w.} \end{cases}$$

Appendix D
Proof of Theorem 2

We compute the average bandwidth requested from the servers ($\overline{S}$). For every $x > 0$ we have:

$$\begin{aligned}\mathbb{P}(S>x) &= \mathbb{P}(S_d > S_{\text{seed}}+x)\\ &= \int_0^\infty \mathbb{P}(S_d > w+x|S_{\text{seed}}=w)\,\mathrm{d}F_{S_{\text{seed}}}(w)\\ &\le \int_0^{\max\{0,d-x\}} C_1\,\mathrm{d}F_{S_{\text{seed}}}(w)\\ &\quad +\int_{\max\{0,d-x\}}^\infty \mathrm{d}F_{S_{\text{seed}}}(w)\\ &\quad \min\{C_3, C_2 e^{-\theta^*(w-d+x)}\}\\ &\le \int_0^{\max\{0,d-x\}} C_1\,\mathrm{d}F_{S_{\text{seed}}}(w)\\ &\quad +\int_0^\infty \mathrm{d}F_{S_{\text{seed}}}(w) C_2 e^{-\theta^*(w-d+x)}\\ &= C_1 F_{S_{\text{seed}}}(\max\{0,d-x\})\\ &\quad +C_2\mathbb{E}[e^{-\theta^* S_{\text{seed}}}]e^{-\theta^*(x-d)}\\ &= C_1 F_{S_{\text{seed}}}(d-x)+C_2\mathbb{E}[e^{-\theta^* S_{\text{seed}}}]e^{-\theta^*(x-d)}\end{aligned}$$

Note that the quantity $F_{S_{\text{seed}}}(\max\{0,d-x\})$ is always equal to $F_{S_{\text{seed}}}(d-x)$: indeed if $d-x\le 0$, then $F_{S_{\text{seed}}}(d-x)=F_{S_{\text{seed}}}(0)=0$, since S_{seed} is a positive random variable. Finally,

$$\begin{aligned}\overline{S} &= \int_0^\infty \mathbb{P}(S>x)\,\mathrm{d}x\\ &\le \int_0^\infty C_1 F_{S_{\text{seed}}}(d-x)\,\mathrm{d}x + \int_0^\infty C_2\mathbb{E}[e^{-\theta^* S_{\text{seed}}}]e^{-\theta^*(x-d)}\,\mathrm{d}x\\ &= \int_{-\infty}^d C_1 F_{S_{\text{seed}}}(y)\,\mathrm{d}y + \int_0^\infty C_2\mathbb{E}[e^{-\theta^* S_{\text{seed}}}]e^{-\theta^*(x-d)}\,\mathrm{d}x\\ &\le C_1 F_{S_{\text{seed}}}(d)d + C_2 e^{\theta^* d}\mathbb{E}[e^{-\theta^* S_{\text{seed}}}]1/\theta^*\end{aligned} \tag{26}$$

Observe that, if $C_2 \gg C_3$, *i.e.*, $w^* \gg d$, the above bound becomes weak. Thus, we obtain a tighter bound in this case using a different approach:

$$\begin{aligned}\overline{S} &= \int_0^\infty \int_{[0,\infty)} P(S_d > w+x)F_{S_{\text{seed}}}(\mathrm{d}w)\mathrm{d}x\\ &= \int_{[0,\infty)} F_{S_{\text{seed}}}(\mathrm{d}w)\int_w^\infty P(S_d>z)\mathrm{d}z\\ &= \int_{[0,d]} F_{S_{\text{seed}}}(\mathrm{d}w)\int_w^\infty P(S_d>z)\mathrm{d}z\\ &\quad +\int_{(d,\infty)} F_{S_{\text{seed}}}(\mathrm{d}w)\int_w^\infty P(S_d>z)\mathrm{d}z\\ &= \int_{[0,d]} F_{S_{\text{seed}}}(\mathrm{d}w)\left[\int_w^d P(S_d>z)\mathrm{d}z + \int_d^\infty P(S_d>z)\mathrm{d}z\right]\\ &\quad +\int_{(d,\infty)} F_{S_{\text{seed}}}(\mathrm{d}w)\int_w^\infty P(S_d>z)\mathrm{d}z\\ &\le \int_{[0,d]} F_{S_{\text{seed}}}(\mathrm{d}w)\left[C_1(d-w)+\int_d^\infty \min\{C_3, C_2\mathrm{e}^{-\theta^*(z-d)}\}\mathrm{d}z\right]\\ &\quad +\int_{(d,\infty)} F_{S_{\text{seed}}}(\mathrm{d}w)\int_w^\infty \min\{C_3, C_2\mathrm{e}^{-\theta^*(z-d)}\}\mathrm{d}z\\ &= C_1 F_{S_{\text{seed}}}(d)d - C_1\int_{[0,d]} wF_{S_{\text{seed}}}(\mathrm{d}w)\\ &\quad +F_{S_{\text{seed}}}(d)\int_d^\infty \min\{C_3, C_2\mathrm{e}^{-\theta^*(z-d)}\}\mathrm{d}z\\ &\quad +\int_{(d,\infty)} F_{S_{\text{seed}}}(\mathrm{d}w)\int_w^\infty \min\{C_3, C_2\mathrm{e}^{-\theta^*(z-d)}\}\mathrm{d}z\\ &= C_1 F_{S_{\text{seed}}}(d)d - C_1\int_{[0,d]} wF_{S_{\text{seed}}}(\mathrm{d}w)\\ &\quad +F_{S_{\text{seed}}}(d)\int_d^{w^*} \min\{C_3, C_2\mathrm{e}^{-\theta^*(z-d)}\}\mathrm{d}z\\ &\quad +F_{S_{\text{seed}}}(d)\int_{w^*}^\infty \min\{C_3, C_2\mathrm{e}^{-\theta^*(z-d)}\}\mathrm{d}z\\ &\quad +\int_{(d,w^*]} F_{S_{\text{seed}}}(\mathrm{d}w)\int_w^\infty \min\{C_3, C_2\mathrm{e}^{-\theta^*(z-d)}\}\mathrm{d}z\\ &\quad +\int_{(w^*,\infty)} F_{S_{\text{seed}}}(\mathrm{d}w)\int_w^\infty \min\{C_3, C_2\mathrm{e}^{-\theta^*(z-d)}\}\mathrm{d}z\\ &= C_1 F_{S_{\text{seed}}}(d)d - C_1\int_{[0,d]} wF_{S_{\text{seed}}}(\mathrm{d}w)\\ &\quad +C_3(w^*-d)F_{S_{\text{seed}}}(d) + C_2\frac{\mathrm{e}^{-\theta^*(w^*-d)}}{\theta^*}F_{S_{\text{seed}}}(d)\\ &\quad +\int_{(d,w^*]} F_{S_{\text{seed}}}(\mathrm{d}w)\int_w^\infty \min\{C_3, C_2\mathrm{e}^{-\theta^*(z-d)}\}\mathrm{d}z\\ &\quad +C_2\frac{\mathrm{e}^{\theta^* d}}{\theta^*}\int_{(w^*,\infty)} \mathrm{e}^{-\theta^* w}F_{S_{\text{seed}}}(\mathrm{d}w)\end{aligned} \tag{27}$$

Note that, if $w<w^*$:

$$\begin{aligned}&\int_w^\infty \min\{C_3, C_2\mathrm{e}^{-\theta^*(z-d)}\}\mathrm{d}z\\ &= \int_w^{w^*} \min\{C_3, C_2\mathrm{e}^{-\theta^*(z-d)}\}\mathrm{d}z\\ &+\int_{w^*}^\infty \min\{C_3, C_2\mathrm{e}^{-\theta^*(z-d)}\}\mathrm{d}z\\ &= C_3(w^*-w)+C_2\frac{\mathrm{e}^{-\theta^*(w^*-d)}}{\theta^*}\end{aligned} \tag{28}$$

Combining (27) e (28) we obtain:

$$\begin{aligned}\overline{S} &\le C_1 F_{S_{\text{seed}}}(d)d - C_1\int_{[0,d]} wF_{S_{\text{seed}}}(\mathrm{d}w) - C_3 F_{S_{\text{seed}}}(d)d +\\ &\quad +C_2\frac{\mathrm{e}^{-\theta^*(w^*-d)}}{\theta^*}F_{S_{\text{seed}}}(w^*) + C_3 w^* F_{S_{\text{seed}}}(w^*)\\ &\quad -C_3\int_{(d,w^*]} wF_{S_{\text{seed}}}(\mathrm{d}w) + C_2\frac{\mathrm{e}^{\theta^* d}}{\theta^*}\int_{(w^*,\infty)}\mathrm{e}^{-\theta^* w}F_{S_{\text{seed}}}(\mathrm{d}w)\end{aligned} \tag{29}$$

Using the bound $\int_0^d(1-F_{S_{\text{seed}}}(w))\,\mathrm{d}w \le d$ and the Chernoff bound $(1-F_{S_{\text{seed}}}(w)) \le \mathbb{E}[e^{\theta^* S_{\text{seed}}}]e^{-\theta^* w}$, the integral $-C_3\int_{(d,w^*]} wF_{S_{\text{seed}}}(\mathrm{d}w)$ in (29) becomes:

$$\begin{aligned}&-C_3\int_{(d,w^*]} wF_{S_{\text{seed}}}(\mathrm{d}w) =\\ &= -C_3\left(w(F_{S_{\text{seed}}}(w)-1)|_{d^-}^{w^*} + \int_d^{w^*}(1-F_{S_{\text{seed}}}(w))\,\mathrm{d}w\right)\\ &= -C_3\Bigg(w^*(F_{S_{\text{seed}}}(w^*)-1) - (F_{S_{\text{seed}}}(d^-)-1)\cdot d\\ &+\int_0^\infty(1-F_{S_{\text{seed}}}(w))\,\mathrm{d}w - \int_0^d(1-F_{S_{\text{seed}}}(w))\,\mathrm{d}w\\ &-\int_{w^*}^\infty(1-F_{S_{\text{seed}}}(w))\,\mathrm{d}w\Bigg)\\ &\le C_3\Bigg(w^* - w^*F_{S_{\text{seed}}}(w^*) + F_{S_{\text{seed}}}(d)d - d - \overline{S}_{\text{seed}}\\ &+d+\int_{w^*}^\infty(1-F_{S_{\text{seed}}}(w))\,\mathrm{d}w\Bigg)\\ &\le C_3\Bigg(w^* - w^*F_{S_{\text{seed}}}(w^*) + F_{S_{\text{seed}}}(d)d - \overline{S}_{\text{seed}}\\ &+\mathbb{E}[e^{\theta^* S_{\text{seed}}}]\int_{w^*}^\infty e^{-\theta^* w}\,\mathrm{d}w\Bigg)\\ &= C_3\Bigg(w^* - w^*F_{S_{\text{seed}}}(w^*) + F_{S_{\text{seed}}}(d)d - \overline{S}_{\text{seed}}\\ &+\mathbb{E}[e^{\theta^* S_{\text{seed}}}]e^{-\theta^* w^*}/\theta^*\Bigg)\end{aligned}$$

Noting that in (29) $-C_1\int_0^d wF_{S_{\text{seed}}}(\,\mathrm{d}w) \le 0$, and that by definition of w^*, $C_2 e^{-\theta^*(w^*-d)} = C_3$, we have:

$$\begin{aligned}\overline{S} &\le (C_1-C_3)F_{S_{\text{seed}}}(d)d + C_3 F_{S_{\text{seed}}}(w^*)/\theta^*\\ &\quad +C_3\left(w^* + F_{S_{\text{seed}}}(d)d - \overline{S}_{\text{seed}} + \mathbb{E}[e^{\theta^* S_{\text{seed}}}]\frac{e^{-\theta^* w^*}}{\theta^*}\right)\\ &\quad +C_2\frac{\mathrm{e}^{\theta^* d}}{\theta^*}\int_{(w^*,\infty)}\mathrm{e}^{-\theta^* w}F_{S_{\text{seed}}}(\mathrm{d}w)\\ &\le C_1 F_{S_{\text{seed}}}(d)d + C_3\left(\tfrac{1}{\theta^*} + w^* - \overline{S}_{\text{seed}} + \mathbb{E}[e^{\theta^* S_{\text{seed}}}]\frac{e^{-\theta^* w^*}}{\theta^*}\right)\\ &\quad +C_3\frac{\mathrm{e}^{\theta^* w^*}}{\theta^*}(\mathrm{e}^{-\theta^* w}F_{S_{\text{seed}}}(w)|_{w^{*-}}^\infty + \theta^*\int_{w^*}^\infty \mathrm{e}^{-\theta^* w}F_{S_{\text{seed}}}(w)\mathrm{d}w)\\ &\le C_1 F_{S_{\text{seed}}}(d)d + C_3\left(\tfrac{1}{\theta^*} + w^* - \overline{S}_{\text{seed}} + \mathbb{E}[e^{\theta^* S_{\text{seed}}}]\frac{e^{-\theta^* w^*}}{\theta^*}\right)\\ &\quad +C_3\frac{\mathrm{e}^{\theta^* w^*}}{\theta^*}(-e^{-\theta^* w^*}F_{S_{\text{seed}}}(w^{*-}) + \theta^*\int_{(w^*,\infty)}\mathrm{e}^{-\theta^* w}\mathrm{d}w)\\ &= C_1 F_{S_{\text{seed}}}(d)d + C_3\Bigg(\tfrac{1}{\theta^*} + w^* - \overline{S}_{\text{seed}} + \mathbb{E}[e^{\theta^* S_{\text{seed}}}]\frac{e^{-\theta^* w^*}}{\theta^*}\\ &\quad -F_{S_{\text{seed}}}(w^{*-})/\theta^* + \tfrac{1}{\theta^*}\Bigg)\\ &\le C_1 F_{S_{\text{seed}}}(d)d + C_3\left(\tfrac{2}{\theta^*} + w^* - \overline{S}_{\text{seed}} + \mathbb{E}[e^{\theta^* S_{\text{seed}}}]\frac{e^{-\theta^* w^*}}{\theta^*}\right)\end{aligned}$$

Appendix E
Derivation of the exact solution in (22)

In this section we derive the solution of (21) for the case in which the bandwidths U are exponentially distributed. The

same approach can be extended to the case in which the bandwidth U has a general phase-type distribution.

When the bandwidth is exponentially distributed we have: $\mathrm{d}F_Z(\alpha \mid 1) = \frac{1}{\overline{U}} e^{-\frac{d-\alpha}{\overline{U}}} \mathbb{I}_{\alpha \le d}\, \mathrm{d}\alpha$. Thus, from (20), we have

$$\begin{aligned} F_Z(w \mid k) &= \int_{-\infty}^{w} F_Z(w-\alpha \mid k-1) \tfrac{1}{\overline{U}} e^{-\frac{d-\alpha}{\overline{U}}} \mathbb{I}_{\alpha \le d}\, \mathrm{d}\alpha \\ &= \int_{-\infty}^{\min\{w,d\}} F_Z(w-\alpha \mid k-1) \tfrac{1}{\overline{U}} e^{-\frac{d-\alpha}{\overline{U}}}\, \mathrm{d}\alpha \end{aligned} \tag{30}$$

If $w < d$, making the substitution $y = w - \alpha$ in the integral, we obtain:

$$F_Z(w \mid k) = \frac{e^{\frac{w-d}{\overline{U}}}}{\overline{U}} \int_0^{\infty} F_Z(y \mid k-1) e^{-\frac{y}{\overline{U}}}\, \mathrm{d}y.$$

Note that the integrand function does not depends on w, thus the whole integral can be regarded as a constant. Moreover, notice that

$$\frac{1}{\overline{U}} \int_0^{\infty} F_Z(y \mid k-1) e^{-\frac{y}{\overline{U}}}\, \mathrm{d}y = F_Z(d|k). \tag{31}$$

Thus, it holds:

$$F_Z(w \mid k) = e^{\frac{w-d}{\overline{U}}} F_Z(d \mid k) \qquad \forall w \le d \tag{32}$$

Now consider equation (30) when $d \le w \le 2d$:

$$\begin{aligned} F_Z(w \mid k) &= \frac{e^{\frac{w-d}{\overline{U}}}}{\overline{U}} \int_{w-d}^{\infty} F_Z(y \mid k-1) e^{-\frac{y}{\overline{U}}}\, \mathrm{d}y \\ &= \frac{e^{\frac{w-d}{\overline{U}}}}{\overline{U}} \int_0^{\infty} F_Z(y \mid k-1) e^{-\frac{y}{\overline{U}}}\, \mathrm{d}y \\ &\quad - \frac{e^{\frac{w-d}{\overline{U}}}}{\overline{U}} \int_0^{w-d} F_Z(y \mid k-1) e^{-\frac{y}{\overline{U}}}\, \mathrm{d}y. \end{aligned} \tag{33}$$

Note that the first term in the sum is equal to $e^{\frac{w-d}{\overline{U}}} F_Z(d \mid k)$. For the second integral, since $d \le w \le 2d$, variable y is such that $0 < y < w - d < d$. Thus, using (31) we can write:

$$\begin{aligned} &\frac{e^{\frac{w-d}{\overline{U}}}}{\overline{U}} \int_0^{w-d} F_Z(y \mid k-1) e^{-\frac{y}{\overline{U}}}\, \mathrm{d}y \\ &= \frac{e^{\frac{w-d}{\overline{U}}}}{\overline{U}} \int_0^{w-d} e^{\frac{y-d}{\overline{U}}} F_Z(d \mid k-1) e^{-\frac{y}{\overline{U}}}\, \mathrm{d}y \\ &= \frac{(w-d)}{\overline{U}} e^{\frac{w-2d}{\overline{U}}} F_Z(d \mid k-1). \end{aligned} \tag{34}$$

Thus, for $d \le w \le 2d$, we obtain

$$F_Z(w \mid k) = e^{\frac{w-d}{\overline{U}}} F_Z(d \mid k) - \frac{(w-d)}{\overline{U}} e^{\frac{w-2d}{\overline{U}}} F_Z(d \mid k-1).$$

Now considering $2d \le w \le 3d$, we can still use (33) to express $F_Z(w \mid k)$ in terms of $F_Z(y \mid k-1)$ over a domain in which $y \le w - d \le 2d$. Again we know explicitly the expression of $F_Z(w \mid k)$ over the considered domain in terms of the two constants $F_Z(d \mid k-1)$ and $F_Z(d \mid k-2)$. It turns out:

$$\begin{aligned} F_Z(w \mid k) &= e^{\frac{w-d}{\overline{U}}} F_Z(d \mid k) - \tfrac{(w-d)}{\overline{U}} e^{\frac{w-2d}{\overline{U}}} F_Z(d \mid k-1) \\ &\quad + \tfrac{(w-2d)^2}{2\overline{U}^2} e^{\frac{w-3d}{\overline{U}}} F_Z(d \mid k-2) \quad 2d \le w \le 3d \end{aligned} \tag{35}$$

Proceeding in a similar way we can express $F_Z(w \mid k)$ for any $w \le kd$ in terms of the constants $F_Z(d \mid 1) \ldots F_Z(d \mid k)$, while for $w > kd$ we have trivially $F_Z(w \mid k) = 1$.

The constants $F_Z(d \mid k)$ can be obtaining forcing $F_Z(kd \mid k) = 1$. Indeed, by imposing $F_Z(w \mid 1) \mid_{w=d} = 1$ we immediately obtain $F_Z(d \mid 1) = 1$. Imposing $F_Z(w \mid 2) \mid_{w=2d} = 1$ we obtain an algebraic linear equation between $F_Z(d \mid 2)$ and $F_Z(d \mid 1)$, from which we can derive $F_Z(d \mid 2)$. In general imposing $F_Z(w \mid k) \mid_{w=kd} = 1$ we obtain a linear algebraic equation containing all constant $F_Z(d \mid i)$ with $i \le k$. This equation can be exploited to derive $F_Z(d \mid k)$ as function of $F_Z(d \mid i)$ with $i < k$.

Appendix F
Proof of Theorem 4

By virtue of Theorem 2, we have

$$\overline{S} \le C_1 F_{S_{\text{seed}}}(d) d + C_2 e^{\theta^* d} \mathbb{E}[e^{-\theta^* S_{\text{seed}}}]/\theta^* \triangleq \overline{S}_{\text{up},1}. \tag{36}$$

Moreover, the first term in the sum above goes to zero, as $\lambda \to \infty$. Indeed, since $\lim_{\lambda\to\infty} \overline{N}_d = \infty$, we have $C_1 \to 1$ as $\lambda \to \infty$, and the claim follows noticing that $F_{S_{\text{seed}}}(d)$ tends to zero, as $\lambda \to \infty$ (*i.e.*, the mean number of seeds $\overline{N}_{\text{seed}}$ and their offered bandwidth tend to infinity).

We first consider the case $\gamma_p < 1$. If $d \le \overline{U}$, by Theorem 1, ϵ may be freely chosen in the interval $[\overline{U} - d, \overline{U})$. We set $\epsilon = \overline{U} - d$, obtaining $A = 0$. For the second term in the sum (36), note that due to $A = 0$ we have $C_2 = e^{-\overline{N}_d}(e^{\overline{N}_d} - \overline{N}_d - 1) = 1 - e^{-\overline{N}_d}\overline{N}_d - e^{-\overline{N}_d}$. So, using again $\lim_{\lambda\to\infty} \overline{N}_d = \infty$, we deduce that $C_2 \to 1$ as $\lambda \to \infty$. Combining this with the relations:

$$\mathbb{E}[e^{-\theta^* S_{\text{seed}}}] = e^{\lambda \overline{T}_{\text{seed}}(\phi_U(-\theta^*)-1)}$$

and $\phi_U(-\theta^*) - 1 < 0$ (this latter inequality holds since $\phi_U(-\theta^*) = \mathbb{E}[e^{-\theta^* U}] < 1$), we easily have that even the second term in the sum (36) tends to zero as $\lambda \to \infty$. Consequently, for $\gamma_p < 1$ and $d \le \overline{U}$, we get $\lim_{\lambda\to\infty} \overline{S}_{\text{up},1} = \lim_{\lambda\to\infty} \overline{S} = 0$. Now, suppose $d > \overline{U}$. Since U is not constant the equation in z: $e^{-z\epsilon}\mathbb{E}[e^{z(\overline{U}-U_1)}] = 1$ has a unique solution, say $\theta^*(\epsilon)$. The properties of the function $\epsilon \mapsto \theta^*(\epsilon)$ are given in Proposition 1 below. For λ large, consider the sequence $\{\epsilon_\lambda\} \subset (0, \overline{U})$ defined by

$$\epsilon_\lambda := (\theta^*)^{-1}(\lambda^{-1/2}).$$

Note that by Proposition 1 $\epsilon_\lambda \to 0$ and $\theta(\epsilon_\lambda) \to 0$, as $\lambda \to \infty$. Furthermore $\lambda\theta^*(\epsilon_\lambda) \to \infty$ as $\lambda \to \infty$. We neglect again the first term in (36), and we obtain[2] $\overline{S}_{\text{up},1} \sim C_2 e^{\theta^* d}\mathbb{E}[e^{-\theta^* S_{\text{seed}}}]/\theta^* = C_2 e^{\theta^* d} e^{\lambda \overline{T}_{\text{seed}}(\phi_U(-\theta^*)-1)}/\theta^*$. We can say that $\overline{S}_{\text{up},1} \to 0$ if and only if $\log \overline{S}_{\text{up},1} \to -\infty$. Thus, we consider

$$\log \overline{S}_{\text{up},1} \sim \log C_2 - \log\theta^* + \theta^* d + \lambda \overline{T}_{\text{seed}}(\phi_U(-\theta^*) - 1).$$

Using the Taylor expansion (and neglecting the term $e^{-\theta^* A}$ that tends to 1 as $\lambda \to \infty$), we obtain that

$$\log C_2 \le -\theta^* A + \overline{N}_d(e^{\theta^* A} - 1)$$

as $\lambda \to \infty$. Using the Taylor expansion as $\lambda \to \infty$, and the choice of ϵ_λ above (thus $\theta^* A \to 0$ and $\theta^* \ge \lambda^{-1/2}$), we have

[2]With abuse of notation we will use the expression $f \sim g$ to indicate that $f \in \Theta(g)$, *i.e.*, f is bounded both above and below by g asymptotically

$$\begin{aligned}\log \overline{S}_{\text{up,1}} &\leq -\theta^* A + \overline{N}_d(e^{\theta^* A}-1) - \log\theta^* \\ &\quad +\theta^* d + \overline{N}_{\text{seed}}(\phi_U(-\theta^*)-1) \\ &\leq \lambda\overline{T}_d(\theta^* A + o(\theta^* A)) - \log(\theta^*) \\ &\quad +\lambda\overline{T}_{\text{seed}}(-\overline{U}\theta^* + o(\theta^*)) \\ &\sim \lambda\overline{T}_d\theta^* A - \log(\theta^*) - \lambda\overline{T}_{\text{seed}}\overline{U}\theta^* \\ &\leq \lambda\theta^*(\overline{T}_d A - \overline{T}_{\text{seed}}\overline{U}) - 1/2\log\lambda \to -\infty\end{aligned}$$

since in the regime $\gamma_p = \frac{\overline{T}_d d}{\overline{U}\overline{T}} < 1$, the quantity $\overline{T}_d A - \overline{T}_{\text{seed}}\overline{U}$ is negative. Therefore, the theorem follows.

We now consider the case $\gamma_p > 1$. By Theorem 2 we have that

$$\begin{aligned}\overline{S} &\leq C_1 F_{S_{\text{seed}}}(d)d + C_3\Bigg(\tfrac{2}{\theta^*} + w^* - \overline{S}_{\text{seed}} \\ &\quad +\mathbb{E}[e^{\theta^* S_{\text{seed}}}]e^{-\theta^* w^*}/\theta^*\Bigg) \triangleq \overline{S}_{\text{up,2}}.\end{aligned} \tag{37}$$

In (37) we can neglect the term $C_1 F_{S_{\text{seed}}}(d)d \sim 1$. Note that as $\lambda \to \infty$, $C_3 \sim 1$. Thus, we obtain:

$$\overline{S}_{\text{up,2}} = \tfrac{2}{\theta^*} + w^* - \overline{S}_{\text{seed}} + \mathbb{E}[e^{\theta^* S_{\text{seed}}}]e^{-\theta^* w^*}/\theta^*$$

The quantity w^*, as $\lambda \to \infty$ becomes $w^* = (1/\theta^*)\log(C_2/C_3) + d \sim \overline{N}_d A \sim \lambda\overline{T}_d(d-\overline{U})$. Therefore, as $\lambda \to \infty$, it is easy to prove that $\log\left(\mathbb{E}[e^{\theta^* S_{\text{seed}}}]e^{-\theta^* w^*}/\theta^*\right) \to -\infty$; as before, we can conclude that $\mathbb{E}[e^{\theta^* S_{\text{seed}}}]e^{-\theta^* w^*}/\theta^* \to 0$.

Finally, we obtain, for $\lambda \to \infty$,

$$\begin{aligned}\overline{S}_{\text{up,2}} &\sim -\overline{S}_{\text{seed}} + w^* + 1/\theta^* \\ &\sim \lambda(-\overline{T}_{\text{seed}}\overline{U} + \overline{T}_d(d-\overline{U})) + \lambda^{1/2} \\ &\sim \lambda\left(\overline{T}_d(d-\overline{U}) - (\overline{T}-\overline{T}_d)\overline{U}\right) \\ &= \lambda(\overline{T}_d d - \overline{U}\ \overline{T}),\end{aligned}$$

or, equivalently:

$$\lim_{\lambda\to\infty} \frac{\overline{S}_{\text{up,2}}}{(\overline{N}_d d - \overline{U}\ \overline{N})} = 1.$$

Note that the quantity $\overline{N}_d d - \overline{U}\ \overline{N}$ is a lower bound for $\overline{S}$, as described in (9). Therefore, necessarily $\liminf_{\lambda\to\infty} \frac{\overline{S}}{(\overline{N}_d d - \overline{U}\ \overline{N})} \geq 1$. Recalling that $\overline{S} \leq \overline{S}_{\text{up,2}}$, we obtain,

$$1 \leq \lim_{\lambda\to\infty} \frac{\overline{S}}{(\overline{N}_d d - \overline{U}\ \overline{N})} \leq \lim_{\lambda\to\infty} \frac{\overline{S}_{\text{up,2}}}{(\overline{N}_d d - \overline{U}\ \overline{N})} = 1$$

and the theorem follows.

Proposition 1: If $d > \overline{U}$, then the equation in z $\mathbb{E}[e^{z(\overline{U}-U_1-\epsilon)}] = 1$ admits a unique solution for $\epsilon \in (0, \overline{U})$. Furthermore, $\theta^*(\epsilon) = \arg_{z>0}(e^{-z\epsilon}\mathbb{E}[e^{z(\overline{U}-U_1)}] = 1)$ is strictly increasing and C^1 on the interval $(0, \overline{U})$. Finally, it holds $\lim_{\epsilon\to 0}\theta^*(\epsilon) = 0$.

Proof: We define the function $f(z,\epsilon) = \mathbb{E}[e^{z(\overline{U}-U_1-\epsilon)}]$. Observe that $f(z,\epsilon)$ is analytic in the domain $z \geq 0$ and $\epsilon \geq 0$, as immediate consequence of the fact that $\overline{U} - U_1 \leq \overline{U} < \infty$.

Observe also that i) $f(0,\epsilon) = 1$ and $f'(0,\epsilon) = -\epsilon < 0$ for any $\epsilon > 0$; ii) $f(z,\epsilon)$ is convex in z, since $\frac{\partial^2 f(z,\epsilon)}{\partial z^2} = \mathbb{E}[(\overline{U}-U_1-\epsilon)^2 e^{z(\overline{U}-U_1-\epsilon)}] > 0$; iii) $\lim_{z\to\infty} f(z,\epsilon) = \infty$ for any $\epsilon < \overline{U}$. This because, for all $c \in \mathbb{R}$,

$$\begin{aligned}f(z,\epsilon) &= \int_{-\infty}^{\infty} e^{z(\overline{U}-w-\epsilon)}dF_U(w) \\ &\geq \int_{-\infty}^{c} e^{z(\overline{U}-w-\epsilon)}dF_U(w) \geq e^{z(\overline{U}-c-\epsilon)}\Pr(U_1 < c).\end{aligned}$$

Since $\epsilon < \overline{U}$, there exists $a > 1$ such that $(\overline{U}-\epsilon)/a > 0$. Defining $c = (\overline{U}-\epsilon)/a$, we have $e^{z(\overline{U}-c-\epsilon)} \to \infty$ while $\Pr(U_1 < c) > 0$.

As a consequence of i) , ii) and iii), recalling that $f(z,\epsilon)$ is continuous w.r.t. z for any $\epsilon \geq 0$ and $z \geq 0$ there is a unique solution $\theta^*(\epsilon) = \arg_{z\geq 0}(e^{-z\epsilon}\mathbb{E}[e^{z(\overline{U}-U_1)}] = 1)$.

The regularity of $\theta^*(\epsilon)$ with respect to ϵ immediately follows by the implicit function theorem. At last the monotonicity of $\theta^*(\epsilon)$ can be derived again from the implicit function theorem, according to which $\frac{d\theta^*(\epsilon)}{d\epsilon} = -\frac{\frac{\partial f(\theta^*(\epsilon),\epsilon)}{\partial\epsilon}}{\frac{\partial f(z,\epsilon)}{\partial z}|_{z=\theta^*(\epsilon)}}$. Note indeed that $\frac{\partial f(\theta^*(\epsilon),\epsilon)}{\partial\epsilon} < -\theta^*(\epsilon)f(\theta^*(\epsilon),\epsilon) < 0\ \forall\epsilon > 0$, while $\frac{\partial f(z,\epsilon)}{\partial z}\ |_{z=\theta^*(\epsilon)} > 0$ by construction, since $f(z,\epsilon)$ is convex w.r.t. z and $f(z,\epsilon) < 1$ for $0 < z < \theta^*(\epsilon)$ and $f(z,\epsilon) > 1$ for $z > \theta^*(\epsilon)$.

At last, it is immediate to see that also for $\epsilon \to 0$ $\theta^*(\epsilon) \to 0$, in light of the fact that $f(0,0) = 1$, and $f(z,0) > 1$ for $z > 0$. ■

As immediate consequence of the fact that $\theta^*(\epsilon)$ is strictly increasing (and thus invertible) and continuous over the domain $(0, \overline{U})$ with $\lim_{\epsilon\to 0}\theta^*(\epsilon) = 0$, we have that the following proposition holds.

Proposition 2: Provided that $d > \overline{U}$, and U not constant, the image of $\theta^*(\epsilon)$ for $0 < \epsilon < \overline{U}$ is the open interval $(0, \delta)$, with $\delta = \lim_{\epsilon\to\overline{U}}\theta^*(\epsilon)$.

APPENDIX G
PROOF OF THEOREM 5

We first focus on the case $d < \overline{U}$. Using the bound in (17):

$$\overline{S}(t) \leq C_1(t)F_{S_{\text{seed}}}(d)d + C_2(t)\frac{e^{\theta^* d}}{\theta^*}\mathbb{E}[e^{-\theta^* S_{\text{seed}}}] \tag{38}$$

with $A = d - \overline{U} + \epsilon$, $C_1(t) = 1 - e^{-\overline{N}_d(t)}$, $C_2(t) = e^{-\theta^* A}e^{-\overline{N}_d(t)}(e^{\overline{N}_d(t)e^{\theta^* A}} - \overline{N}_d(t)e^{\theta^* A} - 1)$.

Since $d < \overline{U}$, from Corollary 1 we can set $A = 0$, and we get

$$\begin{aligned}C_1(t) &\leq \min\{1, \overline{N}_d(t)\} \\ C_2(t) &= 1 - \overline{N}_d(t)e^{-\overline{N}_d(t)} - e^{-\overline{N}_d(t)} \\ &\leq 1 - e^{-\overline{N}_d(t)} \\ &\leq \min\{1, \overline{N}_d(t)\}\end{aligned}$$

By using the Chernoff bound $F_{\text{seed}}(d) \leq \mathbb{E}[e^{-\theta^* S_{\text{seed}}}]e^{\theta^* d}$ in (38), we obtain

$$\begin{aligned}\overline{S}(t) &\leq \min\{1, \overline{N}_d(t)\}(d + 1/\theta^*)\mathbb{E}[e^{-\theta^* S_{\text{seed}}}]e^{\theta^* d} \\ &= \min\{1, \overline{N}_d(t)\}(d + 1/\theta^*)e^{-\overline{N}_{\text{seed}}(t)(1-\phi_U(-\theta^*))}e^{\theta^* d}\end{aligned} \tag{39}$$

Now we can compute an upper bound to the data volume requested from servers over time as:

$$V \leq (d + \tfrac{1}{\theta^*})e^{\theta^* d}\int_0^\infty \min\{1, \overline{N}_d(t)\}e^{-\overline{N}_{\text{seed}}(t)(1-\phi_U(-\theta^*))}\,\mathrm{d}t \tag{40}$$

In order to compute V, we define T_1 as the temporal threshold such that $T_1 \triangleq \sup\{t : \overline{N}_d(t) > 1\}$, where $T_1 \equiv 0$ if $\overline{N}_d(t) < 1$ for all $t \geq 0$. Thus, we compute

$$\begin{aligned} V_1 &\triangleq \int_0^{T_1} \min\{1, \overline{N}_d(t)\} e^{-\overline{N}_{\text{seed}}(t)(1-\phi_U(-\theta^*))}\,\mathrm{d}t \\ &= \int_0^{T_1} e^{-\overline{N}_{\text{seed}}(t)(1-\phi_U(-\theta^*))}\,\mathrm{d}t \\ &\leq T_1 \end{aligned} \tag{41}$$

and

$$\begin{aligned} V_2 &\triangleq \int_{T_1}^{\infty} \min\{1, \overline{N}_d(t)\} e^{-\overline{N}_{\text{seed}}(t)(1-\phi_U(-\theta^*))}\,\mathrm{d}t \\ &= \int_{T_1}^{\infty} \overline{N}_d(t) e^{-\overline{N}_{\text{seed}}(t)(1-\phi_U(-\theta^*))}\,\mathrm{d}t \\ &\leq \int_{T_1}^{\infty} \overline{N}_d(t)\,\mathrm{d}t \end{aligned} \tag{42}$$

We obtain the bound $V = (d + 1/\theta^*)e^{\theta^* d}(V_1 + V_2) \leq (d + 1/\theta^*)e^{\theta^* d}(T_1 + \int_{T_1}^{\infty} \overline{N}_d(t)\,\mathrm{d}t)$.

Recall that $\overline{N}_d(t) \triangleq \int_{(t-\overline{T}_d)^+}^{t} \Lambda q(z) G_T(t-z)\,\mathrm{d}z$. Now, we define $Q(t) = \int_{(t-\overline{T}_d)^+}^{t} G_T(t-z)q(z)\,\mathrm{d}z$, thus $\overline{N}_d(t) = \Lambda Q(t)$. We know that $Q(t) > 0\ \forall t$ and $Q(t) \to 0$ for $t \to \infty$. By the definition of T_1 we know that $\overline{N}_d(T_1) = 1$, thus we have

$$T_1 = T_1(\Lambda) = Q^{-1}\left(\frac{1}{\Lambda}\right). \tag{43}$$

Since $Q(t) \to 0$ as $t \to \infty$, it can be easily shown that $Q^{-1}(1/\Lambda) = T_1(\Lambda) \to \infty$ as $\Lambda \to \infty$.

We first consider the case $0 < q(t) < Kt^{-(\alpha+1)}$ for $t > T_0$, $K > 0$ and $\alpha > 1$. Thus, for all $t > \max\{T_0, T_d\}$ we have:

$$\begin{aligned} \overline{N}_d(t) &\leq \Lambda \int_{(t-\overline{T}_d)^+}^{t} q(z)\,\mathrm{d}z \\ &\leq \Lambda \int_{(t-\overline{T}_d)}^{t} K z^{-(\alpha+1)}\,\mathrm{d}z \\ &\leq \Lambda K t^{-\alpha} \triangleq \overline{N}_d^*(t) \end{aligned} \tag{44}$$

We define the time instant $T_1^* \triangleq \max\{t : \overline{N}_d^*(t) \geq 1\}$; thus, T_1^* is such that $\overline{N}_d^*(T_1^*) = 1$. We obtain $T_1^* \sim \Lambda^{1/\alpha}$. It is easy to see that since for all $t > \max\{T_0, T_d\}$ it holds $\overline{N}_d(t) \leq \overline{N}_d^*(t)$, we have that $T_1(\Lambda) < T_1^*(\Lambda)$. Moreover, since $\Lambda \to \infty$, it exists Λ_0 such that for all $\Lambda > \Lambda_0$, it holds $T_1(\Lambda) > T_0$. We conclude $T_1(\Lambda) > T_0$, and thus $T_1^*(\Lambda) > T_0$, as $\Lambda > \Lambda_0$.

Finally, from (41) and (42) we obtain an upper bound to the average data volume requested from the servers:

$$\begin{aligned} V &= (d + 1/\theta^*)e^{\theta^* d}(V_1 + V_2) \\ &\leq (d + 1/\theta^*)e^{\theta^* d}(T_1 + \textstyle\int_{T_1}^{\infty} \overline{N}_d(t)\,\mathrm{d}t) \\ &\leq (d + 1/\theta^*)e^{\theta^* d}(T_1 + \textstyle\int_{T_1}^{T_1^*} \overline{N}_d(t)\,\mathrm{d}t + \int_{T_1^*}^{\infty} \overline{N}_d(t)\,\mathrm{d}t) \\ &\leq (d + 1/\theta^*)e^{\theta^* d}(T_1 + (T_1^* - T_1) + \Lambda T_1^{*1-\alpha}/(\alpha - 1)) \\ &= (d + 1/\theta^*)e^{\theta^* d}(\Lambda^{1/\alpha} + \Lambda^{1/\alpha}/(\alpha - 1)) \\ &\sim \Lambda^{1/\alpha}, \quad \Lambda \to \infty. \end{aligned} \tag{45}$$

We consider now the case where $q(t)$ has a finite support, *i.e.*, there exists an instant $T_2 \geq T_d$ such that $q(t) = 0$ for all $t > T_2$. We observe that T_2 is a constant as $\Lambda \to \infty$, otherwise $q(t)$ would not have finite support. We conclude that both T_1 and T_2 are constant in Λ.

Therefore, in this case we obtain the bound

$$\begin{aligned} V &= (d + 1/\theta^*)e^{\theta^* d}(V_1 + V_2) \\ &\leq (d + 1/\theta^*)e^{\theta^* d}(T_1 + \textstyle\int_{T_1}^{T_2 + T_d} \overline{N}_d(t)\,\mathrm{d}t) \\ &\leq (d + 1/\theta^*)e^{\theta^* d}(T_1 + (T_2 + T_d - T_1)) \\ &= (d + 1/\theta^*)e^{\theta^* d}T_2 = \Theta(1) \quad \text{when } \Lambda \to \infty. \end{aligned} \tag{46}$$

We now consider a popularity distribution with an exponential decreasing tail. Thus, $\exists\, T_0$ such that for all $t > T_0$, $q(t) \sim e^{-\alpha t}$, $\alpha > 0$. In this case we have

$$\begin{aligned} \overline{N}_d(t) &= \textstyle\int_{(t-\overline{T}_d)^+}^{t} \Lambda e^{-\alpha z}\,\mathrm{d}z \\ &= \Lambda(e^{-\alpha(t-\overline{T}_d)^+} - e^{-\alpha t})/\alpha \\ &\sim \Lambda e^{-\alpha t}/\alpha. \end{aligned} \tag{47}$$

We obtain again that the quantity T_1 is such that $\overline{N}_d(T_1) = 1$, *i.e.*, $T_1(\Lambda) \sim \log \Lambda$.

Finally, we obtain the following upper bound to the average amount of data requested from the servers:

$$\begin{aligned} V_1 &\sim \log \Lambda \\ V_2 &\sim \textstyle\int_{T_1}^{\infty} \Lambda e^{-\alpha t}\,\mathrm{d}t \\ &= \Lambda e^{-\alpha T_1(\Lambda)}/\alpha \\ &\sim \Lambda e^{-\log \Lambda} = 1 \end{aligned} \tag{48}$$

and thus

$$V \leq V_1 + V_2 \sim \log \Lambda \tag{49}$$

Now we consider the case $d > \overline{U}$. In can be easily shown that $V \sim \Lambda$ under both the considered popularity distributions. We first note that, since the average number of users in the system grows asymptotically linearly with Λ, this linear behavior is a trivial upper-bound for the average data volume requested from the servers. We consider the following lower-bound:

$$\overline{S}(t) \geq \left((d - \overline{U})\overline{N}_d(t) - \overline{UN}_{\text{seed}}(t)\right)^+ \tag{50}$$

Integrating the previous lower-bound, and noting that by definition, $\overline{N}_{\text{seed}}(t) = 0$ for $t < \overline{T}_d$, we obtain

$$\begin{aligned} V &\geq \textstyle\int_0^{\infty} \left((d - \overline{U})\overline{N}_d(t) - \overline{UN}_{\text{seed}}(t)\right)^+ \mathrm{d}t \\ &\geq \textstyle\int_0^{\overline{T}_d} \left((d - \overline{U})\overline{N}_d(t) - \overline{UN}_{\text{seed}}(t)\right)^+ \mathrm{d}t \\ &= \textstyle\int_0^{\overline{T}_d} (d - \overline{U})\overline{N}_d(t)\,\mathrm{d}t \\ &= (d - \overline{U}) \textstyle\int_0^{\overline{T}_d} \Lambda Q(t)\,\mathrm{d}t \\ &\sim \Lambda \quad , \ \Lambda \to \infty \end{aligned} \tag{51}$$

The last line follows noting that the integral is a positive quantity that does not depend on Λ and $d > \overline{U}$ in this case. Since we have proven that V is bounded above and below by a quantity that grows asymptotically as Λ, we can conclude that $V = \Theta(\Lambda)$ when $d > \overline{U}$ for any choice of the shaping function $q(t)$ modeling the video popularity.

REFERENCES

[1] D. Ciullo, V. Martina, M. Garetto, E. Leonardi, and G. L. Torrisi, "Stochastic Analysis of Self-Sustainability in Peer-Assisted VoD Systems," in *IEEE INFOCOM*, 2012.

[2] "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2010–2015," White paper published on Cisco web site, 2011.

[3] C. Huang, J. Li, and K. W. Ross, "Can Internet Video-on-Demand Be Profitable?" in *ACM SIGCOMM*, 2007.

[4] Y. Huang, T. Z. J. Fu, D. ming Chiu, J. C. S. Lui, and C. Huang, "Challenges, Design and Analysis of a Large-scale P2P VoD System," in *ACM SIGCOMM*, 2008.
[5] W. Wu, J. Lui, "Exploring the Optimal Replication Strategy in P2P-VoD Systems: Characterization and Evaluation," in *IEEE INFOCOM*, 2011.
[6] S. Borst, V. Gupta, and A. Walid, "Distributed Caching Algorithms for Content Distribution Networks," in *IEEE INFOCOM*, 2010.
[7] K. Suh, C. Diot, J. Kurose, L. Massoulie, C. Neumann, D. Towsley, and M. Varvello, "Push-to-Peer Video-on-Demand system: design and evaluation," *JSAC*, vol. 25, no. 9, pp. 1706–1716, 2007.
[8] B. Tan and L. Massoulie, "Optimal Content Placement for Peer-to-Peer Video-on-Demand Systems," in *IEEE INFOCOM*, 2011.
[9] T. Bonald, L. Massoulie, F. Mathieu, D. Perino, and A. Twigg, "Epidemic Live Streaming: Optimal Performance Trade-Offs," in *ACM SIGMETRICS*, 2008.
[10] S. Liu, R. Zhang-Shen, W. Jiang, J. Rexford, and M. Chiang, "Performance Bounds for Peer-Assisted Live Streaming," in *ACM SIGMETRICS*, 2008.
[11] S. Asmussen, *Applied Probability and Queues. 2nd ed.* Springer-Verlag, New York, 2003.
[12] X. Yang and G. de Veciana, "Service Capacity of Peer to Peer Networks," in *IEEE INFOCOM*, 2004.
[13] D. Qiu and R. Srikant, "Modeling and Performance Analysis of BitTorrent-Like Peer-to-Peer Networks," in *ACM SIGCOMM*, 2004.
[14] N. Parvez, C. Williamson, A. Mahanti, and N. Carlsson, "Analysis of BitTorrent-like Protocols for On-Demand Stored Media Streaming," in *ACM SIGMETRICS*, 2008.
[15] R. Kumar, Y. Liu, , and K. Ross, "Stochastic Fluid Theory for P2P Streaming Systems," in *IEEE INFOCOM*, 2007.
[16] D. Ciullo, V. Martina, M. Garetto, E. Leonardi, and G. L. Torrisi, "Performance Analysis of Non-stationary Peer-assisted VoD Systems," in *IEEE INFOCOM Mini-Conference*, 2012.
[17] B. Fan, D. Andersen, M. Kaminsky, and K. Papagiannaki, "Balancing Throughput, Robustness, and In-Order Delivery in P2P VoD," in *ACM CoNEXT*, 2010.

Delia Ciullo received the Master degree in Telecommunications Engineering and the Ph.D. degree in Electronics and Communications Engineering, both from Politecnico di Torino in 2007 and 2011, respectively. In 2009, she has been a visiting student at the CNRG group of MIT, under the supervision of Prof. Eytan Modiano. Between 2012 and 2013 she was a post-doc ERCIM fellow at INRIA Sophia Antipolis. She is currently a post-doc researcher at EURECOM Sophia Antipolis, France. Her research interests are in the fields of energy-aware networks, scaling properties in wireless networks, and P2P systems.

Valentina Martina received the Master degree in Mathematical modeling in Engineering and the Ph.D. degree in Electronics and Communication Engineering, both from Politecnico di Torino in 2007 and 2011, respectively. In 2010, she has been a visiting student at the Technicolor Paris Research Lab. She is currently a post-doc at Politecnico di Torino. Her research interests are in the fields of scaling properties in Wireless Networks, mobility models, and P2P systems.

Michele Garetto (M'04) received the Dr.Ing. degree in Telecommunication Engineering and the Ph.D. degree in Electronic and Telecommunication Engineering, both from Politecnico di Torino, Italy, in 2000 and 2004, respectively. In 2002, he was a visiting scholar with the Networks Group of the University of Massachusetts, Amherst, and in 2004 he held a postdoctoral position at the ECE department of Rice University, Houston. He is currently assistant professor at the University of Torino, Italy.

Emilio Leonardi (M'99, SM'09) is an Associate Professor at the Dipartimento di Elettronica of Politecnico di Torino. He received a Dr.Ing degree in Electronics Engineering in 1991 and a Ph.D. in Telecommunications Engineering in 1995 both from Politecnico di Torino. In 1995, he visited the Computer Science Department of UCLA, Los Angeles; in 1999 he joined the High Speed Networks Research Group, at Bell Laboratories/Lucent Technologies, NJ; in 2001, the Electrical Engineering Department of the Stanford University, and finally in 2003, the IP Group at Sprint, Advanced Technologies Laboratories, CA. His research interests are in the field of performance evaluation of wireless networks, P2P systems, packet switching.

Giovanni Luca Torrisi graduated in Mathematics in 1994 at the University of Rome "La Sapienza" and obtained a Ph.D. in Mathematics at the University of Milan. Since December 2001 he has been a researcher at the CNR. He was at the "Laboratoire des Systemes et Signaux" of CNRS (Gif-sur-Yvette) from January 1998 to May 1998, and from October 1999 to December 1999, hosted by Prof. Pierre Brémaud. In 2003 he was at the Department of Mathematical Sciences of Aalborg University hosted by Prof. Jesper Møller. In 2005 he visited the Ecole Normale Superiéure (Paris), hosted by Dr. Charles Bordenave. In 2006 he visited the Microsoft Research Lab (Cambridge,UK), hosted by Dr. Ayalvadi Ganesh. His research interests are in the field of Probability Theory and Applied Probability.