Exact Sojourn Time Distribution in an Online IPTV Recording System

Tobias Hoßfeld[1], Kenji Leibnitz[2], and Marie-Ange Remiche[3]

[1] University of Würzburg, Institute of Computer Science,
Am Hubland, 97074 Würzburg, Germany
hossfeld@informatik.uni-wuerzburg.de
[2] Osaka University, Graduate School of Information Science and Technology
1-5 Yamadaoka, Suita, Osaka 565-0871, Japan
leibnitz@ist.osaka-u.ac.jp
[3] Université Libre de Bruxelles, CoDE/SMG, CP 165/15,
Av. F.D. Roosevelt, 50; B-1050 Bruxelles, Belgium
mremiche@ulb.ac.be

Abstract. In this paper we analytically derive the sojourn time of a user accessing an online IPTV recording service. Basically, the system consists of a server (or server farm) and the bandwidth at which a user can download recorded files depends on the number of other concurrently downloading users. In particular, we consider both cases, one in which the user's download speed is limited by his own access bandwidth, whereas in the other case the server is the bottleneck and all downloading users share the server's upload capacity. Our model includes user impatience which depends on the actual download speed. A user may abort his download attempt if a threshold duration is exceeded. The file size distribution is obtained by measurements of offered files at an existing server.

1 Introduction

With the recent advancements in access technology for end-users, the Internet is becoming an attractive medium for distributing large volume contents. This is one of the reasons that *Internet-based Television* (IPTV) has gained enormous popularity as a means of delivering high-quality video images, especially since it is capable of offering additional interactive features in contrast to conventional TV, such as chats, rankings, or discussion forums of the shows. Unlike television programs that are broadcast as terrestrial, cable, or satellite signals, IPTV transmits the video data as stream of IP packets and permits the user to ubiquitously access his favorite TV show on-demand, whenever, wherever, and on whatever device (TV, PC, portable player). It is, therefore, not really surprising that a wide range of new free and commercial services are currently being offered over the Internet, e.g. *YouTube*, *Joost*, or *Zattoo* which have become extremely popular. For instance, as of December 2007 the free Japanese service *GyaO* counted about 17.4 million registered subscribers [1], of which about 3-4 million are estimated to be active weekly, in spite of the fact that access to the service is only limited from within Japan.

As discussed in [2], these previously mentioned IPTV systems differ significantly in their network architecture (server-based vs. peer-to-peer), as well as in their video delivery mechanism. Basically, there are three major categories of IPTV content distribution methods. First, there is *live TV streaming*, in which the current live TV program is packetized and simply streamed as application-layer multicast to the connected overlay peers, which then share the stream with other peers. Then, there are *Video-on-Demand* (VoD) systems that are not restricted to any broadcasting schedules of TV stations, but permit the user to browse a catalogue of available video files and play the content entirely upon request. For both solutions, respectively, Zattoo and Joost are two popular services, both based on peer-to-peer technology. However, as they operate on proprietary protocols and architectures, an analytical evaluation for such systems can only be done from the edge of the network by studying the user's perceived *quality of service* (QoS) or *quality of experience* (QoE).

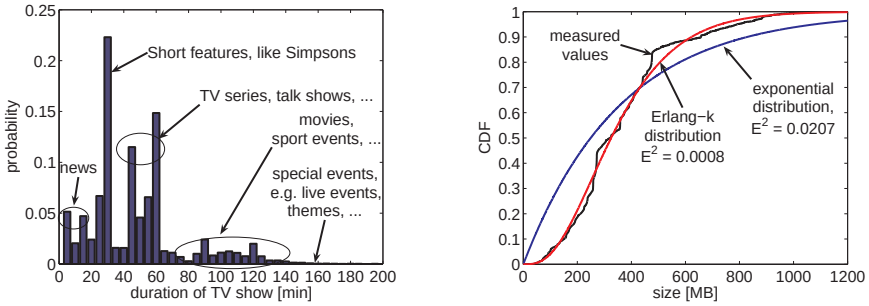
On the other hand, *network-based TV recorders* operate basically in the same way as home hard-disc video recorders with the only difference that the content is recorded and stored at some remote server in the Internet. An example for such a service is *OnlineTVRecorder* (OTR) in Germany [3]. The live TV program is recorded at the OTR server and registered users can download their previously programmed shows and later view them offline on a PC or handheld device. We will describe OTR in more detail in Section 2.

In this paper we extend our previous work in [4] on modeling the performance of an OTR video delivery service by means of a Markov model to derive the stationary sojourn time, which corresponds to the time until a typical user completes the download of a file. Since the file sizes are very large, the download duration may take longer than the user is willing to wait. For this reason, we include the user impatience in our model, where a waiting or downloading user may leave the system before completing the download. Queuing models with user impatience can be found in [5, 6], however, those models can not be applied to our system. For further discussion of the existing literature, see [4].

The remainder of this paper is organized as follows. In Section 2, we describe the basic operation of an OTR server and provide measurements for characterizing the content offered at the server. This is required to approximate the service time in our analytical model, which we define in Section 3. Next, we obtain the stationary sojourn time of a typical customer in Section 4, defined as the time needed to completely download the desired video file. In Section 5 we numerically investigate the influence of the system parameters on the sojourn time and conclude our work in Section 6.

2 Model of the OTR Recording System

The access to OTR is provided by a portal at its main web page [3]. There, a registered user has the possibility to choose programs he wishes to record from an *electronic program guide* (EPG) or can download previously recorded shows. The recordings can be either downloaded directly from the main server, from



(a) Distribution of TV show durations (b) CDF of files sizes

Fig. 1. Measurements of offered files at an OTR server

user-created mirror sites, or alternatively via P2P file-sharing networks (eMule or BitTorrent). However, the majority of clients are using the HTTP-based server download platforms and in this work we will focus on this type of file transfer. In the following, when we refer to OTR server (or simply *server*), we treat the main server and the mirror servers in the same way, since their basic operation is the same with the only difference that mirror sites usually offer only a subset of available recordings after a slight delay.

As mentioned above, the recorded files are offered at a server via HTTP which can be freely accessed over the Internet. In order to safeguard licensing restrictions and prevent unauthorized access, a user may only download files that he had previously recorded. For this reason, the content is offered encoded which can not be played directly, but must be decoded prior to playback. When requesting a file, the user may begin immediately his download if the server has available download slots, else he will be put on a waiting queue. OTR also offers prioritizing certain users who provide donations, but we assume that all users are treated equal in this paper. Eventually, as over time other users will finish their downloads and leave the system, the waiting user will commence his download.

We investigated the actual file sizes of video files offered at an OTR server by measurements, which were taken in April 2007 consisting of 11563 randomly selected file samples from 19 different TV channels. According to the information provided by OTR, standard video files are encoded at a resolution of 512×384 pixels at a video bitrate of about 750 kbps and an audio bitrate of 128 kbps [3]. The measured data contains only standard quality video files and consists of approximately 80% encoded in DivX format and 20% in Windows Media Video (WMV) format.

Fig. 1(a) shows the probability distribution of the TV show durations in minutes. The majority of the files (95%) are discretized in units of 5 min. We can distinguish 4 major categories of TV shows. Most files are short features (e.g. animation series) of about 30 min and shorter files may be for instance news reports. Another peak can be found between 45-60 min, which is the usual dura-

tion of TV dramas or other periodical shows. Movies usually have the duration between 90-120 min and very few larger recordings of special events exist, like broadcasts of live sports events.

In this paper, we are more interested in the file size distribution than the duration of the shows themselves in order to approximate the download time. Fig. 1(b) shows that the actual file size distribution has a mean of 368.31 MB and standard deviation of 196.82 MB. It can be well fitted by a an Erlang-distribution with only small residual mean squared error $E^2 = 0.0008$. We will assume an exponential file size distribution for the sake of analytical tractability in spite of its slightly higher residual error, cf. Fig. 1(b), however, the equations developed here could be extended to the Erlang case, too.

3 Description of the Model

Let us assume an OTR server responsible of managing the demands of a maximum finite number of $\underline{N}$ customers. In the following we will describe the model of the behavior of one of these servers.

We assume first that user requests arrive at the server according to a Poisson process with parameter $\lambda > 0$. When a request arrives and the system has free download slots, the client immediately proceeds with the download. Then, the user becomes a *downloading client* and we also say that the customer is *served*.

We may further assume that the server has a total fixed upload bandwidth of capacity C. In our model, the capacities are normalized by C, thus without loss of generality the normalized server capacity is 1. This bandwidth is equally shared among a maximum of $n^\star$ simultaneously downloading clients. If there were more than $n^\star$ simultaneous clients, the exceeding customers would wait for a downloading slot to become free. We refer to those clients as *waiting clients*. The total number of downloading and waiting clients at the server is thus in this particular setting finite (with a maximum number equal to $\underline{N}$).

When downloading a file, the access bandwidth of the client may be the bottleneck, that is R/C, where R is the maximum physical download bandwidth of the customer and C the real capacity at the server. We assume this bandwidth R is the same for all customers. Clearly, the condition $R/C < 1$ must hold.

The average rate at which clients complete their downloads also depends on the file size. We assume here that the file size is randomly distributed following an exponential distribution of parameter μ reflecting the measurement results described in Section 2 and normalized by the system capacity.

While in the system, a client might become *impatient* and decides to leave the system after a random amount of time. We assume that the average impatience duration depends on the speed of the download. That is, when there are less than $n^\star$ customers in the system, the impatience duration is distributed according to an exponential random amount of time with average θ_1^{-1}. In the other case, i.e., there are more than $n^\star$ customers in the system, the impatience duration for customers being served remains the same, but the average impatience time for waiting customers is $\theta_2^{-1} < \theta_1^{-1}$.

Our objective is to derive the time needed for an arbitrary customer to successfully complete the download of a file. We call this the *sojourn time* of a customer.

4 Derivation of the Stationary Sojourn Time Distribution

The changeover point $N^\star$ reflects whether the user's access bandwidth or the server's capacity is the limiting factor. Let first $N^\star$ be such that

$$\frac{C}{N^\star - 1} > R \qquad \text{and} \qquad \frac{C}{N^\star} \leq R\,. \tag{1}$$

We assume that $N^\star < n^\star$, otherwise the resulting model is trivial.

We are interested in computing the exact sojourn time a customer spends in the system in order to completely download the entire file. The method we will apply actually consists of solving a system of differential equations and is inspired by the work of Sericola et al. in [7] or Masuyama and Takine in [8]. However, our system is more complex since it integrates impatience and different service behaviors according to the actual number of customers in the system. We first derive equations for the remaining sojourn time distribution of an observed customer, then we establish the stationary distribution of the number of customers in the system and finally obtain the stationary distribution of the sojourn time of an arbitrary customer.

4.1 Remaining Sojourn Time

When the system counts less than $N^\star$ customers, any new customer is served at an average speed of $(R/C\mu)^{-1}$. Obviously, in this case the bandwidth is not shared. However, as soon as the system counts more than $N^\star$ clients, say n_d clients, the bandwidth of our observed customer shrinks to $(\mu/n_d)^{-1}$ on average. It is important to know exactly when this happens.

It is clear that as soon as our observed customer is in service, i.e. downloading the file, he will continue until either the file is completely downloaded or the customer becomes impatient and leaves the system before completion. Nevertheless, we still need to take the arriving customers following his arrival into account, even if they do not directly interfere with our observed customer's sojourn time, i.e. if they are waiting customers. Indeed, these waiting customers will eventually become downloading customers. Accordingly, the service rate will remain at a level $\mu/n^\star$, even when a customer in service leaves the system.

Imagine now our customer entering the system counting already more than $n^\star$ clients. Our observed customer becomes, thus, a waiting customer. It is then important to know exactly how many clients were waiting prior to his arrival in order to exactly determine when his service will start. In the same manner we also need to record how many clients arrive after his arrival, in order to determine the subsequent speed of service.

Owing to the necessity to keep track of the actual number of the other clients in the system and, therefore, to differentiate between the system behavior, we define the following three different conditional random variables.

For $K \in \{0, 1, \ldots, n^\star - 1\}$, we define the random variable $W(K, 1)$ as the remaining time a customer needs in order to completely download the file he requested, given that there are K customers in service. For $K \in \{n^\star, \ldots, \underline{N} - 1\}$ and $N \in \{n^\star + 1, \ldots, \underline{N}\}$, we define the two random variables $W(K, 1)$ and $W(K, 0, N)$ according to whether our customer is in service or not. The r.v. $W(K, 1)$ is the remaining sojourn time of the observed customer when the system counts K customers. The r.v. $W(K, 0, N)$ is the remaining sojourn time of the observed customer when there are K customers in the system and our observed client is waiting at position N to be served, i.e. $N - n^\star$ customers need to leave the system before our customer starts downloading the file. Note that these $N - n^\star$ clients have to be clients in service or waiting customers located in front of our observed customer in the queue.

We denote by $\mathcal{E}(x)$ an exponentially distributed random variable with mean $1/x$ and formulate the following theorem, considering all possible cases that can occur.

Theorem 1. *For $K \in \{0, \ldots, N^\star - 1\}$, the remaining sojourn time of a customer $W(K, 1)$ in a system that counts K customers is such that:*

$$W(K,1) = \begin{cases} \mathcal{E}(\Lambda(K)) & w.p.\ \mu R/C(\Lambda(K))^{-1} \\ \mathcal{E}(\Lambda(K)) + W(K-1,1) & w.p.\ K(\mu R/C + \theta_1)(\Lambda(K))^{-1} \\ \mathcal{E}(\Lambda(K)) + W(K+1,1) & w.p.\ \lambda(\Lambda(K))^{-1} \end{cases} \tag{2}$$

where $\Lambda(K)$ is the exponentially distributed rate at which the next event occurs that changes the system state:

$$\Lambda(K) = (K+1)\,(\mu R/C + \theta_1) + \lambda. \tag{3}$$

When K belongs to $\{N^\star, \ldots, n^\star - 1\}$, we have:

$$W(K,1) = \begin{cases} \mathcal{E}(\Lambda(K)) & w.p.\ \mu((K+1)\Lambda(K))^{-1} \\ \mathcal{E}(\Lambda(K)) + W(K-1,1) & w.p.\ (K/(K+1)\mu + K\theta_1)\,(\Lambda(K))^{-1} \\ \mathcal{E}(\Lambda(K)) + W(K+1,1) & w.p.\ \lambda(\Lambda(K))^{-1} \end{cases} \tag{4}$$

where

$$\Lambda(K) = \mu + (K+1)\,\theta_1 + \lambda. \tag{5}$$

When $K \in \{n^\star, \ldots, \underline{N} - 2\}$, the remaining sojourn time of the observed customer, assuming he is already in service, is:

$$W(K,1) = \begin{cases} \mathcal{E}(\Lambda(K)) & w.p.\ \mu(n^\star\Lambda(K))^{-1} \\ \mathcal{E}(\Lambda(K)) + W(K-1,1) & w.p.\ \frac{(n^\star-1)/n^\star\mu + (n^\star-1)\theta_1 + (K+1-n^\star)\theta_2}{\Lambda(K)} \\ \mathcal{E}(\Lambda(K)) + W(K+1,1) & w.p.\ \lambda(\Lambda(K))^{-1} \end{cases} \tag{6}$$

where

$$\Lambda(K) = \mu + n^\star\theta_1 + (K + 1 - n^\star)\,\theta_2 + \lambda. \tag{7}$$

When the observed customer is not in service and assuming he is at position $n^\star + 1$*, we have:*

$$W(K, 0, n^\star + 1) = \begin{cases} \mathcal{E}(\Lambda(K)) + W(K-1, 1) & w.p.\ (\mu + n^\star\theta_1)(\Lambda(K))^{-1} \\ \mathcal{E}(\Lambda(K)) + W(K-1, 0, n^\star + 1) & w.p.\ (K - n^\star)\theta_2(\Lambda(K))^{-1} \\ \mathcal{E}(\Lambda(K)) + W(K+1, 0, n^\star + 1) & w.p.\ \lambda(\Lambda(K))^{-1}. \end{cases} \tag{8}$$

In case the observed customer is at position N *with* $N \in \{n^\star + 2, \ldots, K + 1\}$*, we have:*

$$W(K, 0, N) = \begin{cases} \mathcal{E}(\Lambda(K)) + W(K-1, 0, N-1) & w.p.\ \frac{\mu + n^\star\theta_1 + (N - (n^\star+1))\theta_2}{\Lambda(K)} \\ \mathcal{E}(\Lambda(K)) + W(K-1, 0, N) & w.p.\ (K + 1 - N)\,\theta_2(\Lambda(K))^{-1} \\ \mathcal{E}(\Lambda(K)) + W(K+1, 0, N) & w.p.\ \lambda(\Lambda(K))^{-1}. \end{cases} \tag{9}$$

In both cases described by (8) and (9), the term $\Lambda(K)$ *is used as defined in (7). When* K *equals* $\underline{N} - 1$*, the remaining sojourn time of the observed customer, already in service, is:*

$$W(\underline{N}-1, 1) = \begin{cases} \mathcal{E}(\Lambda(\underline{N} - 1)) & w.p.\ \mu\,(n^\star\Lambda(\underline{N} - 1))^{-1} \\ \mathcal{E}(\Lambda(\underline{N} - 1)) + W(\underline{N} - 2, 1) & w.p.\ \frac{(n^\star - 1)/n^\star\mu + (n^\star - 1)\theta_1) + (\underline{N} - n^\star)\theta_2}{\Lambda(\underline{N} - 1)} \end{cases} \tag{10}$$

where

$$\Lambda(\underline{N} - 1) = \mu + n^\star\theta_1 + (\underline{N} - n^\star)\,\theta_2. \tag{11}$$

When the observed customer is not in service, then assuming he is at position $n^\star + 1$*, we have:*

$$W(\underline{N} - 1, 0, n^\star + 1) = \begin{cases} \mathcal{E}(\Lambda(\underline{N} - 1)) + W(\underline{N} - 2, 1) \\ \qquad w.p.\ (\mu + n^\star\theta_1)(\Lambda(\underline{N} - 1))^{-1} \\ \mathcal{E}(\Lambda(\underline{N} - 1)) + W(\underline{N} - 2, 0, n^\star + 1) \\ \qquad w.p.\ (\underline{N} - (n^\star + 1))\,\theta_2(\Lambda(\underline{N} - 1))^{-1}. \end{cases} \tag{12}$$

In case the observed customer is at position N *with* $N \in \{n^\star + 2, \ldots, \underline{N}\}$*, we have:*

$$W(\underline{N} - 1, 0, N) = \begin{cases} \mathcal{E}(\Lambda(\underline{N} - 1)) + W(\underline{N} - 2, 0, N - 1) \\ \qquad w.p.\ (\mu + n^\star\theta_1 + (N - (n^\star + 1))\,\theta_2)\,(\Lambda(\underline{N} - 1))^{-1} \\ \mathcal{E}(\Lambda(\underline{N} - 1)) + W(\underline{N} - 2, 0, N) \\ \qquad w.p.\ (\underline{N} - N)\,\theta_2(\Lambda(\underline{N} - 1))^{-1}. \end{cases} \tag{13}$$

where the definition of $\Lambda(K)$ *found in Equation (11) is used in (12) and (13).*

Proof. We only establish a formal proof of Equation (8), since the proof for all other equations can be obtained following a similar argument.

We compute the remaining sojourn time of an observed customer, given that the observed customer is in a system counting K other clients with $K \geq n^\star$.

Moreover, we assume that our tagged customer's service has not yet started. However, as soon as one of the $n^\star$ clients already in service leaves the system, our observed customer will begin with his download. We, thus, compute the remaining sojourn time of our observed customer given that he is at position $n^\star + 1$. In this case, because of the memoryless property of the exponential distribution, the next event (arrival or departure) takes place after an exponentially distributed time with parameter $\Lambda(K)$. We have, as stated in Equation (7):

$$\Lambda(K) = \mu + n^\star\theta_1 + (K + 1 - n^\star)\,\theta_2 + \lambda. \tag{14}$$

Indeed, we may observe one of the $n^\star$ clients in service, either finishing their download (at a rate $\mu/n^\star$), or becoming impatient (at a rate θ_1). The remaining customers including our observed customer, thus, those $K + 1 - n^\star$ customers that have not yet started downloading their file, may become impatient at a rate θ_2. Of course, we may still observe the arrival of a new customer at a rate λ.

If the departure of a served customer occurs, then our observed customer will become served. This happens with the probability $(\mu + n^\star\theta_1)\,(\Lambda(K))^{-1}$ and corresponds to the first case in (8). The second case corresponds to where a waiting customer becomes impatient, leaving our observed customer still waiting for service, but the system counts one customer less. This happens with probability $(K - n^\star)\,\theta_2\,(\Lambda(K))^{-1}$. The last case corresponds to the arrival of a new user. ■

Let us remark the following point. Since we are interested in computing the sojourn time of a customer, defined as the total time needed to download his desired file, we only consider successfully completed downloads and the event that our observed customer leaves the system due to impatience is not taken into account in any of the equations in Theorem 1.

Let W be the remaining sojourn time of a typical customer. We define the following conditional probabilities. For $K \in \{0, \ldots, n^\star - 1\}$

$$\begin{aligned} R(y \,|\, K, 1) &= P\,[W > y \,|\, X = K, S = 1] \\ &= P\,[W(K, 1) > y]\,, \end{aligned} \tag{15}$$

thus, the complementary remaining sojourn time distribution of a customer in service ($S = 1$) in a system counting K users ($X = K$). For $K \in \{n^\star, \ldots, \underline{N} - 1\}$

$$\begin{aligned} R(y \,|\, K, 1) &= P\,[W > y \,|\, X = K, S = 1] \\ &= P\,[W(K, 1) > y] \end{aligned} \tag{16}$$

$$\begin{aligned} R(y \,|\, K, 0, N) &= P\,[W > y \,|\, X = K, S = 0, P = N] \\ &= P\,[W(K, 0, N) > y] \end{aligned} \tag{17}$$

where $N \in \{n^\star + 1, \ldots, K + 1\}$ and P is the position of the observed user in the queue of waiting users, since $S = 0$ indicates that our observed customer is not in service, yet.

We now establish the system of differential equations in the next theorem, which is given without proof.

Theorem 2. *The conditional complementary probability distributions $R(y \mid K, 1)$ and $R(y \mid K, 0, N)$ respect the following differential equations.*

If $0 \leq K < N^\star$:

$$\begin{aligned}\partial_y R\left(y \mid K, 1\right) = &- \Lambda(K) R\left(y \mid K, 1\right) + K\left(R/C\mu + \theta_1\right) R\left(y \mid K-1, 1\right) \\ &+ \lambda R\left(y \mid K+1, 1\right).\end{aligned} \tag{18}$$

If $N^\star \leq K < n^\star$:

$$\begin{aligned}\partial_y R\left(y \mid K, 1\right) = &- \Lambda(K) R\left(y \mid K, 1\right) \\ &+ \left(\frac{K}{K+1}\mu + K\theta_1\right) R(y \mid K-1, 1) + \lambda R\left(y \mid K+1, 1\right).\end{aligned} \tag{19}$$

If $n^\star \leq K < \underline{N} - 1$:

$$\begin{aligned}\partial_y R\left(y \mid K, 1\right) = &- \Lambda(K) R\left(y \mid K, 1\right) + \lambda R\left(y \mid K+1, 1\right) \\ &+ \left(\frac{n^\star - 1}{n^\star}\mu + \left(n^\star - 1\right)\theta_1 + \left(K + 1 - n^\star\right)\theta_2\right) R\left(y \mid K-1, 1\right).\end{aligned} \tag{20}$$

$$\begin{aligned}\partial_y R\left(y \mid K, 0, n^\star + 1\right) = &- \Lambda(K) R\left(y \mid K, 0, n^\star + 1\right) + \left(\mu + n^\star\theta_1\right) R\left(y \mid K-1, 1\right) \\ &+ \left(K - n^\star\right)\theta_2 R\left(y \mid K-1, 0, n^\star + 1\right) \\ &+ \lambda R\left(y \mid K+1, 0, n^\star + 1\right)\end{aligned} \tag{21}$$

Moreover, if $n^\star + 1 < N \leq K + 1$:

$$\begin{aligned}\partial_y R\left(y \mid K, 0, N\right) = &- \Lambda(K) R\left(y \mid K, 0, N\right) + \lambda R\left(y \mid K+1, 0, N\right) \\ &+ \left(\mu + n^\star\theta_1 + \left(N - 1 - n^\star\right)\theta_2\right) R\left(y \mid K-1, 0, N-1\right) \\ &+ \left(K + 1 - N\right)\theta_2 R\left(y \mid K-1, 0, N\right).\end{aligned} \tag{22}$$

Finally, we have

$$\begin{aligned}\partial_y R\left(y \mid \underline{N} - 1, 1\right) = &- \Lambda\left(\underline{N} - 1\right) R\left(y \mid \underline{N} - 1, 1\right) \\ &+ \left(\frac{n^\star - 1}{n^\star}\mu + \left(n^\star - 1\right)\theta_1 + \left(\underline{N} - n^\star\right)\theta_2\right) R\left(y \mid \underline{N} - 2, 1\right)\end{aligned} \tag{23}$$

$$\begin{aligned}\partial_y R\left(y \mid \underline{N} - 1, 0, n^\star + 1\right) = &- \Lambda\left(\underline{N} - 1\right) R\left(y \mid \underline{N} - 1, 0, n^\star + 1\right) \\ &+ \left(\underline{N} - 1 - n^\star\right)\theta_2 R\left(y \mid \underline{N} - 2, 0, n^\star + 1\right) \\ &+ \left(\mu + n^\star\theta_1\right) R\left(y \mid \underline{N} - 2, 1\right)\end{aligned} \tag{24}$$

and

$$\begin{aligned}\partial_y R\,(y\,|\,\underline{N}-1,0,N) = &-\Lambda\,(\underline{N}-1)\,R\,(y\,|\,\underline{N}-1,0,N)\\ &+(\underline{N}-N)\,\theta_2 R\,(y\,|\,\underline{N}-2,0,N)\\ &+(\mu+n^\star\theta_1+(N-1-n^\star)\,\theta_2)\,R\,(y\,|\,\underline{N}-2,0,N-1)\end{aligned} \tag{25}$$

when $n^\star+1<N\leq\underline{N}$.

For $n^\star\leq i<\underline{N}$, we define the vectors $\mathbf{R}(y,i)$ of size $(i-n^\star+2)\times 1$ as follows:

$$\mathbf{R}(y,i)=\begin{pmatrix}R(y\,|\,i,0,n^\star+1)\\ R(y\,|\,i,0,n^\star+2)\\ \cdots\\ R(y\,|\,i,0,i+1)\\ R(y\,|\,i,1)\end{pmatrix}. \tag{26}$$

Accordingly, we define the vector $\mathbf{R}(y)$ as composed as follows:

$$\mathbf{R}(y)=\begin{pmatrix}R(y\,|\,0,1)\\ R(y\,|\,1,1)\\ R(y\,|\,n^\star-1,1)\\ \mathbf{R}(y,n^\star)\\ \mathbf{R}(y,n^\star+1)\\ \cdots\\ \mathbf{R}(y,\underline{N}-1)\end{pmatrix} \tag{27}$$

which has the dimension $\frac{1}{2}\left((n^\star)^2-(2\,\underline{N}+1)\,n^\star+3\,\underline{N}+\underline{N}^2\right)$.

Theorem 2 can now be written as

$$\partial_y\mathbf{R}(y)=A\mathbf{R}(y) \tag{28}$$

$$\mathbf{R}(0)=\mathbf{1}, \tag{29}$$

where $\mathbf{1}$ is a vector of appropriate size consisting of 1. The matrix A is defined as composed of the following blocks:

$$A=\begin{pmatrix}C_0 & A_0 & 0 & \dots & 0\\ B_1 & C_1 & A_1 & \dots & 0\\ 0 & B_2 & C_2 & \dots & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & 0 & \dots & C_{\underline{N}-1}\end{pmatrix}, \tag{30}$$

where for $0\leq i<n^\star$

$$C_i=-\Lambda(i) \tag{31}$$

$$A_i=\lambda \tag{32}$$

$$B_i=\begin{cases}i\,(R/C\mu+\theta_1) & \text{if } 0<i<N^\star\\ i/\,(i+1)\,\mu+i\,\theta_1 & \text{if } N^\star\leq i<n^\star\end{cases} \tag{33}$$

When i belongs to $\{n^\star, \ldots, \underline{N}-1\}$, we have $C_i = -\Lambda(i)I$ with I the identity matrix of appropriate size $(i-n^\star+2)\times(i-n^\star+2)$. We have

$$B_{n^\star} = \begin{pmatrix} \mu + n^\star\theta_1 \\ \frac{n^\star-1}{n^\star}\mu + (n^\star-1)\,\theta_1 + \theta_2 \end{pmatrix} \tag{34}$$

$$B_{n^\star+1} = \begin{pmatrix} \theta_2 & \mu + n^\star\theta_1 \\ \mu + n^\star\theta_1 + \theta_2 & 0 \\ 0 & \frac{n^\star-1}{n^\star}\mu + (n^\star-1)\,\theta_1 + 2\,\theta_2 \end{pmatrix} \tag{35}$$

For $2 \le i \le \underline{N} - 1 - n^\star$ the matrix $B_{n^\star+i}$ is of size $(i+2)\times(i+1)$ and such that:

$$\begin{array}{lll} B_{n^\star+i}(j,j) & = (i-(j-1))\,\theta_2 & \text{for } 1 \le j \le i \\ B_{n^\star+i}(j+1,j) & = \mu + n^\star\,\theta_1 + j\,\theta_2 & \text{for } 1 \le j \le i \\ B_{n^\star+i}(1,i+1) & = \mu + n^\star\,\theta_1 & \\ B_{n^\star+i}(i+2,i+1) & = \frac{n^\star-1}{n^\star}\mu + (n^\star-1)\,\theta_1 + (i+1)\,\theta_2, & \end{array} \tag{36}$$

and other elements being equal to 0. For $0 \le i \le (\underline{N}-1) - n^\star$, the matrix A_i is a matrix of size $(i+2)\times(i+3)$, whose elements are

$$\begin{array}{rl} A_{n^\star+i}(j,j) = \lambda & \text{for } 1 \le j \le i+1 \\ A_{n^\star+i}(i+2,i+3) = \lambda & \end{array} \tag{37}$$

and others elements being equal to 0.

The system of differential equations (28) with initial conditions (29) can be easily solved and we get:

$$R(y) = \exp(Ay). \tag{38}$$

4.2 Stationary Distribution of the Number of Users in the System

Our objective is to compute the stationary distribution of the time a customer needs in order to completely download a file. When a customer enters the system, he may find the system already occupied with $K < \underline{N}$ customers. This section aims at computing the stationary distribution of the number of customers in the system at the arrival instant of the observed customer. Due to the PASTA property, this stationary distribution is simply equal to the stationary distribution of the number of customers in the system at any time.

Let $\{X(t);\, t \in \mathbb{R}^+\}$ be the Markov process counting the number of customers in the system. As previously mentioned, the corresponding stationary random variable is given by X. We denote by the vector π the corresponding stationary distribution, that is

$$\pi(K) = P\left[X = K\right], \tag{39}$$

with $K \in \{0, \ldots, \underline{N}\}$. We define Q as the generator associated to the process $\{X(t);\, t \in \mathbb{R}^+\}$. The elements of Q are:

$$\begin{array}{lll} Q(i,i+1) = \lambda & & \text{for } 0 \le i \le \underline{N}-1 \\ Q(i,i-1) = i\,R/C\mu + i\,\theta_1 & & \text{for } 1 \le i \le N^\star \\ \qquad\quad = \mu + i\,\theta_1 & & \text{for } N^\star+1 \le i \le n^\star \\ \qquad\quad = \mu + n^\star\theta_1 + (i-n^\star)\,\theta_2 & & \text{for } n^\star+1 \le i \le \underline{N} \end{array} \tag{40}$$

The diagonal elements of Q are such that $Q\mathbf{1} = \mathbf{0}$, with $\mathbf{0}$ being a vector consisting of entries 0 and of appropriate size. Other elements of Q are zeros. Clearly, the process $\{X(t);\, t \in \mathbb{R}^+\}$ is a birth-and-death process. Accordingly, let ρ_i be

$$\begin{aligned} \rho_i &= \lambda / \left((i+1)\, R/C\mu + (i+1)\, \theta_1 \right) && \text{for } 0 \le i < N^\star \\ &= \lambda / \left(\mu + (i+1)\, \theta_1 \right) && \text{for } N^\star \le i < n^\star \\ &= \lambda / \left(\mu + n^\star\, \theta_1 + (i+1-n^\star)\, \theta_2 \right) && \text{for } n^\star \le i < \underline{N} - 1. \end{aligned} \tag{41}$$

We obtain the stationary probabilities for $K \in \{1, \ldots, \underline{N}\}$

$$\pi(K) = \pi(0) \prod_{i=0}^{K-1} \rho_i \qquad \text{with} \qquad \pi(0) = \left(1 + \sum_{K=1}^{\underline{N}} \prod_{i=0}^{K-1} \rho_i \right)^{-1}. \tag{42}$$

4.3 Stationary Total Sojourn Time

When a customer enters the system, he may either be immediately served or is placed last in the queue depending on the current number of customers present in the system. The complementary total sojourn time distribution of a customer respects the following equation:

$$P\left[W > y\right] = \sum_{K=0}^{n^\star - 1} \pi(K) R\left(y \,|\, K, 1\right) + \sum_{K=n^\star}^{\underline{N}-1} \pi(K) R\left(y \,|\, K, 0, K+1\right). \tag{43}$$

For $i \in \{0, \ldots, \underline{N} - 1 - n^\star\}$, we define the vectors $\underline{\pi}(i)$ as:

$$\underline{\pi}(i) = \mathbf{e}(i+2, i+1)\, \pi(n^\star + i), \tag{44}$$

where $\mathbf{e}(i,j)$ is a row vector of size i full of 0's except element j which is equal to 1. Accordingly, we define $\mathbf{\Pi}$ as composed of

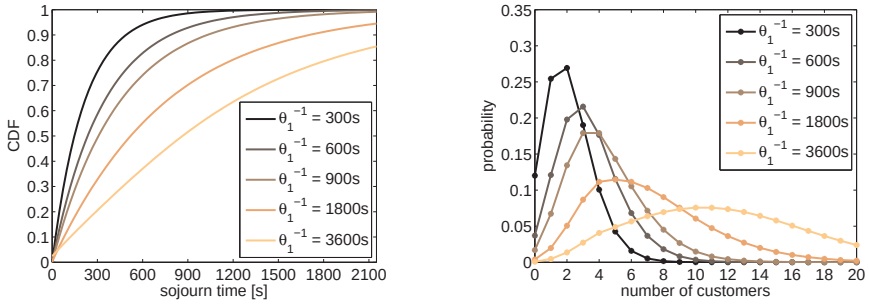
$$\mathbf{\Pi} = \left(\pi(0)\ \pi(1) \ldots \pi(n^\star - 1)\ \underline{\pi}(n^\star) \ldots \underline{\pi}(\underline{N} - 1) \right). \tag{45}$$

Then, using Equation (38), we obtain the complementary distribution of the user's sojourn time as

$$P\left[W > y\right] = \mathbf{\Pi} \exp(Ay) \mathbf{1}. \tag{46}$$

5 Numerical Evaluation

In this section we will provide some numerical results and briefly discuss the influence of some of the important parameters on the system behavior. Let us consider an OTR mirror server site with a total capacity of $C = 20\,\text{Mbps}$. Note that while the analytical model used rates normalized by the server's capacity, we give absolute values here, as we are interested in the actual downloading durations and they are more meaningful for verifying the plausibility of the results. The average file size μ^{-1} is 359.87 MByte and the maximum download rate R of all users is 4 Mbps as specified by the ITU G.992.2 standard for ADSL



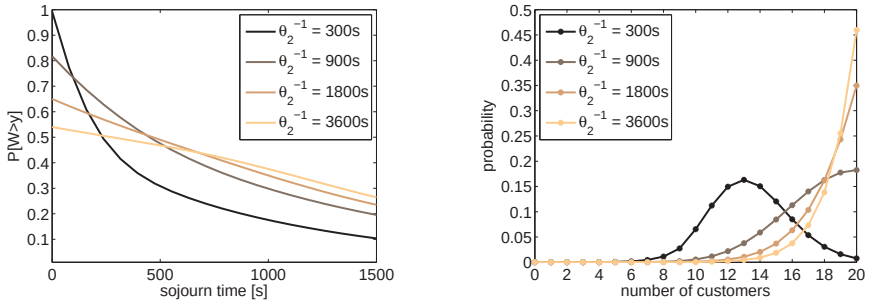
(a) CDF of sojourn time (b) Distribution of number of users

Fig. 2. Influence of the impatience threshold θ_1^{-1}

Lite. Since we assume this is a mirror site operated by a private person, it is reasonable to assume that the maintainer only limits access to a relatively small number of concurrently served downloads, e.g. $n^\star = 10$ and $\underline{N} = 20$, due to the following consideration. Many existing mirror servers indicate the current queue length and the expected waiting time until a download slot will become free. In the above scenario, a user at the first position of the waiting queue would have to wait 20 min and at the last (i.e. 10th) position would require a waiting time of 200 min. A requesting user who would be facing to wait so long before service would obviously select a different mirror site beforehand. As further parameters, if not indicated otherwise, we assume a request arrival rate of $\lambda = 10^{-2}$ requests/s as well as the impatience thresholds of $\theta_1^{-1} = 2\,\mathrm{h}$ and $\theta_2^{-1} = 1\,\mathrm{h}$ for downloading and waiting users, respectively. Furthermore, we verified the accuracy of our numerical implementation by simulations.

5.1 Influence of Impatience During Downloading

Let us first consider the impatience threshold θ_1 and its influence on the sojourn time of a successful customer as shown in Fig. 2. On the left, in Fig. 2(a), the CDF of the sojourn time as computed from Equation (46) is shown with different impatience thresholds for downloading users θ_1^{-1}. Darker lines represent smaller values of θ_1^{-1}. Obviously, the sojourn time increases when the patience threshold increases. This can be explained by the fact that when θ_1^{-1} is small, only small files are actually downloaded and users downloading larger files will have a large tendency to abort their attempts, so on average the sojourn time in the system will be small. This is also suggested when we look at the steady state distribution of the number of users in the system, see Fig. 2(b). For all considered values, in particular when θ_1^{-1} is small, the probability of finding an empty system upon arrival, i.e. $\pi(0)$, is greater than zero. A larger θ_1^{-1} shifts the weight of the distribution toward a larger number of customers. In the case of $\theta_1^{-1} = 3600\,\mathrm{s}$ we can see that $\pi(\underline{N}) > 0$, leading to blocking of potentially new arrivals.



(a) CCDF of sojourn time

(b) Distribution of number of users

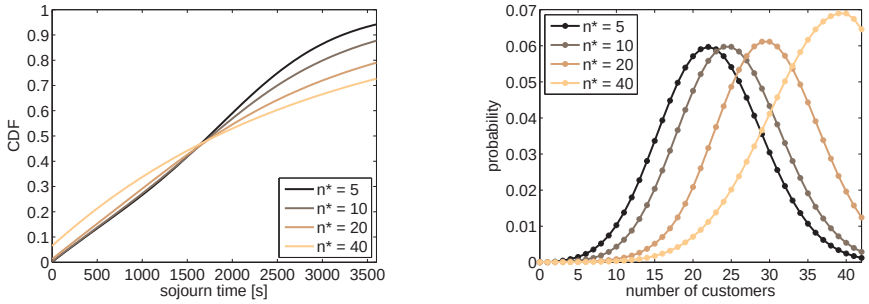
Fig. 3. Influence of the impatience threshold θ_2

5.2 Influence of Impatience during Waiting

We now investigate the influence of the impatience threshold of waiting users θ_2 on the sojourn time. The numerical results are shown in Fig. 3. Note that in contrast to Fig. 2(a) the left figure (Fig. 3(a)) now shows the complementary CDF of the sojourn time. It is remarkable that the probabilities $P[W > 0]$ may be less than 1, if the waiting impatience time θ_2^{-1} is large. Again, we can interpret this result better by looking at the corresponding distributions of customers as shown in Fig. 3(b). When $\theta_2^{-1} = 300\,\mathrm{s}$ the system is already serving $n^\star$ customers and there are on average 3 waiting users in the queue, indicated by the highest probabilities $\pi(i)$ for $i \in \{12, 13, 14\}$. However, since the impatience time is small, queued users quickly leave the system again. On the other hand for larger impatience values of θ_2^{-1}, especially when $\theta_2^{-1} > 900\,\mathrm{s}$, users wait longer in the queue and the probability is very large to find the system fully occupied. Note that $P[W = 0] = P[X = \underline{N}]$, i.e., the probability that the sojourn time is zero is equal to the probability that there are $\underline{N}$ customers in the system.

5.3 Influence of the Number of Available Download Slots

Finally, we investigate how the number of available download slots affects the sojourn time. We now look at a slightly different scenario, with $\lambda = 1/80\,\mathrm{s}^{-1}$, $\underline{N} = 42$ and for values of $n^\star = 5, 10, 20, 40$ to emphasize the effect. Although the curves for the CDF of the sojourn time do not intersect at the same point, cf. Fig. 4(a), the intersection point y' for any two curves lies in a small range between 1600 s and 1700 s. With larger $n^\star$ the download bandwidth decreases and together with a higher patience while downloading than while waiting, the probability for sojourn times $y > y'$ beyond this intersection point increase with $n^\star$. The CDF of the sojourn time displays a similar behavior with blocking for large $n^\star$ as already observed for θ_2. Note that a value of $n^\star = 40$ and $\underline{N} = 42$ means that the maximum waiting queue length is 2.



(a) CDF of sojourn time (b) Distribution of number of users

Fig. 4. Influence of the number of download slots $n^\star$

6 Conclusion

In this paper we analytically derived the sojourn time of an arbitrary user, who successfully completes downloading a file from an IPTV server. Since the data volume of the requested file is very large, it is necessary to also take the user behavior in terms of his impatience into account, as this is a phenomenon frequently observed in actual systems. We investigated the effects of the impatience thresholds of waiting and downloading users, as well as the number of available downloading slots. Further numerical evaluations that we performed, e.g. the impact of the arrival rate λ, were also studied, but left out of this paper due to lack of space. In the future, we wish to perform a more detailed evaluation of the influencing parameters and extend the model to consider a more accurate file size distribution as well heterogeneous users with different access bandwidths.

References

1. USEN Corporation: News release. (`http://www.usen.com/admin/corp/news/pdf/2008/080118.pdf` (*in Japanese*))
2. Hoßfeld, T., Leibnitz, K.: A qualitative measurement survey of popular internet-based IPTV systems. submitted for publication (2008)
3. OnlineTVRecorder. (`http://www.onlinetvrecorder.com/`)
4. Hoßfeld, T., Leibnitz, K., Remiche, M.A.: Modeling of an online TV recording service. ACM SIGMETRICS Performance Evaluation Review **35** (2007) 15–17
5. Brandt, A., Brandt, M.: On the M(n)/M(n)/s queue with impatient calls. Perform. Eval. **35** (1999)
6. Gromoll, H.C., Robert, P., Zwart, B., Bakker, R.: The impact of reneging in processor sharing queues. In: Proc. of SIGMETRICS '06/Performance '06, New York, NY, USA, ACM Press (2006) 87–96
7. Sericola, B., Guillemin, F., Boyer, J.: Sojourn times in the M/PH/1 processor sharing queue. Queueing Systems **50** (2005) 109–130
8. Masuyama, H., Takine, T.: Sojourn time distribution in a MAP/M/1 processor-sharing queue. Operations Research Letters **31** (2003) 406–412