The effectiveness of throughput sampling for capacity management: a queueing approach

Wendy Ellens*†, Michel Mandjes†§
†KdVI, University of Amsterdam, the Netherlands
§CWI, Amsterdam, the Netherlands
Email: wendy.ellens@tno.nl, m.r.h.mandjes@uva.nl

Daniël Worm*, Hans van den Berg*‡
*TNO, Delft, the Netherlands
‡DACS, University of Twente, Enschede, the Netherlands
Email: j.l.vandenberg@tno.nl, daniel.worm@tno.nl

***Abstract*—For effective capacity management in access networks, it is essential to have a good insight in the service quality perceived by the users. As users share the service capacity available, one would want to know how the achieved per-user throughput fluctuates over time. In this paper we present a novel method that assesses the per-user throughput performance on the basis of throughput measurements at equidistant points in time. Our method relies on a queueing-theoretic framework, and allows us to explicitly quantify various statistics concerning the *minimum per-user throughput* obtained in a sample period, given the measured per-user throughput at the end points of that sample period, as well as the measured utilization during the interval.**

In an extensive numerical study we show the impact of important system parameters on these statistics. In addition, using illustrative examples, we demonstrate how the developed method can be used in practice for capacity management, with a specific focus on applications in cable access networks, an application for which our approach is particularly suitable.

***Index Terms*—throughput sampling, capacity management, queueing theory, birth-death processes, processor sharing**

I. Introduction

Capacity management is one of the primary tasks of network operators. Under-provisioned network links will lead to a degradation of the service quality perceived by the users, while severe over-provisioning involves unnecessary or too early investments (CAPEX). Network operators therefore continuously monitor the traffic load on their network links, for example via SNMP (Simple Network Management Protocol) which is widely used and nowadays implemented in most network devices [1]. Besides link loads also more detailed information regarding network performance may be used for capacity management, e.g. packet delays or buffer occupancies [2], [3]. In access networks, in particular, additional information regarding network usage is often obtained via *throughput sampling* ('speed tests', see e.g. [4]), i.e., measuring the throughputs of up- and downloads to/from locally installed servers, through regularly requests generated by PC s that act as fake users. The topic of this paper is the effectiveness of throughput sampling for capacity management in access networks.

Obviously, an important question is up to what extent the per-user throughput samples yield statistically representative results — i.e., adequately reflect the actual access rates obtained by the users, in relation to the access rate specified in the service level agreement (SLA) — and how this depends on, e.g., the sampling frequency. In particular, the challenge is to find a proper trade-off between accuracy of the sampling method and the required measurement/computational effort, see e.g. [5]. This is a difficult task as common statistical techniques to assess the accuracy of sampling methods can mostly not be used in practical network environments due to the non-stationary nature of the involved traffic processes [6]. Therefore, operators usually rely on 'experience' and simple rules of thumb for interpreting the results of their speed tests (and on additionally measured metrics) to assess whether or not link capacities need to be extended.

In the present paper we consider use of per-user throughput samples from a new perspective. In particular, we investigate the effectiveness of throughput sampling by focusing on the *fluctuations* from the sampled per-user throughputs that may have occurred *between* two consecutive samples. Therefore, we analyze the statistical behavior of the traffic (at the flow level) between consecutive samples, *given* the outcomes of the per-user throughput samples, as well as the measured link utilization during the sample interval. This analysis, for instance, quantifies the likelihood of the per-user throughput dropping below a certain critical value during a time frame of given length. Such insights help operators in deciding on a proper sampling frequency and in developing better triggering rules for capacity management.

Our focus in this paper is on developing a *theoretical framework* for analyzing the minimum per-user throughput in a sample period. In our setup, the flow arrival process between two consecutive throughput samples (say samples i and $i+1$) is 'locally' approximated by a homogeneous Poisson process with rate λ_i, while we impose the (common) assumption of relatively stable flow-size characteristics over longer time intervals (many sample periods, that is). As a result, the λ_i can be estimated from the utilization during the sample period. The next step is to realize that during the sample periods the evolution of the number of ongoing flows (active users) behaves as a birth-death process [7]. When few users are active, they all receive their maximum access rate (making the simplifying assumption that all users have the same maximum access rate). In this case the birth-death process corresponds to an M/M/∞ queue [8]. If the number of simultaneously active users increases such that the available link capacity becomes

too small to accommodate their access rates, the link capacity is equally shared by the users. In that situation the birth-death process essentially behaves as an M/M/1 *processor sharing* (PS) model [9].

Note, that modeling (TCP) flow-level dynamics on an access link in the way described above (i.e., a combination of M/M/∞ and M/M/1-PS models) has been widely accepted, see e.g. [10], [11], [12]. However, in those existing studies one usually exploits well-known results for the long-term (or: *steady-state*) behavior, under the assumption of homogeneous Poisson arrivals, to derive per-user throughputs and other performance measures; or, in [13] for example, new steady-state performance results are derived for similar systems with other resource sharing policies reflecting the presence of users with different access rates. In the present paper, as stated before, the focus is on the system's *transient behavior* over relatively short time intervals.

For choosing the network and traffic parameter values in our numerical examples we consider the typical situation in a *cable access network*, in particular a cable segment that connects a large number of subscribers to the Internet via a Cable Modem Termination Station (CMTS) [14]. Note that the proposed modeling approach is particularly suited to such a network environment as bandwidth sharing among active users in cable access networks is explicitly enforced by the CMTS (i.e., it does not rely on bandwidth allocations implicitly realized by TCP's congestion control).

Summarizing, the central contribution of this paper is the development of a novel method for fast transient analysis of the access link model described above. This method facilitates the numerical assessment of statistics related to the per-user throughput during a sample interval, given the realizations of the throughput samples and the measured link utilization during the sample period. One of these statistics is the probability that the experienced per-user throughput during the sample interval has dropped below a certain critical level (or, equivalently, remained above a certain level). In addition, using some illustrative examples, we demonstrate how the developed method can be used for capacity management.

The rest of this paper is organized as follows. Section II introduces the access link model, pointing out how to model the number of active users as a birth-death process. In Section III new results are derived for the conditional minimum of birth-death processes, given its begin and end states. In Section IV typical values for a cable access link are inserted into the model, and the theory is applied to assess the probability that the per-user throughput during a sample interval has dropped below a certain critical level. We first (in Section IV-A) present an extensive numerical study that shows the impact of the key system parameters on this probability; then, in Section IV-B, we demonstrate how the results can be used in practice. The paper is concluded (Section V) by a brief summary and some suggestions for further research.

II. Model Description

We model the number of clients in service as a *birth-death process* X_t, with $t \geq 0$. A birth-death process is a continuous-time Markov process with exponential transition times where transitions only occur between neighboring states [7]. The rate from state n to state $n+1$ (for $n \geq 0$) is denoted by $\lambda_n > 0$, while the rate from state n to state $n-1$ (for $n \geq 1$) is denoted by $\mu_n > 0$. We consider in this paper the following, more specific, model:

1) The state space is $S = \{0, 1, 2, \ldots\}$, where $X_t = n \in S$ means that at time t there are n active users.
2) Arrivals of new user requests follow a Poisson process with rate $\lambda > 0$ (e.g., $\lambda = 100$ s^{-1}, meaning that there are on average 100 new user requests per second).
3) At each request a user downloads a file with an exponentially distributed size, with mean f (e.g., 6 Mb).
4) The link capacity is C (e.g., 800 Mbps).
5) Individual users are allocated at most a limited access rate of $R_{\max}$ (e.g., 80 Mbps), with $R_{\max} \leq C$.
6) The queueing system is based on *processor sharing*: each user obtains the same throughput. Due to the per-user capacity restriction, we have that each user is allocated $R := \min\{R_{\max}, C/n\}$.

It is directly seen that this particular system is of the birth-death type, with arrival rate $\lambda_n = \lambda$ and departure rate $\mu_n = \min\{nR_{\max}/f, C/f\}$. It essentially consists of two regimes: for $n \leq n_0 := \lfloor C/R_{\max} \rfloor$, $\mu_n = nR_{\max}/f$, while for $n > n_0$, $\mu_n = C/f$. As a consequence, it behaves like an M/M/∞ queue in the states $0, 1, \cdots, n_0$, and like an M/M/1 queue in the states $n_0 + 1, n_0 + 2, \cdots$ [8], [9]. Note further that we do not impose a stability condition, since this paper aims at aspects related to the system's transient behavior.

Every t seconds, we measure the 'instantaneous per-user throughput', enabling us to calculate the number of active users. Our objective is to determine the distribution of the minimum per-user throughput (or: maximum number of users) between two subsequent throughput measurements, given the values of the per-user throughput at the start and end of the corresponding interval.

More specifically,

1) We assume that the operator knows the average file size f, and that the link capacity C and the maximum access rate $R_{\max}$ are given. Regularly, say also every t seconds (but the interval may be larger), the utilization ρ (the portion of the link capacity that was in use) is measured. With this information the average arrival rate of user requests during these intervals of t seconds can be calculated using $\rho = \lambda f/C$.
2) We use the throughput measurements and the link capacity to calculate the number of active users at the beginning and the end of a sample period. Actually this is only possible if the per-user throughput is less than $R_{\max}$, but we will see in Section IV-A that our assumption poses no limitations.
3) In Section III we show how to compute the probability

distribution of the maximum of a (general) birth-death process over a certain time period in terms of the arrival and departure rates. Knowing the values of λ, C, f, and $R_{\max}$, this facilitates the computation of the distribution of the maximum number of active users during a sample period of length t (and hence also that of the minimum per-user throughput during this period).

III. Theoretical analysis

The objective of this section is to analyze the so-called *running maximum process* $\bar{X}_t := \sup_{s\in[0,t]} X_s$, jointly with the final value X_t. We set up a technique to evaluate the probability

$$q_{m,i,j,t} := \mathbb{P}(\bar{X}_t \le m \,|\, X_0 = i, X_t = j),$$

i.e., the probability that the maximum of a birth-death process during a fixed time interval remains below a certain threshold m.

The derivation consists of three steps, described in Sections III-A up to III-C. Initially we assume that the sample period does not have the deterministic value t, but is rather an (independent) random variable T that follows an exponential distribution (with mean τ^{-1}). As it will turn out, this 'trick' enables straightforward computation of the probability under study; it is explained below how the result for random sample periods can be translated into one for fixed sample period.

In more detail, this section is organized as follows. In Section III-A we exploit the memoryless property of the exponential distribution to compute the probability

$$r_{m,i} := \mathbb{P}(\bar{X}_T = m \,|\, X_0 = i).$$

Section III-B uses a similar approach to find the joint distribution of the running maximum and the 'final value', by providing a scheme to compute

$$s_{m,i,j} := \mathbb{P}(\bar{X}_T = m, X_T = j \,|\, X_0 = i).$$

Finally, Section III-C describes an elegant approach to use the probability $s_{m,i,j}$ for an *exponential time* T to find an expression for the probability $q_{m,i,j,t}$ of having a maximum m in a *deterministic time* t, given that we know there were i users at the start of the sample period, and j at the end, again with $m \ge \max\{i,j\}$.

A. Running maximum

In this section we detail a procedure to determine $r_{m,i}$; this is the probability that the maximum over an exponential time T is m, given that the process starts in state i. To this end, define the probability that, starting at level n, level $n+1$ is reached before the 'exponential clock' T expires:

$$p_n := \mathbb{P}(\bar{X}_T \ge n+1 \,|\, X_0 = n).$$

It is clear that, relying on the memoryless property of the exponential distribution and the fact that in a birth-death process all intermediate states are visited,

$$r_{m,i} = \begin{cases} (p_i p_{i+1} \cdots p_{m-1}) \cdot \bar{p}_m & \text{if } i \le m, \\ 0 & \text{if } i > m, \end{cases}$$

with $\bar{p}_m := 1 - p_m = \mathbb{P}(\bar{X}_T = m \,|\, X_0 = m)$. In case $m = i$ this expression is to be understood as $\bar{p}_i$.

Hence we are left with devising a procedure to compute p_n. Again relying on the memoryless property, we have the evident relation

$$p_n = \frac{\lambda_n}{\lambda_n + \mu_n + \tau} + \frac{\mu_n}{\lambda_n + \mu_n + \tau}\, p_{n-1} p_n,$$

so that, for $n \ge 1$,

$$p_n = \frac{\lambda_n}{\lambda_n + \tau + \mu_n(1 - p_{n-1})}. \tag{1}$$

Now $r_{m,i}$ can be calculated recursively, since we know that $p_0 = \lambda_0/(\lambda_0 + \tau)$. These p_n will form the basis for further computations in Section III-B.

B. Running maximum jointly with final value

By refining the above analysis of $r_{m,i}$ we now point out how to compute $s_{m,i,j}$, as was defined above. First observe

$$s_{m,i,j} = \begin{cases} (p_i p_{i+1} \cdots p_{m-1}) \cdot \bar{p}_{m,j} & \text{if } i,j \le m, \\ 0 & \text{if } i > m \text{ or } j > m. \end{cases}$$

where $\bar{p}_{m,j} := \mathbb{P}(\bar{X}_T = m, X_T = j \,|\, X_0 = m)$. It now remains to determine $\bar{p}_{m,j}$. By using elementary 'Markovian arguments', it is directly seen that the following relations apply:

$$\bar{p}_{m,j} = \begin{cases} \dfrac{\mu_m}{\lambda_m + \mu_m + \tau} \left(p_{m-1}\bar{p}_{m,j} + \bar{p}_{m-1,j}\right), & \\ & \text{if } j < m; \\ \dfrac{\tau}{\lambda_m + \mu_m + \tau} + \dfrac{\mu_m}{\lambda_m + \mu_m + \tau} p_{m-1}\bar{p}_{m,m}, & \\ & \text{if } j = m; \\ 0, & \\ & \text{if } j > m. \end{cases} \tag{2}$$

Combining (1) and (2) yields $\bar{p}_{m,m} = \tau p_m/\lambda_m$ and

$$\bar{p}_{m,m-1} = \frac{\mu_m \bar{p}_{m-1,m-1}}{\lambda_m + \tau + \mu_m(1 - p_{m-1})} = \frac{\mu_m p_m}{\lambda_m} \bar{p}_{m-1,m-1},$$

which together lead to

$$\bar{p}_{m,m-1} = \frac{\mu_m p_m}{\lambda_m} \frac{\tau p_{m-1}}{\lambda_{m-1}}.$$

Using an analogous line of reasoning, we obtain in general, for $j < m$,

$$\bar{p}_{m,j} = \left(\frac{\mu_m p_m}{\lambda_m} \frac{\mu_{m-1} p_{m-1}}{\lambda_{m-1}} \cdots \frac{\mu_{j+1} p_{j+1}}{\lambda_{j+1}}\right) \cdot \frac{\tau p_j}{\lambda_j},$$

and hence

$$\begin{aligned} s_{m,i,j} = {} & (p_i p_{i+1} \cdots p_{m-1}) \cdot (p_j p_{j+1} \cdots p_m) \\ & \cdot \left(\frac{\mu_{j+1}}{\lambda_{j+1}} \cdots \frac{\mu_m}{\lambda_m}\right) \cdot \frac{\tau}{\lambda_j}. \end{aligned} \tag{3}$$

Now that we know $s_{m,i,j}$, we can use the probability

$$\begin{aligned} \bar{s}_{m,i,j} &:= \mathbb{P}(\bar{X}_T \le m, X_T = j \,|\, X_0 = i) \\ &= \sum_{k=i}^{m} s_{k,i,j} = \sum_{k=j}^{m} s_{k,i,j}, \end{aligned}$$

to compute the probability that the maximum over an exponential time T is m, given that the process starts in state i and ends in state j:

$$r_{m,i,j} := \mathbb{P}(\bar{X}_T \leq m \,|\, X_0 = i, X_T = j) = \frac{\mathbb{P}(\bar{X}_T \leq m, X_T = j \,|\, X_0 = i)}{\mathbb{P}(X_T = j | X_0 = i)} = \frac{\bar{s}_{m,i,j}}{\lim_{n\to\infty} \bar{s}_{n,i,j}}.$$

C. Deterministic times

The probability $r_{m,i,j}$ is the exponential-time counterpart of the probability $q_{m,i,j,t}$ that we wish to evaluate; in this section we point out how this 'translation' can be done. A first way to do this is to rely on *Laplace inversion*. To this end, observe that

$$r_{m,i,j} = \int_0^\infty \tau e^{\tau t} q_{m,i,j,t} \mathrm{d}t.$$

There are fast and reliable numerical algorithms to numerically invert $r_{m,i,j}$ into $q_{m,i,j,t}$ [15], [16].

We now describe an alternative method, that is specifically appropriate for computing probabilities related to running maxima. The main idea is to approximate the deterministic time t by a convolution of exponentially distributed times. More specifically, with T_k denoting the sum of k independent exponentially distributed random variables, each with mean t/k, it is well-known that T_k converges to the deterministic number t; this also follows directly from the limit of the Laplace transform of T_k: for $\alpha \geq 0$,

$$\lim_{k\to\infty} \mathbb{E}e^{-\alpha T_k} = \lim_{k\to\infty} \left(\frac{k/t}{\alpha + k/t}\right)^k = e^{-\alpha t}.$$

We now point out how this observation can be used to evaluate $q_{m,i,j,t}$. To this end, we define the $(m+1)\times(m+1)$ matrix $S_{m,\tau}$ by

$$(S_{m,\tau})_{i,j} = \bar{s}_{m,i-1,j-1}.$$

Realize that the (i,j)-th element of the matrix $(S_{m,\tau})^k$ describes the probability, when starting in value $i-1$, that the end value is $j-1$ while at the same time the maximum remains below m within the run-down of k subsequent exponential clocks, each of which has expiration rate τ.

Now consider the sequence $S^k_{m,k\tau}$, where $\tau = 1/t$, for $k = 1, 2, \ldots$. Based on the above observations, it is clear that it converges, as $k \to \infty$, to a matrix $S^\star_{m,t}$, where

$$\left(S^\star_{m,t}\right)_{i+1,j+1} = \mathbb{P}(\bar{X}_t \leq m, X_t = j \,|\, X_0 = i) =: s^*_{m,i,j,t}.$$

From this, we can compute the probability of our interest:

$$q_{m,i,j,t} := \mathbb{P}(\bar{X}_t \leq m \,|\, X_0 = i, X_t = j) = \frac{\mathbb{P}(\bar{X}_t \leq m, X_t = j \,|\, X_0 = i)}{\mathbb{P}(\bar{X}_t = j | X_0 = i)} = \frac{s^\star_{m,i,j,t}}{\lim_{n\to\infty} s^\star_{n,i,j,t}}.$$

The idea is to evaluate $q_{m,i,j,t}$ by computing $S^k_{m,k\tau}$ for a suitably large k. Realize that for instance $k = 2^{10} = 1024$ requires just 10 matrix multiplications; the distribution of T_k is 'nearly deterministic', as it has mean t (obviously), while its standard deviation is just $t/2^5 = t/32$.

IV. Numerical Analysis of the Per-User Throughput

In this section we numerically assess the per-user throughput, focusing on the minimum throughput, as assigned to an individual user, between two subsequent throughput measurements. The numerics are done using the model described in Section II, by applying the results for maxima of birth-death processes derived in Section III. We calculate the probability that during the time between two throughput measurements the number of active users stays below a certain threshold m, conditional on the number of users present at the beginning and the end of the sample period; this assures that the per-user throughput remains above a corresponding threshold. In Section IV-A we study the impact of the measured throughputs (at the beginning and end of the sample period), the sample frequency, the user arrival rate, and the average file size. Section IV-B shows some examples of charts that can be generated using our approach and that may result useful in network traffic management.

Unless stated otherwise, we have picked the following *default values*: the capacity $C = 800$ Mbps, the maximum access rate $R_{\max} = 80$ Mbps, the arrival rate $\lambda = 100$ s^{-1}, the average file size $f = 6$ Mb, the utilization $\rho = \lambda f/C = 0.75$, and the sample period $t = 2$ s. For these parameters, users start to share the bandwidth when there are more than $C/R_{\max} = 10$ active users. In our experiments, we work with the threshold $m = 20$, meaning that we look at the probability that the number of active users does not exceed 20, implying that the per-user throughput is at least half of the maximum access rate (40 Mbps). In the model's stationary situation, the average number of active users is approximately 8.4, while the most probable state corresponds to 8 active users.

A. Impact of the System Parameters

In Fig. 1 we study the *effect of the measured begin and end states* (i.e., numbers of users at beginning and end of the sample period) on the probability that no more than m have been active simultaneously. A crucial observation is that, when varying the begin state between 0 and 10 — being the point beyond which users start to share the bandwidth — the probability of having at most m users hardly changes; we see the same when varying the end state between 0 and 10. This indicates that, as long as the number of users in the begin or end of the time interval is below $C/R_{\max}$, its precise number hardly affects the probability of exceeding m during the sample period.

It is remarked that in the setting considered in this paper we measure the per-user throughputs, rather than the number of users present. An immediate consequence of the above observation is that if the measured per-user throughput equals $R_{\max}$, then, while we do not know the exact number of users present, we do know that this number is at most $C/R_{\max}$ and that this suffices to get a good estimate of the probability of our interest.

A second important observation is that a begin or end state above 10 *does* significantly affect the probability $q_{m,i,j,t}$ of

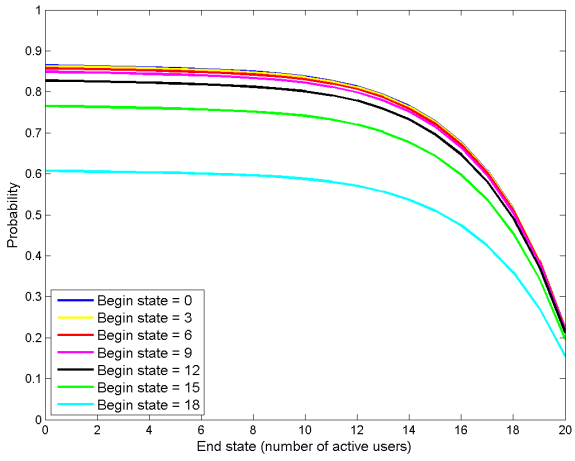


Fig. 1. Probability to have at most 20 active users as a function of the end state for several values of the begin state (both in number of active users)

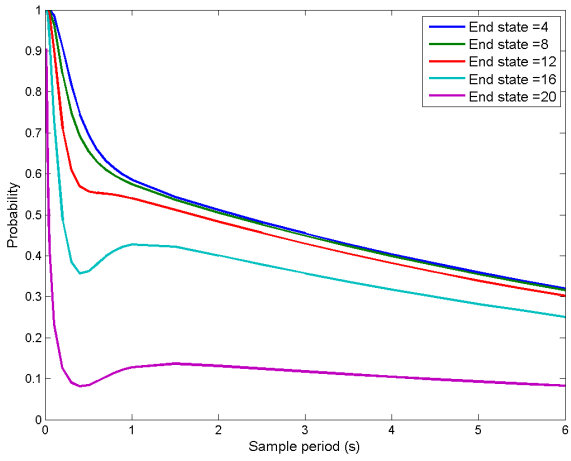


Fig. 2. Probability to have at most 20 active users as a function of the time between measurements for a begin state of 18 active users and several values of the end state (in number of active users)

having at most m users. This entails that the throughput measurements provide us with useful information. An alternative would be to not take into account the begin and end states, and to compute the probability of having maximally m users *in stationarity*, but this figure shows that such a procedure could lead to highly inaccurate estimates.

We now consider the *effect of the sample period* t on $q_{m,i,j,t}$; see Fig. 2. One would intuitively expect that the longer t, the higher the probability of crossing a threshold in that interval. This intuition provides us with the right qualitative behavior for 'moderate' begin and end states i and j, but, interestingly, breaks down when i and j are close to m; in the latter case $q_{m,i,j,t}$ first decreases, then increases, and finally decreases again. One should realize here that there are sound arguments that tell that $\mathbb{P}(\bar{X}_t \leq m \,|\, X_0 = i)$ decreases in t, but these arguments are *not* valid for the probability of our interest (where we condition on the end state as well). This observation further motivates the need for accurate methods to evaluate probabilities of this type.

Increasing the arrival rate λ or increasing the file size f, while keeping the other fixed, increases the utilization, and hence $q_{m,i,j,t}$ decreases. A more interesting experiment concerns the *effect of varying* f *and* λ *simultaneously, while keeping the utilization fixed* (at 75%). In many practical situations the network's capacity is upgraded when the utilization exceeds a certain predefined level, but the question is whether such a procedure is sound. The main conclusion we draw from Fig. 3 is that procedures of this kind are questionable, as the utilization provides us with just partial information about the behavior of the underlying system. Decreasing λ and increasing f increases the probability of our interest, the reason being that this change in parameters corresponds to switching to larger time units and thus decreasing t, which (except for cases in which both i and j are close to the 'overload value' m) leads to an increase of $q_{m,i,j,t}$, as seen before.

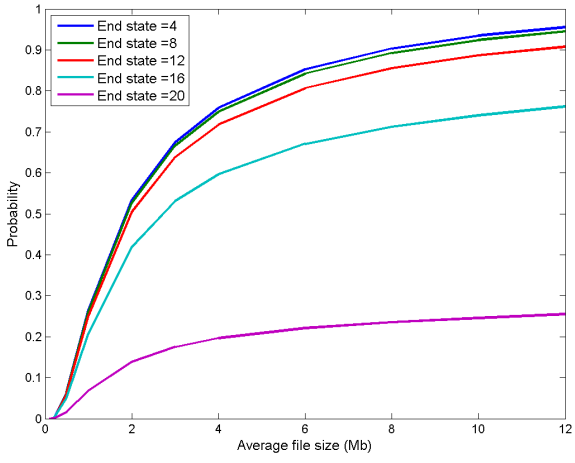


Fig. 3. Probability to have at most 20 active users as a function of the file size for a begin state of 5 active users and several values of the end state (in number of active users). The arrival rate is varied together with the file size in such a way that the utilization remains the same.

Summarizing, we have observed the following: 1) Per-user throughput samples facilitate a better analysis of the minimum per-user throughput in a time interval. 2) Per-user throughput samples are 'sufficient', in that it is not necessary to know how many users are active when this cannot be concluded from the per-user throughput measurement. 3) Typically, the longer the time between throughput samples, the less accurate the samples are (i.e., the larger the probability that the per-user throughput has dropped below a certain threshold). 4) To estimate the minimum per-user throughput within the sample period, it is necessary to know the user arrival rate λ and the average file size f; knowledge of the utilization $\lambda f/C$ does not suffice.

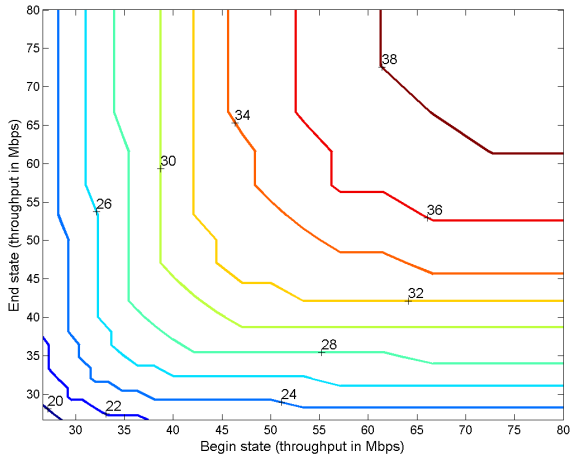
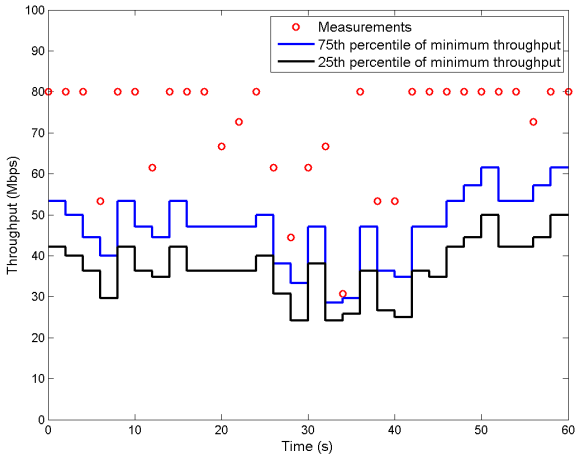
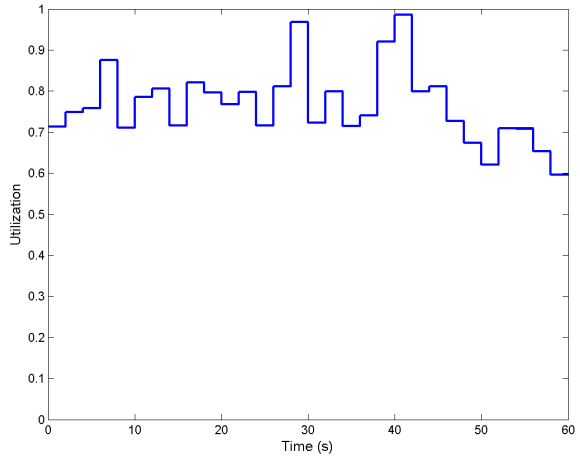


Fig. 4. 25th percentile of the minimum per-user throughput, given the begin and end throughput measurements (on the axes), for a utilization of 75%



(a) Realization of a simulation of the per-user throughput (red circles) and the corresponding (calculated) percentiles of the minimum per-user throughput between those measurements



(b) Realization of a simulation of the utilization, used as input for Fig. 5(a)

Fig. 5. Simulation with a sample period of 2 s.

B. Illustration of Practical Use

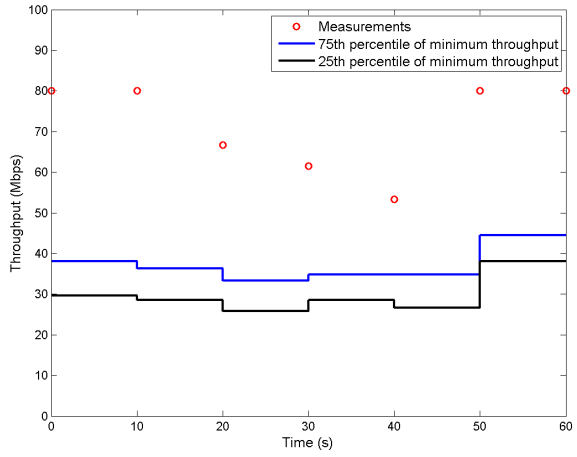
An operator may wish that an alarm is triggered as soon as the per-user throughput drops below a certain critical threshold (of, e.g., 30 Mbps), in line with the conditions agreed upon in the SLA. The approach we have introduced allows us to base this alarm not only on the *per-user throughput measurements themselves*, but also on our computations regarding the *minimum per-user throughput between measurements*. For example, an alarm can be raised if the probability that the per-user throughput has been less than 30 Mbps during the time between two measurements is larger than 25%. With our method, plots like the one presented in Fig. 4 can be produced for different measured utilizations (caused by different arrival rates). After each per-user throughput measurement, it can be found in the plot for which threshold the probability of interest was 25%. Note the plot's symmetry, which is a consequence of the reversibility of the underlying birth-death process.

Fig. 5(a) shows another type of chart that gives the operator insight in the per-user throughput between two subsequent measurements. In this (simulation-based) experiment we consider 31 sampled per-user throughputs, at time $0, 2, \ldots, 60$ s. Every sample period another λ has been used, resulting in the utilizations depicted in Fig. 5(b). The red circles show (simulated) per-user throughput measurements. With these input values (and knowing the values of C, $R_{\max}$, f, and t) we have calculated the probability distribution of the minimum per-user throughput within the corresponding sample periods— the blue and the black lines in Fig. 5(a) represent the corresponding 75th and 25th percentiles.

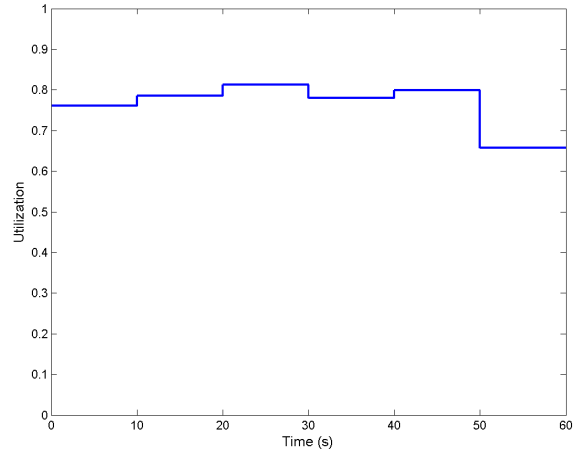
The graph shows that with our approach it is possible to estimate the minimum per-user throughput, without overloading the system with more measurements. The 25th percentile is interesting in practice as it shows the operator that the probability is 75% that all users experience a throughput larger than the given value. The 75th percentile shows that it is important to analyze what happens between the measurements, because the red circles and the 75th percentile curve are relatively far apart, which means that with a rather high probability (i.e., 75%) there is a substantial gap between the minimum per-user throughput during the interval on one hand, and the measured values at the end points on the other hand.

Comparing Figures 5(a) an 5(b) we conclude that the minimum per-user throughput cannot be predicted by either the throughput measurements or the utilization measurements only. It is therefore important to keep track of both performance measures.

Fig. 6 shows the results of the same simulation, now with a longer time between per-user throughput samples: 10 s instead of 2 s. In this figure the percentiles of the minimum per-user throughput are lower than in Fig. 5, showing that more frequent samples form a better estimate of the minimum per-user throughput in the interval.



(a) Realization of a simulation of the per-user throughput (red circles) and the corresponding (calculated) percentiles of the minimum per-user throughput between those measurements



(b) Realization of a simulation of the utilization, used as input for Fig. 6(a)

Fig. 6. Simulation with a sample period of 10 s.

V. Conclusions

This paper presents a novel method to estimate the minimum per-user throughput in access networks, on the basis of throughput samples ('speed tests'). The method relies on a queueing-theoretic description of the arrival of user requests and their processing. We have set up a framework that facilitates the computation of the probability distribution of the maximum number of active users (which is related to the minimum per-user throughput) between two measurements, given its begin and end state. We have illustrated the practical use of the method through a series of examples. These plots point out how the method can help operators to get insight in fluctuations of the throughput within the sample period, without actually performing additional measurements. Also, the impact of the key parameters, such as length of the sample period and the average flow size, is investigated.

A main conclusion is that the numerical experiments indicate that measurements of the link utilization alone do not give full information of the system's performance: regular throughput samples add valuable information that allows an accurate assessment of the minimum per-user throughput. A second observation is that knowledge of (the distribution of) the minimum throughput within the sample periods is highly relevant, as the minimum per-user throughput between throughput measurements may differ significantly from the measured values: substantial fluctuation within the sample period is likely for a broad range of parameter values.

We have developed a promising method to assess the minimum per-user throughput in access networks. Further research is needed to validate the method by comparing the theoretical results to real traffic traces and to assess the sensitivity to Poisson request arrivals and exponential file sizes. Another interesting topic for further research lies in the analysis of the frequency at which the per-user throughput drops below a certain level during the sample period, and if it does, how long such an overload situation persists.

References

[1] J. Schönwälder, "Simple Network Management Protocol (SNMP) Context Engine ID Discovery," RFC 5343, 2008.

[2] M. Mandjes and R. van de Meent, "Resource dimensioning through buffer sampling," *IEEE/ACM Transactions on Networking*, vol. 17, no. 5, pp. 1631–1644, 2009.

[3] B. Anjum, H. Perros, X. Mountrouidou, and K. Kontovasilis, "Bandwidth allocation under end-to-end percentile delay bounds," *International Journal of Network Management*, vol. 21, no. 6, pp. 536–547, 2011.

[4] O. Goga and R. Teixeira, "Speed measurements of residential internet access," in *Proc. Passive and Active Measurement Conference; Lecture Notes in Computer Science, vol. 7192*. Springer, 2012, pp. 168–178.

[5] R. de Oliveira Schmidt, A. Sperotto, R. Sadre, and A. Pras, "Towards bandwidth estimation using flow-level measurements," in *Proc. AIMS 2012, Dependable Networks and Services; Lecture Notes in Computer Science, vol. 7279*. Springer, 2012, pp. 127–138.

[6] F. Baccelli, B. Kauffmann, and D. Veitch, "Inverse problems in queueing theory and internet probing," *Queueing Systems*, vol. 63, no. 1-4, pp. 59–107, 2009.

[7] H. Tijms, *A First Course in Stochastic Models*. Wiley Online Library, 2003, vol. 2.

[8] L. Kleinrock, *Queueing Systems. Volume 1: Theory*. Wiley-Interscience, 1975.

[9] ——, *Queueing Systems. Volume 2: Computer Applications*. Wiley-Interscience, 1976.

[10] S. Ben Fredj, T. Bonald, A. Proutiere, G. Régnié, and J. W. Roberts, "Statistical bandwidth sharing: a study of congestion at flow level," *ACM SIGCOMM Computer Communication Review*, vol. 31, no. 4, pp. 111–122, 2001.

[11] T. Bonald and J. W. Roberts, "Congestion at flow level and the impact of user behaviour," *Computer Networks*, vol. 42, no. 4, pp. 521–536, 2003.

[12] H. van den Berg, M. Mandjes, R. van de Meent, A. Pras, F. Roijers, and P. Venemans, "QoS-aware bandwidth provisioning for IP network links," *Computer Networks*, vol. 50, no. 5, pp. 631–647, 2006.

[13] T. Bonald, J.-P. Haddad, and R. R. Mazumdar, "Congestion in large balanced multirate links," in *Proc. ITC23*, 2011, pp. 182–189.

[14] "Docsis 3.0: Third-generation transmission systems for interactive cable television services IP cable modems (MAC and Upper Layer protocols)," ITU Recommendation J.222.2, 2007.

[15] J. Abate and W. Whitt, "Numerical inversion of laplace transforms of probability distributions," *ORSA Journal on Computing*, vol. 7, no. 1, pp. 36–43, 1995.

[16] P. Den Iseger, "Numerical transform inversion using gaussian quadrature," *Probability in the Engineering and Informational Sciences*, vol. 20, no. 1, pp. 1–44, 2006.