Packet Waiting Time for Multiplexed Periodic On/Off Streams in the Presence of Overbooking

Michael Menth and Stefan Mühleck
University of Wurzburg, Institute of Computer Science, Germany
Email: {menth,muehleck}@informatik.uni-wuerzburg.de

***Abstract*—We present simple approximation formulae for the distribution of the packet waiting time of multiplexed periodic traffic. The multiplexed streams may have different periods and packet sizes. We show by extensive simulations the accuracy of the proposed methods. They are simpler than other existing formulae which make them attractive for engineers, applicable in practice, and easy to implement in switching devices. Packetized speech traffic has a periodic flow structure. Many compression techniques preserve it during talk phases but suppress the generation of packets during silence phases. When such on/off streams are multiplexed, advantage can be taken of their reduced flow rates by overbooking the link bandwidth. We adapt the proposed formulae to cope with on/off traffic and overbooking and validate them by extensive simulations. They can be applied for admission control in networks carrying different types of real-time traffic.**

***Index Terms*—waiting time distribution; on/off streams; multiplexing; overbooking.**

I. Introduction

Realtime applications usually create data packets in short periodic intervals to minimize the packetization delay. For instance, the G.711 voice codec [1] generates packets of 172 bytes every 20 ms leading to a rate of 68.8 kbit/s. Other codecs produce smaller packets at the same period, e.g., the G.729.1 codec [2] produces 38 bytes every 20 ms. The duration of the periods can often be configured so that other periods like 10 or 30 ms also exist. Constant bitrate circuit-switched data emulation services in UMTS also lead to strongly periodic streams. Although these data do not contribute the major traffic in today's communication networks, they still yield the major revenues. However, the major part of the traffic in the terrestrial access network of cellular communication systems like GSM or UMTS has realtime requirements and it is carried over low bitrate links of 1–4 Mbit/s due to the small traffic aggregation level. These links usually are leased lines and expensive so that operators wish to use them efficiently but without degrading the quality of service of the carried traffic. To that end, admission control (AC) is performed on the links and new connections are blocked if they would lead to extensive packet delay due to the multiplexing process on the link. Thus, the AC decision requires queuing formulae that predict the expected delay if another flow is admitted. The AC decision is based on statistical criteria, e.g., 0.01% of the packets may have a delay of 5 ms or longer. Therefore, the complementary cumulative distribution function (CCDF) of the packet waiting time caused by the multiplexing delay is of interest.

In [3, Chapter 15.2] a very accurate approximation for the CCDF of the packet waiting time of multiplexed homogeneous periodic flows is presented as well as a simpler exponential but less accurate approximation. For heterogeneous flows, i.e. for those with different periods or packet sizes, rather complex and numerically demanding expressions are provided in [3, Chapter 15.3] that we failed to implement correctly. As a consequence, we developed simple approximations and validated their accuracy by extensive simulations. They are very simple and easy to implement in switching devices.

Voice over IP (VoIP) applications and also wireless phones use more efficient codecs like iLBC [4], G.723.1 [5], or GSM 06.10 [6] that collect speech samples from periodic intervals and compress them. Most of them use silence detection to avoid the generation of data packets during silence phases. This reduces the flow rate and makes the effective output an on/off stream. The above mentioned formulae are not applicable for multiplexed on/off streams. In [3, Chapter 15.3] a method is presented to calculate the CCDF of homogeneous on/off sources if the sum of their peak rates does not exceed the link bandwidth, thus, it is not applicable in the presence of overbooking. We add some minor modifications to the method of [3, Chapter 15.3] and show by simulations that it leads to reasonable results that can be used for AC purposes. We further extend this approach to heterogeneous periodic on/off flows and again validate them by simulations. Our additions, albeit simple, enable AC for heterogeneous periodic on/off streams in the presence of overbooking which is the most frequent application scenario in practice.

The paper is structured as follows. Section II reviews related work. Section III proposes and validates new, simple approximations to calculate the CCDF of the packet waiting time of multiplexed heterogeneous periodic flows. Section IV presents and validates new methods to calculate the CCDF for multiplexed on/off traffic with a periodic base structure in the presence of overbooking. Section V summarizes this work.

II. Related Work

In this section, we give a short overview of related work considering the multiplex of periodic traffic and on/off traffic.

A. Queuing Formulae for Strictly Periodic Traffic

An excellent summary of formulae for the waiting time of multiplexed periodic traffic is given in [3, Chapter 15.3] that reports many results from [7], [8]. The authors consider so-called $n \cdot D/D/1$ queuing systems where several homogeneous periodic flows are multiplexed onto a common link and calculate the CCDF of the packet waiting time due to queuing. They present the closed-form solution Equation (1) that requires substantial computation time if many sources are multiplexed. Therefore, they also present Equation (2) which serves as a good approximation for relevant scenarios. Both formulae are simple and can be easily implemented. As our work is heavily based on them, we validate their accuracy in Section III-B4. An algorithmic solution to this queuing problem was given in [9] and an exact derivation was provided by [10], [11].

Furthermore, [3] considers the $\sum_{0 \leq i < k} n_i \cdot D_i/D/1$ queuing system where n_i periodic flows of k different classes with different inter-arrival times are multiplexed. Virtamo and Roberts [7] present a fairly complex algorithm calculating the CCDF of the resulting packet waiting times (VR-method). For our comparison we donwloaded the above algorithm from [12]. Unfortunately, for some input values we got inconsistent output from which we conclude that the above algorithm is numerically unstable.

Finally, [3] studies the $\sum_{0 \leq i < k} n_i \cdot D_i^{X_i}/D/1$ queue. That means, periodic streams of k different classes are multiplexed. All packets have the same size (like cells in ATM), but flows may have different periods and batch arrivals of X_i packets. We reformulate this system to $\sum_{0 \leq i < k} n_i \cdot D_i/D_i/1$, i.e., periodic flows from k different classes are multiplexed, each of them being characterized by its own period and packet size. A very complex method is given to calculate the CCDF of the packet waiting time. Again, we failed to implement that method, but in this case we could not obtain an implementation of the algorithm. Therefore, we believe that this approach might be good, but it is not simple and accessible enough to be used in practice.

B. Queuing Formulae for On/Off Traffic

We give a brief overview on work regarding the multiplex of on/off traffic. Some methods are based on fluids and cannot account for packet scale queueing. Others are rather coarse approximations or cannot cope with overbooking.

1) Approximative Solution for Multiplexed On/Off Fluids with Overbooking: The well known Anick-Mitra-Sondhi (AMS) approach [13] yields the waiting time distribution of multiplexed on/off fluids for infinite buffers. Fluids describe continuous flows of information that are not partitioned in packets, i.e., we can think of them as infinitesimally small packets, a bit stream instead of a packet stream. Therefore, AMS cannot describe delay caused by packet scale queuing, but we will use it as a lower bound to validate our solutions in Section IV using the implementation provided at [12].

2) Waiting Time Distribution for Multiplexed On/Off Fluids: Bensaou, Roberts et al. [3], [14] derive an exact formula for the queue length distribution in the presence of an infinite buffer based on the results of Beneš [15] for fluid queues. Since for most practical problems it is not possible to evaluate this exact solution, approximations are developed and the asymptotic behavior is studied. The results for the waiting time are good for larger numbers of connections and moderate link utilization. The approximation does not account for congestion effects that may arise due to long term correlation of on/off traffic.

3) Approximation Based on Modulated Periodic Arrivals $\sum D_i/D/1$ for Systems without *Temporary Overload:* Ramamurthy and Sengupta studied the superposition of periodic sources that arrive according to a Poisson model [16]. The sources have an exponential inter-arrival time distribution with rate $\frac{1}{\lambda}$ and general holding times during which they send periodic traffic. They derive the stationary waiting time distribution for n multiplexed sources, i.e. $\sum_{0 \leq i < n} D_i/D/1$. Then, they calculate the waiting time distribution of the overall system under the assumption that the overall rate of the calls never exceeds the link bandwidth. They achieve that by weighting the packet waiting time distributions for i active flows with the probability p_i for i active flows and the number of transmitted packets i such that the overall packet waiting time distribution is calculated according to Equation (12). However, the meaning of this analysis is different from ours: while we consider a fixed number of multiplexed flows with on/off characteristics, [16] studies a variable number of strictly periodic flows whose overall rate cannot exceed the link bandwidth.

4) More Approximations Based on Different Queuing Models: Baiocchi et al. approximate the arrival process by a Markov modulated Poisson process in [17]. They give an approximation of the cell loss probability for different buffer sizes. Stern and Ganguly [18], [19] calculate the queue length distribution of multiplexed packet streams with finite buffer capacity. Sriram and Whitt [20] approximate the queue length distribution of a packet multiplexer with a special two parameter $GI/GI/1{-}\infty$ approximation tool [21]. Humblet et al. [22] study the effect of an added smoothing delay on the resulting packet waiting time.

III. Packet Waiting Time of Multiplexed Periodic Flows

In this section, we first review existing queuing formulae for multiplexing homogeneous periodic flows and illustrate the queuing behavior of periodic packet arrivals. Then we extend the formulae to heterogeneous flows with equal periods but different packet sizes, different periods but equal packet sizes, and different periods and packet sizes, and validate these methods by simulations.

A. Notation

In the following, we denote periods, i.e. packet inter-arrival times, by a, packet sizes by b, link bandwidths by c, and transmission time of packets by $d = \frac{b}{c}$. Given n multiplexed homogeneous flows, the relative offered load is $\rho = \frac{n \cdot d}{a}$. It is the system utilization when packet loss does not occur. In case of $\rho < 1$, this can be achieved for multiplexed periodic traffic by a moderately large multiplexing buffer.

B. $n \cdot D/D/1$: *Multiplexing Flows with Identical Periods and Packet Sizes*

We review an accurate but computationally rather demanding queuing formula for multiplexed homogeneous flows as well as a simple approximation. We compare them and validate their accuracy by simulations.

1) A Closed-Form Solution for the CCDF of the Waiting Time: When several periodic streams with the same period a are multiplexed onto a single link with sufficiently large capacity c, the multiplexing buffer is emptied at least once within the period a if the system is not overloaded. Hence, the waiting time is smaller than a period ($t < a$) and the waiting time of a packet depends only on the arrival instants of its preceding packets within the last period (i). We call the arrival instant of a flow within an observed period its phase. As the phase of a flow does not change in consecutive periodic intervals, the phase pattern of a superposition of several flows is also periodic (ii). Combining (i) and (ii), it follows that each packet belonging to the same flow faces the same waiting time in each periodic interval. Therefore, the packet waiting time is deterministic and depends on the phase pattern of the flow arrivals within a period. However, different realizations of such a process with the same number of multiplexed streams n lead to different phase patterns. Taking into account all possible phase patterns leads to a distribution function of the packet waiting time. This so-called $n \cdot D/D/1$ queueing system has been studied in [3, Chapter 15.2.1] and can be calculated by the following closed-form solution:

$$\begin{aligned} P(W>t) &= W_{CCDF}^{binom}(n,d,a,t) \\ &= \sum_{\frac{t}{d}<m\le(n-1)} \binom{n-1}{m} \cdot \left(\frac{m\cdot d-t}{a}\right)^{m} \cdot \left(1-\frac{m\cdot d-t}{a}\right)^{n-1-m} \cdot \\ &\quad \frac{a-(n-1)\cdot d+t}{a-m\cdot d+t}. \end{aligned} \quad (1)$$

We call this formula the "binomial closed-form solution". The relative offered load must be smaller than one ($\rho < 1$) and the waiting time t must be smaller than the period ($t < a$). However, this does not limit the applicability of the formula. The formula is computationally expensive for a large number of multiplexed flows n. Note that Equation (1) yields the CCDF of the real waiting time while the CCDF of the virtual waiting time is obtained using the same formula for $n+1$ instead of n customers.

2) An Approximation Formula for the CCDF of the Waiting Time: The following computationally efficient formula is a good approximation for Equation (1) if the relative offered load ρ is high enough [3, Chapter 15.2.2]. We call it the "exponential approximation".

$$\begin{aligned} P(W>t) &= W_{CCDF}^{exp}(n,d,a,t) \\ &\approx \exp\left(\frac{-2\cdot t}{d} \cdot \left(\frac{t}{(n-1)\cdot d} + 1 - \frac{n\cdot d}{a}\right)\right). \end{aligned} \quad (2)$$

3) Simulation of the CCDF of the Packet Waiting Time: The superposition of periodic flows leads to a non-ergodic process which is difficult to simulate. The phase pattern chosen at the initialization of the simulation fully determines all future packet waiting times as the rest of the process is deterministic. Moreover, all packets of each flow experience the same waiting time from the second simulated period on. Therefore, for strictly periodic systems it is enough to collect the waiting time data of the second simulated period. To get a statistically reliable estimate of the CCDF of the packet waiting time, it is essential to collect waiting times from many different phase patterns. Hence, we start many simulation runs (10^7 per reported CCDF) with different seeds, group their results into batches, derive from them CCDFs for the packet waiting time, and calculate confidence intervals from the CCDFs received by the independent batches. The obtained confidence intervals are very small as long as the CCDF values are in the order of magnitude of 10^{-5} or larger. Therefore, we do not show them in our figures for the sake of clarity. When simulating the superposition of periodic flows with different periods, the statistical data must be gained from several simulated periods until the process repeats.

In case of on/off modulated periodic streams (see Section IV), the number of flows in the on-phase mainly influence the packet waiting time distribution. Therefore, many periods must be simulated in a single simulation run to capture the impact of the on/off phases. We perform 50 long runs (10^7 packet arrivals), derive the CCDF of the packet waiting time, and calculate confidence intervals from the CCDFs of the individual runs. The confidence intervals are again small for probability values of 10^{-5} or larger; we omit them in the figures.

4) Validation of the Binomial Closed-Form Solution and the Exponential Approximation: We validate the accuracy of the above presented approximation methods by simulations. Figure 1 shows the CCDF of the packet waiting time for different utilization levels ρ and 10 flows. The x-axis shows the waiting time t as a fraction of the period and the y-axis indicates the probability that packets wait longer than t. The binomial closed-form solution is exact and its curves coincide with those from simulations. The curves of the exponential approximation show small deviations from those of the simulation for moderate utilization levels of up to $\rho = 0.7$. The approximation quality of both formulae increases with the number of flows n and gives a perfect fit for $n > 50$. However, the exponential approximation can be safely used for admission control since it yields an upper bound on the real CCDF.

5) Queuing Behavior of Periodic Traffic: To understand the queuing behavior of $n \cdot D/D/1$ systems, we study the impact of the number of multiplexed flows, the period, and the packet size or service time on the packet waiting times.

For the first experiment we keep the packet service time constant at $d = 1$ ms. We vary the number of the multiplexed homogeneous flows n and their period a such that $\frac{a}{n}$ is constant and choose the bandwidth c of the link such that that its utilization is $\rho = 0.9$. Figure 2(a) shows that the waiting time decreases with a decreasing period because the period is an

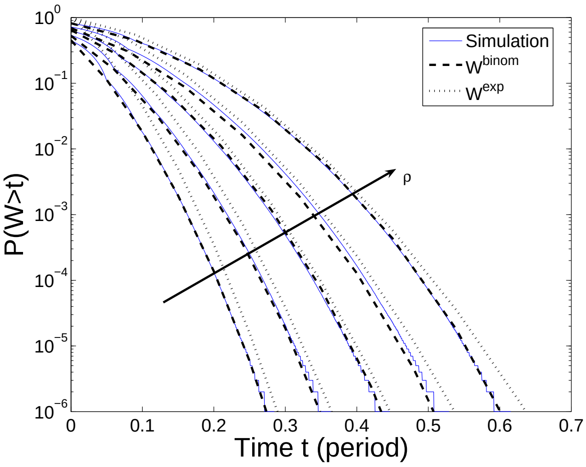


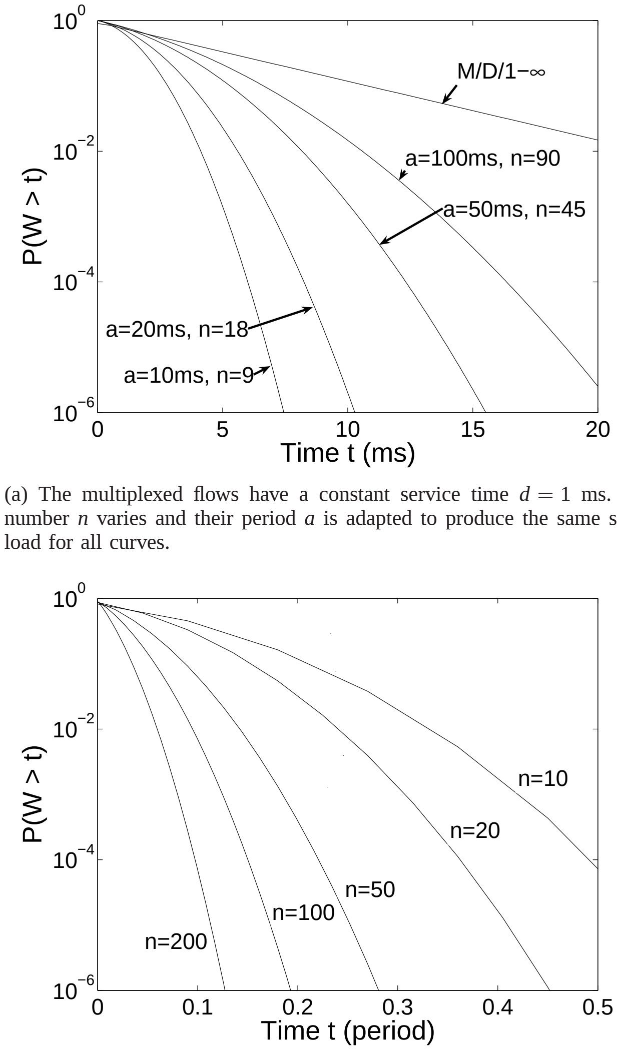
Fig. 1. CCDF of the packet waiting time for 10 multiplexed flows determined by simulation, the binomial closed-form solution, and the exponential approximation; the system utilizations are $\rho \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$; time t is given as a fraction of the period.

upper bound on the maximum waiting time. Furthermore, we observe that the CCDF of the packet waiting time of an $n \cdot D/D/1$ system converges with an increasing period a to the one of an $M/D/1-\infty$ system with the same packet service time $d = 1$ ms and system utilization $\rho = 0.9$. However, we also observe that the $M/D/1-\infty$ system overestimates the waiting time of periodic systems dramatically. Therefore, it is very important to take into account the periodic structure of multiplexed flows to calculate the CCDF of their packet waiting time.

For the second experiment we keep the period constant. We vary the number of the multiplexed homogeneous flows n and the packet service time d such that $n \cdot d$ is constant. We choose the bandwidth c of the link such that that its utilization is $\rho = 0.9$. Figure 2(b) shows the CCDF for the packet waiting time. The waiting time on the x-axis is given as a fraction of the period a. The figure shows that the packet waiting time decreases with an increasing number of flows or in other words, it increases with increasing packet size. There are two reasons for this observation. First, the transmission instants in the experiments with more flows are likely to be distributed more evenly over a period a than the transmission instants in the experiments with fewer flows. Second, the packet service time d in the experiments with more flows is shorter than in the experiments with fewer flows. Both issues effect that the traffic in the experiments with more flows is smoother on the time scale of a period than in the experiments with fewer flows.

C. $\sum_{0 \leq i < k} n_i \cdot D/D_i/1$: *Multiplexing Flows with Identical Periods but Different Packet Sizes*

We consider k classes of flows with the same periods a but different packet sizes b_i, $0 \leq i < k$. The corresponding queuing model is denoted by $\sum_{0 \leq i < k} n_i \cdot D/D_i/1$ for which Equations (1) and (2) are not applicable.



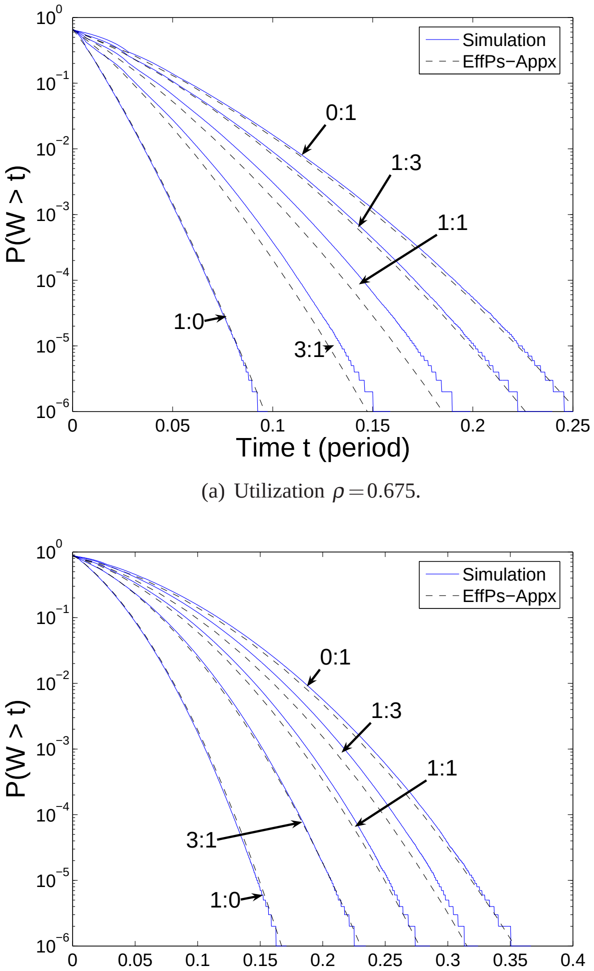
(a) The multiplexed flows have a constant service time $d = 1$ ms. Their number n varies and their period a is adapted to produce the same system load for all curves.

(b) The multiplexed flows have a constant period a. Their number n varies and their packet service time d is adapted to produce the same system load.

Fig. 2. CCDF of the packet waiting time for multiplexed periodic streams for a system utilization of $\rho = 0.9$

1) Approximation Formula Based on Effective Packet Sizes (EffPs): We propose a simple approximation for the packet waiting time of a $\sum_{0 \leq i < k} n_i \cdot D/D_i/1$ queuing system. It is based on the idea of an an effective number of flows n_{eff} and an effective packet size b_{eff}. These parameters are calculated from the characteristics of the multiplexed flows. The CCDF of the packet waiting time is approximated by either Equation (1) or (2) using the effective number of multiplexed flows n_{eff} and the effective packet service time $d_{eff} = \frac{b_{eff}}{c}$ which is derived from the effective packet size b_{eff}.

The most intuitive approach defines the effective number of flows by the number of multiplexed flows $n_{eff}^{intuitive} = \sum_{0 \leq i < k} n_i$ amd the effective packet size by the average packet size $b_{eff}^{intuitive} = \frac{1}{\sum_{0 \leq i < k} n_i} \cdot \sum_{0 \leq i < k} n_i \cdot b_i$. However, $n_{eff}^{intuitive}$ and $b_{eff}^{intuitive}$ significantly underestimate the packet waiting times.



(a) Utilization $\rho = 0.675$.

(b) Utilization $\rho = 0.9$.

Fig. 3. Approximated and simulated CCDFs of packet waiting time of a $\sum_{0 \le i<k} n_i \cdot D/D_i/1$ system; the link bandwidth is 1.138 Mbit/s, the flows have a common period of $a = 20$ ms but different packet sizes of $b_0 = 20$ bytes and $b_1 = 80$ bytes.

We propose a different approach which is given by Equations (3) – (6). It computes the effective packet size b_{eff}^{tmp} by averaging packet sizes of the multiplexed flows weighted by their contribution to the overall multiplexed traffic. The effective number of flows n_{eff} is chosen that n_{eff} homogeneous flows with packet size b_{eff}^{tmp} produce about the same traffic rate as the multiplexed flows in the queue. Finally the effective packet size is adjusted that the same traffic rate is exactly met. The rationale behind this cumbersome computation is the fact that Equations (1) and (2) require integer values for the number of flows.

$$s_i = \frac{n_i \cdot b_i}{\sum_{0 \le j<k} n_j \cdot b_j} \tag{3}$$

$$b_{eff}^{tmp} = \sum_{0 \le i<k} s_i \cdot b_i \tag{4}$$

$$n_{eff} = \left\lfloor \frac{\sum_{0 \le i<k} n_i \cdot b_i}{b_{eff}^{tmp}} \right\rfloor \tag{5}$$

$$b_{eff} = \frac{\sum_{0 \le i<k} n_i \cdot b_i}{n_{eff}}. \tag{6}$$

2) Validation of the EffPs-Approximation: To validate this approach, we choose two classes with the same period $a = 20$ ms but different packet sizes of $b_0 = 20$ bytes and $b_1 = 80$ bytes, and multiplex them onto a link. To facilitate the choice of the experiment parameters, we set the capacity of the link such that 16 flows of 64 kbit/s lead to a resource utilization of $\rho = 0.9$, i.e., we get $c \approx 1.138$ Mbit/s. We consider 5 different traffic mixes $s_0 : s_1 \in \{1:0, 3:1, 1:1, 1:3, 0:1\}$ to assess the accuracy of the formula for a link utilization of $\rho = 0.675$ and $\rho = 0.9$.

Figures 3(a) and 3(b) show the CCDF of the packet waiting time for the considered traffic mixes. Multiplexing flows with only 80 byte large packets leads to the longest packet waiting times while multiplexing flows with only 20 byte large packets leads to the shortest ones. The waiting time for the traffic mixes lies in between. For a utilization of $\rho = 0.9$, the packet waiting time is larger than for $\rho = 0.675$. The quality of the presented approximation is relatively good in all considered scenarios. However, the accuracy of the formula is limited if only a small number of flows are multiplexed, e.g. $n = 10$, or if packet sizes extremely differ, e.g., $b_0 = 20$ bytes and $b_1 = 500$ bytes or larger.

D. $\sum_{0 \le i<k} n_i \cdot D_i/D/1$: Multiplexing Flows with Different Periods but Identical Packet Sizes

We consider k classes of flows with different periods a_i, $0 \le i < k$ but the same packet sizes b. The corresponding queuing model is denoted by $\sum_{0 \le i<k} n_i \cdot D_i/D/1$.

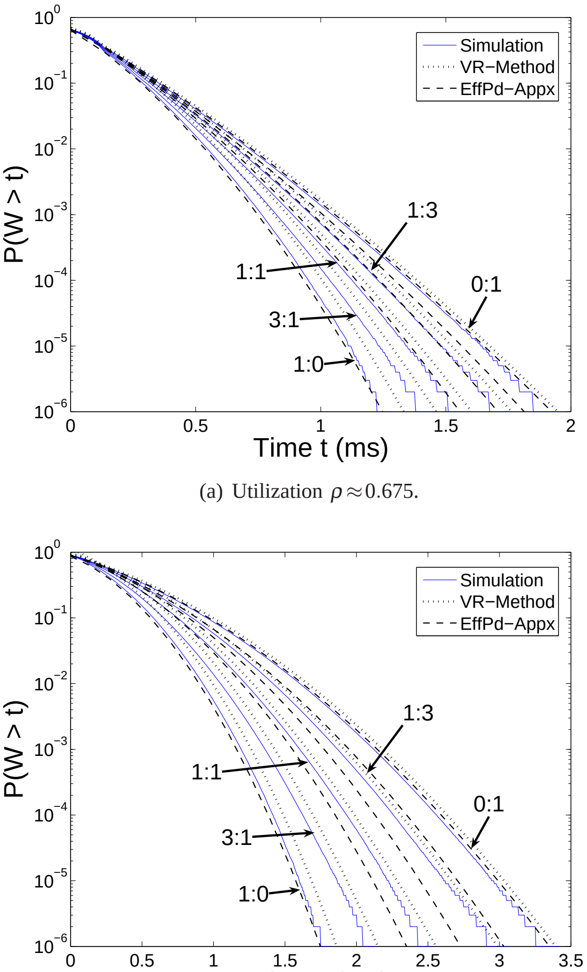
1) Approximation Formula Based on Effective Periods (EffPd): We propose the following simple, new approximation which is based on the effective period a_{eff}. This is calculated using the proportion of the traffic volume of class i with respect to the overall traffic volume: $s_i = \frac{\frac{n_i}{a_i}}{\sum_{0 \le j<k} \frac{n_j}{a_j}}$ and

$$a_{eff} = \sum_{0 \le i<k} s_i \cdot a_i. \tag{7}$$

We use the effective period a_{eff}, the packet service time $d = \frac{b}{c}$, and the number of multiplexed flows $n = \sum_{0 \le i<k} n_i$ as parameters for Equation (1) or (2) to calculate the CCDF of the packet waiting time.

2) Validation of the EffPd-Approximation: To validate this approach, we choose two classes with the same packet size $b = 20$ bytes but different periods $a_0 = 5$ ms and $a_1 = 20$ ms and multiplex them onto a link with a capacity of $c \approx 1.138$ Mbit/s. We consider 5 different traffic mixes $s_0 : s_1 \in \{1:0, 3:$

$1, 1:1, 1:3, 0:1\}$ to assess the accuracy of the formula for a link utilization of $\rho = 0.675$ and $\rho = 0.9$.



(a) Utilization $\rho \approx 0.675$.

(b) Utilization $\rho \approx 0.9$.

Fig. 4. Approximated and simulated CCDFs of packet waiting time of a $\sum_{0 \leq i<k} n_i \cdot D_i/D/1$ system; the link bandwidth is 1.138 Mbit/s, the flows have common packet sizes of $b = 20$ bytes but different periods of $a_0 = 5$ ms and $a_1 = 20$ ms. The traffic mix $s_0 : s_1$ is indicated in the figures.

Figures 4(a) and 4(b) present the CCDF of the packet waiting time derived by simulation, our new EffPd-approximation, and the complex VR-method [3]. Multiplexing homogeneous traffic with only short periods of $a_0 = 5$ ms leads to the shortest packet waiting times while multiplexing traffic with only long periods of $a_1 = 20$ ms leads to the longest ones. The waiting times of the traffic mixes lie in between. The curves of the VR-method are close to the simulation results. Our new approximation EffPd-Appx deviates significantly more from the real values, especially for the traffic mixes 3:1 and 1:1, but it is sufficiently accurate to get a rough estimate of the packet waiting time. As EffPd-Appx yields an upper bound on the CCDF, it may be used for admission control purposes. While the VR-method experiences numerical difficulties for some delay values and is algorithmically and computationally demanding, the new method is simple, fast, and has no known instabilities. The packet waiting time is at a utilization of $\rho = 0.9$ clearly larger than at $\rho = 0.675$. In both cases, the EffPd-approximation and the VR-method yield upper bounds for the CCDFs, therefore, they can be used for admission control purposes. For links with larger bandwidth the accordance of the approximation and the simulation results improves. However, the accuracy of the formula is limited if only a small number of flows are multiplexed, e.g., $n = 10$, or if periods extremely differ, e.g., $a_0 = 10$ ms and $a_1 = 100$ ms or larger.

E. $\sum_{0 \leq i<k} n_i \cdot D_i/D_i/1$: Multiplexing Flows with Different Periods and Packet Sizes

We consider k classes of flows with different periods a_i and different packet sizes b_i, $0 \leq i < k$. The corresponding queuing model is denoted by $\sum_{0 \leq i<k} n_i \cdot D_i/D_i/1$.

1) Approximation Formula Based on Effective Packet Sizes and Periods (EffPsPd): We propose the following new approximation which is based on the concepts of effective packet size and period. First, the effective period is calculated according to Equation (7). Then, the contribution s_i of class i to the overall multiplexed traffic is calculated by

$$s_i = \frac{\frac{n_i \cdot b_i}{a_i}}{\sum_{0 \leq j<k} \frac{n_j \cdot b_j}{a_j}}. \tag{8}$$

The rough estimate b_{eff}^{tmp} of the effective packet sizes is calculated based on Equation (4). It is used together with the effective period a_{eff} to compute the effective number of flows by

$$n_{eff} = \left\lfloor \sum_{0 \leq i<k} n_i \cdot \frac{b_i}{a_i} \cdot \frac{a_{eff}}{b_{eff}^{tmp}} \right\rfloor. \tag{9}$$

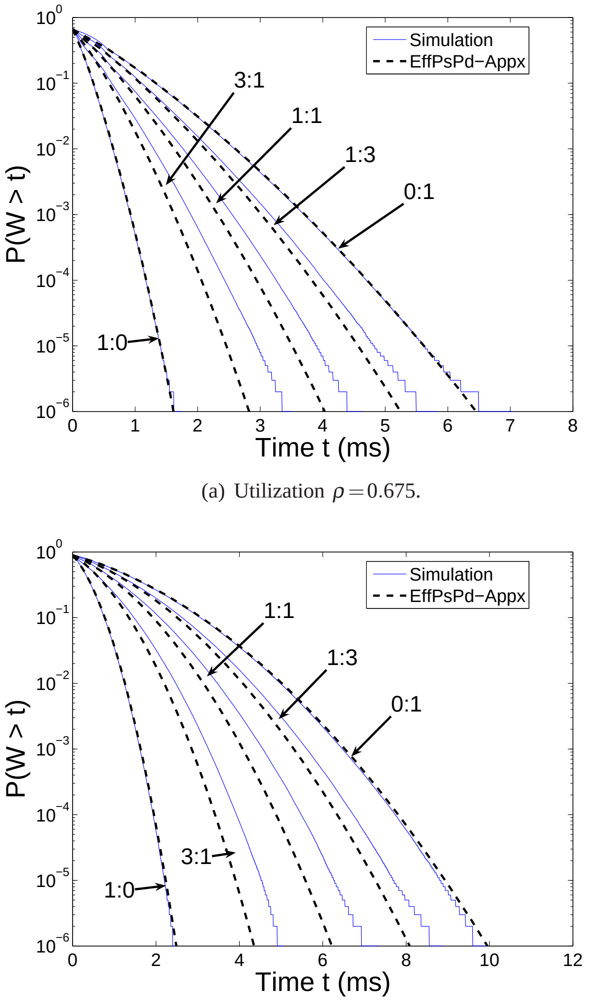
The effective packet size is adjusted to produce the same system load by

$$b_{eff} = \sum_{0 \leq i<k} n_i \cdot \frac{b_i}{a_i} \cdot \frac{a_{eff}}{n_{eff}}. \tag{10}$$

Finally, we use the effective number of multiplexed flows n_{eff}, the effective packet service time $d_{eff} = \frac{b_{eff}}{c}$, and the effective period a_{eff} as parameters for Equation (1) or (2) to calculate the CCDF of the packet waiting time.

2) Validation of the Proposed Approximation: To validate this approach, we choose two flow types with the same bitrate. They have different periods $a_0 = 10$ ms and $a_1 = 40$ ms and different packet sizes $b_0 = 20$ bytes and $b_1 = 80$ bytes. We multiplex them onto a link with a capacity of $c \approx 1.138$ Mbit/s. We consider 5 different traffic mixes $s_0 : s_1 \in \{1:0, 3:1, 1:1, 1:3, 0:1\}$ to assess the accuracy of the formula for a link utilization of $\rho = 0.675$ and $\rho = 0.9$.

Figures 5(a)–5(b) present the CCDF of the packet waiting time derived by simulation and our new EffPsPd-approximation. Both traffic types have different queueing properties. Multiplexing traffic with longer periods and largeer packet sizes yields longer waiting times. The simulation curves



(a) Utilization $\rho=0.675$.

(b) Utilization $\rho=0.9$.

Fig. 5. Approximated and simulated CCDFs of packet waiting time of a $\sum_{0\leq i<k} n_i\cdot D_i/D_i/1$ system; the link bandwidth is 1.138 Mbit/s, the flows have periods of $a_0=10$ ms and $a_1=40$ ms and packet sizes of $b_0=20$ bytes and $b_1=80$ bytes. The traffic mix $s_0:s_1$ is indicated in the figures.

are above the approximated curves which means that EffPsPd-Appx cannot be used as an upper bound on the real CCDF. However, for practical problems they may be useful to get an estimate of the waiting time given the fact that no alternative exists and that the $M/D/1$ queue heavily overestimates the waiting time. This observation holds for both moderate and high utilizations of $\rho=0.675$ and $\rho=0.9$. The limitations of EffPsPd-Appx are inherited from EffPs-Appx und EffPd-Appx.

IV. Extensions for Multiplexing Periodic On/Off Traffic in the Presence of Overbooking

In this section, we review quantitative models for on/off traffic with a periodic base structure. When such flows are multiplexed, overbooking can be applied to save bandwidth. We modify the equations presented above to calculate the CCDF of the packet waiting time under these conditions.

A. Modelling Compressed Voice Traffic

Many voice codecs like G.723.1 [5] or GSM 06.10 [6] use silence detection and suppress the generation of packets when the speaker is silent. Their output stream is basically periodic with interruptions leading to an on/off stream which is also called an on/off modulated periodic stream.

We have parameterized a simple two state on/off model for G.723.1 traffic in [23]. The packet payload is 24 bytes and packets are sent every 30 ms. The duration of the on/off phases are geometrically distributed with a mean of $E[D_{on}]=10.43$ s and $E[D_{off}]=13.09$ s which leads to a flow activity probability of $\alpha=\frac{E[D_{on}]}{E[D_{on}]+E[D_{off}]}=0.44332$. Thus, silence detection reduces the flow rates on average to 44%.

The model in [23] differs from other models in literature by longer on/off phases. They capture the packet generation on the time scale of sentences whereby some missing packets are just disregarded. In contrast, mean durations of only $E[D_{on}]=0.352$ s and $E[D_{off}]=0.650$ s are reported in [20]. Such values are obtained when phase lengths are determined by strictly contiguous on/off phases. However, [23] shows that the queuing behavior of the source model with the long phase durations approximates the one of compressed voice traces better than the source model with short phase durations. In particular, long phase durations describe the autocorrelation of the series of generated and suppressed packets within flow traces better than short phase durations.

because it captures the length of whole sentences that may contain small pauses. This results in only a few missing packets during an on-phase. In contrast, mean durations of only $E[D_{on}]=0.352$ s and $E[D_{off}]=0.650$ s are reported in [20] which are obtained when phase lengths are determined by strictly contiguous on/off phases. However, the queuing behavior of the source model with the long phase durations approximates the one of compressed voice traces better than the source model with short phase durations [23] because long phase durations describe the autocorrelation of flow traces better than short phase durations.

When several such on/off streams are multiplexed onto a single link and if the sum of the peak rates of the multiplexed flows do not exceed the link bandwidth, the average link utilization is at most $\alpha=0.44332$. Therefore, the link bandwidth may be overbooked which introduces the risk of temporary overload when the number of active flows varies. Then, a queue arises and packets can face significantly larger waiting times than in a multiplexing system with constant bitrate sources. Therefore, the second objective of this work is to describe the CCDF of the packet waiting time for this scenario by a simple approximation formula.

B. Modulated $n\cdot D/D/1$: Multiplexing On/Off Periodic Flows with Identical Periods and Packet Sizes

We review a method from [3] to calculate the CCDF of the packet waiting time for multiplexed on/off flows without overbooking. We extend it to the case with overbooking, validate the new method by simulations, and compare its results with the fluid approximation by Anick-Mitra-Sondhi (AMS) [13].

1) CCDF for Multiplexed On/Off Flows without Overbooking: The modulated $n \cdot D/D/1$ queue in [3, Chapter 15.2.4] addresses the superposition of exactly n on/off sources with a voice activity factor of α. They essentially calculate the probability p_m of m active flows by

$$p_m = P(n,m,\alpha) = \binom{n}{m} \cdot \alpha^m \cdot (1-\alpha)^{n-m} \tag{11}$$

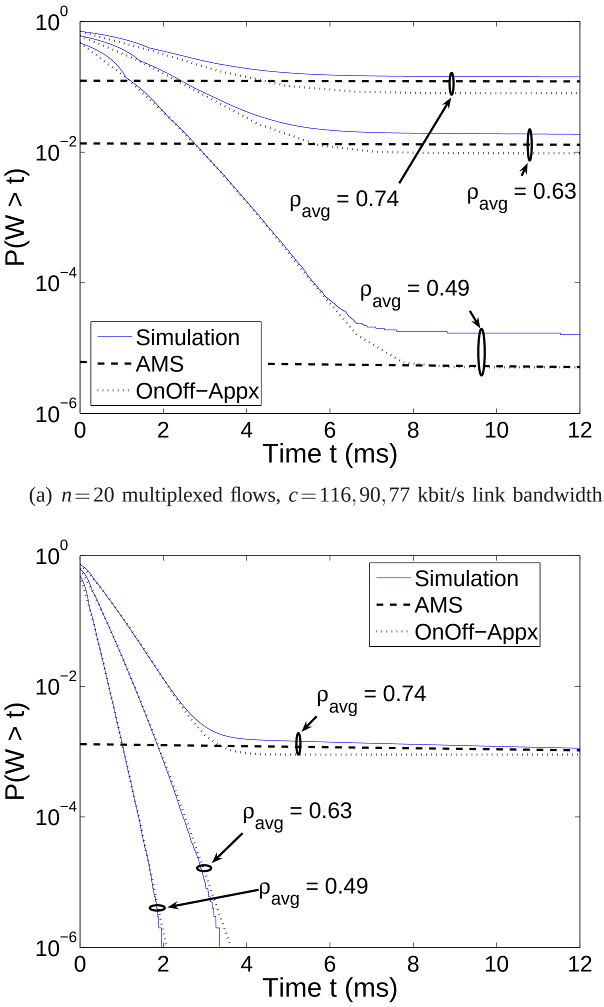
and compute $P(W > t)$ by

$$P(W > t) = \frac{\sum_{0 \leq m \leq n} m \cdot p_m \cdot W_{CCDF}(m,d,a,t)}{\sum_{0 \leq m \leq n} m \cdot p_m}. \tag{12}$$

For W_{CCDF} we can use either Equation (1) or (2). The approach presented in [3, Chapter 15.2.4] requires that the maximum offered relative traffic load $\rho_{max} = \frac{n \cdot d}{a}$ is smaller than 1; otherwise, $W_{CCDF}(m,d,a,t)$ is not defined. Hence, it is not possible to calculated the CCDF of the packet waiting time of multiplexed on/off traffic in the presence of overbooking.

2) Approximative Solution for Multiplexed On/Off Flows with Overbooking (OnOff-Appx): To approximate the CCDF of the packet waiting time for multiplexed on/off flows with overbooking, we propose to modify the CCDF $W_{CCDF}(m,d,a,t)$ for the packet waiting time for periodic flows by $W_{CCDF}(m,d,a,t) = 1$ for $\rho_{max} = \frac{n \cdot d}{a} \geq 1$. Then, we apply Equation (12) also when $\frac{n \cdot d}{a} \geq 1$ provided that the average load of the system does not exceed its capacity.

3) Validation of the Proposed Approximation: We validate the approximation method of Section IV-B2 by comparing its resulting CCDF with the one from simulations and from the AMS method introduced in Section II-B1. To that end, we multiplex on/off flows with a period of $a = 30$ ms, a packet size of $b = 24$ bytes, and an activity factor of $\alpha = 0.44332$. Thus, the flow rates are no longer constant and, therefore, we deal with average rates. We dimension the capacity of the link such that its average utilization is $\rho_{avg} \in \{0.49, 0.63, 0.74\}$. This value ρ_{avg} is rather a long-term average and differs from varying short-term averages. To obtain them, the link bandwidth is set to $c \in \{116, 90, 77\}$ kbit/s for $n = 20$ multiplexed flows and to $c \in \{579, 450, 383\}$ kbit/s for $n = 100$ multiplexed flows. This setting covers a wide range of long-term utilization values that are of interest when transmission capacities are overbooked by on/off traffic.



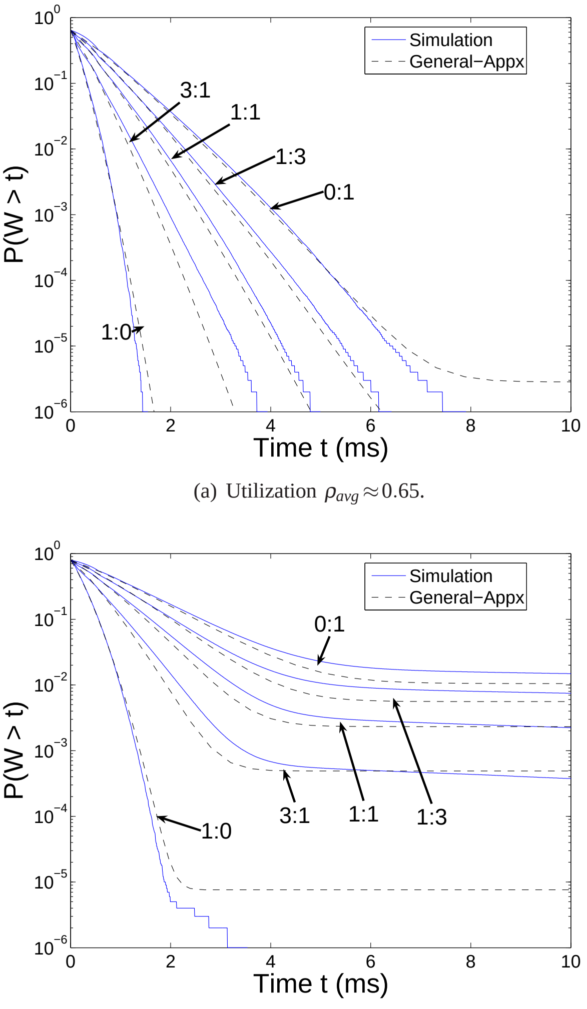
(a) $n = 20$ multiplexed flows, $c = 116, 90, 77$ kbit/s link bandwidth.

(b) $n = 100$ multiplexed flows, $c = 579, 450, 383$ kbit/s link bandwidth.

Fig. 6. Approximated and simulated CCDFs of packet waiting time of an on/off-modulated $n \cdot D/D/1$ system with G.723.1 traffic; the multiplexed flows have common periods of $a = 30$ ms, packet sizes of $b = 24$ bytes, and a voice activity factor of $\alpha = 0.44332$.

Figures 6(a) and 6(b) show the CCDFs of the packet waiting times. They look different from those for multiplexed strictly periodic traffic. Instead of decaying quickly, they converge to a certain probability value which is about the probability that overload occurs on the overbooked link. This probability is also approximated by the AMS method, but our approach also well captures the course of the real curves to that threshold. This threshold increases with increasing average utilization since the links in our experiment have then less bandwidth to carry the same traffic, i.e., the likelihood for overload becomes larger. For $n = 20$ flows, the new approximation captures the behavior of the CCDFs qualitatively, but significant deviations to the simulated curves are visible. For $n = 100$ flows, the approximation results of our new method are already quite accurate as they almost coincide with the results from simulations. The overload probability for $\rho_{avg} = 0.49$ and $\rho_{avg} = 0.63$ is now so small that we do not observe the horizontal line, i.e., the overbooked system is not likely to run into severe congestion. The approximation result is better than the one from the AMS methods as this disregards packet scale queuing completely and yields, therefore, too short waiting times, most of them are zero. However, AMS serves as lower bound for the real CCDF.

We choose the rather small values of the link capacities to show that our approximation works well also for a small multiplexing degree, which in general limits the accuracy of these approximation formulae. For more than $n=20$ multiplexed flows and larger bandwidths the approximation accuracy improves. The applicability of this method is also limited to sufficiently long on/off periods, i.e., very short on/off periods do not lead to excessive queuing delay. However, the formula is applicable when the traffic has the characteristics of compressed speech.



(a) Utilization $\rho_{avg}\approx 0.65$.

(b) Utilization $\rho_{avg}\approx 0.8$.

Fig. 7. Approximated and simulated CCDFs of packet waiting time of an on/off modulated $\sum_{0\le i<k} n_i\cdot D_i/D_i/1$ system; the link bandwidth is 1.138 Mbit/s, the flows have periods of $a_0=10$ ms and $a_1=40$ ms, packet sizes of $b=20$ bytes and $b=80$ bytes, and activity factors of $\alpha_0=0.75$ and $\alpha_1=0.375$. The traffic mix $s_0:s_1$ is indicated in the figures.

C. Modulated $\sum_{0\le i<k} n_i\cdot D_i/D_i/1$: Multiplexing On/Off Periodic Flows with Different Periods and Packet Sizes

In a more general case, we multiplex k different classes of on/off modulated periodic flows. Streams of different classes i send with different periods a_i, different packet sizes b_i, and different activity factors α_i, $0\le i<k$.

1) Approximation Formula Based on the Pattern of Active Flows (General-Appx): We extend Equation (12) and its modification in Section IV-B2 to this problem by considering all possible patterns of active on/off flows $(m_0,\dots,m_{k-1})$ together with their likelihood $p(m_0,\dots,m_{k-1})=\Pi_{0\le i<k}P(n_i,m_i,\alpha_i)$ (see Equation (11)). For each such pattern we can calculate the CCDF of the packet waiting time $W_{CCDF}(m_0,\dots,m_{k-1})$ according to the approximation presented in Section III-E1. Then, the CCDF of the packet waiting time for multiplexed on/off traffic can be approximated by $P(W>t)=\frac{X_{num}}{X_{denom}}$ where the expressions for the numerator and the denominator are

$$X_{num}=\sum_{0\le m_0<n_0}\dots\sum_{0\le m_{k-1}<n_{k-1}}\Big(p(m_0,\dots,m_{k-1})\cdot\big(\sum_{0\le i<k}m_i\big)\cdot W_{CCDF}(m_0,\dots,m_{k-1})\Big)$$

$$X_{denom}=\sum_{0\le m_0<n_0}\dots\sum_{0\le m_{k-1}<n_{k-1}}\Big(p(m_0,\dots,m_{k-1})\cdot\big(\sum_{0\le i<k}m_i\big)\Big).\quad(13)$$

2) Validation of the Proposed Approximation: We validate the General-Appx approach in a similar way as in Section III-E. We choose two classes with different periods $a_0=10$ ms and $a_1=40$ ms, packet sizes $b_0=20$ bytes and $b_1=80$ bytes, activity factors $\alpha_0=0.75$ and $\alpha_1=0.375$, and multiplex them onto a link with a capacity of $c\approx 1.138$ Mbit/s. We consider 5 different traffic mixes $s_0:s_1\in\{1:0,3:1,1:1,1:3,0:1\}$ to assess the accuracy of the formula for link utilizations of $\rho_{avg}=0.55$ and $\rho_{avg}=0.8$.

Figures 7(a)–7(b) show the CCDFs for all considered traffic mixes. Traffic mix $1:0$ has the shortest packet waiting time and traffic mix $0:1$ has the longest one, the others are in between. We observe a similar behavior regarding horizontal lines as in Figures 6(a)–6(b). However, there, the different overload probabilities, i.e. the y-value of the horizontal lines are caused by different average utilizations while the differences in Figures 7(a)–7(b) are caused by the different activity factors α_0 and α_1. For small utilization values, the overload is unlikely and Appx-General captures the packet waiting time quite well. The accordance of the results from approximation and simulation is rather good for all investigated traffic mixes and utilization values. The match of the simulated and approximated curves even improves for larger link bandwidths. Appx-General inherits its limitations from EffPsPd-Appx and OnOff-Appx, i.e., the CCDFs are inaccurate for a small number of flows with extremely large differences in packet sizes and periods or if the durations of on/off phases are rather short, e.g. 1 s or smaller.

V. CONCLUSION

The first contribution of this work are new simple formulae to calculate the CCDF of the packet waiting time of multiplexed periodic flows with different periods and packet sizes. We showed by extensive simulations that they are sufficiently accurate if periods and packet sizes differ by an order of magnitude. Complex formulae already exist for that objective, but they are hard to implement while our methods are simple and, therefore, well applicable in practice.

The second contribution of this work is the extension of the formulae mentioned above to on/off traffic. Such a formula already existed for homogeneous on/off flows whose traffic rates cannot exceed the link bandwidth. We extended this approach towards overbooking and studied its accuracy which is good if on/off phases are sufficiently long. In particular, the method works well if the traffic has the on/off characteristics of typical compressed speech [23].

The presented methods are useful in an environment with overbooked transmission lines to support admission decisions for new flows in order to avoid extensive packet loss and delay [24].

ACKNOWLEDGEMENTS

The authors would like to thank Brahim Bensaou, John Daigle, Paul Kühn, Jim Roberts, Bhaskar Sengupta, Phuoc Tran-Gia, Jorma Virtamo, and Ward Whitt for valuable pointers, personal and email discussions.

REFERENCES

[1] ITU-T, “G.711: Pulse Code Modulation (PCM) of Voice Frequencies,” Nov. 1988.

[2] ——, “G.729: Coding of Speech at 8 kbit/s Using Conjugate-Structure Algebraic-Code-Excited Linear Prediction (CS-ACELP),” Jan. 2007.

[3] J. Roberts, U. Mocci, and J. Virtamo, *Broadband Network Teletraffic - Final Report of Action COST 242*. Berlin, Heidelberg: Springer, 1996.

[4] S. Andersen, A. Duric, H. Astrom, R. Hagen, W. Kleijn, and J. Linden, “RFC3951: Internet Low Bit Rate Codec (iLBC),” Dec. 2004.

[5] ITU-T, “G.723.1: Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 And 6.3 kbit/s,” May 2006.

[6] European Telecommunications Standards Institute (ETSI), “European digital cellular telecommunications system (phase 1); full rate speech; transcoding (gsm 06.10),” http://webapp.etsi.org/workprogram/Report_WorkItem.asp?WKI_ID=399, Feb. 1992.

[7] J. W. Roberts and J. T. Virtamo, “The Superposition of Periodic Cell Arrival Streams in an ATM Multiplexer,” *IEEE Transactions on Communications*, vol. 39, no. 2, pp. 298 – 303, Feb. 1991.

[8] I. Norros, J. W. Roberts, A. Simonian, and J. T. Virtamo, “The Superposition of Variable Bit Rate Sources in an ATM Multiplexer,” *IEEE Journal on Selected Areas in Communications*, vol. 9, no. 3, pp. 378 – 387, Apr. 1991.

[9] A. E. Eckberg, “The Single Server Queue with Periodic Arrival Process and Deterministic Service Times,” *IEEE Transactions on Communications*, vol. 27, no. 3, pp. 556–562, Mar. 1979.

[10] A. Dupuis, A. Gravey, P. Boyer, and J.-M. Pitié, “The Output Process of the Single Server Queue with Periodic Arrival Process and Deterministic Service Time,” *Modelling and Performance Evaluation Methodology, Lecture Notes in Control and Information Sciences*, vol. 60, pp. 397–401, 1983.

[11] A. Gravey, “Temps d’attente et nombre de clients dans une file $nD/D/1$,” *Annales de l’Institut Henri Poincaré, section B*, vol. 20, no. 1, pp. 53–73, 1984.

[12] S. Aalto, E. Hyytiä, J. Lakkakorpi, I. Norros, A. Pirhonen, V. Timonen, and J. Virtamo, “The qlib library,” http://www.netlab.tkk.fi/qlib/.

[13] D. Anick, D. Mitra, and M. M. Sondhi, “Stochastic Theory of a Data Handling System with Multiple Sources,” *Bell Systems Technical Journal*, vol. 61, no. 8, pp. 1871–1894, 1982.

[14] B. Bensaou, J. Guibert, J. W. Roberts, and A. Simonian, “Performance of an ATM Multiplexer Queue in the Fluid Approximation Using the Beneš Approach,” *Annals of Operations Research*, vol. 49, pp. 137–160, 1994.

[15] V. E. Beneš, *General Stochastic Processes in the Theory of Queues*. Addison-Wesley, 1963.

[16] G. Ramamurthy and B. Sengupta, “Delay Analysis of a Packet Voice Multiplexer by the $\sum D_i/D/1$ Queue,” *IEEE Transactions on Communications*, vol. 39, no. 7, pp. 1107–1114, Jul. 1991.

[17] A. Baiocchi, N. Melazzi, M. Listanti, A. Roveri, and R. Winkler, “Loss Performance Analysis of an ATM Multiplexer Loaded with High-Speed On-Off Sources,” *IEEE Journal on Selected Areas in Communications*, vol. 9, no. 3, pp. 388–393, Apr. 1991.

[18] T. E. Stern, “A Queuing Analysis of Packet Voice,” in *IEEE Globecom*, San Diego, CA, USA, Nov. 1983.

[19] S. Ganguly and T. E. Stern, “Performance Evaluation of a Packet Voice System,” *IEEE Transactions on Communications*, vol. 37, no. 12, pp. 1394–1397, Dec. 1989.

[20] K. Sriram and W. Whitt, “Characterizing Superposition Arrival Processes in Packet Multiplexers for Voice and Data,” *IEEE Journal on Selected Areas in Communications*, vol. 4, no. 6, pp. 833–846, Sep. 1986.

[21] W. Whitt, “The Queuing Network Analyzer,” *Bell Systems Technical Journal*, vol. 62, no. 9, Nov. 1983.

[22] P. Humblet, A. Bhargava, and M. G. Hluchyj, “Ballot Theorems Applied to the Transient Analysis of nD/D/1 Queues,” *IEEE/ACM Transactions on Networking*, vol. 1, pp. 81–95, Feb. 1993.

[23] M. Menth, A. Binzenhöfer, and S. Mühleck, “Source Models for Speech Traffic Revisited,” *IEEE/ACM Transactions on Networking*, vol. 17, no. 4, pp. 1042–1051, Aug. 2009.

[24] M. Menth and S. Mühleck, “Admission Control for Speech Traffic in the Presence of Overbooking,” *Praxis der Informationsverarbeitung und Kommunikation (PIK)*, vol. 30, no. 4, pp. 227–233, Dec. 2007.