Admission Control for Speech Traffic in the Presence of Overbooking

Michael Menth and Stefan Muehleck

University of Würzburg, Institute of Computer Science
Am Hubland, D-97074 Würzburg, Germany
Phone: (+49) 931-888 6644, Fax: (+49) 931-888 6632
{menth,muehleck}@informatik.uni-wuerzburg.de

Abstract. Voice codecs use silence detection to reduce their average output rate and produce on/off packet streams. When several of such on/off packet streams are multiplexed onto a single link, the packets face both packet scale and burst scale congestion. If the sum of the peak rates of the flows does not exceed the link bandwidth, the distribution of the packet waiting time resulting from packet scale queuing can be calculated by a weighted $n \cdot D/D/1$ formula [1, 2]. However, advantage can be taken of the reduced flow rates by overbooking the link bandwidth in the sense that the sum of their mean rates is below the link bandwidth, but the sum of their peak rates exceeds the link bandwidth. Then, the well-known formula cannot be applied, but this is the most frequent application scenario. Therefore, we change the weighted formula by a simple adaptation and show by means of simulation that it well approximates the distribution of the packet waiting times in an overbooked system. We also clarify the relation of our new method to the well-known AMS [3] solution for on/off fluid flows. The simplicity of our approximation makes it attractive for engineers and applicable for admission control purposes in switching devices.

1 Introduction

Voice over IP (VoIP) applications use vocoders like GSM 06.10 [4] or G.723.1 [5] 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 makes the effective output of a single vocoder an on/off stream.

In the terrestrial radio access network of UMTS (UTRAN) several such on/off voice connections are multiplexed and carried on a single line from the NodeB to the radio network controller (RNC), respectively. This is depicted in Figure 1. The bandwidth of these lines typically ranges from 2 to 8 Mbit/s and, therefore, the waiting time of the packets may suffer from packet and burst scale queueing delay. The quality of service (QoS) requirements are usually formulated in a probabilistic way, i.e., the probability to exceed a given delay budget of, e.g., 5 ms should not be exceeded with a probability of more than 10^{-4} per node. To prevent excessive packet delay, admission control blocks new calls if the number of existing calls is large. However, to keep the blocking probability low, the links must be provided with sufficient capacity. Since rural low bit rate lines are rather expensive, it is attractive to overbook them which is possible due to the

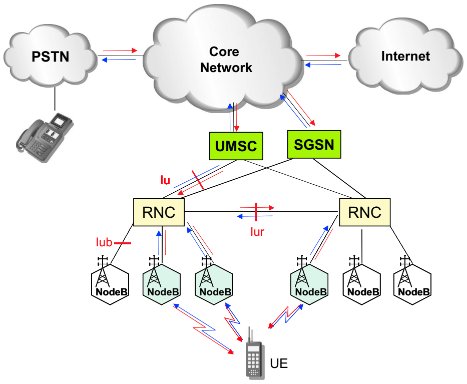


Fig. 1. The terrestrial radio access network of UMTS (UTRAN).

on/off nature of the compressed streams. That means that the sum of the peak rates of the carried flows may exceed the link bandwidth, but the system should be engineered that this is not likely to occur.

We present a simple formula to determine the distribution of the packet waiting time in such an overbooked system. It can be use to determine the number of calls that may be admitted by the admission control. It is based on a weighted sum of waiting time distributions from systems with k truly periodic packet streams without on/off modulations ($k \cdot D/D/1$). This has been done in a similar way in [1, 2] for systems without overload. However, the mentioned formula cannot be applied directly as the overbooking of the link bandwidth makes temporary overload possible. In this work, we adapt the weighted formula and show by simulations that it is sufficiently accurate. The approximation formula is simple, easy to apply, and intended for use by traffic engineers. Therefore, we finally present an application example concerning admission control in the presence of compressed voice traffic.

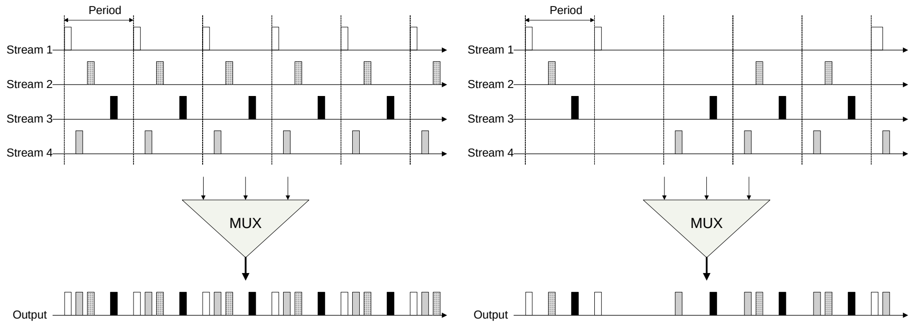
The paper is structured as follows. Section 2 explains the problem and the traffic model in more detail. In Section 3 we present existing approaches from the literature to calculate the packet waiting time distribution and develop our new formula. In Section 4 we compare the accuracy of the new approach by simulations of various parameter ranges. In Section 5 we briefly illustrate the application of the new method in practice. Section 6 summarizes our work and draws conclusions.

2 Multiplexing Speech Traffic

In this section, we describe the nature of speech traffic. The multiplexing of constant bit rate (CBR) flows can be modelled by an $n \cdot D/D/1$ queuing system. The waiting time based on this model cannot be approximated by similar queuing systems with Poisson arrivals or by fluid approximations. Codecs with silence detection yield quasi-periodic on/off streams for which we review quantitative models. When such flows are multiplexed, overbooking can be applied to save bandwidth.

2.1 Modelling Constant Bit Rate Streams by Periodic Arrivals

The G.711 [6] vocoder packetizes speech samples in periodic intervals of $t_{iat} = 20$ ms. Thus, it produces a periodic stream of UDP/IP packets with 172 bytes payload, i.e. the overall size of the packets is $b = 200$ bytes. When several of such streams are multiplexed onto a single link with sufficiently large capacity c, the multiplexing buffer is emptied at least once within the period t_{iat} if the system is not overloaded. Therefore, it needs only a limited size to avoid packet loss and the waiting time of a packet depends only on the arrival instants of its preceding packets within the last period (1). 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 periods, the phase pattern of the superposition of several flows is also periodic (2). This is shown in Figure 2(a).



(a) Constant bit rate sources.

(b) On/off sources.

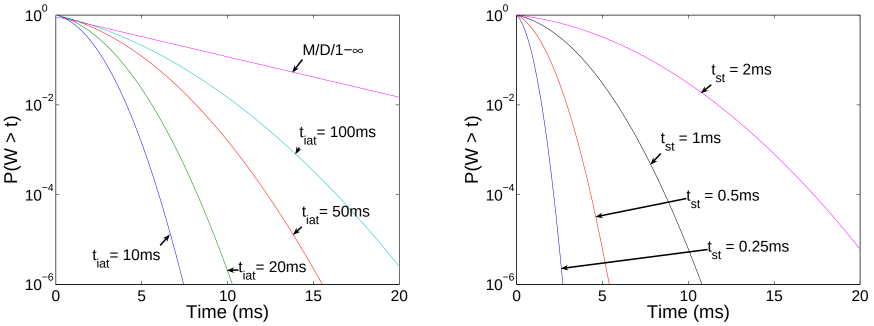
Fig. 2. Multiplexing (quasi-)periodic streams onto a single link.

Combining (1) and (2) follows that each packet belonging to the same flow faces the same waiting time in each period. 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 streams n may 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 [2] which proposes the following simple approximation formula (15.2.4):

$$P(W > t) = W_{CDF}(n, t_{st}, t_{iat}, t) \approx \exp\left(\frac{-2 \cdot t}{t_{st}} \cdot \left(\frac{t}{n \cdot t_{st}} + 1 - \frac{n \cdot t_{st}}{t_{iat}}\right)\right) \quad (1)$$

where $t_{st} = \frac{b}{c}$ is the service time for a single packet. The formula is only applicable for a system utilization of $\rho = \frac{n \cdot t_{st}}{t_{iat}} < 1$. The waiting time must be shorter than the period, i.e. $t < t_{iat}$, otherwise we set $P(W > t) = 0$. The simple approximation formula (1) produces good results for large utilization values ρ. Its accuracy is improved by the more complex Equation (15.3.15) in [2] that allows the calculation of the waiting time distribution of an $\sum D_i/D/1$ queuing system, i.e., flows may have different periods but the same packet size. To understand the queuing behavior of $n \cdot D/D/1$ systems, we study the complementary cumulative distribution function (CCDF) of the packet waiting times in different experiments.

First, we assume that periodic flows are multiplexed onto a link with a capacity such that the packet service time is $t_{st}=1$ ms. We vary the length of the period t_{iat} and adapt the number n of multiplexed flows such that a link utilization of 90% is achieved. Figure 3(a) shows that with an increasing period, the CCDF of the packet waiting time approximates the one of an $M/D/1-\infty$ system. However, we also observe that the $M/D/1-\infty$ system overestimates the waiting time of periodic systems with the same link utilization and packet sizes dramatically. Therefore, it is important to respect the periodic structure of multiplexed voice calls for the calculation of their packet waiting time.



(a) The packet service time $t_{st}=1$ ms is constant, the length of the period t_{iat} varies, and the number of flows n is adapted to achieve 90% link utilization.

(b) The number of flows $n=20$ is constant, the packet service time t_{st} varies, and the length of the period t_{iat} is adapted to achieve 90% link utilization.

Fig. 3. CCDF of the packet waiting time for multiplexed periodic streams.

We now assume that periodic flows are multiplexed onto the same link, but we keep the number of flows $n = 20$ constant. We vary the duration of the service time t_{st} and adapt the length of the period t_{iat} proportionally to keep the link utilization at 90%. As the link utilization is below 100%, the period t_{iat} is an upper bound on the packet waiting time and, therefore, the average waiting time decreases with a decreasing service time, too. This is illustrated in Figure 3(b) for different values of t_{st}. Hence, for very small service times t_{st}, the waiting times also become very short. In particular, fluid flows do not face any waiting time when the link load is always below 100% since their pseudo-packets have infinitesimal length. As a consequence, fluid approximations cannot account for packet scale queuing.

Hence, the waiting time of periodic packet streams can neither be derived based on Poisson arrivals nor on fluid models.

2.2 Modelling the Output of Vocoders with Silence Detection by On/Off-Modulated Periodic Arrivals

In contrast to the G.711 vocoder, the GSM 06.10 vocoder uses silence detection and does not generate packets when the speaker is silent. Consecutive periods during which packets are either generated or not, can be modelled by on- and off-phases. Thus, the output of this vocoder is quasi-periodic since it yields on/off modulated periodic packet streams.

We have developed and parameterized a simple two state on/off model for GSM 06.10 traffic in [7]. The packet payload is 33 bytes such that the overall packet size including the IP header is $b=61$ bytes and packets are sent every 20 ms such that the peak rate of a flow is $c_{peak}^{flow} = \frac{b}{t_{iat}} = 24.4$ kbit/s. The duration of the on/off phases are geometrically distributed with a mean of $E[D_{\mathit{on}}]=14.04$ s and $E[D_{\mathit{off}}]=18.39$ s which leads to a flow activity probability of $p_{active} = \frac{E[D_{\mathit{on}}]}{E[D_{\mathit{on}}]+E[D_{\mathit{off}}]} = 0.43293$. The model in [7] differs from other models in literature by longer on/off phases 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 $E[D_{\mathit{on}}]=0.352$ s and $E[D_{\mathit{off}}]=0.650$ s are reported in [8] 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 [7] because long phase durations describe the autocorrelation of flow traces better than short phase durations.

Figure 2(b) illustrates the multiplexing of several on/off streams onto a single link. If the sum of the peak rates of the multiplexed flows equals the link bandwidth, the link utilization is only $p_{active}=0.43293$. Therefore, the link bandwidth may be overbooked which introduces the risk of temporary overload when the number of active flows is exceptionally large. Then, a queue arises and packets face significantly larger waiting times than in a multiplexing system with the same average link utilization but continuously sending constant bit rate sources. The objective of this work is to describe the CCDF of the packet waiting time for this scenario by a simple approximation formula.

3 Approximation Methods for the Waiting Time of Multiplexed On/Off Processes

In the following, we give an overview of related work. Some methods are based on fluids and cannot account for packet scale queueing. Others are rather coarse approximations or cannot cope with overbooked systems. Finally, we present our new method.

3.1 Waiting Time Distribution for Multiplexed On/Off Fluids

The well known Annick-Mitra-Sondhi (AMS) approach [3] yields the waiting time distribution of multiplexed on/off fluids. Fluids describe continuous flows of information that are not partitioned into packets, i.e., we can think of them rather as a bit streams than as packet streams. As a consequence, the multiplexed information is immediately served as long as the capacity of the link is not exceeded. Therefore, the AMS solution cannot capture effects caused by packet scale queueing, but it well describes the impact of temporary overload caused by exceptionally many active sources. Thus, it is not suitable to calculate the waiting time distribution of individual multiplexed packets, but we use it as a lower bound to validate our solution in Section 4. To that end, we use the implementation provided at [9].

Bensaou, Roberts et al. [10, 2] derive an exact formula for the queue length distribution in the presence of an infinite buffer based on the results of Beneš [11] for fluid queues. The number of flows is assumed to be very large. Since for most practical problems it is not possible to evaluate this exact solution, approximates are developed and the asymptotic behavior is studied. The results for the waiting time are good for larger numbers of connections and not too high loads. The approximation does not allow for congestion effects that may arise because of the long term correlation of on/off traffic.

3.2 Measurement-Based Admission Control for On/Off Traffic

Brandt et al. [12] suggest measurement-based admission control for on/off traffic with exponential phase length durations. The rate of admitted flows is continuously measured and if it exceeds a certain threshold, no more flows can be admitted. The flows are carried over a link with a given transmission rate and packets exceeding this rate are dropped. The paper finds suitable values for the admission control threshold with a good tradeoff between packet loss and flow blocking probabilities. The analytical approach is based on a fluid flow model and calculates only packet loss while packet waiting times are not considered.

3.3 Approximation Based on a Markov Modulated Poisson Process

Baiocchi et al. study the superposition of on/off packet sources in [13]. They approximate the arrival process by a Markov modulated Poisson process and give an approximation of the cell loss probability for different buffer sizes. This is not applicable to our problem as voice traffic has a basic periodic structure which cannot be modelled by a Poisson arrival process.

3.4 Approximation Based on a Semi-Markov Process

Stern [14] and Ganguly and Stern [15] calculate the queue length distribution of multiplexed packet streams with finite buffer capacity. The number of sources in the on-phase, which determines whether the queue length is increasing or decreasing, is modelled as a continuous-time Markov chain called the phase process. It captures the long-term behavior of on/off sources. The overall multiplexing system is modelled as a semi-Markov process. Daigle and Langford [16] divided the states of this phase process in an underload and an overload region and embedded a Markov chain at the transition instants between these regions to approximate the waiting time distribution. Stern remarks that his method produces reasonable results for loss systems in the case of small buffers while it may fail for large buffers. Daigle reports that his approximation deviates from simulation results if the number of multiplexed connections is large (about $n = 30$).

3.5 Approximation Based on $GI/GI/1\!-\!\infty$

Sriram and Whitt [8] approximate the queue length distribution of a packet multiplexer with finite and infinite buffer. They apply a utility called queuing network analyzer (QNA) [17] which implements a special two parameter $GI/GI/1\!-\!\infty$ approximation. The on/off sources are modelled as renewal processes determined by two parameters, i.e. the mean arrival rate and the squared coefficient of variation. An elaborate method is developed to analyze the on/off traffic and to fit the squared coefficient of variation used for the renewal processes in the $GI/GI/1\!-\!\infty$ approximation. To represent the dependence of subsequent inter-arrival times of packets in the superposition of on/off processes, they propose the use of the index of dispersion (IDI). Let $S_k = X_1 + ... + X_k$ denote the sum of k consecutive i.i.d. (independent and identically distributed) inter-arrival times. The IDI is $I_X(k) = \frac{k \cdot Var[S_k]}{(E[S_k])^2} = \frac{VAR[\sum_{i=1}^{k} X_i]}{k \cdot E[X]^2}$ with $E[X]$ denoting the mean and $VAR[X]$ the variance of the X_i. The sequence of IDI values then forms the base for the calculation of the squared coefficient of variation for the QNA. This method works well for many connections, i.e. between tens to several hundreds depending on the link utilization.

3.6 Approximation Based on Modulated Periodic Arrivals $\sum D_i/D/1$ for Systems *without* Temporary Overload

Ramamurthy and Sengupta study the superposition of periodic on/off sources that arrive according to a Poisson model. That means that the sources have an exponential inter-arrival time distribution with rate $\frac{1}{\lambda}$ and general holding times. The multiplexer has an infinite buffer size [1]. They derive the stationary waiting time distribution for k multiplexed sources $\sum D_i/D/1$. 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 k active flows with the probability p_k for k active flows and the number of transmitted packets k such that the overall packet waiting time distribution is

$$P(W > t) = \frac{\sum_{0 \leq k \leq n} k \cdot p_k \cdot W_{CDF}(k, t_{st}, t_{iat}, t)}{\sum_{0 \leq k \leq n} k \cdot p_k} \tag{2}$$

This formula is investigated only under the assumption that the rate of k active flows does not exceed the link bandwidth c because only then $W_{CDF}(k, t_{st}, t_{iat}, t)$ exists. Temporary overload is not considered. The application of this method is only discussed for multiplexing a varying number k of strictly periodic flows. However, we also propose its application for multiplexing a fixed number of flows n out of which a varying number k is active.

The modulated $n \cdot D/D/1$ queue in 15.2.4 of [2] addresses the superposition of exactly n on/off sources with a voice activity factor of p_{active}. They essentially calculate p_k by

$$p_k = \binom{n}{k} \cdot (p_{active})^k \cdot (1 - p_{active})^{n-k} \tag{3}$$

and then compute $P(W > t)$ in the same way as in Equation (2). Again, an exact solution requires that the overall rate of the n sources never exceeds the link bandwidth and the formula has not been considered for temporarily overloaded systems. A similar approach has been presented in [18] to calculate loss probabilities.

3.7 Extending the Approximation Based on Modulated Periodic Arrivals $\sum D_i/D/1$ to Systems *with* Temporary Overlaod

Vocoders use silence detection to reduce flow rates whenever possible to increase the number of flows that can be transported over a link compared to without silence detection. Given a voice activity factor p_{active}, up to $1/p_{active}$ more traffic can be transmitted. However, when this kind of overbooking is realized, Equation (2) does not work since it requires that the system is never overloaded.

We also apply the simple approximation Equation (2) to calculate the expected waiting time for systems with capacity overbooking with the following modifications. In case of overload, i.e. if $\frac{k \cdot t_{st}}{t_{iat}} > 1$ holds, we simply approximate the waiting time by infinity, i.e. $W_{CDF}(k, t_{st}, t_{iat}, t) = 1$ for all t. This approximation produces reasonable results if the duration of an on-phase is very long. Then, an overloaded system remains overloaded for quite a while even if the number of flows that contribute to this on-phase is large. As a consequence, a large queue arises which justifies the coarse approximation of the waiting time by infinity.

4 Validation of the Approximation Accuracy

We validate the use of Equation (2) together with our modification in Section 3.7 for the approximation of the CCDF of packet waiting times when on/off sources are multiplexed in the presence of overbooking. First, we illustrate the influence of the average link load ρ_{avg} and the number of flows n on the instantaneous link load ρ_{inst}. Then, we compare the CCDFs calculated by the new approximation, by simulations based on the theoretical on/off model reviewed in Section 2.2, and by the AMS method [3] reviewed in Section 3.1. For all experiments in this section we use the parameters for the GSM 06.10 vocoder reported in Section 2.2.

4.1 Impact of Overbooking on the Instantaneous Link Load

In our experiments, the average link load is $\rho_{avg} = \frac{n \cdot c_{peak}^{flow} \cdot p_{active}}{c_l}$ and we dimension the link capacity c_l to obtain a certain target value of ρ_{avg}. We assume n independent flows in the system. Then, the number of active flows k is binomially distributed with parameters n and p_{active}. The arrival rate for the link depends on this number such that the instantaneous link load is $\rho_{inst} = \frac{k \cdot c_{peak}^{flow}}{c_l}$. In contrast to the link utilization it may take values larger than 1.0. Figure 4 presents the CCDF of ρ_{inst} for different average link load ρ_{avg} and different numbers of flows n. Most instantaneous link loads ρ_{inst} are centered around their mean value ρ_{avg} and their deviation from this value decreases with increasing n. The link is overbooked for all considered values of ρ_{avg}

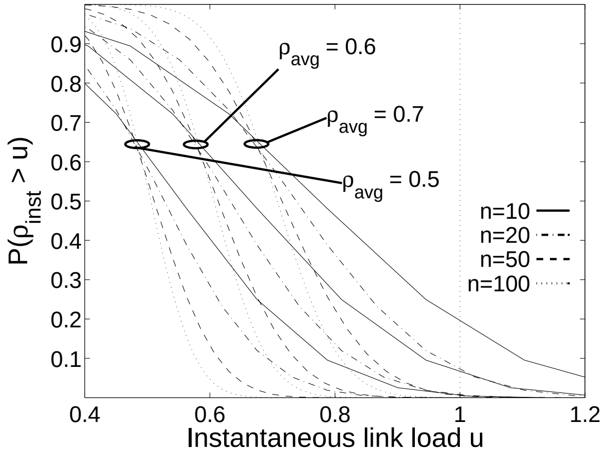


Fig. 4. CCDF of the instantaneous link load ρ_{inst} for different numbers of flows n and different average link loads ρ_{avg}.

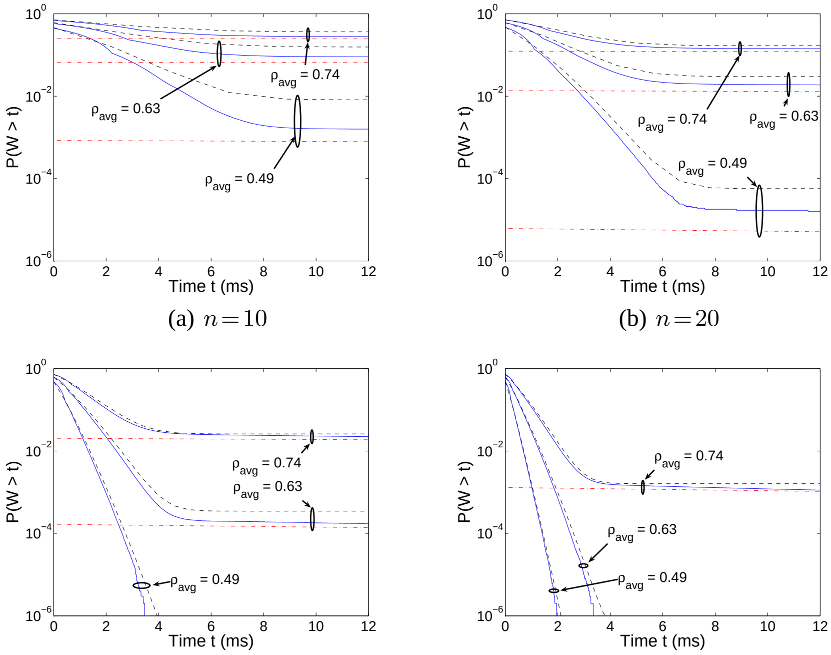
and the probability $P(\rho_{inst} > 1.0)$ that the instantaneous link load ρ_{inst} exceeds 1.0 is larger than zero. This is not always visible in the figure since these probabilities are sometimes rather small. However, they are quite high if the average link load is large, e.g. $\rho_{avg} = 0.7$, or if the number of flows is small, e.g. $n = 10$.

4.2 Validation of the New Approximation of the Waiting Time Distribution

We consider different numbers n of multiplexed flows and different average link loads of $\rho_{avg} = 0.49$, 0.63, and 0.74 which correspond to an overbooking of 110.8%, 142.4%, and 166.2%. We choose these values to meet the integer constraints of the formula for the calculation of the exact waiting time in $n \cdot D/D/1$ systems. We simulate the on/off modulated periodic traffic over $\frac{10^6}{n}$ different phase patterns for 1000 periods and cut off an initial warmup phase of 100 periods. Figures 5(a)–5(d) show the approximated CCDFs for the packet waiting time based on the adapted approximation of Equation (2), on simulations, and on the AMS fluid model.

The solid curves show the simulation results and illustrate the waiting time in the real system. Looking at $n = 100$ multiplexed flows we observe that the CCDFs quickly decrease and that long waiting times are not likely. However, if the average load ρ_{avg} is large, e.g. $\rho_{avg} = 0.74$,, the system is significantly overbooked (166.2%). Thus, the probability that the traffic arriving within a single period cannot be served within that time is rather large. As a consequence, a quite long queue arises leading to long waiting times under these conditions. Such a situation is more likely to occur for a small number n of multiplexed calls and for large values of average resource utilization ρ_{avg}.

The fluid approximation by the AMS method cannot account for packet scale delay, but it yields a lower bound for the waiting time W. It is rather accurate for large values



(a) $n = 10$

(b) $n = 20$

(c) $n = 50$

(d) $n = 100$

Fig. 5. Comparison of the CCDF of the packet waiting time computed by the approximation formula, simulations, and the AMS method.

of W because the AMS method accounts for the burst scale delay resulting from the on/off behavior of the periodic sources. Our new approximation captures the behavior of the simulated curves. Its accuracy is rather approximative for a small number of multiplexed calls ($n = 10$), but it leads to estimates usable in practice for a medium number of calls ($n = 20, 50$), and to fairly good matches for a large number of calls ($n = 100$). We observe that the new approximation is conservative in the sense that it overestimates the CCDF of the waiting times compared to the simulation results.

Hence, when admission control is performed based on this approach, the number of admissible calls is not overestimated. However, a small probability remains for temporary link overload. A system in practice has only a limited buffer which also leads only to a limited delay due to packet drops. This can be better tolerated by many realtime applications than excessive delay for all packets.

5 Application of the Formula for Admission Control

We apply Equation (2) together with our modification to perform admission control with safe overbooking in the presence of on/off traffic. We assume that packets must meet a waiting time of at most 5 ms with a probability of at least $p_q = 99.99\%$. We determine the number of flows n_{p_q} that can be carried on a link with capacity c_l without violating this constraint. To that end, we calculate the p_q-quantile of the waiting time by $t^W_{p_q} = \inf_t(t : P(W \leq t) \geq p_q) = \inf_t(t : P(W > t) < 1-p_q)$. We consider three different link bandwidths with $c_l = 1$, 2, and 4 Mbit/s. We perform our study for the GSM 06.10 codec, i.e., flows have a peak rate of $c^{flow}_{peak} = \frac{b}{t_{iat}} = \frac{61 \text{ bytes}}{20 \text{ ms}} = 24.4$ kbit/s and an average rate of $c^{flow}_{avg} = c^{flow}_{peak} \cdot p_{active} = 24.4$ kbit/s $\cdot\ 0.43293 = 10.56$ kbit/s. Based on the average rate, at most n_{max} =95, 189, and 379 flows can be carried over the links. However, this corresponds to an overbooking of $\frac{1}{p_{active}} = 2.31$ when the peak rates of the flows are considered. The question in practice is how much overbooking can be allowed without violating the delay criterion?

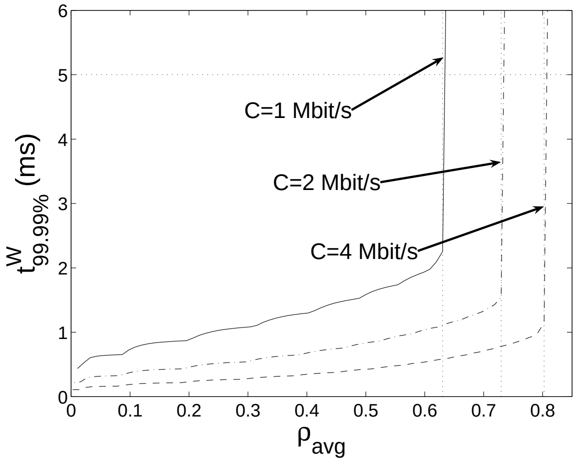


Fig. 6. 99.99%-quantile of the packet waiting time depending on the average link load ρ_{avg}.

A number of k active flows corresponds to an average load of $\rho_{avg} = k \cdot \frac{c^{flow}_{avg}}{c_l}$ and we calculate for each $0 \leq k \leq n_{max}$ the 99.99%-quantile of the waiting time using Equation (2) with our modifications. The results are presented in Figure 6. The quantiles $t^W_{99.99\%}$ increase with the average link load ρ_{avg} and they are smaller for larger link bandwidths. The figure shows that the links with 1, 2, and 4 Mbit/s may be utilized by at most $\rho^{99.99\%}_{avg}$ =0.63, 0.73, and 0.80 in order to prevent that the 99.99%-quantile of the waiting time exceeds 5 ms. This corresponds to $n_{99.99\%} =$ 60, 138, and 303 flows. Thus, the links can be overbooked to 141.5%, 164.7%, and 181.2% while meeting the delay criterion.

This example shows the usefulness of the derived approximation formula for admission control in systems with homogeneous on/off flows. In practice, however, multiplexed flows have distinct periods and packet sizes. Therefore, we intend to extend and validate this method in future work.

6 Conclusion

In this work, we considered compressed voice traffic being multiplexed onto a single link. Voice compressed by vocoders with silence detection results in a periodic packet stream with an on/off structure. This effectively reduces the average flow rate since packets are generated only with probability p_{active}. Thus, the link can be overbooked in the sense that the sum of the peak rates of all flows exceed the link capacity, but the sum of the peak rates of the flows in the on-state does not exceed the link capacity. However, as the number of active flows changes, there is still a chance for link overload and excessive delay.

A method from the literature $[1, 2]$ can calculate the distribution function of the packet waiting time in such a system when the system is not overbooked. We have extended this approach that it can also be applied for systems with overbooking. Our solution is only an approximation, but extensive simulation studies have shown that it produces useful results for 20 or more multiplexed calls and the approximation accuracy is excellent for 100 calls or more.

Finally, we illustrated the use of the extended formula for admission control. Our approach is simple such that it is likely to be used by engineers in practice, e.g. in access networks of mobile communication systems like UMTS. In future work, we plan to extend our solution towards traffic mixes of constant bit rate and on/off flows with periodic base structure, different periods, and packet sizes.

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 as well as personal and email discussions.

References

1. Ramamurthy, G., Sengupta, B.: Delay Analysis of a Packet Voice Multiplexer by the $\sum D_i/D/1$ Queue. IEEE Transactions on Communications **39**(7) (1991) 1107–1114
2. Roberts, J., Mocci, U., Virtamo, J.: Broadband Network Teletraffic - Final Report of Action COST 242. Springer, Berlin, Heidelberg (1996)
3. Anick, D., Mitra, D., Sondhi, M.M.: Stochastic Theory of a Data Handling System with Multiple Sources. Bell Systems Technical Journal **61**(8) (1982) 1871–1894
4. 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 (1992)
5. ITU-T: (G.723.1: Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 And 6.3 kbit/s)
6. CounterPath Solutions, Inc.: X-Lite 3.0. (http://www.xten.com/)
7. Binzenhöfer, A., Menth, M., Mühleck, S.: Source Models for Speech Traffic. In: currently under submission, http://www3.informatik.uni-wuerzburg.de/staff/menth/Publications/Menth07-Sub-13.pdf. (2007)
8. Sriram, K., Whitt, W.: Characterizing Superposition Arrival Processes in Packet Multiplexers for Voice and Data. IEEE Journal on Selected Areas in Communications **4**(6) (1986) 833–846

9. Aalto, S., Hyyti, E., Lakkakorpi, J., Norros, I., Pirhonen, A., Timonen, V., Virtamo, J.: The qlib library. (http://www.netlab.tkk.fi/qlib/)
10. Bensaou, B., Guibert, J., James W. Roberts, A.S.: Performance of an ATM Multiplexer Queue in the Fluid Approximation Using the Beneus Approach. Annals of Operations Research **49** (1994) 137–160
11. Beneš, V.E.: General Stochastic Processes in the Theory of Queues. Addison-Wesley (1963)
12. Brandt, A., Brandt, M., Rugel, S., Weber, D.: Admission Control for Realtime Traffic: Improving Performance of Mobile Networks by Operating on Actual Throughput. In: IEEE Wireless Communications and Networking Conference (WCNC). (2005)
13. Baiocchi, A., Melazzi, N., Listanti, M., Roveri, A., Winkler, R.: Loss Performance Analysis of an ATM Multiplexer Loaded with High-Speed ON-OFF Sources. IEEE Journal on Selected Areas in Communications **9**(3) (1991) 388–393
14. Stern, T.E.: A Queuing Analysis of Packet Voice. In: IEEE Globecom, San Diego, CA, USA (1983)
15. Ganguly, S., Stern, T.E.: Performance Evaluation of a Packet Voice System. IEEE Transactions on Communications **37**(12) (1989) 1394–1397
16. Daigle, J.N., Langford, J.D.: Models for Analysis of Packet Voice Communications Systems. IEEE Journal on Selected Areas in Communications **4**(6) (1986) 847–855
17. Whitt, W.: The Queuing Network Analyzer. Bell Systems Technical Journal **62**(9) (1983)
18. Jacobsen, S.B., Moth, K., Dittmann, L., Sallberg, K.: Load control in ATM networks. In: International Switching Symposium. Volume 5. (1990) 131–138