Determining the tolerable load generated by a set of packet-based phones on a multiplexing node

D. De Vleeschauwer[a], A. Van Moffaert[a], M.J.C. Büchli[a], J. Janssen[a], G.H. Petit[a],
B. Steyaert[b], H. Bruneel[b]

[a]Alcatel Bell, Network Strategy Group, Network Architecture Team,
Francis Wellesplein 1, B-2018 Antwerp, Belgium[*]

[b]Ghent University, Department TELIN, SMACS Research Group,
Sint-Pietersnieuwstraat 41, B-9000 Ghent, Belgium[*]

Abstract. To determine the voice load that can be supported by a packet-based node, we study the MMBP/D/1 model. Using this load the number of phones that can be supported by that node is calculated. We investigate the case of homogeneous and heterogeneous traffic and phones with and without silence suppression, and study the impact of the link rate, the activity grade and the codec bit rate. We also compare the number of phones that can be supported by the node according to the MMBP/D/1 model with the number of phones that can be supported according to the M/D/1 model.

1. INTRODUCTION

Packetized voice transport (e.g., Voice over IP (VoIP)) is more flexible than circuit-switched voice transport as an active phone does not occupy a trunk (i.e., a fixed portion of the capacity) for the entire duration of a phone call, but capacity is dynamically shared with all other active phones (and possibly with data sources). Unlike the (circuit-switched) Public Switched Telephone Network (PSTN), which switches 64 kb/s voice streams encoded in the G.711 format, a packet-based voice network can also carry any codec format that both communicating phones support. Moreover, in contrast to on a circuit-switched network, silence suppression can be exploited on a packet-based network, reducing the average codec bit rate even further.

However, a voice call routed over a packet-based network risks experiencing a degraded quality, because more delay and distortion are likely to be introduced with respect to a call switched over the PSTN. The bounds on the delay and the distortion are known and reported in [2]. The remaining issue is how to dimension the network elements (the packetizer, the network nodes, the dejittering mechanism, etc.) such that these bounds are met. This paper concentrates on the delay introduced in one node in the network.

---

[*] This work was carried out within the framework of the project LIMSON sponsored by the Flemish Institute for the Promotion of Scientific and Technological Research in the Industry (IWT).

The problem tackled here has some similarity with the dimensioning of a PSTN switch. Consider (circuit-switched) phones (that generate voice streams in the G.711 format) with an activity grade p. The activity grade p is determined by the average call duration and the average passive period. Both, the active and passive period, are assumed to be exponentially distributed. Each active phone requires a trunk in the switch. If no trunk is available when a phone attempts a call, the call is blocked. The (inverse) Engset formula [3] calculates the number of phones N that can be supported by a switch with N_{PSTN} trunks (i.e., of capacity $R_{link} = N_{PSTN}$ 64 kb/s) given a certain tolerated blocking probability. Here, we also consider phones with an activity grade p, but they are packet-based and not necessarily encode the voice in the G.711 format. Calls are (in principle) never blocked, but there is a restriction on the queuing delay introduced in the node. In fact, the number of phones N that can be supported by one packet-based multiplexing node of capacity R_{link} is calculated given that (a quantile of) the queuing delay has to be bounded by a certain value.

The paper is organized as follows. In Section 2 we describe the essential stages in the packetized transport of voice signals. We concentrate on the transport stage, where we focus on one network node. We estimate the delay budget that can be consumed in that single node. Section 3 describes the MMBP/D/1 model that we will use to calculate the queuing delay in a node. We do not develop the mathematics in detail, since the model was already studied in full detail in [5]. In Section 4 we use the MMBP/D/1 model to calculate the number of phones that can be supported by one node using the delay budget determined in Section 2. We investigate the impact of several parameters (the codec bit rate, the link rate and the activity grade) for homogeneous and heterogeneous traffic and for sources with and without silence suppression. We also compare the number of phones that can be supported according to the MMBP/D/1 model with the number of phones that can be supported according to the M/D/1 model. In the related literature, it is often assumed that due to the specific characteristics of voice traffic (i.e., relatively low bit rates and bursts of moderate lengths), an aggregate of a large number of these sources will more or less behave as a Poisson process. In this contribution we investigate to what extent and for which traffic parameters this assumption is valid. Finally, in the last section the main conclusions of the paper are summarized.

2. PACKETIZED VOICE TRANSPORT

In the packetized transport of digital voice there are three essential stages (see Figure 1). Although in this paper we consider VoIP, the theory developed here is applicable for any kind of packet-based network, albeit with different amounts of overhead per voice payload.

In the first stage, the digital voice signal is encoded and packetized. We consider three codec bit rates R_{cod} = 64 kb/s, 32 kb/s and 16 kb/s. To be able to easily compare the multiplexing behavior of these different codecs, we always take a payload size of 160 bytes. This means that the packetization delay is 20 ms, 40 ms and 80 ms, respectively. A packetization delay of 20 ms for the 64 kb/s is quite reasonable, but a packetization delay of 80 ms for the 16 k/s codec is a bit on the high side. However, it is not the aim of this paper to determine the optimal packetization delay. Because in VoIP the header consists of 20 IP bytes, 8 UDP bytes and 12 RTP bytes, the size of all voice packets is 200 bytes. As a result, when a phone is active (and talking), it produces a flow of IP packets of 200 bytes with a fixed inter-packet time, i.e., each period of the duration of this inter-packet time exactly one packet is produced. The inter-packet time is equal to 20 ms, 40 ms and 80 ms, for the 64 kb/s, 32 kb/s and 16 kb/s codec, respectively.

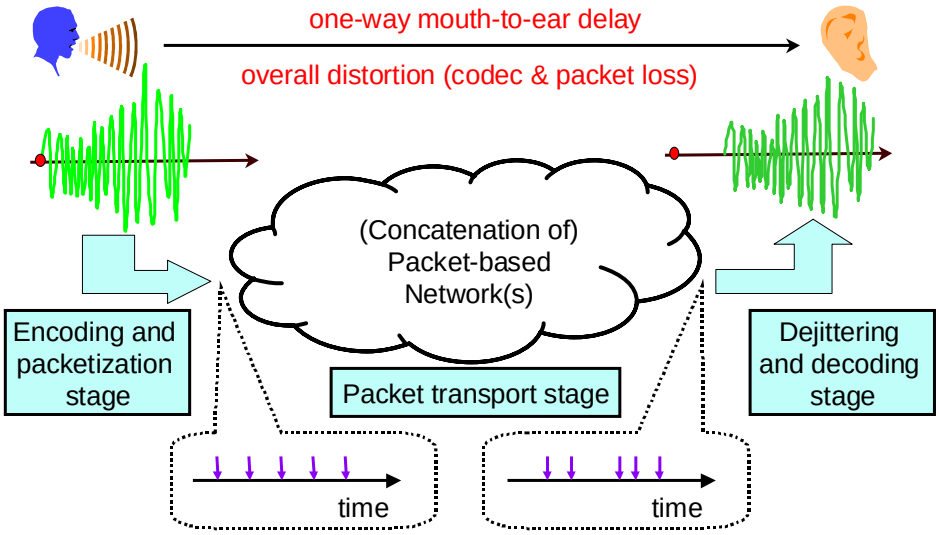

Figure 1: Three essential stages in the packetized transport of voice.

In the second stage, this flow of packets is transported over an IP network consisting of several access and backbone nodes. In the transport of the voice flow over this network some delay is incurred. The network delay can be split into two parts: a deterministic part, referred to as the minimal network delay, and a stochastic part, referred to as the total queuing delay. The minimal network delay mainly consists of the propagation delay (of 5 µs per km), the sum of all serialization delays, the route look-up delay, etc. The total queuing delay is the sum of the queuing delay in each node. The queuing delay in one network node is due to the competition of the packets of several flows for the available resources in that node. The total queuing delay is responsible for the jitter introduced in the voice flow. The aim of this paper is to study the queuing delay the packets incur in one of the network nodes. We consider link rates R_{link} from 512 kb/s up to 10.24 Mb/s, which are typical link rates in an access network. Moreover, as will become clear from the results below, at a link rate of 10.24 Mb/s the limiting behavior is almost reached. If there is only one bottleneck node on the mouth-to-ear path, the queuing delay incurred in this node practically solely determines the total queuing delay. If there are more nodes with considerable contribution to the total queuing delay, the individual delay statistics need to be combined. How to combine these individual statistics of delays incurred in several nodes (e.g., whether the queuing delays in consecutive nodes are statistically independent or not) is beyond the scope of this paper.

In the last stage the jittered packet flow is dejittered and decoded. Since the decoder needs the packets at a constant rate, dejittering is absolutely necessary. Dejittering a voice flow consists of retaining the fastest packets in the dejittering buffer to allow the slowest ones to catch up. The fastest packets are the ones that do not have any queuing delay in any of the nodes. So, in principle, the fastest packets have to be retained for a time equal to the maximal total queuing delay in the dejittering buffer. Because voice codecs can tolerate some packet loss and because waiting for the slowest packet frequently introduces too much delay, often the fastest packets are retained in the dejittering buffer for a time equal to the (1-P)-quantile of the total queuing delay. This means that a fraction P of the packets will be lost, because they arrive too late. It depends on the codec how large a packet loss can be tolerated. Typical values lie in the interval $[10^{-5},10^{-2}]$ (see [2]). Here, we take a value of $P = 10^{-4}$.

In this paper we determine up to which value ρ a node can be loaded for the $(1\text{-}10^{-4})$-quantile of the queuing delay to reach a specific value T_d. In most cases we use T_d = 6.25 ms, but we also consider the effect of relaxing T_d to 12.5 ms. The choice for these values is reasonable, although somewhat conservative. It is well known that if the echo is perfectly controlled, the bound on the mouth-to-ear delay to ensure an interactive call is 150 ms, although this bound is not strict and may be exceeded under certain circumstances [2]. If we subtract the main components (i.e., the codec delay, the packetization delay, the sum of all serialization delays, the propagation delay, etc.) from this 150 ms budget, and take into account that possibly more than one node can contribute to the total queuing delay, we end up with a queuing delay budget for one node in the neighborhood of the values we consider in this paper.

3. THE QUEUING MODEL

3.1. Modeling the phones

In order to assess the delay incurred in one network node, we study the following discrete-time queuing model. The time unit (i.e. slot) is taken equal to the time to put a voice packet on the link. Hence, the service time for each packet is one time unit. Remark that the time unit depends on the link rate R_{link} and the packet size (always equal to 200 bytes here), and hence, ranges between 156.25 μs and 3.125 ms for R_{link} = 10.24 Mb/s and 512 kb/s, respectively.

The buffer space in the network node is assumed to be infinite.

We consider N phones and model the traffic generated by one phone as a Markov Modulated Bernoulli Process (MMBP). A phone is in one of several states. If the phone is passive (or silent), it sends no packets at all. When the phone is in the active (and talking) state, it behaves as a Bernoulli source, i.e., it sends a packet with a probability $1/I_a$ in each slot, with I_a the inter-packet time expressed in slots. In reality an active (and talking) phone produces packets with constant inter-packet time I_a. This Bernoulli behavior makes that the queuing delay that we calculate with this MMBP/D/1 model is likely to be (slightly) higher than the queuing delay incurred in reality. We make a distinction between phones that do not use silence suppression and phones that do.

3.1.1. A phone without silence suppression

A phone without silence suppression can be either active or passive. The phone is in the active state “A” with probability p (and in the passive state “P” with probability 1-p). The average duration of a call is 120 s. In the following the average call duration T_A and the average passive period T_P are expressed in the time unit (slot) defined above, and hence are dimensionless. The average passive period, i.e., the sojourn time in the passive state “P”, is then given by

$$T_P = \frac{1-p}{p} T_A \quad . \tag{1}$$

Both the active and the passive sojourn times are geometrically distributed. The state transition diagram of such a MMBP source is depicted in Figure 2.

The transition probabilities are solely determined by the average sojourn times and are calculated as

$$p_{11} = 1 - \frac{1}{T_A}, \quad p_{12} = 1 - p_{11}, \tag{2}$$

$$p_{22} = 1 - \frac{1}{T_P}, \quad p_{21} = 1 - p_{22} \quad . \tag{3}$$

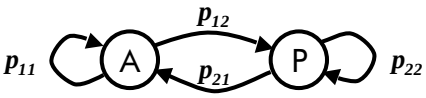

Figure 2: The state transition diagram for a phone without silence suppression.

Notice that a phone that does not use silence suppression is completely described by three parameters: the (average) inter-packet time I_a, the average call duration T_A and the activity grade p.

The (average) load one such phone places on the node is

$$\rho_1 = \frac{p}{\phi} \frac{R_{cod}}{R_{link}} \tag{4}$$

with ϕ the filling factor of the packets, i.e., the payload size divided by the packet size, which in this paper is always equal to 0.8 (=160/200). Hence, the number of phones associated with (average) load ρ is

$$N = \rho\gamma \frac{\phi}{p} N_{PSTN} \tag{5}$$

with $N_{PSTN} = R_{link}/64$ the number of 64 kb/s trunks that a PSTN switch of link rate R_{link} would have and $\gamma = 64/R_{cod}$ the gain due to the use of a low bit rate codec.

Notice that a switch with N_{PSTN} trunks can support a lot more phones than just N_{PSTN}. Consider e.g. a link rate of $R_{link} = 5.12$ Mb/s (i.e. $N_{PSTN} = 80$) and phones with an activity grade $p = 0.1$ (e.g., an average call duration $T_A = 120$ s and a call attempt rate of 3 calls per hour). In that case according to the (inverse) Engset formula, $N = 540$ phones can be supported with a call blocking probability of 10^{-4}.

3.1.2. A phone with silence suppression

A phone that uses silence suppression can also be in an active "A" or a passive "P" state, but in the active state there are two sub-states. An active phone with silence suppression can either be "talking" or "listening". In the former state, referred to as "T", the phone generates packets as a Bernoulli source with the same rate as before, while in the latter state, referred to as "L", as in the passive state "P", no packets are generated at all.

Again, all sojourn times "T", "L", and "P", are geometrically distributed. According to [7] the average talking and listening period is about 1 s and 1.5 s, respectively. The state transition diagram of such a MMBP source is depicted in Figure 3.

The transition probabilities are given by

$$p_{11} = 1 - \frac{1}{T_T}, \quad p_{12} = 1 - p_{11}, \quad p_{13} = 0 \quad , \tag{6}$$

$$p_{22} = 1 - \frac{1}{T_L}, \quad p_{23} = \frac{1}{(1-\alpha)T_A}, \quad p_{21} = 1 - p_{22} - p_{23} \quad , \tag{7}$$

$$p_{33} = 1 - \frac{1}{T_P}, \quad p_{31} = 1 - p_{33}, \quad p_{32} = 0 \quad , \tag{8}$$

with the silence suppression factor α defined as $T_T/(T_T + T_L)$ and equal to 0.4 in this paper.

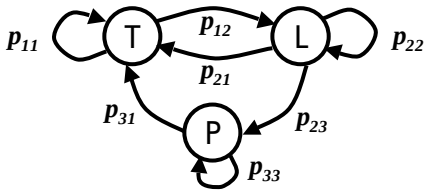

Figure 3: The state transition diagram for a phone with silence suppression.

Notice that a phone that uses silence suppression is completely described by five parameters: the (average) inter-packet time I_a, the average call duration T_A, the activity grade p, the average duration of a talk spurt T_T and of a silence period T_L.

The (average) load one phone with silence suppression places on the network, is

$$\rho_1 = \alpha \frac{p}{\phi} \frac{R_{cod}}{R_{link}} \quad . \tag{9}$$

Hence, the number of phones associated with (average) load ρ is

$$N = \rho\gamma \frac{1}{\alpha} \frac{\phi}{p} N_{PSTN} \quad . \tag{10}$$

3.2. Determining a quantile of the queuing delay

In this paper we consider the case of homogeneous and heterogeneous traffic. In the homogeneous case all phones that compete for the available resources in one node use the same codec, and either all use silence suppression, or none do. In the heterogeneous case we consider a mix of traffic generated by phones that use one of two types of codecs. Each component of the mix has its own associated load. We consider mixes of several proportions.

The homogeneous MMBP/D/1 model (in fact, an even more general model) was extensively studied in [5]. In particular, the probability generating function (pgf) of the queuing delay was found. We do not repeat the mathematical theory here, but refer the interested reader to that paper. We concentrate on the tail distribution of the queuing delay of the MMBP/D/1 model. In [5] it is demonstrated that finding a close approximation for the tail distribution of the queuing delay boils down to finding a number of dominant poles of the pgf. The heterogeneous model was studied in [6].

We determine the load ρ (or loads in case of heterogeneous traffic) for which the $(1\text{-}10^{-4})$-quantile of the queuing delay (in the MMBP/D/1 node) has a specific value T_d. Once this tolerable load ρ is identified, the number of phones that can be supported by this MMBP/D/1 node can be calculated with eq. (5) and eq. (10) for phones without and with silence suppression, respectively.

For the homogeneous case, we also compare the tolerable load according to the MMBP/D/1 model with the one derived from the M/D/1 model. Remark that once the tolerable load is identified, eq. (5) and eq. (10) can still be used to calculate the number of phones that can be supported by the M/D/1 node for phones without and with silence suppression, respectively.

In [1] a continuous-time fluid-flow model, where during the active period a fluid at constant rate (instead of packets governed by a Bernoulli process) is generated, is discussed. While this model captures the competition at the time scale of a burst (i.e., an on-period), it cannot accurately describe the competition at the time scale of a slot [4]. Moreover, [1] only discusses the buffer occupancy, not the incurred delay.

3.3. Burstiness

To show the fundamental difference between an MMBP source and a Bernoulli source, we consider the arrival process at a coarser time scale. Therefore, we define the mean e_K of the number of packets a single source generates over K consecutive slots as

$$e_K \hat{=} \frac{1}{K}\sum_{k=1}^{K} e_k \quad , \tag{11}$$

where e_k is the number of arrivals (0 or 1) in slot k for a source in steady state.

We first consider the case without silence suppression. The average and variance of the random variable e_K can be readily calculated as

$$E[e_K] = \frac{p}{I_a} \quad , \tag{12}$$

$$Var[e_K] = \frac{1}{K^2}\left\{K\,E[e_K](1-E[e_K]) + \frac{2E[e_K](1-p)}{I_a}\,\frac{(K-1)\lambda - K\lambda^2 + \lambda^{K+1}}{(1-\lambda)^2}\right\} \quad , \tag{13}$$

with $\lambda = 1 - 1/((1\text{-}p)T_A)$ the eigenvalue of the transition matrix, different from 1. Remark that this eigenvalue is in absolute value smaller than 1. Under normal circumstances, we have that $(1\text{-}p)T_A >> 1$, certainly if the link rate R_{link} is high enough or equivalently if the slot length is small enough. On top of that for large values of K we can neglect the terms of the order $O(K^{-2})$ in eq. (13). Then the variance is approximated by

$$Var[e_K] \approx \frac{1}{K}\left\{E[e_K](1-E[e_K]) + 2E[e_K](1-p)^2\left(\frac{T_A}{I_a}\right)\right\} \quad . \tag{14}$$

It is in the second term that an MMBP source differs from a Bernoulli source. This second term is the reason that one MMBP source is burstier than one Bernoulli source. When the second term of eq. (14) is significant for one source, it remains significant for an aggregate of sources. Consequently, an aggregate of MMBP sources is also burstier than an aggregate of Bernoulli sources. The latter is closely approximated by a Poisson process, if the number of sources of the aggregate is large enough. Hence, an aggregate of MMBP sources is not very well approximated by a Poisson process. Therefore, we expect the tolerable load obtained with the M/D/1 model to give an upper bound of the one obtained with the MMBP/D/1 model. Eq. (14) indicates that the M/D/1 model, where the traffic is only characterized by one parameter, i.e., the total load, is only able to capture the behavior of an aggregate of on-off

sources, if the activity grade p is close to 1. The burstiness of an aggregate of MMBP sources stems from the fact that these sources are active only a fraction p of the time. Notice also from eq. (14) that the burstiness increases, if I_a decreases or equivalently (because the packet size is always 200 bytes in this paper) if R_{cod} increases.

The case with silence suppression can be treated in a similar way, and leads to

$$E[e_K] = \frac{\alpha p}{I_a} \quad , \tag{15}$$

$$Var[e_K] \cong \frac{1}{K}\left\{ E[e_K](1 - E[e_K]) + 2E[e_K](1-p)^2 \alpha \frac{T_A}{I_a} + 2E[e_K](1-\alpha)^2 \frac{T_T}{I_a} \right\}. \tag{16}$$

It is in the second and third term that an MMBP source differs from a Bernoulli source. Additional burstiness stems from the use of silence suppression.

4. RESULTS

4.1. General

All results below calculate the load ρ (or combination of loads in the heterogeneous case) that can be tolerated, so that the $(1\text{-}10^{-4})$-quantile of the queuing delay reaches a specific value T_d. The influence of various parameters is investigated. The number of phones that can be supported depends on the tolerable load, but is also proportional (see eq. (5)) to the gain γ due to the use of a low bit rate codec and to the filling factor ϕ, if no silence suppression is used. On top of that, it is also proportional (see eq. (10)) to the gain $1/\alpha$ due to silence suppression, in case this is used.

4.2. The influence of the codec bit rate R_{cod}

Figure 4 illustrates the effect of the codec bit rate R_{cod}. This figure shows that the lower the bit rate of the codec, the larger the tolerable load on the node is. This is corroborated by eq. (14), which shows that the smaller the codec rate (i.e. the larger I_a), the closer an aggregate of MMBP sources approximates an aggregate of Bernoulli sources, and hence, a Poisson process. Notice that the M/D/1 model overestimates the tolerable load, and hence the number of phones that can be supported, in some cases even by a factor 3.

The multiplexing gain is small when the link rate R_{link} is low. Only when the link rate is a couple of Mb/s there is a reasonable multiplexing gain.

Eq. (5) shows the gain γ due to the use of a low bit rate codec: the lower the codec bit rate, the larger the number of phones that can be supported. On top of this gain γ, an additional gain is achieved, because the node can be loaded up to a higher value if the codec bit rate decreases. For example, for T_d = 6.25 ms and p = 0.1 (see Figure 4), the tolerable load for a link rate of 5.12 Mb/s is 0.538 and 0.692 for the 64 kb/s codec and the 16 kb/s codec, respectively. The number of phones that can be supported for the 16 kb/s codec, is 1770, while it is only 344 for the 64 kb/s codec. This results in a total gain of 5.15 (=4*(0.692/0.538)), when the same delay quantile (of 6.25 ms) has to be respected.

We have to bear in mind, however, that the packet size for each codec bit rate was chosen the same (i.e. 200 bytes). This means that the packetization delay for the 16 kb/s codec is 4 times higher than the one for the 64 kb/s codec. If the packetization delay would have been chosen the same for codecs with different bit rate, the filling factor ϕ for the 64 kb/s codec

would have been (a lot) larger than for the 16 kb/s codec. From eq. (5) we see that this filling factor ϕ too has a direct impact on the number of phones that can be supported.

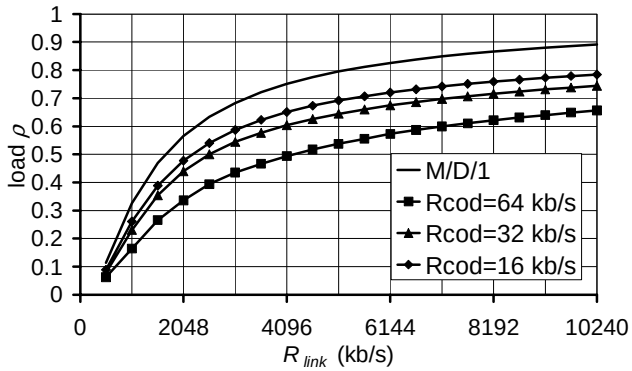

Figure 4: Load that can be supported in case of phones without silence suppression, T_d = 6.25 ms and p = 0.1.

Since it is beyond the scope of this paper to identify the optimal packetization delay, it is not quantitatively investigated here what the impact is of an increase in queuing delay budget T_d due to the fact that less packetization delay is consumed while keeping the mouth-to-ear delay budget equal to 150 ms. Remark that also the time unit and serialization delay change, if the packet size changes. A comparison of Figure 4 with Figure 5 merely gives a qualitative trend. Comparing these figures illustrates that if the delay constraint is relaxed (i.e., T_d is increased from 6.25 ms to 12.5 ms), the tolerable load, and hence, the number of sources, increases.

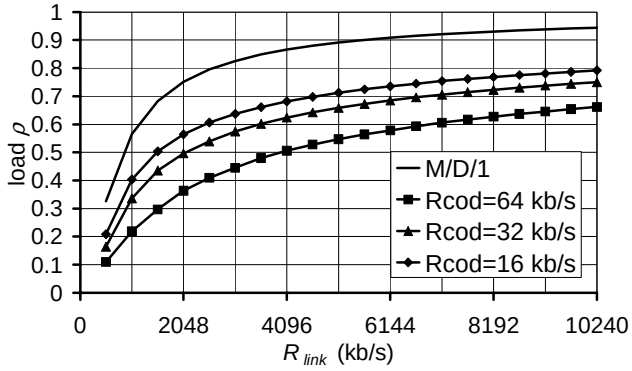

Figure 5: Load that can be supported in case of phones without silence suppression, T_d = 12.5 ms and p = 0.1.

4.3. Heterogeneous Traffic

In Figure 6 we consider the case of heterogeneous traffic. Again we have set the activity grade p = 0.1. In this figure we have plotted the tolerable load combinations of two traffic classes that are multiplexed in a network node, for various values of the link rate R_{link}. Packets of the first traffic class are generated by 64 kb/s codecs, while packets of the second traffic class are generated by 16 kb/s codecs. We have taken the extremes (16 kb/s and 64 kb/s) of the codec bit rate considered in this paper to maximally illustrate the effect of mixing traffic. No silence suppression is used. Since the packet size is the same (i.e. 200 bytes) for both codecs, a 64 kb/s codec generates packets every 20 ms on average, while a 16 kb/s codec generates packets every 80 ms on average during an active period. The tolerable load (ρ_{64} and

ρ_{16} respectively) of both classes is again calculated by requiring that the fraction of packets having a queuing delay larger than 6.25 ms may not exceed 10^{-4}.

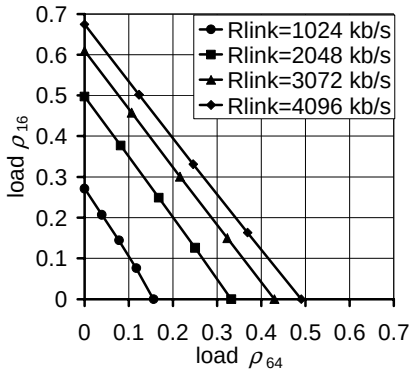

Figure 6: Allowable mix of 64 kb/s and 16 kb/s phones without silence suppression, T_d = 6.25 ms and p = 0.1.

From this figure we again deduce that little multiplexing gain is to be achieved when the link rate R_{link} is low. Only when R_{link} is at least a couple of Mb/s, the tolerable load of both classes reaches acceptable values. Secondly, it becomes once more clear that the lower bit rate codecs have a higher multiplexing gain (at the expense of a larger packetization delay) as explained before. Most importantly, in view of the quasi-linear relation between the tolerable load combination of both traffic classes, these curves could also be easily derived from combining the results in the homogeneous case for 16 kb/s and 64 kb/s codecs, respectively.

4.4. The influence of the activity grade *p*

The influence of the activity grade p for phones without silence suppression can be observed in Figure 7. It can be concluded that the higher the activity grade p, the better the M/D/1 model is. For p = 1 and α = 1 (i.e., all phones are active and talking all of the time) an MMBP source is a Bernoulli source. Since an aggregate of Bernoulli sources is closely approximated by a Poisson process, the M/D/1 and MMBP/D/1 model are (nearly) equivalent in this case.

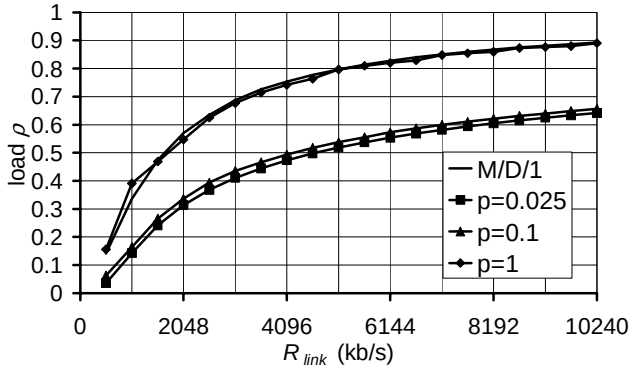

Figure 7: Load that can be supported in case of phones without silence suppression, T_d = 6.25 ms and R_{cod} = 64 kb/s.

There is not a lot of difference in the tolerable load when the activity grade p decreases from 0.1 to 0.025. Hence, supporting a number of phones that are active 10 % of the time is more or less equivalent to supporting 4 times this number of phones that are active only 2.5 % of the time. Hence, if the activity grade is low enough, the number of phones that can be

supported by the node is inversely proportional to the activity grade. However, if the activity grade is close to 100 % (in the case no silence suppression is used), the number of phones that can be supported by the node is larger than suggested by this inverse proportionality rule.

4.5. The influence of silence suppression

First, we consider the case of the activity grade $p = 0.1$. When we compare the relevant curves on Figure 8, we observe that in this case the tolerable load is about the same irrespective of the fact whether silence suppression is used or not. Hence, (compare eq. (10) with (5)) there is a gain of $1/\alpha$ (i.e., 2.5 in this paper) in terms of the number of phones that can be supported, when silence suppression is switched on for all phones. The reason is that one source is already bursty, because it is active only 10 % of the time and passive for the rest of the time. The additional burstiness silence suppression introduces has only a marginal effect.

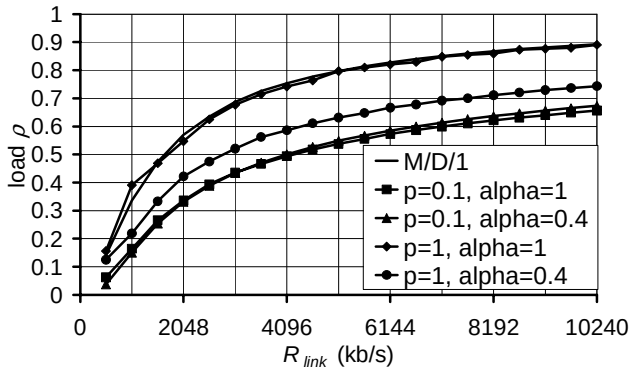

Figure 8: Load that can be supported in case of phones with ($\alpha = 0.4$) and without silence suppression ($\alpha = 1$), T_d = 6.25 ms and R_{cod} = 64 kb/s.

Next, we consider the case of phones that are always active, i.e. $p = 1$. By comparing the third term of eq. (16) with the second term of eq. (14), we notice that in the case $p = 1$ with silence suppression (the second term in eq. (16) vanishes), α has a similar role as p in the case without silence suppression. Comparing the relevant curves on Figure 8 leads to the conclusion that switching on silence suppression does have an effect on the tolerable load in this case. When silence suppression is switched on, the network cannot be loaded up to the same value as in the case without silence suppression. For example for a link rate R_{link} = 5.12 Mb/s the node can be loaded up to 0.79 when no silence suppression is used. When silence suppression is used the tolerable load drops to 0.63. Hence, the gain due to silence suppression drops from the naively expected value of 2.5 (=$1/\alpha$) to 2.

5. CONCLUSIONS

In this paper we studied the MMBP/D/1 model to determine the number of packet-based phones that can be supported by a multiplexing node. The fact that not all phones are active all of the time can be exploited in a packet-based node in much the same way as in a circuit-switched node. For a circuit-switched case this leads to the Engset formula. In the packet-based case there is an additional gain due to silence suppression. We considered the case of homogeneous traffic generated by phones that either all use silence suppression or none of them do, as well as the case of heterogeneous traffic where the traffic is a mix generated by phones that use one of two possible bit rates. We investigated the impact of the activity grade,

the codec bit rate and the link rate on the number of phones that can be supported by the packet-based node. We came to the following conclusions.

The M/D/1 model overestimates (in some cases even by a factor of 3) the number of sources that can be supported, even for large link rates where lots of phones can be supported. Apparently, an aggregate of MMBP sources is burstier than a Poisson process.

If the activity grade is considerably smaller than 100 %, the tolerable load is more or less independent of the activity grade, which leads to the rule that the number of phones that can be supported by the node is inversely proportional to the activity grade. However, if the activity grade approaches 100 % (in the case no silence suppression is used), the M/D/1 model is a better approximation, the tolerable load increases, and hence, the number of phones that can be supported by the node is larger than suggested by this inverse proportionality rule.

The use of a low bit rate codec leads to a gain that is larger than just the codec gain, because an aggregate (of the same total bit rate and of the same packet size) becomes less bursty as the codec bit rate of the contributing sources decreases. However, we have to keep in mind that more packetization delay is introduced for a low bit rate codec.

When mixing traffic the combinations of tolerable loads of the individual traffic classes lie practically on a straight line. Hence, it is easy to calculate the tolerable load combinations in the case of heterogeneous traffic, if the tolerable loads in case of homogeneous traffic are known.

Finally, the bit rate reduction introduced by silence suppression largely outweighs the possible decrease in tolerable load introduced by the additional burstiness of the aggregate traffic.

REFERENCES

1. D. Anick, D. Mitra, M.M. Sondhi, “Stochastic Theory of a Data-Handling System with Multiple Sources”, The Bell System Technical Journal, Vol. 61, No. 8, pp. 1871-1894, October 1982.
2. D. De Vleeschauwer, J. Janssen, G. H. Petit, F. Poppe, “Quality Bounds for Packetized Voice Transport”, Alcatel Telecom Review, First quarter 2000, pp. 19-23, January 2000.
3. A. Myskja, “An Introduction to Teletraffic”, Telektronikk, Vol. 91, No. 2/3, pp. 3-41, 1995.
4. I. Norros, J.W. Roberts, A. Simonian, T.T. Virtamo, “The Superposition of Variable Bit Rate Sources in an ATM mulitplexer”, IEEE Journal on Selected Areas in Communications, Vol. 9, No. 3, pp. 378-387, April 1991.
5. B. Steyaert, H. Bruneel, G.H. Petit, D. De Vleeschauwer, “A Versatile Queueing Model Applicable in IP Traffic Studies”, COST257 MC Meeting, Document COST 257TD(00)02, Barcelona (Spain), 20 January 2000.
6. B. Steyaert, Y. Xiong, H. Bruneel, “An Efficient Solution Technique for Discrete-Time Queues Fed by Heterogeneous Traffic”, International Journal of Communication Systems, Vol. 10, pp. 73-86, 1997.
7. “Objective Measurements of Active Speech Level”, ITU-T Recommendation P.56, March 1993.