DETERMINING THE NUMBER OF PACKET-BASED PHONES THAT CAN BE SUPPORTED BY ONE ACCESS NODE

Danny De Vleeschauwer, Annelies Van Moffaert, Jan Janssen,
Maarten Büchli, Guido H. Petit
Alcatel Bell, Network Strategy Group
Francis Wellesplein 1
B-2018 Antwerp, Belgium
E-mail: {danny.de_vleeschauwer, annelies.van_moffaert,
jan.janssen, maarten.buchli, guido.h.petit}@alcatel.be
Phone: +32(0)32408196. Fax: +32(0)32404888
and
Bart Steyaert, Herwig Bruneel
Ghent University, Department TELIN, SMACS research group
Sint-Pietersnieuwstraat 41
B-9000 Ghent, Belgium
E-mail: {bs,hb}@telin.rug.ac.be
Phone: +32(0)92643411. Fax: +32(0)92644295

ABSTRACT

To determine the number of phones that can be supported by a packet-based multiplexing node, we study the MMBP/D/1 queuing model. Similar as in the circuit-switched case, where the Engset formula is used to determine the number of phones that can be supported by a switch, the fact that not all phones are active and talking all of the time can be exploited in the packet-based case. We investigate the impact of the link rate, the activity grade, the codec bit rate and the silence suppression factor on the number of phones that can be supported. 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. If the activity grade of the phones is low, the M/D/1 model overestimates the number of phones that can be supported, even for high link rates. Only in the case where the phones are active all the time, the M/D/1 model gives accurate results. Reducing the codec bit rate yields a larger gain than expected because aggregating more sources of lower bit rate decreases the burstiness. Activating silence suppression on each phone normally results in the expected gain, because the additional burstiness introduced by silence suppression is negligible, if the activity grade of the phones is not too high.

1. Introduction

Packetized voice transport (e.g. Voice over IP (VoIP)) has several advantages over circuit-switched voice. In general packetized voice is more flexible than circuit-switched voice, because 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). A packet-based voice network is not bound to a certain codec as the (circuit-switched) Public Switched Telephone Network (PSTN) is to the 64 kb/s G.711 codec. Indeed, each codec that both communicating phones support, can be used. This is particularly an advantage, because the codec technology is still evolving to ever-smaller bit rates. 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.

The remaining issue is how to offer the same quality for a voice call routed over packet-based network as for a call switched over the PSTN. For a voice call routed over a packet-based network more delay and distortion (due to the use of a low bit rate codec and due to packet loss) are likely to be introduced. The bounds on the delay and the distortion are known and reported in [2,3]. The question is how to dimension the network elements (the packetizer, the network nodes, the dejittering mechanism, etc.) such that these bounds are met [1,8]. This paper studies the delay contribution of one node in the network.

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. It is assumed that both, the active and passive period, are 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 [5] calculates the number N of phones that can be supported by a switch with N_{PSTN} trunks (i.e., a switch of capacity $R_{link} = N_{PSTN}$ x 64 kb/s) given a certain tolerated blocking probability. The (inverse) Erlang B formula [4] can be used instead of the (inverse) Engset formula, if the number N of phones is large enough.

In this paper 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 N of phones 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.

It might be argued that on the advent of terabit networks, capacity will be for free, i.e., that future nodes will have so much capacity that they can support much more voice calls than they actually will have to. Although this might be the case for backbone networks, for access networks, resources will remain very much limited in the near future. Even for backbone nodes that will transport voice and data, voice traffic will be transported in (one of) the high priority classes and it is likely that not more than what is strictly necessary will be reserved for voice traffic. Although we concentrate on the converse problem (i.e., determine N given R_{link}) here, the same methodology developed in this paper, can be used to determine the minimal capacity R_{link} to support a given number of phones (i.e., determine R_{link} given N).

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 [7]. 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, the activity grade and the silence suppression factor). 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 high 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).

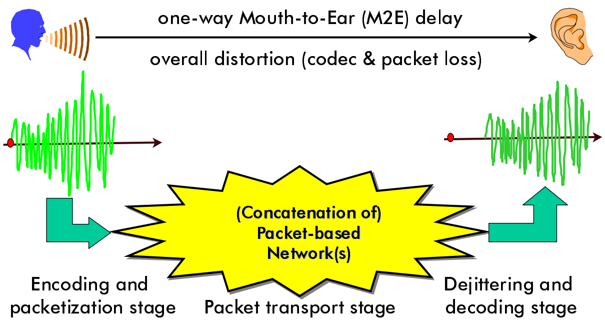
one-way Mouth-to-Ear (M2E) delay
overall distortion (codec & packet loss)
(Concatenation of) Packet-based Network(s)
Encoding and packetization stage
Packet transport stage
Dejittering and decoding stage

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

In the first stage, the digital voice signal (i.e., voice that is sampled at e.g. 8 kHz and quantized with a uniform e.g. 13-bit quantizer) 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 kb/s codec is perhaps 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 inter-packet time I_a, i.e., each period of duration I_a exactly one packet is produced. The inter-packet time I_a is equal to 20 ms, 40 ms and 80 ms, for the 64 kb/s, 32 kb/s and 16 kb/s codec, respectively.

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 node is due to the competition of several flows for the available resources in the queue of that network 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. 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 each 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 in the codec how large a packet loss can be tolerated. Typical values lie in the interval [10^{-5},10^{-2}]. Here, we take a value of P=10^{-4}.

The above description is the ideal operation of the dejittering mechanism. In a packet-based environment this ideal operation cannot be reached, because when the first packet of a voice flow is received, it is not known whether this packet is a fast or slow one. Adaptive dejittering mechanisms are able to reach the ideal operation in the long run, see e.g. [6]. The study of dejittering mechanisms is beyond the scope of this paper.

From the above explanation it is clear that the delay incurred in one network node contributes to the mouth-to-ear delay. This paper considers the delay introduced in one node. More specifically, we determine up to which value ρ the node can be loaded for the (1-10^{-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 the bound on the mouth-to-ear delay to ensure a reasonable interactive call is 150 ms, if the echo is perfectly controlled [2,3]. 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 homogeneous 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 1 time unit. Remark that the time unit depends on the link rate R_{link} and the packet size (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.

We always consider homogeneous scenarios in this paper, i.e., all N phones that compete for the available resources in one node use the same codec, and either all use silence suppression, or none do.

Packets generated by the N phones that cannot be served immediately are stored in a FIFO queue with infinite buffer space.

We 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 (where I_a is dimensionless, as it is expressed in the time unit defined above). This latter assumption is worst case. 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 with constant inter-packet times during active periods.

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 can be either active or passive. We assume that a phone is in the active state "A" with probability p (and in the passive state "P" with probability 1-p) and that the average duration T_A of a call is 2 minutes (120 s). 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 an MMBP source is depicted in Figure 2.

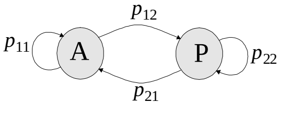

Figure 2: The state transition diagram for a phone without silence suppression.

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

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

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

Remark that in the above equations the time T_A and T_P are dimensionless, as they are expressed in the time unit (slot) defined above.

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 that can be supported if the node is loaded up to load ρ is

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

with

$$N_{PSTN} = \frac{R_{link}}{64} \tag{6}$$

the number of 64 kb/s trunks that a PSTN switch of link rate R_{link} would have, and

$$\gamma = \frac{64}{R_{cod}} \tag{7}$$

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 [9] 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.

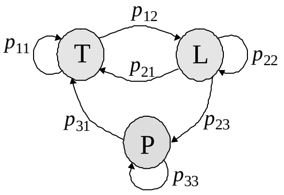

Figure 3: The state transition diagram for a phone with silence suppression.

The transition probabilities are given by

$$p_{11} = 1 - \frac{1}{T_T}, \quad p_{12} = 1 - p_{11}, \quad p_{13} = 0 \quad , \tag{8}$$

$$p_{22} = 1 - \frac{1}{T_L}, \quad p_{23} = \frac{1}{(1-\alpha)T_A}, \quad p_{21} = 1 - p_{22} - p_{23} \ , \tag{9}$$

$$p_{33} = 1 - \frac{1}{T_P}, \quad p_{31} = 1 - p_{33}, \quad p_{32} = 0 \quad , \tag{10}$$

with the silence suppression factor α defined as

$$\alpha = \frac{T_T}{T_T + T_L} \tag{11}$$

and equal to 0.4 (=1/(1+1.5)) in this paper.

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

$$\rho_1 = \alpha\frac{p}{\phi}\frac{R_{cod}}{R_{link}} \quad . \tag{12}$$

Hence, the number of phones that can be supported if the node is loaded up to load ρ is

$$N = \rho\gamma\frac{1}{\alpha}\frac{\phi}{p}N_{PSTN} \quad . \tag{13}$$

3.2. Determining a quantile of the queuing delay

The MMBP/D/1 model (in fact, an even more general model) was extensively studied in [7]. In particular, the probability generating function (pgf) of the buffer occupancy and the queuing delay were 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 [7] 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.

We determine to which value ρ the MMBP/D/1 node can be loaded so that the (1-10^{-4})-quantile of the queuing delay reaches 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. (13) (combined with eq. (6)) for the case without and with silence suppression, respectively.

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. (13) can still be used to calculate the number of phones that can be supported by the M/D/1 node for the case without and with silence suppression, respectively. 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, since an aggregate of MMBP sources is burstier than a set of Bernoulli sources. This burstiness stems from the fact that MMBP sources are active only a fraction p of the time and from the use of silence suppression (in case the latter is used). The M/D/1 model, where the traffic is only characterized by one parameter, i.e., the total load, is not able to capture the behavior of an aggregate of such on-off sources.

In [8] the N*D/D/1 model is compared with the M/D/1 model. There, the worst case assumption that all phones are always active, is taken.

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 number e_K of packets a single source generates over K consecutive slots as

$$e_K \hat{=} \frac{1}{K}\sum_{k=1}^{K} e_k \quad , \tag{14}$$

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

We first consider the case without silence suppression. Using the moment generating function, which can be derived with a technique similar as in [7], the average and variance of the random variable e_K can be readily calculated:

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

$$Var[e_K] = \frac{1}{K^2}\Bigg\{ K\,E[e_K](1-E[e_K]) + \frac{2E[e_K](1-p)}{I_a}\,\frac{(K-1)\lambda - K\lambda^2 + \lambda^{K+1}}{(1-\lambda)^2}\Bigg\}, \tag{16}$$

with

$$\lambda = 1 - \frac{1}{(1-p)T_A} \tag{17}$$

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-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. (16). Then the variance is approximated by

$$Var[e_K] \approx \frac{1}{K}\left\{E[e_K](1-E[e_K]) + 2E[e_K](1-p)^2\left(\frac{T_A}{I_a}\right)\right\} . \tag{18}$$

It is in the second term that an MMBP source differs from a Bernoulli source. This second term is significant, even for R_{link} tending to infinity. It is the reason that one MMBP source is burstier than one Bernoulli source. Consequently, an aggregate of MMBP sources is also burstier than an aggregate of Bernoulli sources, which is itself 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. Notice that the second term in eq. (18) increases if I_a decreases or equivalently if R_{cod} increases.

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

$$E[e_K] = \frac{\alpha p}{I_a} \quad , \tag{19}$$

$$Var[e_K] \cong \frac{1}{K}\Bigg\{E[e_K](1-E[e_K]) + 2E[e_K](1-p)^2\alpha\frac{T_A}{I_a} + 2E[e_K](1-\alpha)^2\frac{T_T}{I_a}\Bigg\} \quad . \tag{20}$$

It is in the second and third term that an MMBP source differs from a Bernoulli source.

4. Results

4.1. General

In order to readily compare the results of the MMBP/D/1 model and the M/D/1 model, we choose the tolerable load as performance parameter. The number of phones that can be supported by a node is related to this tolerable load by eq. (5) or eq. (13). Remark that this number of phones that can be supported does not only depend 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. (13)) to the gain $1/\alpha$ due to silence suppression, in case this is used.

Figure 4 to Figure 9 show the load ρ that can be tolerated, so that the (1-10^{-4})-quantile of the queuing delay reaches a specific value T_d, as a function of the link rate R_{link}. The influence of various parameters is investigated next.

Notice that, as expected, in each case considered here the tolerable load according to the M/D/1 model is higher than the one according to the MMBP/D/1 model. In some cases (i.e., low link rate and high codec rate) the overestimation factor can be quite high (even up to 3 and very often close to 2). Only in the case where the activity grade p equals 1 and no silence suppression is used (see Figure 7) the tolerable load is just about equal for both the M/D/1 model and the MMBP/D/1 model. Remark that in this case the MMBP sources reduce to Bernoulli sources, and hence, Figure 7 is just an illustration of the well-known fact that an aggregate of Bernoulli sources closely approximates a Poisson process (see eq. (18)).

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

Figure 4 to Figure 9 also illustrate the effect of the codec bit rate R_{cod}. Remember that because of 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.

The figures show that the lower the bit rate of the codec, the larger the tolerable load on the node is. This is corroborated by eq. (18), which shows that the smaller the codec rate (i.e. the larger I_a), the closer an aggregate of MMBP source approximates an aggregate of Bernoulli sources, and hence, a Poisson process.

Hence, on top of the gain γ achieved by using a low bit rate codec, an additional gain is possible, because the node can be loaded up to a higher value (see eq. (5) and eq. (13)). 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 codec with bit rate 16 kb/s, is 1770, while it is only 344 for the codec with bit rate 64 kb/s. 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 equal. This means that the packetization delay for the codec with bit rate 16 kb/s is 4 times higher than the one for the codec with bit rate 64 kb/s. 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) and eq. (13) we see that this filling factor ϕ too has a direct impact on the number of phones that can be supported.

Since it is beyond the scope of this paper to identify the optimal packet size, it is not quantitatively investigated here what the impact is of a increase in queuing delay budget T_d due to the fact that less packetization delay is consumed and the mouth-to-ear delay budget remains 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. This figure shows 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.

4.3. The influence of the activity grade *p*

The influence of the activity grade p can be observed by comparing Figure 4, Figure 6 and Figure 7. It can be concluded that the higher the activity grade p, the better the M/D/1 model is. It was already observed in Section 4.1 that for $p = 1$ and $\alpha = 1$ (i.e., all phones are active and talking all of the time) an aggregate of MMBP sources is closely approximated by a Poisson process and that in that case the M/D/1 and MMBP/D/1 model are (nearly) equivalent.

When we compare Figure 4 with Figure 6, we conclude that 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 1 (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.

4.4. The influence of silence suppression

First, we consider the case of the activity grade $p = 0.1$. When we compare Figure 4 with Figure 8, we observe that the tolerable load is about the same irrespective of the fact whether silence suppression is used or not. Hence, (compare eq. (13) with (5)) there is a gain of $1/\alpha$ (i.e., 2.5 (=1/0.4) 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.

Next, we consider the case of phones that are always active, i.e. $p = 1$. When we compare Figure 7 with Figure 9 we observe that switching on silence suppression does have on effect on the tolerable load. When silence suppression is switched on, the network cannot be loaded up to the same value as in the case without silence suppression. The loss in tolerable load increases as the codec bit rate increases. For example for a link rate R_{link} of 5.12 Mb/s the node can be loaded up to 0.79 when no silence suppression is used (see Figure 7). When silence suppression is used (see Figure 9), the tolerable load drops to 0.73, 0.70 and 0.63 for the codec with bit rate 16 kb/s, 32 kb/s and 64 kb/s, respectively. Hence, the gain due to silence suppression drops from the naively expected value of 2.5 (=1/α) to 2.38, 2.22 and 2.00, respectively.

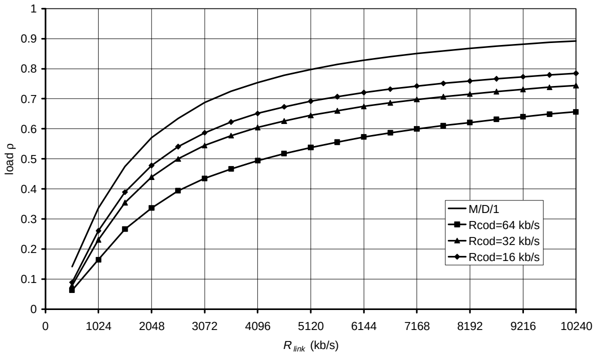

Figure 4: Load that can be supported in case of phones without silence suppression, T_d = 6.25 ms and p = 0.1.

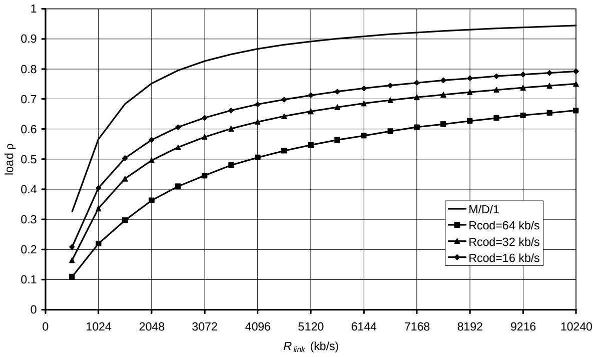

Figure 5: Load that can be supported in case of phones without silence suppression, T_d = 12.5 ms and p = 0.1.

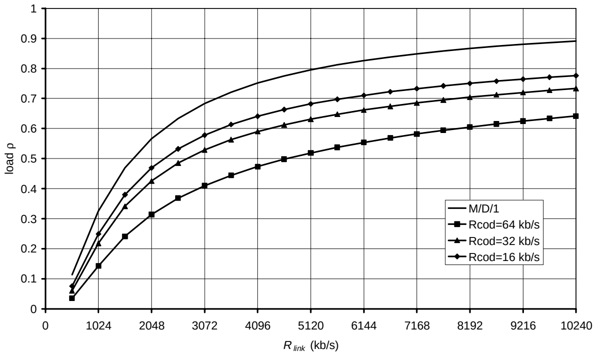

Figure 6: Load that can be supported in case of phones without silence suppression, T_d = 6.25 ms and p = 0.025.

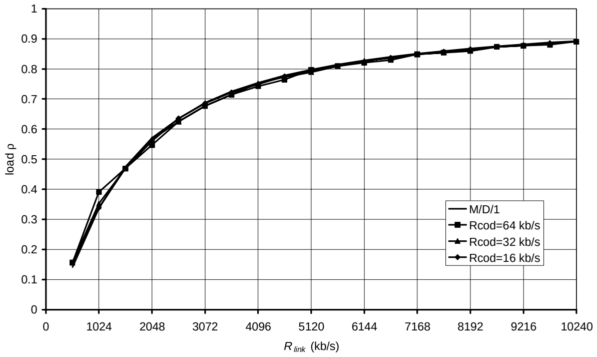

Figure 7: Load that can be supported in case of phones without silence suppression, T_d = 6.25 ms and p = 1.

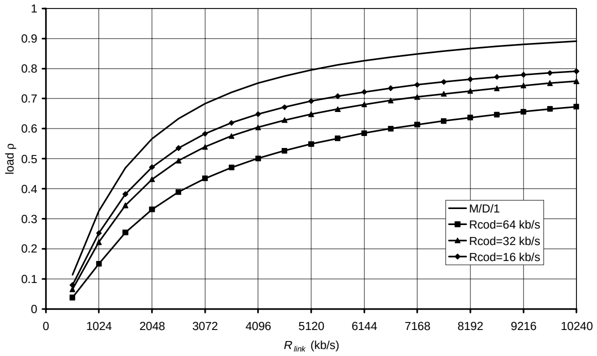

Figure 8: Load that can be supported in case of phones with silence suppression, T_d = 6.25 ms and p = 0.1.

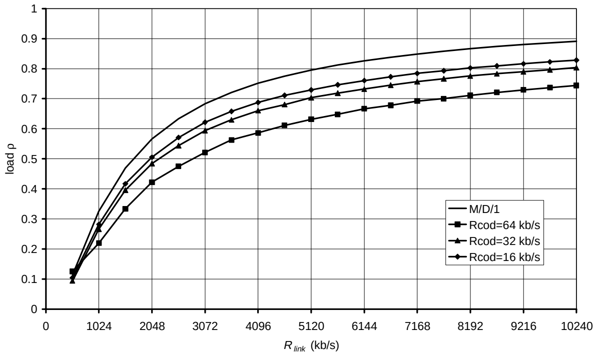

Figure 9: Load that can be supported in case of phones with silence suppression, T_d = 6.25 ms and p=1.

4.5. Call Blocking

Remark that, in contrast to a circuit-switched node, where calls may be blocked because of a lack of available trunks, a packet-based node does not strictly require call acceptance control. Indeed, with the method developed here it is ensured that if the calls arrive as a Poisson process and if the activity grade p and the silence suppression factor α are correctly anticipated, then all packets (except for a fraction of 10^{-4}) generated by any phone incur a queuing delay of maximally T_d in the node. If the phones do not behave properly, i.e., if calls do not arrive as a Poisson process (as during e.g. a calamity) or if the activity grade p or the silence suppression factor α are larger than expected, then a considerable part of the packets may have a queuing delay larger than T_d. Suppose that the number of phones that can be supported by the node was calculated under the assumption that the phones have an activity grade $p = 0.1$ and use silence suppression, i.e. $\alpha = 0.4$. Increasing the activity grade of each phone up to $p = 0.2$ or switching the silence suppression algorithm off on each phone, i.e. $\alpha = 1$, will have a drastic impact on the load, and hence, on the queuing delay. It may even lead to packet loss, if the buffer in the network node is finite.

Hence, we conclude that although call acceptance control might seem superfluous under normal circumstances for calls transported over a packet-based network, it still is useful to protect the network against wrong assumptions about the behavior of the phones. It may however be sufficient that this call acceptance control mechanism runs only on edge nodes.

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 it leads to the Engset formula. In the packet-based case there is an additional gain due to silence suppression.

We investigated the impact of the activity grade, the codec bit rate, the silence suppression factor and the link rate on the number of phones that can be supported by the packet-based node.

If the activity grade is considerably smaller than 1, 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 1 (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.

The use of a low bit rate codec and silence suppression always leads to a considerable gain in number of phones that can be supported by the network node. In case a low bit rate codec is used, the gain is larger than just the codec gain. An aggregate of 16 kb/s codecs is less bursty than an aggregate of 64 kb/s codecs for the same total bit rate and packet size. However, we have to keep in mind that more packetization delay is introduced for a low bit rate codec.

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

We also compared 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. The conclusion is that 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.

Acknowledgement

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).

REFERENCES

[1] D. De Vleeschauwer, J. Janssen, G.H. Petit, “Voice over IP in Access Networks”, Proceedings of the 7th IFIP Workshop on Performance Modelling and Evaluation of ATM/IP Networks (IFIP ATM ‘99), Antwerp (Belgium), 28-30 June 1999.

[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] J. Janssen, D. De Vleeschauwer, G.H. Petit, Delay and Distortion Bounds for Packetized Voice Calls of Traditional PSTN Quality”, Proceedings of the First IP Telephony Workshop (IPTEL2000), GMD Report 95, pp. 105-110, Berlin (Germany), 12-13 April 2000.

[4] L. Kleinrock, “Queueing Systems, Volume I: Theory”, John Wiley & Sons, 1975.

[5] A. Myskja, “An Introduction to Teletraffic”, Telektronikk, Vol. 91, No. 2/3, pp. 3-41, 1995.

[6] R. Ramdjee, J. Kurose, D. Towsley, H. Schulzrinne, “Adaptive Playout Mechanisms for Packetized Audio Applications in Wide-Area Networks”, Proceedings of IEEE Infocom 94, Toronto (Canada), pp. 680-688, June 1994.

[7] 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.

[8] K. Van Der Wal, M. Mandjes, H. Bastiaansen, “Delay Performance Analysis of the New Internet Services with Guaranteed QoS”, Proceedings of the IEEE, Vol. 85, No. 12, pp. 1947-1957, December 1997.

[9] “Objective Measurements of Active Speech Level”, ITU-T Recommendation P.56, March 1993.