A Rate-Based Overload Control Method for the Radio Channel in PCN

Nikos I. Passas*
passas@di.uoa.gr

Lazaros F. Merakos
merakos@di.uoa.gr

Department of Informatics - University of Athens
TYPA Buildings - Panepistimiopolis
Zografou, 157 71 Athens - GREECE
TEL: +30 1 7211119 (× 323)
FAX: +30 1 7219561

Abstract

Third-generation wireless digital communication systems, currently being developed, are intended to integrate all the existing wireless systems and cover a wide range of services, including voice, video and multimedia. A difficult problem towards this direction is the efficient use of the limited available bandwidth. Although considerable improvements have been made recently in transmitter and receiver technology, the capacity of the air interface is still considerably smaller compared to other media such as fiber optics. Accordingly, traffic congestion is an important problem, especially for bandwidth demanding applications (e.g., video), leading to poor quality-of-service (QoS). This paper presents an overload control method to temporarily reduce the source rate requirements to a sustainable level, in order to avoid a sudden degradation in QoS. The control is activated when the aggregate rate crosses a predefined threshold that identifies congestion. To ensure fairness, the selection of the sources whose rate will be reduced is performed in co-operation with a priority-based scheduling technique. The performance of the system under the proposed method is analyzed and system parameter values are optimized. It is shown that the method attains considerable improvement in the loss probability performance.

Technical Subject Category: Personal Communications

---

*designated presenter

1 Introduction

Wireless personal communications are considered as one of the most rapidly developing areas in telecommunications. The term "wireless" encompasses cordless systems like CT-2, cellular radio like GSM, personal communication systems like DCS1800, wireless LANs like HIPERLAN, even satellite-based mobile systems [1]. There are no more than a few years where second-generation digital wireless systems were designed and introduced and now they are enjoying full market acceptance. For example, GSM is now succesfully used in more that 60 countries, while IS-54 is a popular system in the U.S.

The spreading of the above systems shows how anxious the market is for new "wireless" services. Of course the development of second-generation technologies will continue, but their capabilities cannot cover all the demands of the end user. These demands include:

- integration with fixed networks, especially ATM/B-ISDN,
- efficient support of multimedia applications, including voice, data, video, and images,
- incorporation of Intelligent Network (IN) architecture concepts to provide advanced network services,
- advanced authentication and security functions.

New systems must be developed to fulfill these requirements. These systems are referred as third-generation systems and much research is currently being done towards their development [2].

An important area of research for the support of multimedia applications is traffic scheduling. Until recently wireless systems were focused on supporting only constant bit rate voice sources. Now they have to handle many different types of traffic, including variable bit rate and real-time traffic. This necessitates the development of advanced scheduling techniques to control the allocation of the available radio bandwidth. Such a technique, based on the leaky bucket regulator known from fixed ATM networks, has been proposed in [3]. Furthermore, the available bandwidth will continue to be limited in the air interface, despite the considerable

improvements in transmitter and receiver technology. This means that bandwidth demanding sources, like video and image, can readily overload the system, leading to congestion.

In this paper, we propose an overload control method to temporarily reduce the source rate requirements to a sustainable level, in oder to avoid a sudden degradation in QoS. We call this method "graceful degradation method" because it actually acts for the benefit of the sources that reduce their requirements. For the duration of congestion, they willingly pass to a lower level of service with acceptable quality, instead of staying at the higher level with poor QoS. The method can be easily applied to video sources resulting in a reduction of resolution or a color-to-black&white conversion, which can, in many cases, be acceptable by the end user, at least for a fraction of the time. The transition from normal to sustainable bit rate can be implemented with a two-level coder in every source. Similar methods have been proposed for fixed ATM networks (e.g., [5], [6])

The rest of the paper is organized as follows. Section 2 describes the proposed method and its interaction with the sources. In Section 3, a performance analysis is presented, focused on the construction of a two-dimensional Markov chain. In Section 4, numerical results, based on the previous analysis, are presented and analysed to reveal the effectiveness of the method. Finally, Section 5 contains our conclusions.

2 The Proposed Method

The method is used in conjunction with the scheduling technique proposed in [3], and aims at improving its efficiency for better utilization of the radio medium. In brief, the leaky-bucket-based scheduling technique of [3] is applied in a TDMA-based cellular system, where each TDMA frame is subdivided into a number of request slots and a number of data slots. Request slots are used by the sources to inform the base station (BS) about the data slots they need in the next frame. The BS decides how to allocate the data slots, based on a priority mechanism similar to the well known leaky-bucket regulator proposed for fixed ATM networks. Every source is assigned a token pool, located at the BS, where tokens are generated with fixed rate equal to the mean packet rate declared by the source at connection setup. The priority of

each source depends on the number of its tokens, compared to its requests. The more tokens a source has in its pool, the greater priority it has for allocation of slots, since it is below its declarations made at connection setup. For every slot allocated a token is removed from the appropriate pool.

During connection setup a source declares some parameters which characterize the kind of traffic that it is going to present to the network and the expected QoS. These parameters are used by the call admission control mechanism in order to determine if the network can support the new connection. Usually they include the anticipated mean bit rate R_n and deviation σ_n, and the acceptable loss probability P_{loss}. For the graceful degradation method a source must also declare the fraction of reduction r $(0 < r \leq 1)$ that it is willing to accept (i.e., the sustainable rate) and the fraction of the time that it agrees to transmit with sustainable rate. This fraction is equal to the acceptable probability of finding the source in sustainable state, referred as sustainable state probability $(P_{sustain})$. Since the reduction is considered uniform, the anticipated mean bit rate in sustainable state is $R_s = (1-r)R_n$ and the deviation $\sigma_s = (1-r)\sigma_n$. The method monitors the requests from all sources and decides which of them should transmit with normal and which with sustainable rate. When the sum of requests exceeds a predefined threshold some sources are forced to decrease their rates to sustainable. The number of these sources is the minimum needed to decrease the aggregate rate below the threshold. The inverse procedure of rate increase is performed when it does not cause crossing of the threshold. When the aggregate number of requests is below the threshold, the maximum number of sources that will keep the total rate below the threshold is permitted to transmit with normal rate.

Let $S_1, S_2, \cdots, S_k$ be the sources that request slots in a frame and let $Req(i)$ be the number of requests of source S_i. If $\sum_{i=1}^{k} Req(i) > threshold$ then some sources must reduce their rates, from normal to sustainable, leading to reduction of the aggregate rate below the threshold; the question is how many and which sources will be affected. If, from the k sources, l are already in sustainable state, from previous calls of the method, then the minimum number of sources that must reduce their rates is found as follows: assumming that $S_1, S_2, \cdots, S_{k-l}$ are the sources transmitting with normal rate and $T_1 \leq T_2 \leq \cdots \leq T_{k-l}$ are the corresponding token variables,

the reduction will be performed in sources S_1 through S_m where:

$$m = \begin{cases} j & \text{if there exists an integer } j \in [1, k-l] \text{ satisfying (1) and (2)} \\ k-l & \text{if there exists no integer } j \in [1, k-l] \text{ satisfying (1) and (2)} \end{cases}$$

$$\sum_{i=1}^{j}(1-r_i)Req(i) + \sum_{i=j+1}^{k} Req(i) \leq threshold \quad (1)$$

$$\sum_{i=1}^{j-1}(1-r_i)Req(i) + \sum_{i=j}^{k} Req(i) > threshold \quad (2)$$

where r_i is the reduction fraction of source S_i.

The selection based on the state of the token pools is performed for two reasons. First it is fair since at this moment the selected sources have consumed more bandwidth compared to their declarations. Second, according to the leaky-bucket scheduler, the selected sources have the lowest priority of getting slots allocated in the next frames. This means that with large probability these sources will experience a temporary degradation of the offered QoS, and that it is preferable for them to reduce their requirements until the traffic load is relaxed. This is a central point in the philosophy of the proposed method: It is preferred for a source to reduce its rate than experience more losses during overload. For example, it is usually preferred for the end user to turn a video source from color to black&white mode than leave it in color mode with poor image quality.

As soon as the BS decides which sources must reduce their rates, it uses a special field in the down-link channel to send a control signal to them. To enforce the acceptance of its decision the BS also reduces the token generation rate of the selected sources from normal to sustainable.

The inverse procedure of increasing the bit rate of some sources is performed when it does not cause crossing of the threshold. Assume that $S_1, S_2, \cdots, S_k$ are the sources requesting slots in one frame and $\sum_{i=1}^{k} Req(i) < threshold$. If l of them are in sustainable state, then the maximum number of sources that can increase their rates without crossing the threshold is found as follows. Let $S_1, S_2, \cdots, S_l$ be the sources in sustainable state and $T_1 \geq T_2 \geq \cdots \geq T_l$ the corresponding token variables. The increase will be performed in sources S_1 through S_n,

where:

$$n = \begin{cases} j & \text{if there exists an integer } j \in [1,l] \text{ satisfying (3) and (4)} \\ l & \text{if there exists no integer } j \in [1,l] \text{ satisfying (3) and (4)} \end{cases}$$

$$\sum_{i=1}^{j} \frac{1}{1-r_i} Req(i) + \sum_{i=j+1}^{k} Req(i) \leq threshold \tag{3}$$

$$\sum_{i=1}^{j+1} \frac{1}{1-r_i} Req(i) + \sum_{i=j+2}^{k} Req(i) > threshold \tag{4}$$

where r_i is the reduction fraction of source S_i. Accordingly, the generation rate of the corresponding tokens is increased from sustainable to normal as well.

The selection procedure described above is fair and selects the sources with the highest priority of allocating slots in the next frames, which deserve a rate increase.

3 Performance Analysis

In this section, we analyze the performance of the method in a cellular environment where many sources communicate with the BS of their cell. For simplicity, the sources are considered identical, although the same analysis can be used for different sources. A well accepted model representing the aggregate bit rate of a number of independent sources is the birth-death Markov model described in [7]. In this model, state sojourn times are exponentially distributed and transmitions occur only between neighboring states, which represent rate quantization levels (Figure 1). N sources are represented by $M \gg N$ minisources, where each minisource can be in one of two states: OFF, transmitting 0 packets/frame, or ON, transmitting A packets/frame. The transitions between the two states occur with rate α from OFF to ON and β from ON to OFF. According to [7]:

$$\beta = a/(1 + \frac{E^2(\lambda_N)}{MC_N(0)}) \qquad \alpha = a - \beta \qquad A = \frac{C_N(0)}{E(\lambda_N)} + \frac{E(\lambda_N)}{M}$$

$E(\lambda_N)$ is the aggregate mean rate and $C_N(t) = N\sigma^2 e^{-at}$ the aggregate autocovariance, where a is $3.9sec^{-1}$, a satisfactory value for many types of sources, especially video. The number M of minisources must be selected large enough for accurate results. Usually a value of $M = 20N$ is sufficient [7].

Assuming that feedback delay for notification of the sources is exponentially distributed [8], we develop a two-dimensional continuous-time Markov chain, such as the one shown in Figure 2. Let the mean feedback delay be $1/\gamma$. The number of rows, referred to as levels, the values of A_i, α_i, β_i for each level, and the direction of the transition arrows of γ depend on the value of the threshold.

Below we describe the construction of the two-dimensional Markov chain which models the proposed method. The chain is constructed gradually, starting from level-1, where all sources are transmitting with normal rate, and terminating to a level where one of the following two conditions holds: i) all sources are transmitting with sustainable rate and therefore no more rate reduction can be performed, or ii) the threshold is so high that cannot be exceeded and therefore there is no need for rate reduction.

Next we describe the procedure of producing A_i, α_i, β_i and determining the direction of the arrows of γ. In each level-i, $E_i(\lambda_N)$ is the aggregate mean rate, $C_{Ni}(t)$ the aggregate auto-covariance, N_{Ni} is the number of sources in normal state and N_{Si} the number of sources in sustainable state.

Level-1

Since all sources are in normal state:

$$E_1(\lambda_N) = N_{N1}R_n = NR_n$$

$$C_{N1}(0) = N_{N1}C(0) = N\sigma_n^2$$

$$A_1 = \frac{C_{N1}(0)}{E_1(\lambda_N)} + \frac{E_1(\lambda_N)}{M} = \frac{N\sigma_n^2}{NR_n} + \frac{NR_n}{M} = \frac{\sigma_n^2}{R_n} + \frac{NR_n}{M}$$

$$\beta_1 = a/(1 + \frac{E_1^2(\lambda_N)}{MC_N(0)}) = a/(1 + \frac{NR_n^2}{M\sigma_n^2})$$

$$\alpha_1 = a - \beta_1$$

If T is the threshold, the transition to the next level will be performed when the aggregate rate is above T. This means that, according to the model, the threshold is exceeded when κ_1 or more minisources are transmitting simultaneously (Figure 2), where κ_1 is the smallest integer

satisfying $\kappa_1 A_1 > T$. This leads to:

$$\kappa_1 = \lceil \frac{T}{A_1} \rceil$$

where $\lceil x \rceil$ denotes the smallest integer greater than x.

If $\kappa_1 \leq M$, then level-2 exists and we must proceed to the next step. If $\kappa_1 > M$ then, according to this model, the threshold cannot be crossed, even if all M minisources are transmitting simultaneously, i.e., there is no need for rate reduction and the Markov chain is actually one-dimensional. This is a special case where the proposed method has no effect because the predefined threshold is very high to be reached.

Level-i

Assuming $\kappa_{i-1} \leq M$, level-i exists and we must compute the values of A_i, α_i, β_i and κ_i. To do so we must determine the number of sources that are in normal and sustainable state. Let, as in level-1:

$$E_i(\lambda_N) = N_{Ni} R_n + N_{Si} R_s$$

$$C_{Ni}(0) = N_{Ni}\sigma_n^2 + N_{Si}\sigma_s^2$$

$$A_i = \frac{C_{Ni}(0)}{E_i(\lambda_N)} + \frac{E_i(\lambda_N)}{M} = \frac{N_{Ni}\sigma_n^2 + N_{Si}\sigma_s^2}{N_{Ni}R_n + N_{Si}R_s} + \frac{N_{Ni}R_n + N_{Si}R_s}{M} \tag{5}$$

Using $N_{Ni} = N - N_{Si}$ in (5) yields:

$$A_i = f(N_{Si}) \doteq \frac{N\sigma_n^2 - (\sigma_n^2 - \sigma_s^2)N_{Si}}{NR_n - (R_n - R_s)N_{Si}} + \frac{NR_n - (R_n - R_s)N_{Si}}{M} \tag{6}$$

$$\beta_i = a/(1 + \frac{E_i^2(\lambda_N)}{MC_{Ni}(0)}) = a/(1 + \frac{(NR_n - (R_n - R_s)N_{Si})^2}{M(N\sigma_n^2 - (\sigma_n^2 - \sigma_s^2)N_{Si})}) \tag{7}$$

$$\alpha_i = a - \beta_i \tag{8}$$

We must calculate the minimum value of N_{Si} that reduces the total rate below the threshold T. In level-i the aggregate rate, when κ_{i-1} minisources are transmitting, must not exceed T, meaning:

$$A_i \leq \frac{T}{\kappa_{i-1}} \tag{9}$$

It is proven in the Appendix that the function $f(x)$, defined in (6), is strictly decreasing for $x \in [0, N]$. We distinguish two cases.

Case 1: $\frac{T}{\kappa_{i-1}} < f(N)$

In this case, it is impossible for the product $\kappa_{i-1} \cdot A_i$ to fall below the threshold T, even if all sources are transmitting with sustainable rate, since N is the maximum value that N_{Si} can take. Nevertheless, all sources are transferred to sustainable state to relax the heavy traffic load. Thus, level-i is the last level of the chain where all sources transmit with sustainable rate. From equations (6), (7), (8) we can compute the parameters of level-i by setting $N_{Si} = N$.

Case 2: $f(N) \leq \frac{T}{\kappa_{i-1}} \leq f(0)$

According to the analysis presented in the Appendix, there is only one solution of $f(x) = T/\kappa_{i-1}$ in the interval $[0, N]$, referred as x_s. Thus, the minimum N_{Si} satisfying (9) is $N_{Si} = \lceil x_s \rceil$. Having N_{Si} we can compute, according to (6), (7), (8), the parameters of level-i. The number κ_i is found as in level-1:

$$\kappa_i = \lceil \frac{T}{A_i} \rceil$$

If $\kappa_i \leq M$ and $N_{Si} < N$, we continue the procedure with level-$(i+1)$. If $\kappa_i > M$, threshold T cannot be crossed, and if $N_{Si} = N$ then all sources are in sustainable state. In both cases, level-i is the last level of the chain.

If, after the above procedure, the Markov chain has L levels, the arrows of γ have the following direction:

$$\begin{array}{llllll} \text{From} & nA_i & \text{to} & nA_{i+1} & \text{if } i < L \text{ and } n \geq \kappa_i \\ \text{From} & nA_{i+1} & \text{to} & nA_i & \text{if } i < L \text{ and } n < \kappa_i \end{array}$$

The transition rates in this chain are then defined as:

$$t_{nA_i, mA_j} = \begin{cases} (M-n)\alpha & \text{if } m = n+1 \text{ and } i = j \text{ and } n < M \\ n\beta & \text{if } m = n-1 \text{ and } i = j \text{ and } n > 0 \\ \gamma & \text{if } n = m \text{ and } j = i+1 \text{ and } i < L \text{ and } n \geq \kappa_i \\ \gamma & \text{if } n = m \text{ and } j = i-1 \text{ and } i > 1 \text{ and } n < \kappa_j \\ 0 & \text{otherwise} \end{cases}$$

Each level in the above analysis is equivalent with a model in which a buffer receives packets from a finite number of statistically independent and identical information sources that asynchronously alternate between exponentially distributed periods in the ON and OFF states. While ON, a source transmits at a uniform rate. The buffer is connected to an output channel with known capacity [9]. The number of sources is equal to the number of minisources M, the mean off time is equal to $1/\alpha_i$, the mean on time is $1/\beta_i$ and the bit rate in on state is A_i. The output rate is equal to the number of slots per frame, referred to as B. Assuming time-of-expiry for each packet between two and three frames, there is no time for retransmissions [3]. Thus, the buffer size in the equivalent model is equal to 0.

According to the analysis in [9] and [10], we can compute the expected overflow probability of the buffer in the equivalent model, which is equal to the loss probability in our model, as follows. Let $T(p,q)$ be the (p,q) element of rate transition matrix T of the system Markov chain; we have

$$\mathbf{T}(p,q) = \begin{cases} t_{nA_i,mA_j} & p \neq q, \quad p \leftrightarrow nA_i, \quad q \leftrightarrow mA_j \\ -\sum\limits_{mA_j} t_{nA_i,mA_j} & p \neq q, \quad p \leftrightarrow nA_i, \quad mA_j \in V_{nA_i} \end{cases}$$

where V_{nA_i} is the set of states from where state nA_i can be reached in one transition.

The drift matrix $\mathbf{D}$ is:

$$\mathbf{D}(p,q) = \begin{cases} n \cdot A_i - B & p = q, \quad p \leftrightarrow nA_i \\ 0 & p \neq q \end{cases}$$

Then the loss probability is equal to [10]:

$$P_{loss} = \sum_{nA_i} \sum_{j:z_j<0} a_j \cdot \Phi_j^{nA_i}$$

where the pairs $(z_j, \mathbf{\Phi}_j)$ are solutions to the eigenvalue problem:

$$\mathbf{D}z\mathbf{\Phi} = \mathbf{T}^T\mathbf{\Phi}$$

and the coefficients a_j are obtained from boundary conditions.

The sustainable state probability is given by:

$$P_{sustain} = \sum_{i=1}^{M} \sum_{j=1}^{L} P(iA_j) \cdot \frac{N_{Sj}}{N}$$

where L is the number of levels of the Markov chain, M the number of minisources, N_{Sj} the number of sources transmitting with sustainable rate in level-j and $P(iA_j)$ the equilibrium probability of state iA_j.

Finally the channel utilization is given by:

$$U_{CH} = \frac{E[v] \cdot (1 - P_{loss})}{B} \tag{10}$$

where $E[v] = \sum_{nA_i} n \cdot A_i \cdot P(nA_i)$ is the mean aggregate rate.

4 Numerical Results and Discussion

To show the improvement that the proposed method can attain, we consider a TDMA-based cellular environment with channel capacity of 1.92 Mb/sec and frame size of 12msec [11]. For compatibility with fixed ATM networks we use a slot size of 53 bytes (48 payload) to fit an ATM cell [3]. In this system we consider variable bit rate sources with mean bit rate 128 Kb/sec and deviation 64 Kb/sec. These characteristics can fit to low-resolution compressed video sources. The acceptable reduction fraction for every source is 0.5.

It is clear from the analysis in the previous section that the most important decision in the proposed method is the threshold setting. It depends on the agreed values for loss and sustainable state probability and the number of sources. A low threshold will increase the sustainable state probability, leading to poor utilization, while a high threshold will increase the loss probability and make the proposed method less effective. From the analysis we can compute the interval of threshold values that satisfy the agreements for both P_{loss} and $P_{sustain}$. As shown in Figure 3, the interval depends on the number of sources n, and it is between the value that corresponds to the maximum acceptable sustainable state probability, referred to as $t_s(n)$, and the value that corresponds to the maximum acceptable loss probability, referred to as $t_l(n)$. Clearly, we must have $t_s(n) \leq t_l(n)$. If $t_s(n) > t_l(n)$, then there is no value for the threshold that can simultaneously satisfy the agreed values for P_{loss} and $P_{sustain}$. For a small number of sources the interval is considerably large but, as more sources are entering the system, the interval gets smaller until $t_s(n) > t_l(n)$. In this way, we can compute the maximum

number of sources, referred to as N_{max}, that can be supported by the system. The algorithm that can be used is described below with a simple pseudo-program:

$for\ n = 1\ to\ \infty\ do$

$if\ (t_s(n) \leq t_l(n))\ \ AND\ \ (t_s(n+1) > t_l(n+1))$

$N_{max} = n$

$break$

$else$

$continue$

$endif$

$endfor$

As shown in Figure 3, the system under study can support up to seven sources, since $t_s(7) < t_l(7)$ and $t_s(8) > t_l(8)$ and the optimal threshold is in the interval $[t_s(7), t_l(7)]$. In the above algorithm it is assumed that for every number of sources the values $t_s(n)$ and $t_l(n)$ are available. The values of $t_s(n)$ and $t_l(n)$ needed in the algorithm described above are obtained using the analysis in the previous section.

Figure 4 shows the gain attained in loss probability when a constant threshold of 45 requests/frame is used for different numbers of sources. As we can see the gain increases as more sources are entering the system. This is explained as follows. When the number of sources is small (≤ 3) the particular threshold is hard or impossible to be reached, so the proposed method is never or seldomly activated and the gain of its use is imperceptible. As the load increases the threshold is more often reached and a selected number of sources each time are forced to transmit with sustainable rate. This has a dual effect: first it increases the gain in loss probability since the sources have reduced requirements, and second it produces a proportional increase of the sustainable state probability. So there is a trade-off to be made in order to keep a balance between the gain in loss probability and the increase in sustainable state probability.

In Figure 5, we can observe the maximum gain attained in loss probability, provided that the sustainable state probability has a guaranteed constant value equal to the maximum acceptable value agreed at connection setup. To do this we use different thresholds for different numbers of sources. For a given number of sources the threshold used is the one that leads to the

guaranteed value of sustainable state probability. As shown in Figure 3, this threshold leads to the minimum obtainable loss probability, thus to the maximun attainable gain for a given number of sources. We should mention that the gain shown in Figure 5 is quite big considering that the price paid by the sources is the probability to transmit with sustainable rate for a fraction of 10^{-5} of their total transmission time. It is obvious that if this number was bigger, say 10^{-3}, the gain in loss probability would have been more impressive. Note also in Figure 5 that the biggest improvement occurs at medium loads. For light loads (number of sources ≤ 3) there is very little space for improvement in loss probability, since it is already very low. Such improvement is probably imperceptible and meaningless for the end user. On the other hand, when the load is very heavy, the network is congested, regardless of the bit rate of the sources (normal or sustainable). Additionally, the threshold value cannot be further reduced to give better improvement in loss probability because this would increase the sustainable state probability as well.

Figure 6 shows the utilization of the available bandwidth for different number of sources and different values of the threshold. Observe that for the same number of sources, the utilization for zero threshold is more than half of the utilization for infinite threshold, although the reduction fraction is 0.5. For example, for eight sources, these two values are 0.31 and 0.56 respectively. This is because in the infinite threshold case all sources are constantly transmitting with normal rate leading to high loss probability. On the contrary, in the zero threshold case the loss probability is much lower since all sources are constantly transmitting with sustainable rate and this reduces the impact of reduction in utilization.

Another interesting observation in the same figure is that the utilization is not a monotonically increasing function of the threshold; depending on the number of sources, there is an interval over which the utilization decreases. This can be explained as follows: According to equation (10), two are the main parameters that determine the value of the utilization: the mean aggregate rate and the loss probability. As the threshold value increases, both these parameters increase as well. In a particular interval the increase of the loss probability is sharper than the increase of the mean aggregate rate, leading to decrease of the utilization. As we can observe, this interval moves to the right as more sources are entering the system.

5 Conclusions

We have presented a method for congestion control in the radio interface of wireless personal communication networks (PCN) where multimedia traffic is considered. The method is based on the concept of graceful degradation, where a source may accept a dynamic variation in service quality by accepting a temporary reduction in quality rather than being rejected, provided that sufficient bandwidth is not available. In this way, more sources can be supported simultaneously in a single cell. Furthermore, under heavy load conditions, a controlled reduction of requirements may be preferred by the sources compared to sudden quality degradations.

The method monitors the total bit rate of all active sources and, when this rate exceeds a predefined threshold, forces some sources to reduce their bandwidth requirements in order to reduce the total rate below the threshold. Under light load conditions the method permits some sources to increase their bit rates, provided that the total rate is kept below the threshold. The selection of the sources to decrease/increase their rates is performed in a way that guarantees fairness.

With an extensive performance analysis we have quantified the improvement that the method can attain in loss probability, at the expense of a controlled quality reduction for an acceptable fraction of time. This result is considered very important for variable bit rate, bandwidth demanding sources (e.g., video). We have also found the optimal threshold values in order to maximize system performance. Additionally, we have proposed an algorithm to calculate the maximum number of sources with known requirements that can be supported.

Appendix

Let

$$f(x) = \frac{A - Bx}{\Gamma - \Delta x} + E - Zx$$

where

$$A = N\sigma_n^2 \geq 0 \qquad B = \sigma_n^2 - \sigma_s^2 \geq 0 \quad \Gamma = NR_n \geq 0$$

$$\Delta = R_n - R_s \geq 0 \quad E = NR_n/M \geq 0 \quad Z = (R_n - R_s)/M \geq 0 \quad (11)$$

We will prove that $f(x)$ is strictly decreasing. The first derivative of $f(x)$ is:

$$f^{'}(x) = \frac{(A - Bx)^{'}(\Gamma - \Delta x) - (A - Bx)(\Gamma - \Delta x)^{'}}{(\Gamma - \Delta x)^2} - Z \rightsquigarrow f^{'}(x) = \frac{A\Delta - B\Gamma}{(\Gamma - \Delta x)^2} - Z \quad (12)$$

It holds that $A\Delta - B\Gamma \leq 0$ (13) since:

$$A\Delta - B\Gamma \leq 0 \leftrightarrow \sigma_s^2 R_n - \sigma_n^2 R_s \leq 0 \leftrightarrow$$

$$(1 - r)^2 \sigma_n^2 R_n - \sigma_n^2 (1 - r) R_n \leq 0 \leftrightarrow$$

$$(1 - r)^2 \sigma_n^2 R_n \leq (1 - r)\sigma_n^2 R_n \leftrightarrow$$

$$1 - r \leq 1 \leftrightarrow r \geq 0$$

which is obvious since r is the reduction fraction. From (11), (12), (13) we conclude that $f^{'}(x) \leq 0$. The quality holds for the special case of $r = 0$. Assuming $r > 0$, $f^{'}(x) < 0$ for every real x, which means that $f(x)$ is strictly decreasing.

$f(x)$ has an asymptote at $x = N/r$, which is the root of the denominator. There is one more asymptote, which is found as follows:

$$\lim_{x \to \infty} \frac{f(x)}{x} = -\frac{R_n r}{M}$$

$$\lim_{x \to \infty} [f(x) + \frac{R_n r}{M} x] = \frac{M\sigma_n^2(2 - r) + NR_n^2}{MR_n}$$

Thus the line

$$y = -\frac{R_n r}{M} x + \frac{M\sigma_n^2(2 - r) + NR_n^2}{MR_n}$$

is an asymptote of $f(x)$. The graph of function $f(x)$ is shown in Figure 7.

References

[1] Donald C. Cox, "Wireless Communications: What is it?", IEEE Personal Communications, April 1995.

[2] Joao S. DaSilva, Bosco E. Fernandes, "The European Research Program for Advanced Mobile Systems", IEEE Personal Communications, Feb. 1995.

[3] Nikos Passas, Nikos Loukas, Lazaros Merakos, "A Leaky-Bucket-Based Scheduling Technique for Wireless Personal Communication Networks", submitted to "1996 International Zurich Seminar on Digital Communications".

[4] Mischa Schwartz, "Network Management and Control Issues in Multimedia Wireless Networks", IEEE Personal Communications, June 1995.

[5] Belkacem Kraimeche, Mischa Schwartz, "Analysis of Traffic Access Control Strategies in Integrated Service Networks", IEEE Trans. on Communications, vol.COM-33, no.10, Oct. 1985.

[6] Nanying Yin, Michael G. Hlychyj, "On Closed-Loop Rate Control for ATM Cell Relay Networks", IEEE INFOCOM'94.

[7] Basil Maglaris et al., "Performance Models of Statistical Multiplexing in Packet Video Communications", IEEE Trans. on Communications, vol.36, no.7, July 1988.

[8] Nanying Yin, Michael G. Hlychyj, "A Dynamic Rate Control Mechanism for Source Coded Traffic in a Fast Packet Network", IEEE INFOCOM'91.

[9] D.Anick, D.Mitra, M. M. Sondhi, "Stochastic Theory of a Data-Handling System with Multiple Sources", Bell System Technical Journal, vol.61, no.2, Oct. 1982.

[10] L.K. Reiss, L.F. Merakos, "Adaptive Bandwidth Reservation for Traffic Streams Sharing an ATM Virtual Path", IEEE GLOBECOM'93.

[11] Dipankar Raychaudhuri, Newman D. Wilson, "ATM-Based Architecture for Multiservices Wireless Personal Communication Networks", IEEE JSAC, vol.12, no.8, Oct. 1994.

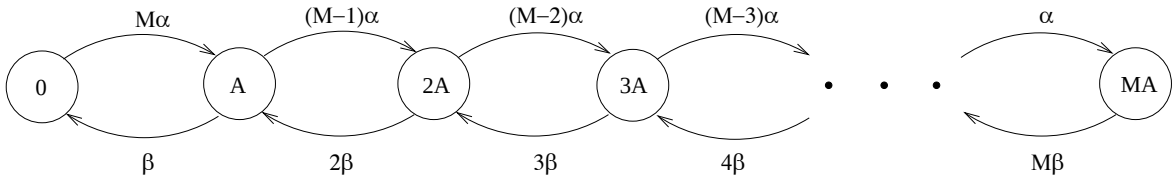

Figure 1: Birth-death Markov chain

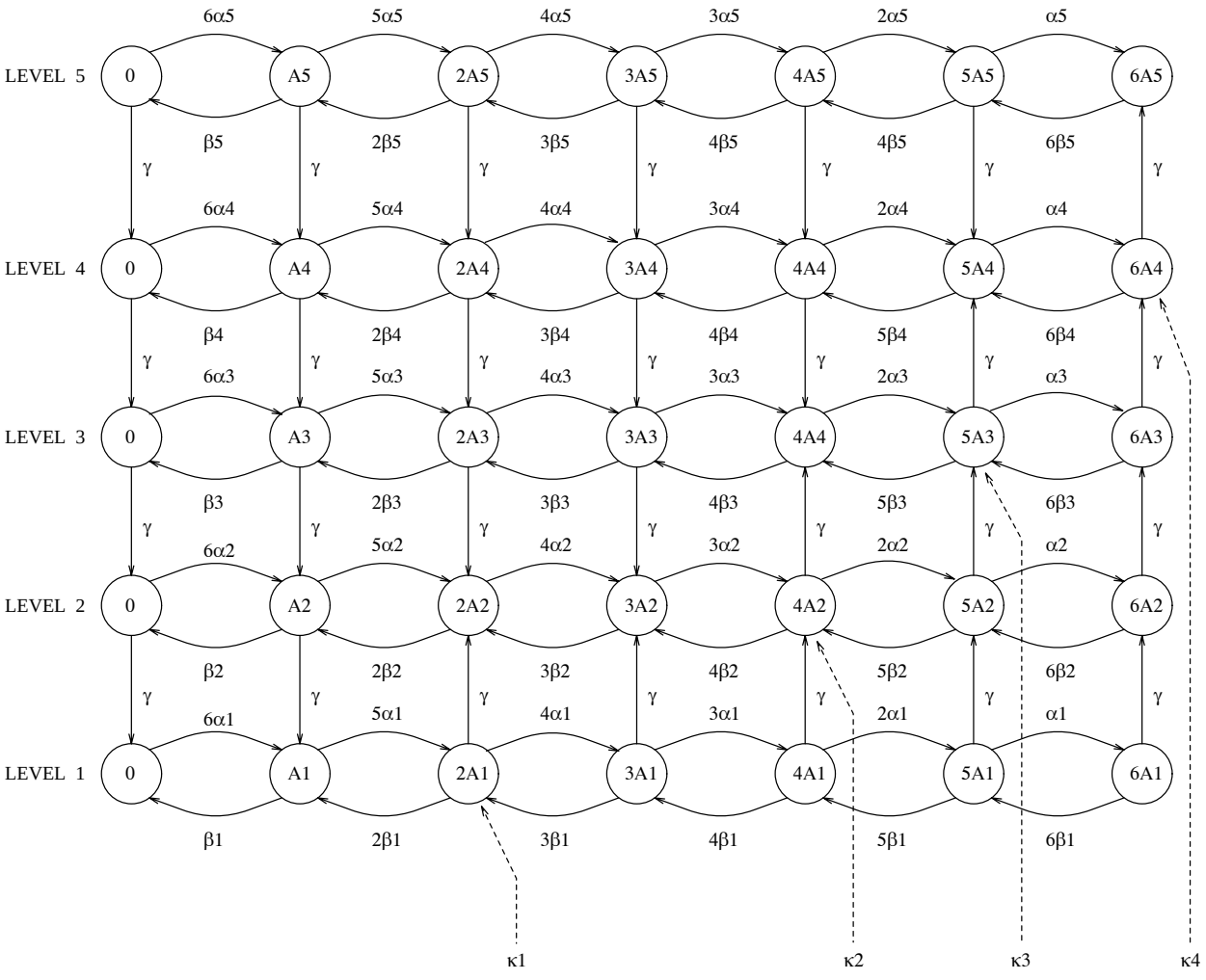

Figure 2: Two-dimensional Markov chain representing the proposed method

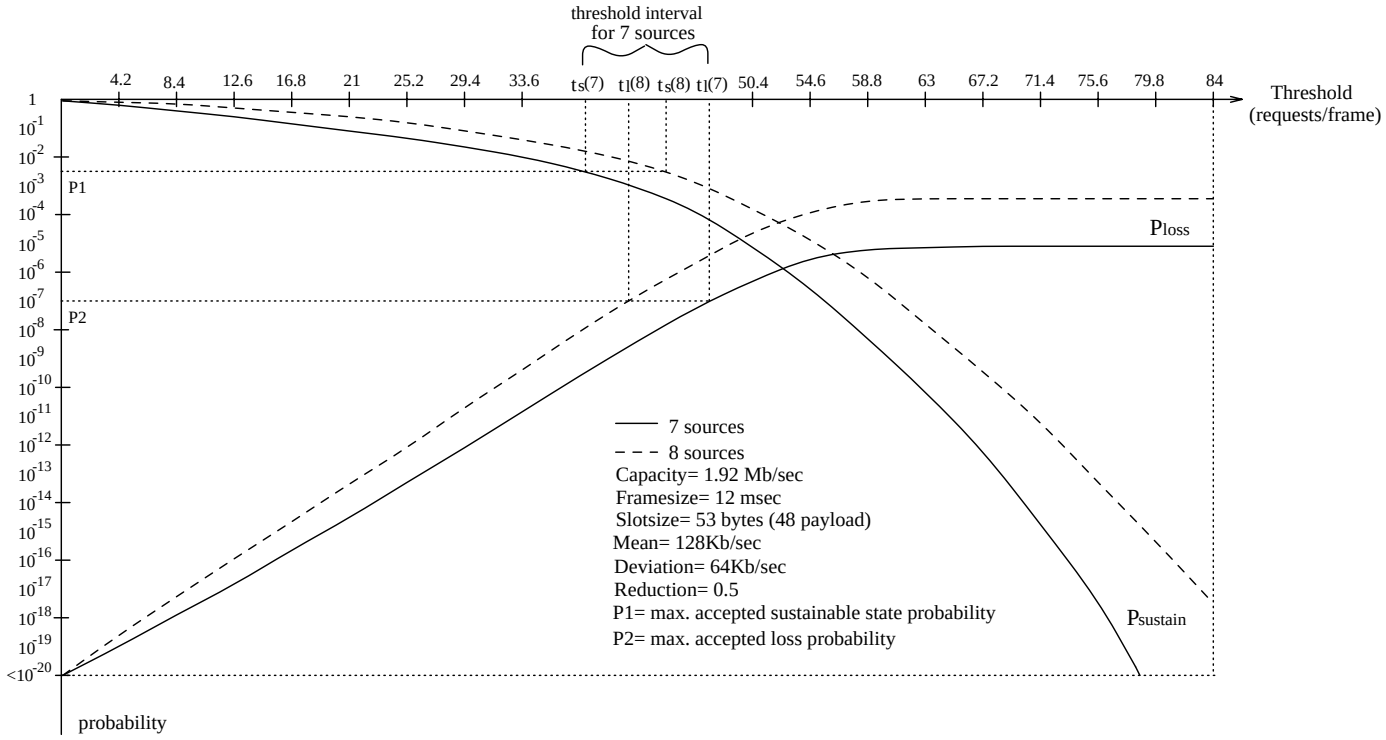

Figure 3: Loss and sustainable state probability for different values of the threshold

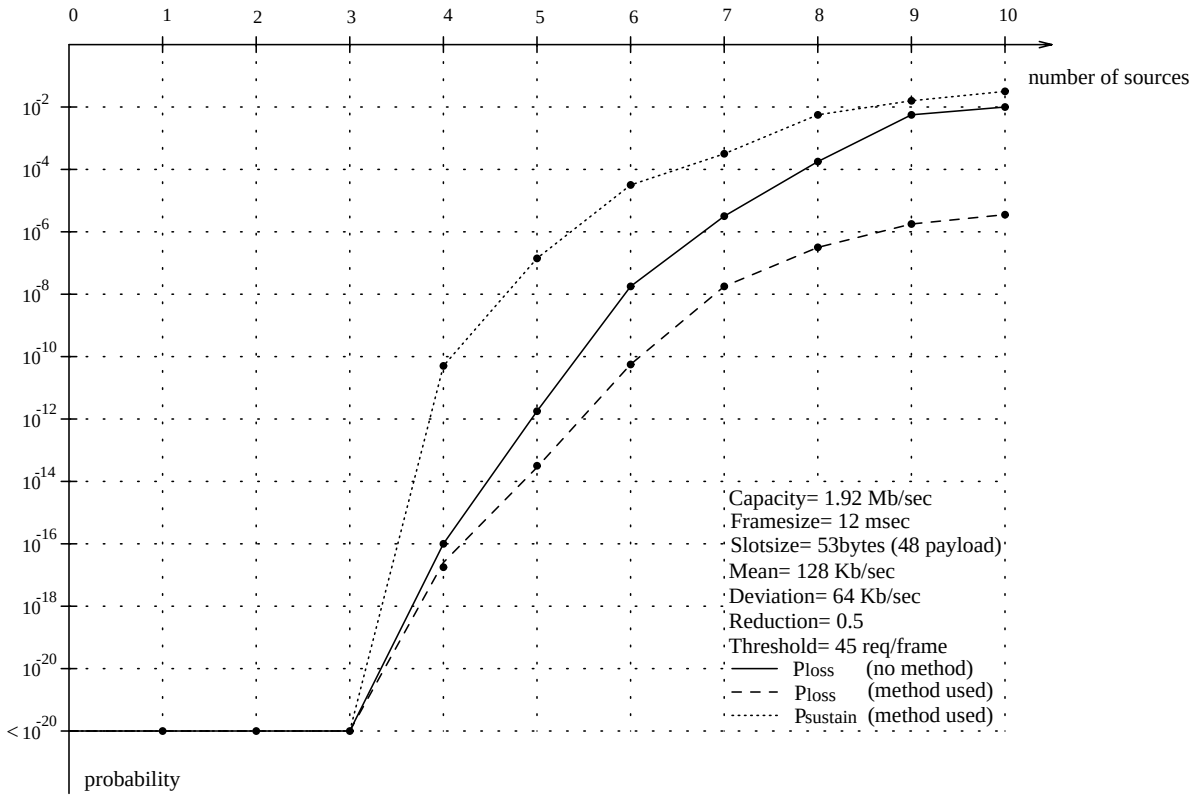

Figure 4: Loss and sustainable state probability for constant threshold

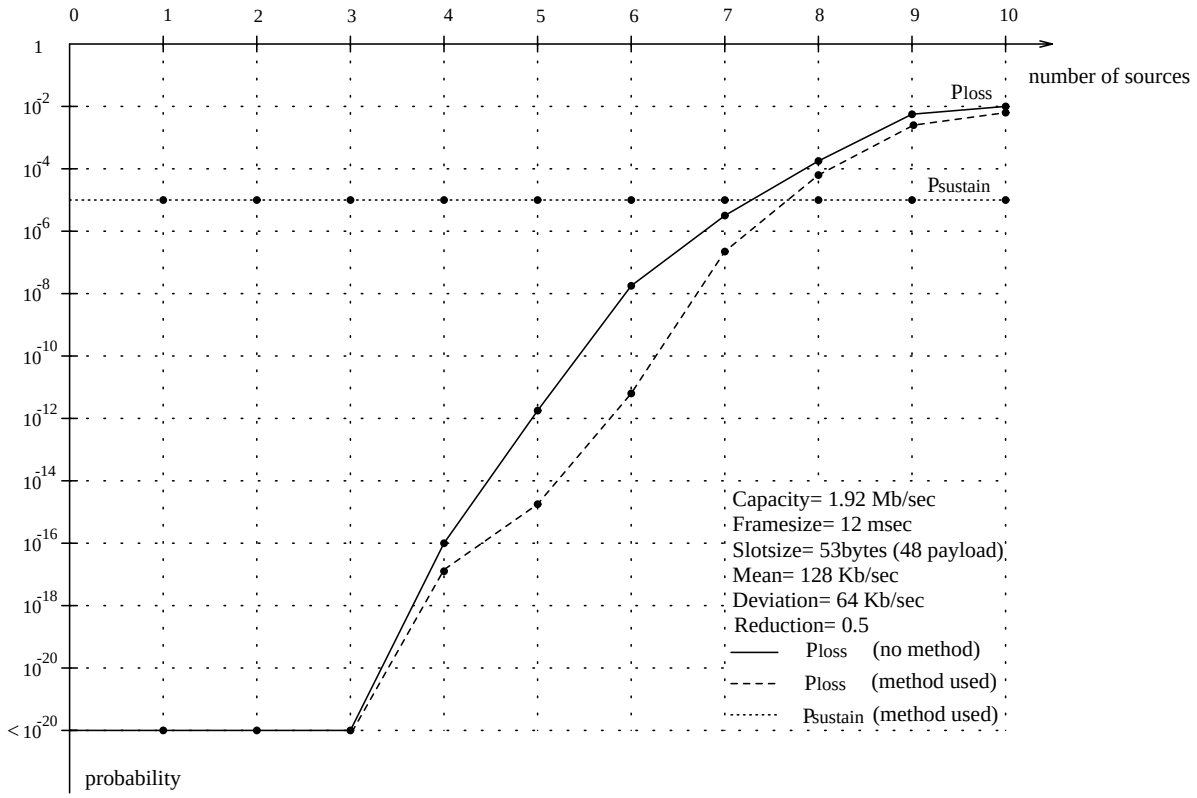

Figure 5: Loss and sustainable state probability for variable threshold

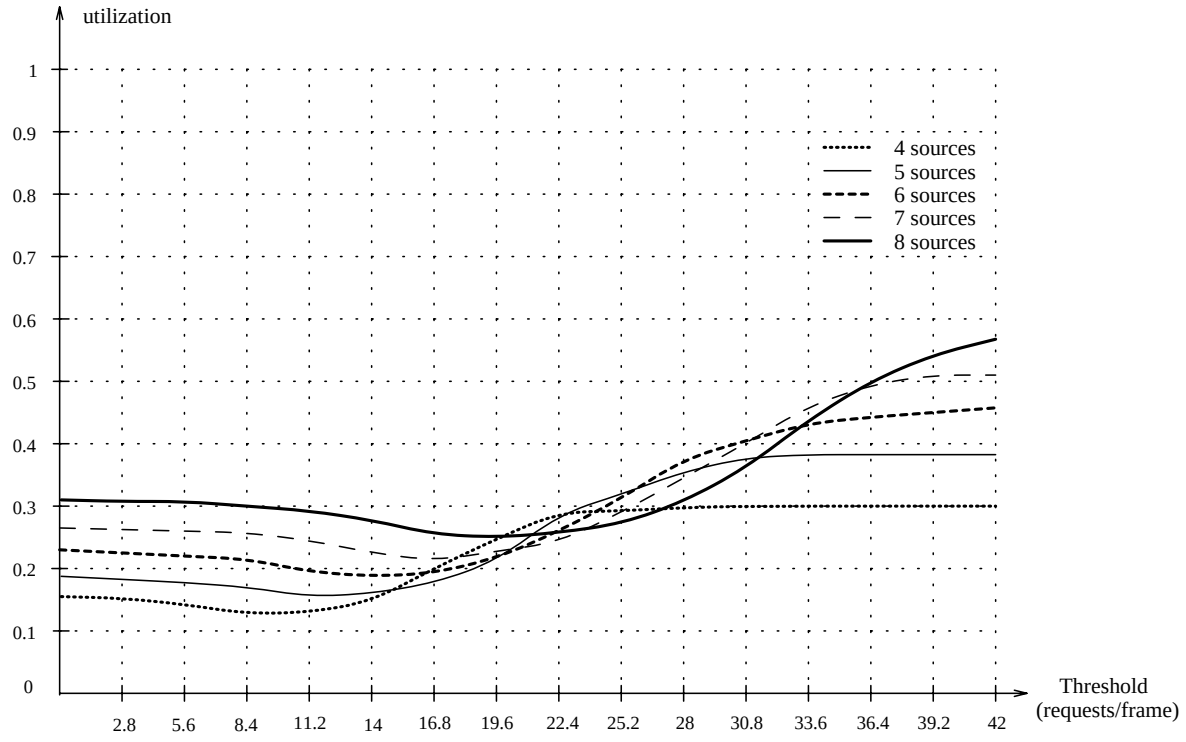

Figure 6: Utilization for different number of sources and variable threshold

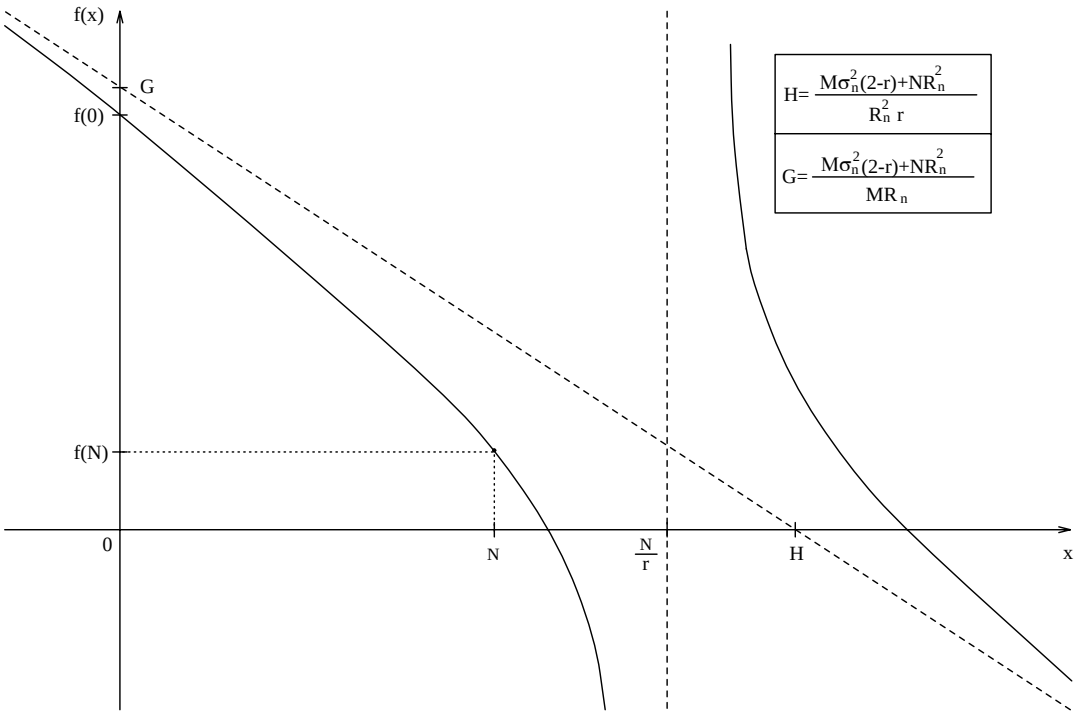

Figure 7: The graph of function f(x)