TOKEN BUCKET DIMENSIONING FOR AGGREGATED VOIP SOURCES

R. Bruno, R. G. Garroppo and S. Giordano
{bruno@netserv., garroppo@, giordano@}iet.unipi.it

Department of Information Engineering
University of Pisa
Via Diotisalvi 2
56126 Pisa - Italy
Tel. +39 050 568511, Fax +39 050 568522

Abstract

The paper presents an analytical study aimed to establish a dimensioning procedure for the *token bucket algorithm* when a stochastic model for the multiplexed traffic flow is available. The framework of this study is the *Differentiated Services Architecture* for supporting QoS (Quality of Service), where we utilise the token bucket filter as a metering algorithm. The proposed analysis, based on an equivalent queueing system, has been carried out considering an aggregation of fluidic On-Off processes with exponentially distributed sojourn times in each state, used to model telephone sources with VAD (Voice Activity Detection). For testing the goodness of our approach we have carried out discrete events simulations, which have highlighted the accuracy of the proposed dimensioning procedure of token bucket algorithm in a VoIP (Voice over IP) scenario.

1. Introduction

A key challenge of the current telecommunication era is represented by the developing of new architecture models for Internet aimed to satisfy the recent QoS requirements of innovative IP-based services (for example IP telephony and videoconferencing). The relevance of this issue is related to the transformation of Internet to a commercial infrastructure able to provide differentiated services to users with widely different service requirements. In this scenario, where the need of an architecture model able to provide different QoS classes is rapidly becoming as important as increasing the bandwidth availability, the *DiffServ* (differentiated services) approach is the most promising for implementing scalable service differentiation in Internet. The scalability is achieved by considering the aggregated traffic flows and conditioning the ingoing traffic at the edge of the network. Hence, it proposes to provide QoS employing a small, well-defined, set of building blocks enabling a large variety of services [BLAKE][BERNET]. These building blocks include a small set of per-hop forwarding behaviours, packet classification and traffic conditioning functions such as metering, marking and shaping.

A packet entering in a *DiffServ* domain is classified by a classifier, which establishes the particular service that the network should offer to the packet, considering the content of different fields of his header (e.g. source and destination address, source and destination port and so on). The service distinction is obtained setting the DS field of the packet header with a particular sequence of bits. The router in the core network schedules the forwarding of each packet using the DS information: in this manner each node has not to maintain a state per each microflow, but only a static behaviour table from which it chooses the corresponding PHB (Per-Hop Behaviour) for each packet. After the classification, each packet is passed to a meter that measures the temporal properties of the packets stream selected by the classifier against a traffic profile specified in a TCS (Traffic Conditioning Specification). Moreover, the meter determines if a packet is *in-profile* or *out-profile* according to whether it is conforming or not to the TCS. In-profile packets can enter in the domain without other conditioning, while out-profile packets could be refused by the network (dropping), delayed to enforce the respect of TCS (shaping) or aggregated to another service class with less strict requirements (remarking).

In this framework, the selection of a proper traffic descriptor (which permits to specify the TCS) is fundamental to achieve an efficient allocation of resources among the users, and to design correctly the scheduling algorithms permitting to guarantee differentiated QoS. After the service provider reaches an agreement (SLS, Service Level Specification) with the customer about the overall features and performances of the assured level of service, it has to traduce this agreement in detailed performance parameters of service and in a traffic profile. If the customer offers out-profile traffic, he will not receive the expected QoS. Most of researchers agree in considering the token bucket algorithm as an efficient meter and in describing the traffic profile by means of the parameters obtained by the LBAP (Linear Bounded Arrival Processes) traffic characterisation [CRUZ]. In particular, for a source modelled by a LBAP it is assumed that, in any interval of length t, the number of transmitted bits is upper bounded by $\lambda t+b$, where λ and b represent the two parameters used to characterise the source. The question, which we would answer in this paper, is how to calculate the operating parameters of the LBAP description if we know the stochastic model of traffic.

In Section 2, we mathematically define the token bucket algorithm, pointing out the reasons that make it one of the most interesting metering schemes. Then, we propose a procedure for analytical analysis and dimensioning of the token bucket filter when a stochastic model of the ingoing traffic is available. We apply our approach considering the aggregation of fluidic On-Off sources, each one corresponding to a model for voice traffic generated by a PCM codec with VAD. At the end, we test the correctness of proposed procedure by means of discrete events simulations driven by the considered packet voice sources.

2. The Token Bucket Filter

As discussed briefly in the introduction a network client must characterise its traffic load offered to the network using the parameters of LBAP model. The traffic is then metered by a token bucket filter based on LBAP parameters (λ and b): the filter logically consists of a pseudo-buffer (bucket) of size b, which is filling up by tokens at rate λ. Every time a packet of size p enters in the node, if in the bucket p tokens are available the packet is considered in-profile and p tokens are removed. Otherwise the packet is considered out-profile. A traffic source is in keeping with a token bucket filter (λ, b) if there are always enough tokens in the bucket whenever a packet arrives. In a more formal manner, defining with t_i and p_i respectively the arrival time and size of packet *i-th*, a source is conforming to a token bucket filter (λ, b) if the sequence n_i, defined by $n_0=b$ and $n_i=MIN[b, n_{i-1}+(t_i-t_{i-1})\lambda-p_i]$, is non negative for all i (the variable n_i represents the number of tokens in the bucket filter after the leaving of *i-th* packet) [CLARK]. Hence, for a given traffic source, we can define the non-increasing function $b(\lambda)$ as the minimal value such that the process is in keeping with a $(\lambda, b(\lambda))$ filter. The traffic filtered by a token bucket will never send more than $b+\tau\lambda$ tokens of data in an interval τ and the long-term mean transmission rate will not exceed λ. This is simply obtained with a counter of tokens in the bucket and a timer producing the new tokens at prefixed rate.

2.1 Equivalent Queueing Model

In this subsection, supposing to know the stochastic model of traffic process, we deal with the problem related to the determination of *(λ, b)* parameters such as the probability that a packet is not conforming, denoted as P_{nc}, is bounded to a fixed value.

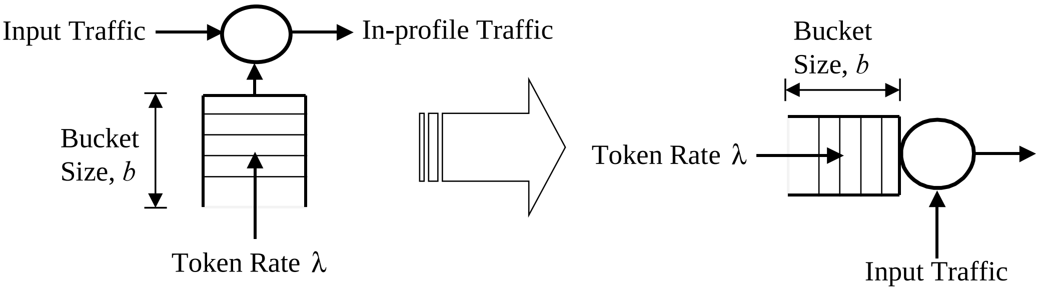


Figure 1 – Token bucket filter and equivalent queueing model

For evaluating P_{nc}, it is necessary to determine the probability to have a number of tokens in the bucket less than the packet size, when a packet arrives. Considering the formal definition presented above, the evaluation of this probability is strictly related to the stochastic features of n_i representing the bucket occupancy at time t_i. For this aim, we propose an equivalent queueing system permitting, by means of the stationary buffer state, to determine the marginal distribution of n_i and, hence, P_{nc}. The equivalent queueing model is a single server system, whose service process is represented by the metered traffic, the buffer size is equal to *b* tokens and the arrival process is deterministic with rate λ (see figure 1). In this modelling structure, if there are p_i tokens in the buffer, these are served at t_i (arrival time of *i-th* packet in the real system), otherwise no tokens are served. Hence, the event of packet not conforming is equivalent to have less than p_i tokens at the time t_i.

2.2 Case study

The proposed approach, based on the study of the equivalent queueing system, can be applied using different traffic models, although practically only in few cases the solution of the problem is easily achievable. In particular, we consider the problem to characterise with the LBAP approach an aggregation of independent VoIP sources, each one modelled as a fluidic On-Off source, where the sojourn times in each state is exponentially distributed. The On-Off assumption can be easily related to the typical behaviour of a voice source with VAD (Voice Activity Detection): it is active or inactive depending on the talker is speaking or silent. Assuming that no compression is applied to voice signal, during active periods the source transmits at the constant bit rate of ν=64 Kb/sec (this corresponds to a standard PCM codec with silence suppression). Moreover, in-depth analyses of this traffic source, shown in literature, have emphasised that the distribution of active and inactive periods lengths can be approximated by an exponential function [DAIGLE]. Recalling the theory developed in [MITRA][ANICK], we can use an efficient computational procedure to calculate *(λ, b)* parameters of aggregated traffic. We assume that each source has a transition matrix

$$\mathbf{G} = \begin{pmatrix} -\alpha & \alpha \\ \beta & -\beta \end{pmatrix} \tag{2.1}$$

where 1/α and 1/β are the mean Off and On periods lengths respectively. The traffic offered to the token bucket is represented by the sum of N statistically independent and identical of such On-Off sources. The number of sources active at time *t*, indicated as Σ_t, is a birth-and-death Markov Chain, where it is possible to jump from state *i* into state *i-1* with rate *iβ*, and into state *i+1* with rate *(N-i)α*. Hence, the *(i,j)-th* element of transition matrix is

$$[\mathbf{M}]_{ij} = \begin{cases} i\beta & j = i-1 \\ -\left(i\beta + (N-i)\alpha\right) & j = i \\ (N-i)\alpha & j = i+1 \\ 0 & otherwise \end{cases} \tag{2.2}$$

Denoting by S={0, 1, ..., N} the space of aggregated sources states, we recall that if $\Sigma_t = i$, $i \in S$, then the rate of the fluidic transmission is *iν*. Moreover, let X_t be the total fluid content of the bucket at time *t*, the steady state distribution of the system is given by

$$\pi_i(x) = \lim_{t\to\infty} \Pr[\Sigma_t = i; X_t \le x] \quad (i \in S, 0 \le x \le b) \tag{2.3}$$

Indicating with π(x)=[π_0(x) π_1(x) ... π_N(x)] and using the approach presented in [MITRA, p. 652], the dynamic of the system can be described by a set of differential equations:

$$\frac{\partial \pi(x)}{\partial t} + \frac{\partial \pi(x)}{\partial x}\mathbf{D} = \pi(x)\mathbf{M}, \; t \ge 0 \;\; and \;\; 0 < x < b \tag{2.4}$$

where

$$\mathbf{D} = diag\{\lambda, \lambda - \nu, \cdots, \lambda - N\nu\} \tag{2.5}$$

is referred as *drift matrix*. The *i-th* element on the diagonal of the matrix **D** is the rate (or drift) of bucket content changing (away from the boundaries) when the Markov Chain is in the state *i*.

The steady state distribution, π(x), can be derived considering the time-independent part of (2.4), i.e.

$$\frac{d}{dx}\pi(x)\mathbf{D} = \pi(x)\mathbf{M} \quad (0 < x < b) \tag{2.6}$$

Considering the conditions to be satisfied at *x=0* and *x=b* (shown in the Appendix, where we recall the procedure presented in [MITRA] for the solution of the above equation), the solution of (2.6) gives the steady distribution π(x) that permits to evaluate P_{nc}. This quantity, corresponding to the probability of having the bucket empty when at least a source is active, is equal to

$$P_{nc} = \Pr[\mathrm{X}_t = 0 \mid \Sigma_t \geq 1] = \frac{\sum_{i=1}^{N}\pi_i(0)}{1 - \Pr[\Sigma_t = 0]} \tag{2.7}$$

3. Simulation Analysis

In this Section we present the test of the analysis previously described, carried out by means of discrete events simulations. During the simulations, each traffic source is assumed packet-based and its behaviour is On-Off. The distribution of active and inactive periods lengths is assumed exponential with mean values equal respectively to 1/β=350 ms and 1/α=650 ms (these values are chosen using the results presented in [DAIGLE]). Moreover, during active periods the source generates packet at regular intervals (i.e. with constant interarrival times), corresponding to a bit rate of 64 Kb/sec. Considering the packet-based nature of real source, the main approximation introduced by the fluidic analysis presented in the previous Section is evident. Therefore, a first set of simulations is aimed to evaluate the errors introduced by this approximation considering a reasonable value of the packet size, i.e. 84 bytes, which corresponds to a filling time of about 10 ms. On the other hand, also tokens are not fluidicly generated and, considering the constant packet size, it is assumed that each packet consumes a single token of the bucket. The aggregated traffic is obtained by a multiplexing module that approximates the behaviour of Markov Chain (see relation (2.2)) used to model the number of active sources. The output of multiplexer is then passed to a classifier, which inserts the DS field in the packets, and to the token bucket regulator that drops every out-profile packet. The simulation environment is OPNET network simulator, release 6.0, and the scenario related to the aggregation of 15 traffic sources is shown in figure 2. In each scenario, the results are obtained repeating the simulation five times with different values of the *seed*. Moreover, each session has an average length of about 5.000.000 packets.

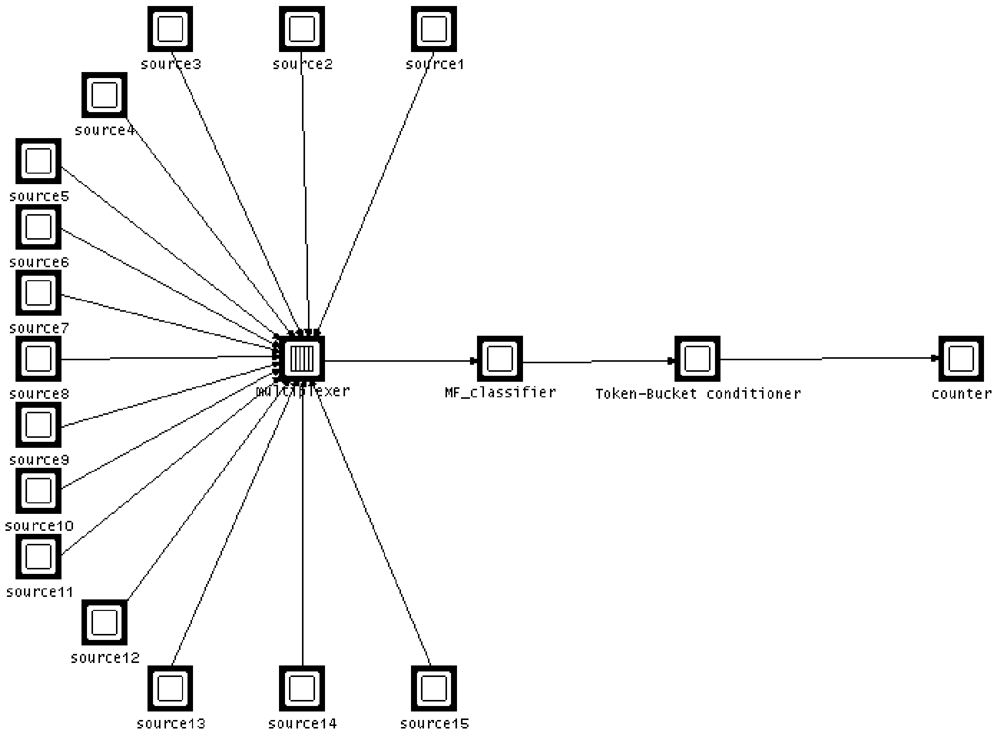
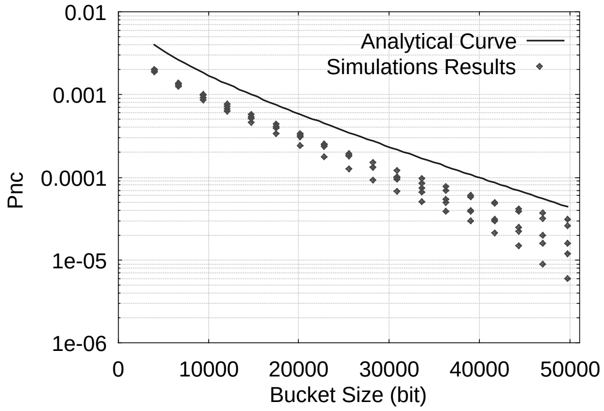


Figure 2 - Simulation scenario with 15 On-Off sources



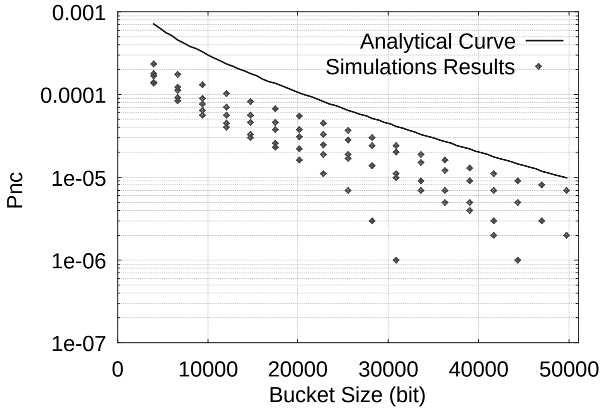


Figure 3 – Comparison of analytical and simulation results for R=61% and N=15

Figure 4 - Comparison of analytical and simulation results for R=55% and N=45

The simulations results are shown in figures 3 and 4 for two different values of multiplexed sources, i.e. N equal to 15 and 45 respectively. The curves represent the estimated P_{nc} as a function of bucket size for a fixed value of R, which is the token rate expressed as a percentage of the peak rate of aggregated traffic, equal to Nν (the token rate λ is equal to 0.01*RNν).

In the two figures the values of R have been chosen to maintain the P_{nc} around 10^{-4}, which permits to have a relative short simulation time and, on the other hand, an acceptable value of packet loss probability for voice services (the last statement is related to the assumption that each non-conforming packet is lost). The results analysis emphasises the "correctness" of the proposed approach, as deducible observing that the analytical curve behaves in figures 3 and 4 as a tight upper bound of simulations results.

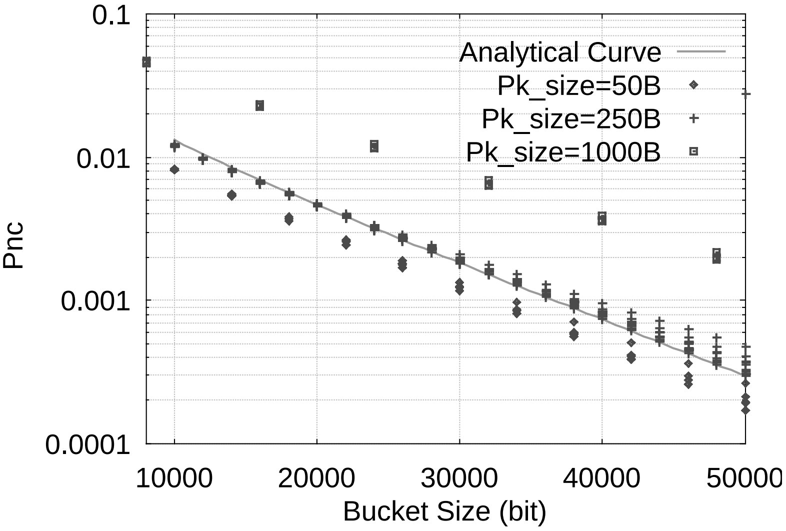


Figure 5 – Comparison of analytical and simulation results for different packet size - R=65% and N=5

On the other hand, figure 5 shows the impact of the packet size on the evaluation errors introduced by the fluidic approximation. In particular in the figure the simulations results are plotted for R=65%, N=5 and three different values of packet length (50, 250 and 1000 bytes). The obtained curves show worse or better behaviour with respect to the analytical results depending on the packet length is large or short (worse/better behaviour indicates a P_{nc}, obtained by simulations, higher/lower than that estimated using the analytical approach). In particular we can note that for packet size equal to 50 and 250 bytes the simulation results are very close to the analytical ones, while when a packet size of 1000 bytes is considered the analytical curve is quite optimistic with respect to the simulations results, highlighting, in this case, high evaluation errors.

3.1 Analysis of relations among Token Bucket parameters and P_{nc}

The first set of simulations has shown the accuracy of our analytical approach when a VoIP service is considered, but several open issues remain. Main problems are related to have different arrival process models that do not permit a relative simple analytical solution, and further tests are needed for evaluating the accuracy of the equivalent queueing approach with different traffic models. However, the proposed case study is of paramount relevance considering the increasing interest toward the VoIP service. After the test of the accuracy of the proposed approach, we further investigate the dependency of P_{nc} on the bucket size *b*, showing a set of numerical results of the queueing analysis presented in Section 2.2.

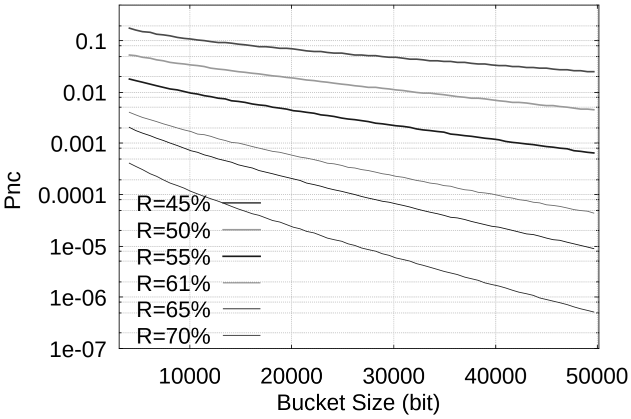


Figure 6 - P_{nc} vs. *b*, with N=15

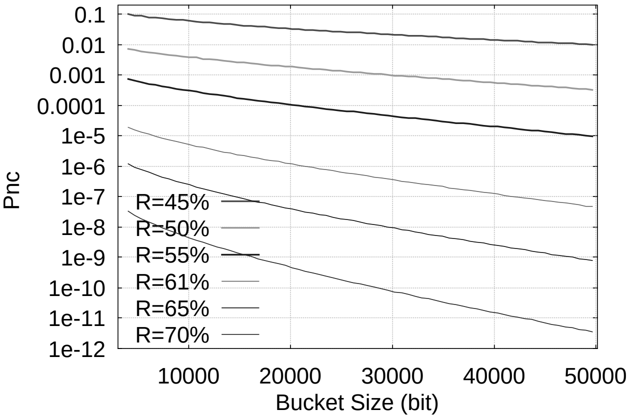


Figure 7 - P_{nc} vs. *b*, with N=45

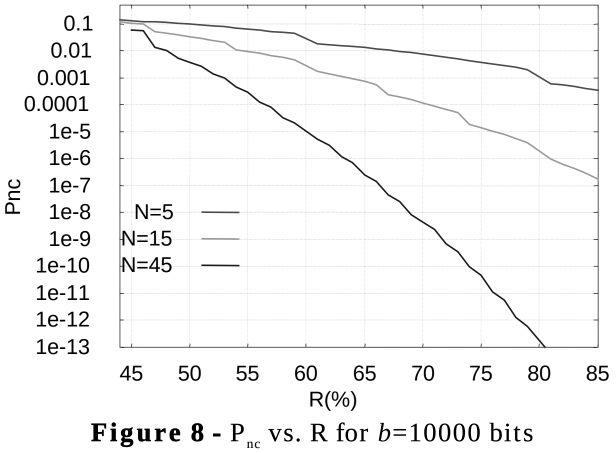


Figure 8 - P_{nc} vs. R for *b*=10000 bits

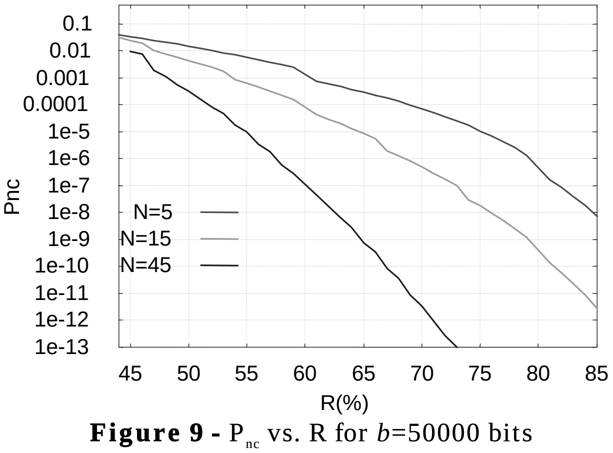


Figure 9 - P_{nc} vs. R for *b*=50000 bits

The curves, shown in figures 6 and 7, present the probability of out-profile packets as function of *b* and for different values of R. The two figures are related to two values of multiplexed sources, N equal to 15 and 45 respectively. The results analysis permits to evaluate the multiplexing gain achievable by means of the sources aggregation. Indeed, fixed the target P_{nc}, this is assured with a token rate that is a lower percentage of the peak rate when the number of aggregated sources increases, maintaining the same value of the ratio b/λ. For example, fixing a $P_{nc}=10^{-5}$, this is assured with a set of *(R,b)* equal to (70%, 26850) and (55%,49700), for N=15 and 45 respectively. In both cases the ratio b/λ is almost the same, indeed it is equal to 40 ms and 31 ms for N equal to 15 and 45 respectively. Summarising, from this example we can deduce that to maintain the same b/λ ratio, which impacts on the estimation of the end-to-end delay evaluated using the worst case analysis (see [PAREKH] where the analysis of a Generalized Processor Sharing network is presented), as higher is the number of multiplexed sources as lower is the token rate, expressed as a percentage of the peak rate, needed to have the same P_{nc}.

A second set of curves, aimed to investigate the dependency of P_{nc} on the token rate λ, confirms the previous results, as it can be noted in figures 8 and 9, which present the probability of out-profile packets as function of R and for two different values of bucket size *b*; the three curves plotted in each figure are related to different number of aggregated sources, N.

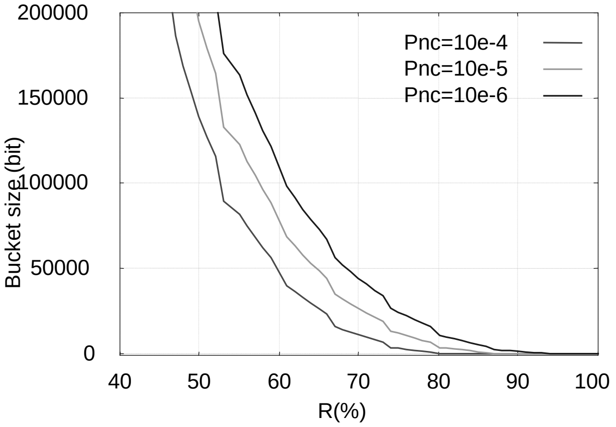


Figure 10 - (R, b(R)) curves for different P_{nc} and N=15

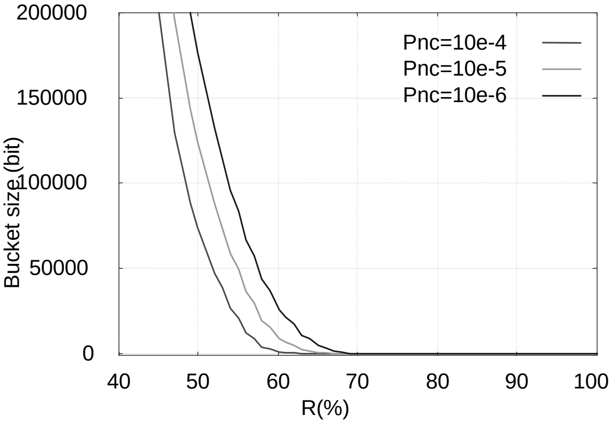


Figure 11 – (R, b(R)) curves for different P_{nc} and N=45

Figures 10 and 11 present the *(R, b(R))* curves for different P_{nc} and considering N equal to 15 and 45 respectively. These figures permit to observe as fixing a given bucket size, we can reduce the token rate simply accepting higher P_{nc}. Moreover from the comparison of the two figures we can observe the smoothing effects due to the aggregation process. Indeed, the curves in figure 11 related to N=45 have a higher slope for low values of R than the ones related to N=15. For example, as it can be deduced from the plot, with R equal to 60% and assuming $P_{nc}=10^{-4}$, the burstiness of traffic for N=45, measured by means of the necessary bucket size, is negligible. On the other hand, for N=15 and the same values of R and P_{nc}, the necessary *b* is equal to about 50000 bits, as deducible from figure 10.

To highlight the gain obtained by the sources aggregation, we present figure 12, where the curves related to $P_{nc}=10^{-6}$, N equal to 15 and 45 are plotted. Moreover in the figures are indicated the two points corresponding to a b/λ ratio equal to 20 ms (indicated as FD=20ms_N=15 and FD=20ms_N=45 for N equal to 15 and 45 respectively). To obtain the same b/λ ratio, in the case N=15 the necessary token rate is about the 79% of the peak rate, while for N=45 a token rate of about the 58% of the peak rate is sufficient to guarantee the same P_{nc}.

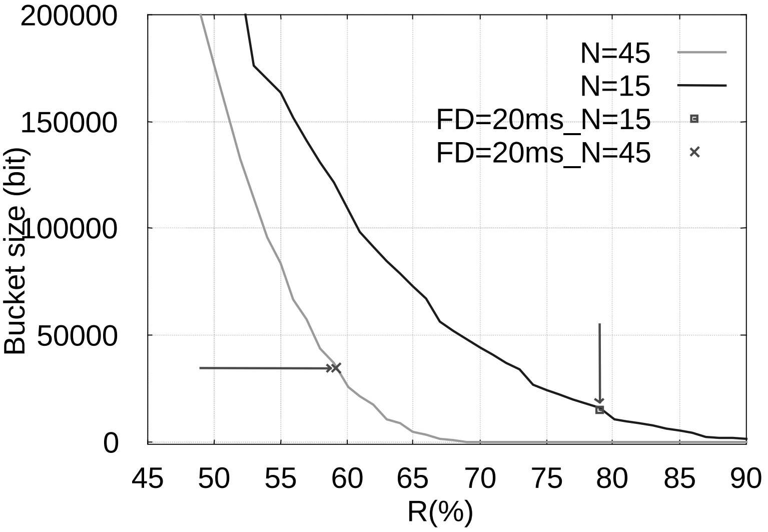


Figure 12 – Comparison of (R, b(R)) curves for $P_{nc}=10^{-6}$ and N equal to 15 and 45

4. Conclusions

The equivalent queueing system approach presented in the paper permits the estimation of the parameters of the LBAP characterisation using the stochastic model of the considered traffic flow.

The relevance of this approach is related to the necessity to have the LBAP parameters in an on-line manner considering the stochastic nature of traffic, while in literature the methods for the evaluation of the *(λ, b)* curve are often measurement-based. As described in the proposed approach, accepting a “well controlled” value of the probability of non-conforming packets and using the information on the stochastic nature of traffic it is possible to optimise the LBAP characterisation (i.e. to reduce the limit values of necessary token rate and bucket size) with respect to the formal definition of the token bucket [CLARK], which assumes P_{nc} equal to zero.

Moreover, as shown in the paper considering the relevant case study of the VoIP scenario, starting from a stochastic model for the single source, it is possible to estimate the LBAP parameters of the traffic flow offered by the aggregated homogeneous sources considering only the number of running voice applications. The simulation study reported in the paper has highlighted the accuracy of the proposed approach for low values of packet size that is relevant when the VoIP service is considered. Moreover, the study of the analytical results has highlighted the multiplexing gain achievable by the aggregation of VoIP sources.

Appendix A

The solution of equation (2.6) gives the distribution π(x) at all points except for the boundaries, i.e. *x=0* and *x=b*, since it may have jumps at these points. To completely specify the distribution, the following boundaries conditions are needed

$$\Pr[X_t = 0; \Sigma_t = i] = \pi_i(0) \qquad \Pr[X_t = b; \Sigma_t = i] = \Pr[\Sigma_t = i] - \pi_i(b) \tag{A.1}$$

Note that the problem of solving for π(x) cannot be specified as an initial value problem since the initial conditions are not known a priori. In the section A.2 we treat the solution of (2.6) as a two-points boundary-values with conditions to be satisfied at *x=0* and *x=b*. Considering the transformed space, the solution of (2.6) can be found by means of the corresponding eigenvalue problem

$$z\phi\,\mathbf{D} = \phi\,\mathbf{M} \tag{A.2}$$

and, if the relation $\sum_{i=0}^{N} \Pr[\Sigma_t = i][D]_{ii} \neq 0$ holds, they are given by

$$\pi(x) = \sum_{i=0}^{N} a_i \phi_i \exp(z_i x) \tag{A.3}$$

where $\{z_i, \phi_i\}$ are the left eigenvalues and eigenvectors, evaluated resolving (A.2) [MITRA, p.655].

A.1 Eigenvalues and eigenvectors

Denote with $\Phi(x)$ the generating function of eigenvector ϕ, i.e $\Phi(x) \hat{=} \sum_{i=0}^{N} \phi_i x^i$ where ϕ_i is the *i-th* components of ϕ. From (A.2) and supposing $\phi_i=0$ if $i\notin[0, N]$, we find that for $0\leq i\leq N$

$$z(\lambda - i\upsilon)\phi_i = (N-(i-1))\alpha\phi_{i-1} - (i\beta + (N-i)\alpha)\phi_i + (i+1)\beta\phi_{i+1} \qquad \text{(A.4)}$$

By multiplying (A.4) by x^i and summing over i we obtain the following equation

$$\frac{\Phi'(x)}{\Phi(x)} = \frac{-\lambda z - N\alpha + N\alpha x}{\alpha x^2 + (-\nu z + \beta - \alpha)x - \beta} \qquad \text{(A.5)}$$

The variable x in the denominator of the right-hand side has two real and distinct roots, say σ_1 and $\sigma_{2,}$, where $\sigma_1>0>\sigma_2$:

$$\sigma_{1,2} \mathrel{\hat{=}} \frac{-(-\nu z + \beta - \alpha) \pm \sqrt{Q(z)}}{2\alpha} \text{ where } Q(z) \mathrel{\hat{=}} (-\nu z + \beta - \alpha)^2 + 4\alpha\beta \qquad \text{(A.6)}$$

Hence (A.5) may be written as

$$\frac{\Phi'(x)}{\Phi(x)} = \frac{k}{x-\sigma_1} + \frac{N-k}{x-\sigma_2} \qquad \text{(A.7)}$$

where the residue k is

$$k = \frac{1}{\sqrt{Q(z)}}\left\{-\lambda z - \frac{N}{2}(-\nu z + \beta + \alpha) + \frac{N}{2}\sqrt{Q(z)}\right\} \qquad \text{(A.8)}$$

Finally, the solution of (A.5) is

$$\Phi(x) = (x-\sigma_1)^k\,(x-\sigma_2)^{N-k} \qquad \text{(A.9)}$$

For compatibility with the definition of Φ(x), the right-hand quantity must be a polynomial in x of degree N, and hence the quantity k must be an integer in *[0,N]*. Manipulating (A.8), we can introduce

$$f(z\,;k) \mathrel{\hat{=}} \left(k - \frac{N}{2}\right)\sqrt{Q(z)} + \frac{N}{2}(-\nu z + \alpha + \beta) + \lambda z \qquad \text{(A.10)}$$

For each integer value of k, with $0\leq k\leq N$, the real zeros of *f(z;k)* are eigenvalues [ANICK, p. 1877]. With simple manipulations we can simplify the problem of finding the zeros of *f(z;k)* eliminating computational complexities related to the square root function $\sqrt{Q(z)}$. To this aim, rearranging *f(z;k)* and squaring as necessary, we obtain a family of polynomials *P(z;k)* whose zeros coincide with those of *f(z;k)*. Denoting

$$L(z) \mathrel{\hat{=}} \frac{N}{2}(-\nu z + \beta + \alpha) + \lambda z \qquad \text{(A.11)}$$

and maintaining the definition of *Q(z)*, we obtain

$$P(z\,;k) = \begin{cases} \left(\frac{N}{2} - k\right)^2 Q(z) - L^2(z) & \text{if} \quad k \neq \frac{N}{2} \\ L(z) & \text{if} \quad k = \frac{N}{2} \end{cases} \qquad \text{(A.12)}$$

The following properties can be easily proven:

(i) *P(z;k)=P(z;N-k)*;

(ii) for each *k<N/2, P(Z;k)* has two simple and real zeros. When *N* is even and *k=N/2* the corresponding polynomial has a real zero;

(iii) there is always a null zero; it can be found considering the polynomial *P(z,0)*;

(iv) the set of eigenvalues coincides exactly with the set of zeros of polinomials *P(z;k)* with *0≤k≤N/2*.

Given the eigenvalues, using (A.6) we can calculate the quantities σ_1, σ_2, which are used for evaluating the coefficients of polynomial $\Phi(x)$ corresponding to the eigenvectors components. In particular for each *k=0, 1, …, N*, considering the corresponding σ_1 and σ_2 we calculate the *i-th* components of the *k-th* eigenvectors ϕ by [ANICK, p. 1879]

$$\phi_i = (-1)^{N-i} \sum_{j=0}^{k} \binom{k}{j} \binom{N-k}{i-j} \sigma_1^{k-j} \sigma_2^{N-k-i+j} \qquad 0 \le i \le N \tag{A.13}$$

A.2 Boundary conditions and problem solution

As stated before, the coefficients a_i of the solution (A.3) are obtained from boundary conditions. In particular, denoting with S_D and S_U the set of aggregated sources states which respectively give a downward and upward drift to the bucket, i.e

$$S_D = \{ i \mid [\mathbf{D}]_{ii} < 0 \} \text{ and } S_U = \{ i \mid [\mathbf{D}]_{ii} > 0 \}, \tag{A.14}$$

we can observe that, if $i \in S_D$, the event {number of active source equal *i* and bucket full} is impossible, except for isolated points in time, i.e. in the right-relation of (7) we have $\Pr[X_t = b; \Sigma_t = i] = 0$ when $i \in S_D$. Hence

$$\pi_i(b) = \Pr[\Sigma_t = i] \;\; for \;\; i \in S_D \tag{A.15}$$

Similarly, if $i \in S_U$ the event {number of active source equal *i* and bucket empty} is impossible, except for isolated points in time, i.e. in the left-relation of (A.1) when $i \in S_U$ we have $\Pr[X_t = 0; \Sigma_t = i] = 0$, which implies

$$\pi_i(0) = 0 \;\; for \;\; i \in S_U \tag{A.16}$$

Resolved the linear system derived by conditions (A.15) and (A.16), we have a complete knowledge of steady distribution π(x).

References

[ANICK] D. Anick, D. Mitra, M. M. Sondhi "**Stochastic Theory of a Data-Handling System with Multiple Sources**", Bell System Technical Journal, 61, pp. 1871-1894 (1982)

[BERNET] Y. Bernet, J. Binder, S. Blake, M. Carlson, E. B. Carpenter, S. Keshav, E. Davies, B. Ohlman, D. Verna, Z. Wang e W. Weiss "**A Framework for Differentiated Services**" Internet Draft, February 1999

[BLAKE] S. Blake, D. Blake, M. Carlson, E. Davies, Z. Wang e W. Weiss "**An Architecture for Differentiated Services**" RFC 2475, December 1998

[CLARK] D. Clark, S. Shenker, L. Zhang "**Supporting Real-Time applications in an Integrated Services Packet Networks: Architecture and Mechanism**" Proc. ACM SIGCOMM '92, pp. 14-26, August 1992

[CRUZ] R.L. Cruz "**A Calculus for Network Delay and a Note on Topologies of Interconnection Networks**" Ph.D. Thesis, Univ. Of Illinois, issued as Report UILU-ENG-87-2246, July 1987

[DAIGLE] J. N. Daigle, J. D. Langford "**Models for Analysis of packet Voice Communications Systems**", IEEE Journal on Selected Areas in Communications, 6, pp 847-855(1986)

[MITRA] D. Mitra "**Stochastic Theory of a Fluid Model of Producers and Consumers Coupled by a Buffer**", Advanced Application in Probability, 20, pp 646-676 (1988)

[PAREKH] A. K. Parekh, R. G. Gallager "**A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks: The Multiple Node Case**", IEEE/ACM Transactions on Networking, Vol. 2, N. 2, pp. 137-150, April 1994