Les Cahiers du GERAD ISSN: 0711–2440

Effective bandwidth and admission control: A tutorial

A. Girard

G–2023–55

November 2023

La publication de ces rapports de recherche est rendue possible grâce au soutien de HEC Montréal, Polytechnique Montréal, Université McGill, Université du Québec à Montréal, ainsi que du Fonds de recherche du Québec – Nature et technologies.

Dépôt légal – Bibliothèque et Archives nationales du Québec, 2023
– Bibliothèque et Archives Canada, 2023

The publication of these research reports is made possible thanks to the support of HEC Montréal, Polytechnique Montréal, McGill University, Université du Québec à Montréal, as well as the Fonds de recherche du Québec – Nature et technologies.

Legal deposit – Bibliothèque et Archives nationales du Québec, 2023
– Library and Archives Canada, 2023

GERAD HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada H3T 2A7

Tél. : 514 340-6053
Téléc. : 514 340-5665
info@gerad.ca
www.gerad.ca

Effective bandwidth and admission control: A tutorial

André Girard [a, b, c]

[a] *GERAD, Montréal (Qc), Canada, H3T 1J4*

[b] *Département de génie électrique, Polytechnique Montréal, Montréal (Qc), Canada, H3T 1J4*

[c] *Centre Énergie Matériaux Télécommunications, INRS, Varennes (Qc), Canada, J3X 1P7*

andre.girard@gerad.ca

November 2023
Les Cahiers du GERAD
G–2023–55

Abstract : We review the development of the concept of effective bandwidth from its origin in the planning and management of ATM networks. We start with the extension of the telephone network to multi-rate circuit-switched networks and then to bufferless networks with variable rate traffic. We also show how the concept can be extended to IP-type networks with buffered queues.

We use two simple traffic models that are amenable to exact computation to show the accuracy of two definitions of effective bandwidth, one based on a single-queue model and the other on the Chernoff bound of the distribution of a variable. We use this to illustrate the notion of admission region and its relation to effective bandwidth. We discuss the accuracy of the two definitions and the scope of applications.

Keywords : Traffic engineering, effective bandwidth, admission control, queuing

Résumé : Nous reprenons ici le développement du concept de bande passante effective comme outil pour la planification et la gestion des réseaux ATM à partir de la transformation des réseaux téléphoniques en réseaux multi-débit et de leur évolution vers des réseaux écoulant du trafic à débit variable et sans mémoires tampon. Nous montrons ensuite comment ce concept s'applique aux réseaux IP avec files d'attente.

Nous utilisons deux modèles de trafic homogène permettant un calcul exact de la performance, à partir desquels nous pouvons évaluer la précision de deux définitions de la bande passante effective: la première basée sur un modèle de file d'attente mono-traffic et la seconde, sur la borne de Chernoff. Nous utilisons ces résultats pour illustrer le concept de région d'admission et sa relation à la bande passante effective. Nous évaluons ainsi la précision des deux définitions et leur domaine d'application.

Mots clés : Gestion du trafic, bande passante effective, région d'admission, files d'attente

1 Introduction

The notions of effective bandwidth and admission control have recently been put forward [2, 27] to improve the quality of wireless communications and in other areas. A short survey of simple queuing models can be found in [25]. Practical applications of the concept can be found in the dynamic management of spectrum access [3, 22], voice traffic in satellite networks [13], video-on-demand servers [24], wireless sensor networks [18], reliable communication networks [9], multimedia WLAN networks [16], call acceptance control for CDMA networks in [4], time slot allocation in OFDM/TDMA systems [26], allocating resource blocks in LTE networks [1], short packet transmission in IoT networks [17], virtual network embedding in 5G networks [6], slicing in wireless networks [28], cross-layer resource allocation [19], joint optimization of small base station power control and service placement [7].

The concept of effective bandwidth was developed in the context of ATM, a technology very different from IP switching. To understand how different they are, consider first the classical telephone network.

1. The network is designed to carry 64 kb/s voice calls exclusively
2. A call is allocated this bandwidth on all the transmission channels required to connect the call origin to its destination
3. This bandwidth is allocated to a user for the whole duration of the call irrespective of its actual use
4. User requests for a circuit are generated randomly and a connection is allocated to a user as long as there is an available circuit. If there is no circuit available, the call is turned down and said to be lost
5. The quality of service is the call loss probability

It became obvious in the 1980's that new services were appearing which needed a large bandwidth but also had a variable bit rate. The large bandwidth meant that the standard 64 kb/s channel was no longer suitable and the variation in the bit rate also meant that allocating a fixed bandwidth for the whole call duration could be very inefficient.

The proposed solution was ATM which could alleviate some of these problems.

1. The network was connection-oriented
2. Data was transported on small 48-byte cells
3. These cells used statistical multiplexing for the transmission, which provided much better efficiency for variable rate services
4. The output buffers were very small so that there was no queuing delay to speak of
5. The network guaranteed that the cell loss probability would be of the order of 10^{-9} which was the quality of service constraint
6. Connections could be rejected if the quality of service could not be maintained

Because the theory of effective bandwidth was developed as a tool to analyze and plan ATM networks, we present in this tutorial the historical development of the theory starting with the extension of classical telephony to multi-rate circuit switching in Section 2 where we show that our intuition about the behavior of queues may not be valid in these cases. The first application to the management of short-term bursts through fast circuit-switching in discussed Section 3 where we introduce the concept of admission region and a first definition of the effective bandwidth of a source. We also introduce the concept of the Chernoff bound that has been extensively studied for bufferless systems. We then go on in Section 4 to buffered queues where the quality of service is measured by the average waiting time showing how we can define an effective bandwidth in some cases. This is followed by a discussion of rate control in Section 5 and some examples where the concept of effective bandwidth may not work very well in Section 6. The tutorial end with a short summary of some important features of the theory and its limitations.

Finally, note that it is not clear how it can be applied to the current IP-based network:

1. The IP network is a best-effort network with no quality of service guarantee of any kind
2. Flow control is implemented by the congestion detection mechanism of the TCP protocol but has no specific measure of quality
3. Some definitions of effective bandwidth are based on the asymptotic behavior of the distribution of some random variable which does not directly apply to quality measures based on some other distribution parameters such as the mean or variance.

2 Multi-rate circuit switching

Multi-rate circuit switching is an extension of classical telephone networks that would allow the network to carry services very different from standard voice calls. It is at the root of the notion of effective bandwidth and for this reason, we review the basic results for this kind of network.

2.1 Classical telephony

Classical telephone networks were designed to carry voice calls on 64, or in some cases, 32 kb/s channels. A link connecting two central offices was made up of a number of these channels and each call would use one of these for the whole time of the call. If no circuit was available when a new call arrived,it was lost. This system can be modeled as an $M/M/N/N$ queue where

- Calls arrive randomly with an exponential inter-arrival time distribution of parameter λ
- There are N channels available, all identical. These are called servers in queuing terminology.
- A call arriving to the queue will try to use any one of the currently unused servers
- If there is such a server, the call starts service immediately
- All calls have the same exponential service time distribution with parameter μ
- If there is no available server, the call leaves and does not return. There is no call buffering

Under these assumptions, the loss probability P is given by the Erlang B function

$$P = E(A, N) = \frac{A^N/N!}{\sum_{i=0}^{N} A^i/i!} \tag{1}$$

where A is called the *offered* traffic and is given by

$$A = \frac{\lambda}{\mu}.$$

This function has two important properties. It is monotone increasing in A, as seen in Figure 1, and is also a quasi-convex function of A. It is also monotone decreasing convex in N, as seen in Figure 2.

2.2 Multi-rate loss probability

This simple model could not be used to analyze new services and had to be extended to handle them. First, even though the application service times distributions are still exponential, their service time parameters μ can be very different. For a superposition of independent Poisson arrival processed, the service time distribution is a hyperexponential. We can then make use of the insensitivity property [23] which states that the stationary distribution of the number of busy servers in the Erlang blocking system M/G/N/N) depends on the service-time distribution only through its mean so that we can still use (1) albeit with a modified average service time as described in Section B.

A more radical change had to be made to take into account the fact that new applications needed very different bandwidths. For this, one has to assume that a customer can use more than a single

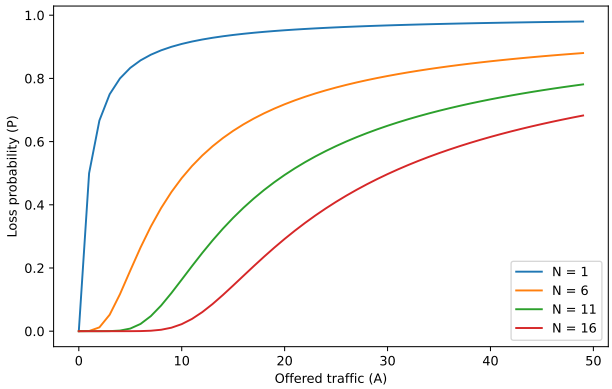

Figure 1: Erlang-B function vs A

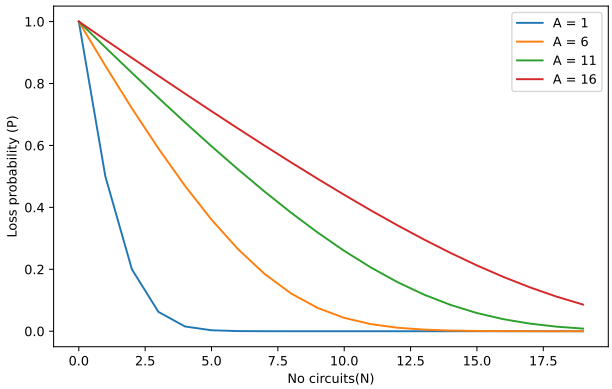

Figure 2: Erlang-B function vs N

server at a time. This has led to the multi-rate circuit-switching model that has been the basis for the development of the concept of effective bandwidth.

In circuit-switched networks, calls that do not find a free server when they arrive are lost. For this reason, the quality of service for multi-rate circuit switching is defined also defined as the probability that an arriving call will be lost. For applications with different bandwidths and service times, the loss probability can be computed under the following assumptions:

- There are N servers, also called *circuits*, available
- There are K call classes
- The calls of class k arrive according to a Poisson process with rate λ_k.
- A call of class k needs b_k circuits. This is an a discrete random variable, often a constant.
- The service time τ_k of a call of class k has a distribution with rational Laplace transform and mean $1/\mu_k$.
- If there are at least b_k servers available when a call arrives, it will immediately start using b_k servers.
- The servers need not be contiguous in any way.
- If there are less than b_k servers available, the call is rejected and does not return.
- For convenience, we also define the traffic rate $A_k = \lambda_k/\mu_k$.

2.2.1 Blocking calculation

In the general case of arbitrary b_k and μ_k, one can compute the distribution of busy servers by a simple relation if one assumes a complete sharing policy, i.e., all servers are available to all calls. Note that this may not be possible in some systems, e.g., in wireless networks, where a number of sub-channels can be allocated to a call only if their bandwidths are contiguous, or in systems where servers are aggregated into blocks.

Let $q(j)$ be the probability that there are j servers busy at some time. We then have [10, 21]

$$\sum_{k=1}^{K} A_k b_k q(j-b_k) = jq(j) \qquad j = 0 \dots N \tag{2}$$

$$q(j) = 0 \qquad \text{if } j < 0$$

$$\sum_{j=0}^{N} q(j) = 1. \tag{3}$$

The system (2–3) can easily be solved first by setting $q(0)$ to some convenient value, e.g., $q(0) = 1$, solving the recurrence relation and then using (3) to normalize the probabilities. From this, we can get the class loss probabilities

$$P_k = \sum_{k=0}^{b_k-1} q(N-k) \qquad k = 1 \dots K \tag{4}$$

and the total loss probability P given by

$$P = \frac{\sum_{k=1}^{K} P_k A_k}{\sum_{k=1}^{K} A_k}. \tag{5}$$

We now examine the behavior of $P(\mathbf{A}, N)$ and $P_k(\mathbf{A}, N)$ as functions of the system parameters $\mathbf{A}$ and N. We show that the multi-rate system is very different from the classical Erlang-B system and that the loss function is *not* monotone in either one of the two parameters. In certain cases, increasing the number of servers can actually *increase* the blocking and increasing some traffic can *decrease* the blocking. This behavior is contrary to our intuition of queuing systems and can lead to unexpected results.

2.2.2 Equivalent Erlang-B model

A simplified solution to the multi-rate blocking problem is to define an equivalent Erlang-B model that is often used as an approximation to the multi-rate system. In this model, a call of class k that needs b_k circuits is replaced by b_k independent calls that need only one circuit each. In that case, the equivalent traffic is given by $A_e = \sum_k b_k A_k$ so that the loss probability can be computed (1) by the Erlang-B function $E(A_e, N)$.

2.2.3 Loss vs N

The behavior of the loss function as a function of N is much more complex than for the simple Erlang-B function as can be seen from Figures 3 and 4. On these figures, we show P, P_1 and P_2 as a function of N. We also show the loss probability found by the equivalent Erlang-B model where the input traffic is given by $A = \sum_k b_k A_k$ so that the loss probability can be computed by the Erlang-B function.

The loss functions shown on Figure 3 are monotone decreasing with N, similar to the Erlang-B function. Note that in this case, the product $A_1 b_1 = A_2 b_2$ so that the two traffic types put a similar load on the system. Also, the equivalent model gets more accurate as N gets significantly larger than the largest b.

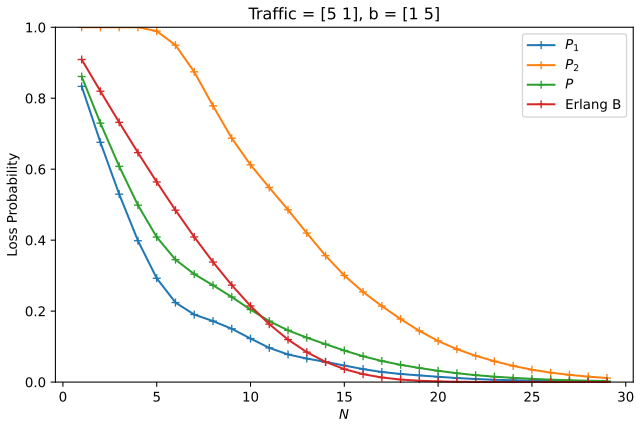

Figure 3: Loss probability, equal loads

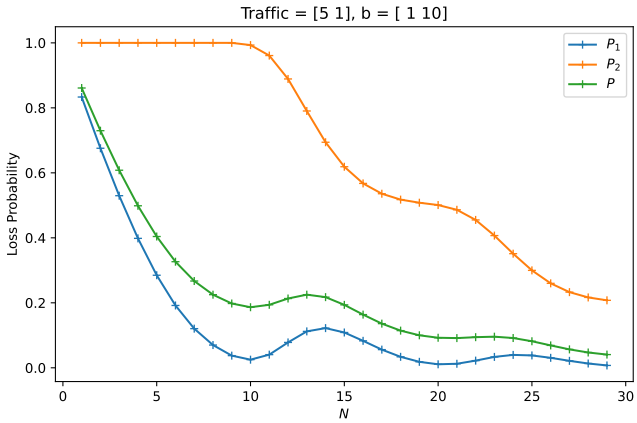

Figure 4: Loss probability, different loads

This is not the case for Figure 4 where the two types have a different load. The most striking difference is that P is *not* monotone decreasing with N, as is the case for the Erlang-B model. In fact, increasing N can actually *increase* both P and P_1. This can be explained as follows. First, note that this happens first when $N = 10$ which is the value of b_2. For lower values of N, only type 1 traffic, with a bandwidth of 1, can be present in the queue which behaves like an Erlang-B system. As soon as $N \geq 10$, some type 2 calls may be present when a type 1 call arrives which can cause it to be lost, hence the increasing probability.

2.2.4 Loss vs A

The behavior of the loss as a function of traffic is shown on Figure 5 which is also not monotone. Here we see that near $A_1 = 0$, *increasing* A_1while keeping A_2 constant *decreases* P_1 while P increases but as A_1 gets larger, it will decrease *both* P_1 and P. The explanation can be found from the behavior of P_2 which is seen to increase as A_1 increases. At $A_1 = 0$, the only traffic comes from type 2 calls which require 10 transmission units each. In that case, $P_2 = E(A_2, 1) = 0.5$. If we increase A_1 slightly, some type 1 calls may be present when a type 2 call arrives and this call will be lost. Because each type 2 call uses 10 transmission units, preventing one of these calls from entering will leave more room for other type 1 calls to enter the system, thus reducing P_1. As A_1 gets larger, this effect becomes more important and P gets very small around $A_1 = 5$.

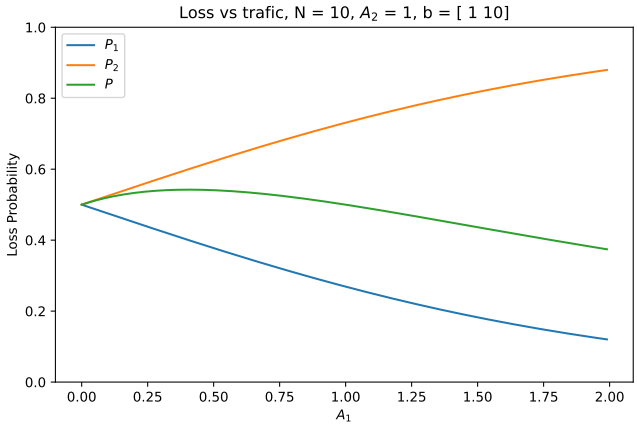

Figure 5: Loss vs traffic, low values

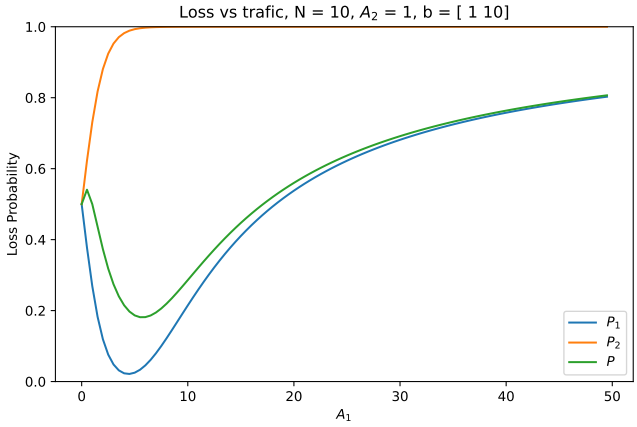

Figure 6: Loss vs traffic, large values

This is explained by noting that on Figure 6, $P_2 \approx 1$ when A_1 gets larger than about 5. At this point, virtually *all* type 2 calls are lost and the system behaves like a single-class $M/M/N/N$ queue. In that case, the loss probability $P = E(5, 10) = 0.018$ which is much smaller than the probability for $A_1 = 0$. As can be seen from Figure 6, from this point on, the system behaves as a single-class system where $P \approx P_1$ and $P_2 \approx 1$.

The range of loss probabilities shown in Figures 5 and 6 is fairly high and would not occur in a real network operating around a few percent loss. We can see from Figures 7 and 8 that the loss function is still non-convex when the loss probability is in a more realistic range of a few percent.

The multi-rate model does not allow much control over the queue. The only option is whether to accept a user or not and this is determined by the condition

$$\sum_k b_k n_k \leq N \tag{6}$$

which must be maintained at all times, where n_k is the number customers of class k present in the queue.

The development of the effective bandwidth technique was basically an extension of the multi-rate circuit-switching model to take into account the variable bit rates of applications and thus improve

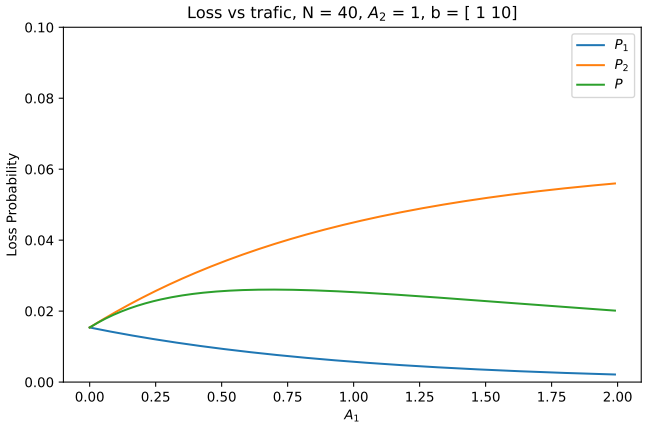

Figure 7: Loss vs traffic, low range, small traffic

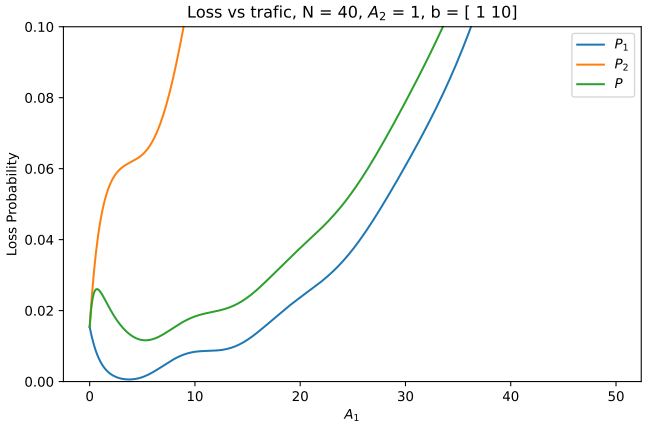

Figure 8: Loss vs traffic, low range, large traffic

the efficiency of the network. This was done for bufferless queues and the QoS was defined as the loss probability. The idea was to find an acceptance rule similar to (6) and then use (5) to compute the blocking probability.

3 Effective bandwidth for bursty traffic

The work of [8] seems to be the first use of the concept of effective bandwidth to manage session admission for systems with stochastic buffer requirements. It has also been the model of much of the theory that was developed in the 1990's.

3.1 System description

The model assumes that a session of type k produces bursts of traffic during which the source transmits data at some given fixed rate and is then idle for some random time. There is a number of identical servers which can be used to transmit the bursts and there is no buffering.

The bursts of a session of type $k, k = 1 \dots K$ are defined by the following parameters as shown on Figure 9.

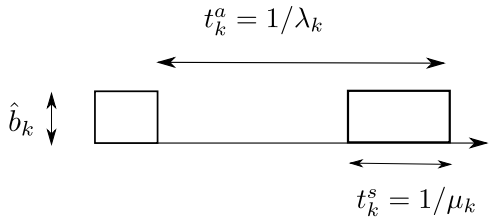

Figure 9: Bursty source

λ_k The burst arrival process has exponential inter-arrival times with parameter λ_k.

μ_k The burst service rate. The time during which the source is transmitting during the burst is $1/\mu_k$. This is an exponential process.

$\hat{b}_k$ The rate of the source when it is transmitting a burst.

ρ_k We also define the traffic of type k as $\rho_k = \lambda_k/\mu_k$.

In a multi-rate circuit-switched system, a session is allocated a number of servers, corresponding to a fixed bandwidth b_k, for the whole duration of the session. This of course could be used to carry the bursty traffic using $b_k = \hat{b}_k$ but would be wasteful since the session is active only a fraction of the time.

We can get a more efficient system if we allocate the circuits only during a burst transmission instead of allocating it for the whole call duration. Because bursts are much shorter than sessions, we will be able to re-allocate bandwidth more efficiently among sessions.

3.2 Admission region

Allocating bandwidth to the bursts has a number of consequences. First, there is no longer the notion of a session bandwidth which is replaced by that of a burst bandwidth. If a session has no bandwidth, we no longer have a capacity bound for the number of admitted sessions. In principle, we could accept any number of sessions so that there could be a large number of bursts offered to the system at some time. Because the bandwidth is allocated to the bursts, it could very well happen that there is not enough bandwidth available. In that case, some bursts will be dropped which will degrade the quality of the communication.

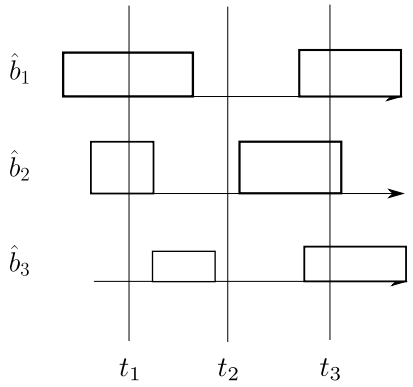

Figure 10: Superposition of bursts

Suppose now that at a given time we have a state vector $\mathbf{n} = [n_1, n_2, \ldots n_K]$ where n_k is the number of sessions of type k currently present in the system. At that time, the total transmission rate is the sum of the rates of the active bursts as can be seen from Figure 10. In that figure, the total rate is $\hat{b}_1 + \hat{b}_2$ at $t = t_1$, zero at $t = t_2$ and $\hat{b}_1 + \hat{b}_2 + \hat{b}_3$ at $t = t_3$. Because the bursts will be lost whenever the total rate at some instant is larger than the transmission rate of the server, the QoS measure is the

probability that the total burst rate is larger than the system capacity. The admission rule is then to admit sessions as long as condition (7) is met:

$$P\{\text{A burst is dropped}\} \leq P_0$$
$$P\left\{\sum_k n_k \hat{b}_k > C\right\} \leq P_0. \tag{7}$$

We can see that this system is really nothing more than multi-rate circuit-switching described in Section 2 with bursts replacing circuits. We can then compute the burst loss probability with (2) where Λ_k, the total burst arrival rate for this type, is given by $\Lambda_k = n_k \lambda_k$ since the arrival process is Poisson.

From this, we define the *admission region* as the set of vectors $\mathbf{n}$ such that the admission rule (7) is met. We can then plot the boundary of the admission region as shown on Figure 11 for $K = 2$, where the admission region is the area below the curve.

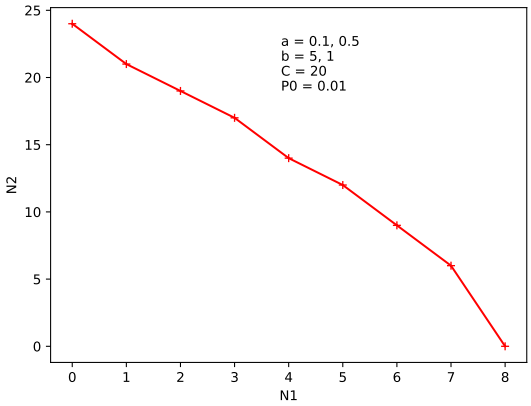

Figure 11: Admission region for burst loss constraint

In the present case, the K types are similar: all arrivals are Poisson and all service times are exponential. The admission rule (7) can thenbe evaluated quickly for any $\mathbf{n}$. This is not always the case when the sessions types have very different characteristics, e.g., when one source is bursty and the other has a more regular traffic. In these cases, computing the admission region boundary may be hard, if not impossible.

For this reason, the notion of *effective bandwidth* was introduced to deal with these cases. In general, an effective bandwidth b_k is a parameter assigned to a particular class k of sessions with the following conditions. First, it can be computed on the parameters of type k only and does not depend on the parameters of the other types. This is much simpler that computing the actual admission region where each point depends on *all* the parameters of *all* the other types. The second point is that we want to use it as a *linear* admission region of the form

$$\sum_{k=1}^{K} b_k n_n \leq C'$$

where C' is a quantity that can be computed from the system parameters and the type.

3.3 Effective bandwidth for bursty sessions

The concept of effective bandwidth was introduced as a way of replacing the complex non-linear admission boundary from (7) by a *linear* boundary similar to the one used for multi-rate circuit-

switching (6). A set of *feasible* effective bandwidths is defined as a set of values b_k such that any vector $\mathbf{n}$ inside a *linear* admission region of the form

$$\sum_k n_k b_k \leq C \tag{8}$$

meets the condition on the *burst* level given by (7).

We can define an effective bandwidth in a number of ways. An obvious choice is to use the burst rate so that

$$b_k = \hat{b}_k$$

which is simply the multi-rate model. Another possibility is to use the *average* rate of the session which yields

$$b_k = \hat{b}_k \frac{t_k^s}{t_k^a} = \hat{b}_k \frac{\lambda_k}{\mu_k} \tag{9}$$

so that the ratio $b_l/\hat{b}_k$ will always be between 0 and 1. This means that defining the effective bandwidth as the mean value will allow more sessions in the system than using the peak value since the sessions require fewer ressources in the first case than in the second. The down point is that there is no guarantee that the QoS constraint (7) will be met since the definition (9) does not involve P_0.

We can try to take the QoS into account using the following argument. Assume that a single type k is offered to the system. The maximum number of bursts n_k^* that can be present in the system is given by

$$n_k^* = \left\lfloor \frac{C}{\hat{b}_k} \right\rfloor.$$

The value of n_k^* can then be used as the number of servers in a $M/M/N/N$ queue and we can compute $\overline{n}_k$, the maximum number of customers that can be admitted, as the solution of the equation

$$E\left(\overline{n}_k A_k, n_k^*\right) = P_0. \tag{10}$$

If we can admit $\overline{n}_k$ bursts, we can consider that each burst needs $C/\overline{n}_k$ units of bandwidth, which gives us the effective bandwidth

$$b_k = \frac{C}{\overline{n}_k} \tag{11}$$

which is simply the two intercepts with the axes of the admission curve of Figure 11. In the following, we will call these the *end-point* effective bandwidths.

We can see on Figure 12 the effective bandwidth of a given session type produced by the average and end-point definitions as a function of the load ρ of a session. For the end-point definition, we show the curve for two values of the burst loss probability P_0 while there is a single curve for the average definition since it does not involve P_0. We can see that the end-point values are always larger than the average and increase with a more stringent burst loss constraint.

Clearly, a particular definition of the effective bandwidth will result in a different linear admission region. We can see this on Figure 13 where we show the admission regions for the average, peak and end-point definitions along with the real admission region. It is clear that the average value yields an infeasible effective bandwidth since there is a large number of points above the real value. Similarly, the definition based on the peak value is feasible but very conservative and is leaving out a number of feasible configurations. The reason of course is that neither definition takes into account the burst level QoS constraint. We can see that we can get a much better approximation of the real region when we do take this into account in the definition of the end-point effective bandwidth. For this reason, we will consider in the following only definitions such as (11) that take the burst QoS into account.

We can see on Figure 14 a detailed view of the admission regions produced by the end-point effective bandwidth and the real region. In this case, the end-point effective bandwidth is feasible.

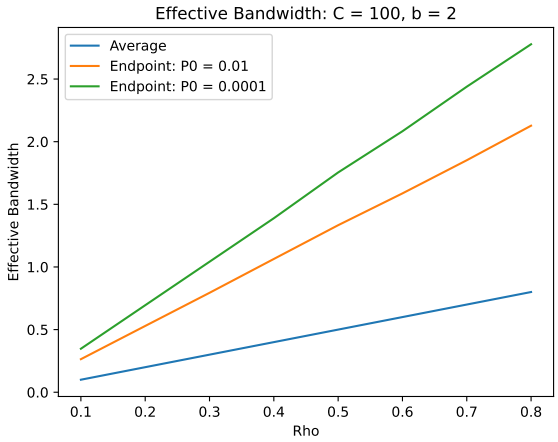

Figure 12: Effective bandwidth: Average and end point

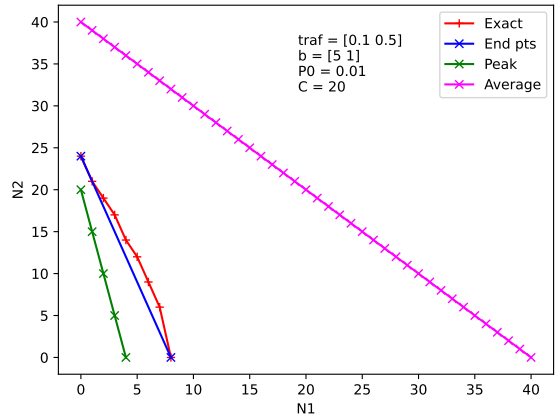

Figure 13: Effective bandwidth: Exact, end-point, peak, average

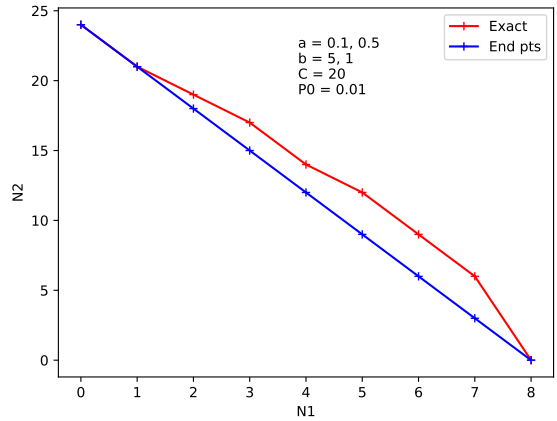

Figure 14: Effective bandwidth: Feasible end-point

Unfortunately, this is not always the case, as can be seen from Figure 15 with a different set of values where the end-point curve can lie above the real one. For larger values of n_1, the end-point effective bandwidth will admit too many connections and the burst QoS constraint will not be met. This is a generic problem with the end-point effective bandwidth which is guaranteed to be feasible only at the end points. There is no guarantee elsewhere in the domain, which can be a problem if this is implemented since it can lead to the admmission of sessions which will produce a burst loss larger than the required QoS.

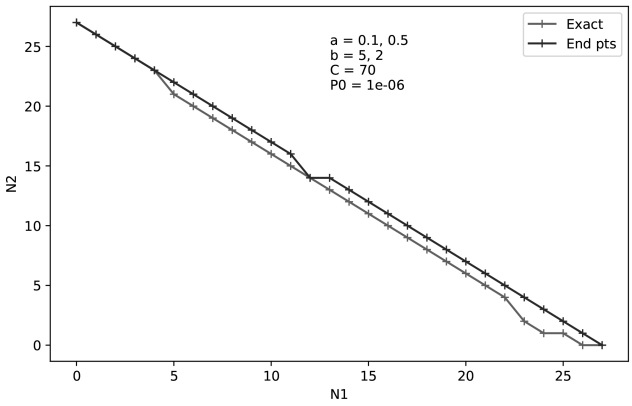

Figure 15: Effective bandwidth: Infeasible end point

3.4 Effective bandwidth based on the Chernoff bound

The problem of dealing with infeasible regions has been the subject of much work right from the work of [8]. The standard approach has been to use the Chernoff bound to produce a feasible region. This has been summarized in [11] from which this discussion is taken.

The Chernoff bound of a distribution is defined as follows. If X is a random variable, its distribution can be bounded by

$$\begin{aligned} P\{X \geq a\} &\leq E\left\{e^{s(X-a)}\right\} \\ &\leq e^{-sa} E\left\{e^{sX}\right\} \end{aligned} \tag{12}$$

for any $s \geq 0$ and where $E\left\{e^{s(X-a)}\right\}$ is the moment-generating function of X. This is often called the *large deviation* bound since it gets exponentially better as a gets larger. The parameter s is sometimes called the *space* parameter in the literature.

Note that we need only the moment-generating function of X in order to compute the bound. In the following, we will denote this function as $G_X(s)$. We can then use the following property. Let $X_1, X_2 \ldots X_K$ be independent variables and define $X = \sum_k X_k$. Using (29), we have

$$G_X(s) = \prod_{k=1}^{K} G_{X_k}(s).$$

We then have the bound

$$\begin{aligned} P\{X \geq a\} &\leq e^{-sa} G_X(s) \\ &\leq e^{-sa} \prod_k E\left\{e^{sX_k}\right\} \\ &\leq e^{-sa} \prod_k G_{X_k}(s) \end{aligned} \tag{13}$$

where we have used the fact that the X_k are independent to get the last inequality. We can get a linear relationship if we define the *logarithmic moment-generating function*, denoted by $M_X(s)$, as

$$M_X(s) = \log G_x(s) = \log E\left\{e^{sX}\right\}. \tag{14}$$

From (13), we get

$$\begin{aligned}\log P\{X \geq a\} &\leq \log e^{-sa} \prod_{k=1}^{K} E\left\{e^{sX_k}\right\} \\ &\leq \sum_k M_{X_k}(s) - as. \end{aligned} \tag{15}$$

We can now use these results to define an effective bandwidth for the system described in Sec. 3 where we have K different types of sessions, each with parameters a_k, b_k, and n_k sessions of type k active at some time t. Define the random variable $X_{k,i}$ as the rate of session i of type k. The total rate *offered* to the server is then

$$X = \sum_{k=1}^{K} \sum_{i=1}^{n_k} X_{k,i}.$$

The system is in a blocking state if $X \geq C$ at that time so that the loss probability is the probability that $X \geq C$. Using (15), we have

$$\begin{aligned}\log P\{X \geq C\} &\leq \sum_k \sum_i M_{X_{k,i}} - sC \\ &\leq \sum_k n_k M_{X_k}(s) - sC \end{aligned} \tag{16}$$

where we have used the fact that all the sessions of type k are iid and thus have the same $M(s)$. Recall also that the bound is valid for any $s \geq 0$ so that we may want to use the best possible bound given by

$$\log P\{X \geq C\} \leq \inf_{s \geq 0} \sum_k n_k M_{X_k}(s) - sC. \tag{17}$$

Denote the best value of s by s^*. Note that this value depends on the vector $\mathbf{n}$ from (16).

Suppose now that the QoS requirement is the total loss probability P_0. We want

$$\log P\{X \geq C\} \leq \log P_0.$$

Using (17), this condition can be met if

$$\log P\{X \geq C\} \leq \sum_k n_k M_{X_k}(s^*) - s^* C \leq \log P_0.$$

The right-hand side inequality holds if

$$\begin{aligned}\sum_k n_k M_{X_k}(s^*) &\leq s^* C + \log P_0 \\ \sum_k n_k \frac{M_{X_k}(s^*)}{s^*} &\leq C + \frac{\log P_0}{s^*}. \end{aligned} \tag{18}$$

This defines a linear admission region with an effective bandwidth b_k defined as

$$b_k = \frac{M_k(s^*)}{s^*} \tag{19}$$

and a reduced capacity

$$C' = C + \frac{\log P_0}{s^*}.$$

The linear admission region is the set of vectors $\mathbf{n}$ such that

$$\sum_{k=1}^{K} \alpha_k^* n_k \leq C'.$$

This definition of the best effective bandwidth requires a one-dimensional minimization due to (17) which may require a numerical solution unless the form of M is simple enough to solve for the minimum analytically. Also, this depends on the vector $\mathbf{n}$ so that it must be re-computed at each point which is not practical.

One may use a simpler approach using the fact that the admission region defined by (16) is feasible for any $s \geq 0$. One can then pick an arbitrary value for s and use this to compute the region. The question is then how sensitive is the effective bandwidth defined in this way to different values of s.

Note that the right-hand side of (18) can be negative for sufficiently small s since $\log P_0 \leq 0$. There is a lower value for s such that there is a solution given by

$$C + \frac{\log P_0}{s} > 0$$
$$s \geq -\frac{\log P_0}{C}.$$

One advantage of the Chernoff bound is that is is always feasible. Another strong point is that the effective bandwidth computed from (19) can be done independently for each type k, which is much simpler than computig the actual boundary.

3.5 Chernoff bound for bursty sources

We now give an example of the Chernoff bound for the case of bursty sources described in Section 3. For this, we consider the *arrival* process of a given stream. We assume that the customers are offered to an infinite number of servers and we are interested in the occupation of this server. First, define $Y(t)$ the number of *customers* present in the server at time t. This is a random variable with a Poisson distribution where the stationary distribution of customers present in the infinite system is given by

$$p_n = \frac{A^n}{n!} e^{-A}$$
$$A = \frac{\lambda}{\mu}$$

where λ is the arrival rate and μ the service rate. From this, we get the moment generating function and its log

$$G_Y(s) = e^{A(e^s - 1)}$$
$$M_Y(s) = A(e^s - 1).$$

Because each customer uses b units, we are really interested in the number of *units* in use as opposed to simply the number of customers. For this, we define the random variable

$$X = bY$$

and we get

$$M_X(s) = A(e^{bs} - 1).$$

If we have n_k sessions of type k, we define X_k as the total number of units used by these sessions and get

$$M_{X_k}(s) = n_k A_k(e^{b_k s} - 1).$$

Next ,define $X_{k,i}$ as the rate of session i of type k. The total number of units used by all types is then

$$M_X(s) = \sum_{k=1}^{K} n_k A_k(e^{b_k s} - 1) \tag{20}$$

where X is the total number of units in use in the infinite group. This is also the total amount of traffic *offered* to the finite group at some instant and if this exceeds the number of servers, there will be some losses.

We can see on Figure 16 the effective bandwidth for each of the two classes computed from the Chernoff bound as a function of s. This is also compared with the corresponding end-point values. We can see that the value of the Chernoff effective bandwidth can depend quite strongly on the value of s.

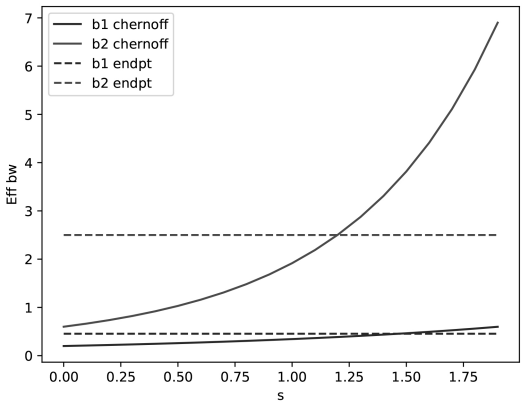

Figure 16: Effective bandwidth from Chernoff bound

If we don't want to compute the optimal bound at each point, as in (17), we need to pick some value of s, which raises the question of how tight the bound will be. Because the strong dependence on s, we examine the sensitivity of admission region as a function of s in Figure 17. One conclusion is that there is no uniformly best value for s. For large n_1, $s = 0.4$ is best while $s = 0.7$ is best at low n_1 and $s = 0.5$ is the best choice for intermediate values. The other conclusion is that in this case, the bound is very conservative, much more than the end-point effective bandwidth. This is due to the fact that the loss probability constraint $\overline{P}_0$ is fairly large while the Chernoff bound is more accurate for the tail distribution of loss probability.

3.6 Other measures of QoS

The shape of the admission region depends in general on all the parameters of the system. As an example, consider the case where we now have separate loss constraints P_k for each flow k. This could be used for instance to provide better service to one of the two flows. We have

$$\begin{aligned} A_k &= [5, 1] \\ b &= [1, 5] \\ C &= 100 \\ P_k &= [0.01, 0.1] \\ P_0 &= 0.01 \end{aligned}$$

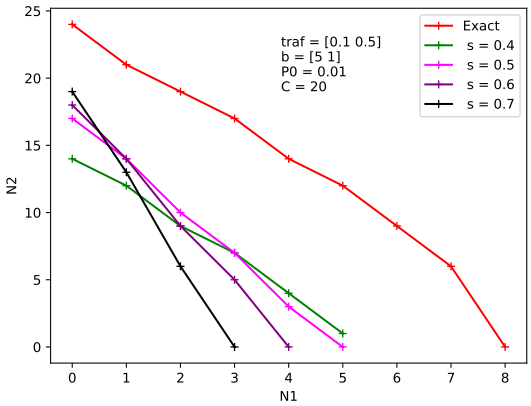

Figure 17: Admission region vs s for the Chernoff bound

We can see from Figure 18 that the admission region with separate loss constraints is somewhat larger than when we use the same constraint for all flows.

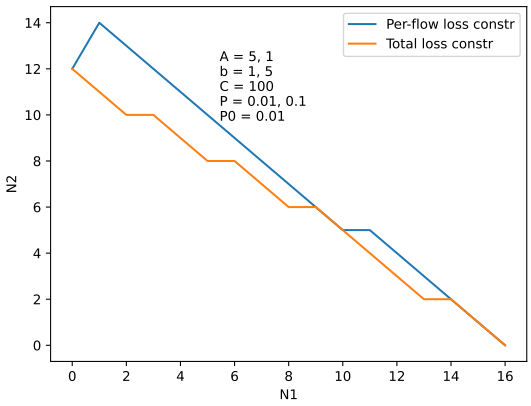

Figure 18: Admission region, per-flow vs total loss constraint

4 Effective bandwidth for buffered systems

The theory of effective bandwidth was developed at a time where ATM was the switching technology put forward by the telephone operators. These networks were supposed to operate with a very low cell loss probability which was well approximated by bufferless systems.

This is of course no longer the case with IP networks where a more realistic model for a network link is a single-server queue with a buffer, generally assumed to be infinite. We now give an example to show how the effective bandwith concept can be used in that context.

4.1 Superposition of Poisson processes

Consider now the case of many different Poisson traffic sources. The arrival rate of source i is λ_i and the service rate is μ_i with an arbitrary distribution. The sources are multiplexed on a single server with an infinite buffer. Given that the individual arrival processes are Poisson, their superposition is

also a Poisson process with a total arrival rate $\lambda = \sum_i \lambda_i$ [29]. This can be modelled as an M/G/1/∞ queue and we can then use the Pollaczek-Khintchine theorem to compute the average waiting W and sojourn T times [15] as shown in Section C.

4.2 Heterogeneous M/H/1 sources

We now consider the case where the service times are exponentially distributed. When a packet arrives to the server, the probability p_i that the packet is of type i is given by

$$p_i = \frac{\lambda_i}{\lambda}.$$

The distribution of service times is thus an hyper-exponential with parameters μ_i, p_i and the server is an $M/H/1$ queue. We can then compute the values of the average service time $\overline{S}$ and the coefficient of variation C_b using the results of Section B and use the values of Section C to get the average time $\overline{T}$.

4.3 Admission control

We can plot the admission region for two connection types where the QoS constraint at the packet level is the total average delay $T \leq \overline{T}$. A typical result is shown by the red curve of Figure 19. The curve labeled "Exact" is that of the real admission region based on the $M/H/1$ queue. The curve labeled "Eff bw" is for the admission region based on the end-point effective bandwidth computed for each one of the two flows separately. In these cases, we have a simple $M/M/1$ queue and n_i^*, the maximum number of connections of type i that can be admitted is given by

$$n_i^* = \left\lfloor \frac{1}{\rho_i} \frac{\overline{s}_i}{\overline{T}} \right\rfloor.$$

where $\overline{s}_i$ is the average service time for type i and $\rho_i = \lambda_i \overline{s}_i$.

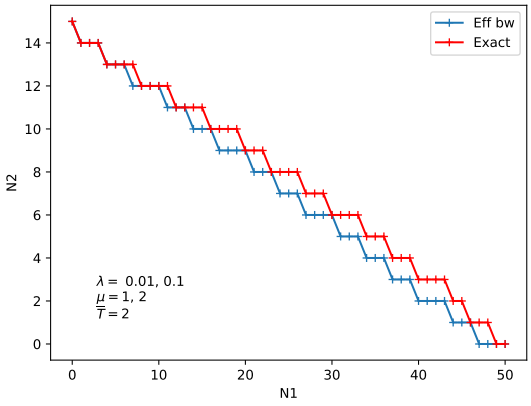

Figure 19: Admission region for two Poisson sources

4.4 Chernoff bound

For bufferless queues, we have seen that then end-point effective bandwidth may not yield a feasible region and this may be the case here as well. We also know that an effective bandwidth based on the Chernoff bound guarantees a feasible region for bufferless queues and this might be the solution here as well.

This is not quite as simple since the bound applies to the tail of the *distribution* of a random variable, as in (12) so that we cannot use the bound directly if the QoS measure is the *average* of some random variable, as in (37).

We can still use the Chernoff bound if we choose some maximum value T^* for the packet delay and a probability P_0 and define the admission region through the constraint

$$P\{T \geq T^*\} \leq P_0. \tag{21}$$

For a suitable choice of the parameters T^* and P_0, we can hope that the average value of T will be sufficiently close to some required value $\overline{T}$.

Using the Chernoff bound on (21), we get

$$\begin{aligned} P\{T \geq T^*\} &\leq E\left\{e^{t(T-T^*)}\right\} \\ &\leq e^{-tT^*} E\left\{e^{tT}\right\} \\ &\leq e^{-tT^*} G_T(t) \end{aligned} \tag{22}$$

where $G_T(t)$ is the moment-generating function of T. We can then guarantee that the constraint (21) is met if we impose the condition

$$\begin{aligned} e^{-tT^*} G_T(t) &\leq P_0 \\ G_T(t) &\leq e^{tT^*} P_0. \end{aligned} \tag{23}$$

From the results of Section C, we know that

$$G_T(t) = \frac{t(1-\rho)}{t+\lambda-\lambda G_S(t)} G_S(t) \tag{24}$$

where $G_S(t)$ is the moment-generating function of the service time S. Replacing (24) in (23)

$$\begin{aligned} \frac{t(1-\rho)}{t+\lambda-\lambda G_S(t)} G_S(t) &\leq P_0 e^{tT^*} \\ t(1-\rho) G_S(t) &\leq P_0 e^{tT^*} \left(t+\lambda-\lambda G_s(t)\right) \\ t(1-\rho) G_S(t) + \lambda G_S(t) P_0 e^{tT^*} &\leq (t+\lambda) P_0 e^{tT^*} \\ G_S(t) &\leq \frac{(t+\lambda) P_0 e^{tT^*}}{t(1-\rho) + \lambda P_0 e^{tT^*}}. \end{aligned} \tag{25}$$

To simplify notation define

$$H(t) = \frac{(t+\lambda) P_0 e^{tT^*}}{t(1-\rho) + \lambda P_0 e^{tT^*}}$$

which is a function of t with all the other parameters known from the definition of the sources and QoS. We rewrite (25) as a function of H and take the logarithm on both sides to get

$$\log\left(G_S(t)\right) = M_S(t) \leq \log\left(H(t)\right) \tag{26}$$

For bufferless queues, the feasibility condition (18) neatly separated into two terms, one that depends on $\mathbf{n}$ and the other, on the system parameters. In the present case, the right-hand term also depends on $\mathbf{n}$ through λ. We don't have a direct definition of an effective bandwidth but (26) can be seen as the definition of an acceptance region that is guaranteed to be feasible. The advantage of using (26) instead of (21) is that the right-hand term depends only on $\mathbf{n}$ and the system parameters and that the left-hand term requires the calculation of $G_{S_i}(t)$ separately for each class i.

If $G_s(t)$ is a hyperexponential variable, we can use the result of (30) and (36) to get

$$\sum_{i=1}^{N} n_i M_{S_i}(t) \leq \log\left(H(t)\right) \tag{27}$$

from which we get the effective bandwidth

$$b_i(t) = M_{S_i}(t) = \frac{\mu_i}{\mu_i - p_i t}.$$

This expression also imposes the condition that

$$t < \min_i \left\{ \frac{\mu_i}{p_i} \right\}.$$

We can get a more intuitive value for the effective bandwidth if we assume that packets of class i have an exponential lenght with parameter L_i and the server, a capacity C. The average service time $S_i = L_i/C$ is a scaled exponential as well and we can rewrite

$$b_i(t) = C\left[\frac{1}{1 - p_i S_i t}\right]$$

which shows that the effective bandwidth of class i is a fraction of the total capacity that increases as the service time of the class

4.5 Equivalent model

We can also examine the admission region produced by a simplified model. In the case where we have many Poisson sources all with the same service rate, we can replace these sources by a single Poisson one with an arrival rate that is the sum of the rates of the original sources. When the service rates are different, we could make the same approximation and represent the aggregate stream by a Poisson process with parameters

$$\lambda = \lambda_1 + \lambda_2$$
$$\overline{S} = \frac{\lambda_1}{\lambda}\mu_1 + \frac{\lambda_2}{\lambda}\mu_2$$

where the service rate is the average rate defined in (33). This yields a M/M/1 queue and we can see the corresponding admission region on Figure 20. The admission region for the equivalent model is completely different from the actual region and that the model will accept far too many type 2 connections.

5 Rate control

The theory of effective bandwidth was developed as an admission control technique for sessions. It can also be viewed as a technique for controlling the rate of sessions already in place. This is possible for sessions that do not need a fixed bit rate, e.g., video streaming where the only requirement is that the playback buffer should not become empty. In case of congestion, one could reduce the rate of the sender for a short time determined by the capacity of the playback buffer.

In this section, we assume that there is a given number of sessions of a given type present in the queue and that this will not change over the interval of time over which we want to control the rate. A “session” could be a single user or the aggregation of a number of similar users.

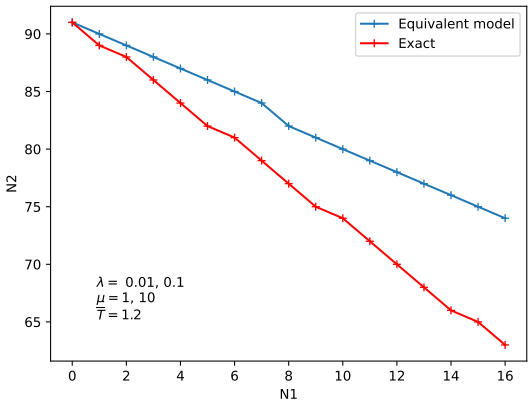

Figure 20: Admission region for the equivalent model

5.1 Bursty sources

The multi-rate loss model of Section 3 can be viewed as a technique for controlling the rates of sessions already in place. Suppose we have a given number of sources K where each source k can adjust its rate A_k. There is a set $\mathcal{S}$ of $\mathbf{A}$ vectors such that $P(\mathbf{A}, \mathbf{b}, N) \leq \overline{P}$ for $\mathbf{A} \in \mathcal{S}$. This is called the *feasible* region for the system. As long as the sources rates $\mathbf{A}$ remain in $\mathcal{S}$, we are guaranteed that the loss probability will be less than $\overline{P}$. The queue management is thus to force the source to remain feasible, which is called *rate* control.

The feasible region is then the region inside the contours of the loss function for some value of $\overline{P}$, shown as the value on the contour curve. This is defined as the set of values (A_1, A_2) such that $P(A_1, A_2, N) \leq \overline{P}$, e.g., the set of traffic values such that the burst loss probability P is no larger than some prescribed value $\overline{P}$. We show on Figure 21 the contour corresponding to Figure 6. We can see that it is basically the same plot with scales axes, as expected.

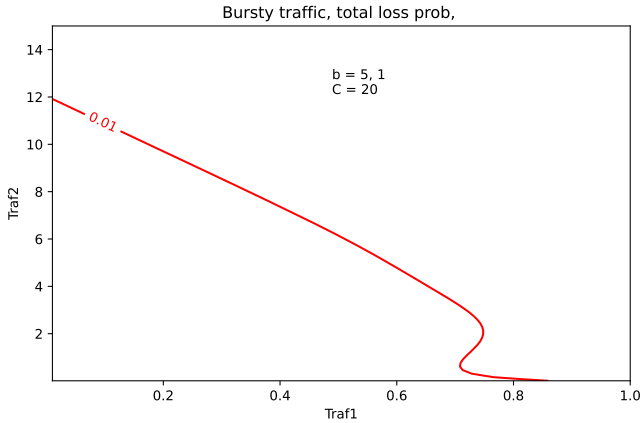

Figure 21: Contours of P, $N = 10$

5.2 Buffered queues

We get similar results for buffered queues. We plot on Figure 22 the admission region as a function of the two traffic rates and we get the same scaled form as Figure 19.

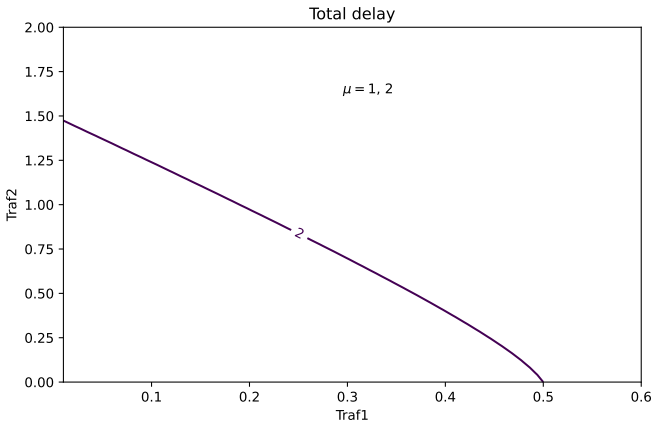

Figure 22: Feasible region, sojourn time

6 Non-convex boundaries

The calculation of an effective bandwidth using the end-point technique seems quite straightforward. Consider again the case shown on Figure 14. The sources produce two types of bursty traffic, denoted as 'large' for type 1 and 'small' for type 2, as described in Section 3. Here, the QoS constraint is the total burst loss probability.

To make the figure, we first set $N_2 = 0$ and compute the end point on the N_1 axis by increasing the value of N_1 from 0 until the first value N_1^* such that the QoS constraint is violated. Similarly, we can compute the end-point N_2^* for the N_2 axis. We then make the assumption that for any value $N_2 > 0$, all the points $N_1 > N_1^*$ will be infeasible. This rests on the "obvious" assumption that if we cannot have more than N_1^* sessions of type 1 when there are no sessions of type 2, adding type 2 sessions will not improve things. The same argument goes for the N_2 axis and we end up with the conclusion that the only points to be examined lie in the rectangle delimited by the points $(0, 0)$, $(N_1^*, 0)$, $(0, N_2^*)$ and (N_1^*, N_2^*). This also assumes that the boundary and the effective bandwidth will be a monotone decreasing function of the horizontal axis.

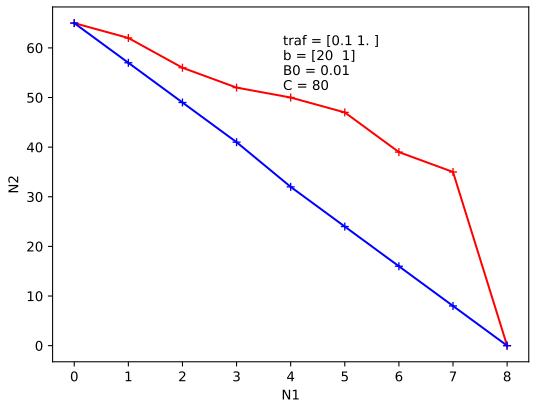

Figure 23: End points admission region, bursty sources

Unfortunately, this reasoning is incorrect in some cases. Consider the plot of Figure 23 which was built using the procedure above. Note that the upper curve corresponding to the real admission

region decreases regularly until $N_1 = 7$. Using the argument above, we never have to compute points $(N_1 = 8, N_2 > 0)$ since we "know" that they are infeasible. This is why the last segment of the upper curve is connected to the point $(8, 0)$ which is not realistic.

The answer to this can be seen from Figure 24 which shows the actual admission region. This plot is produced by actually computing the QoS for all points in some given large rectangle in the (N_1, N_2) plane and showing only the points where the QoS is met. The contour plot as a function of continuous traffic variables is shown on Figure 25. In other words, the assumption that we need not look beyond N_1^* is wrong.

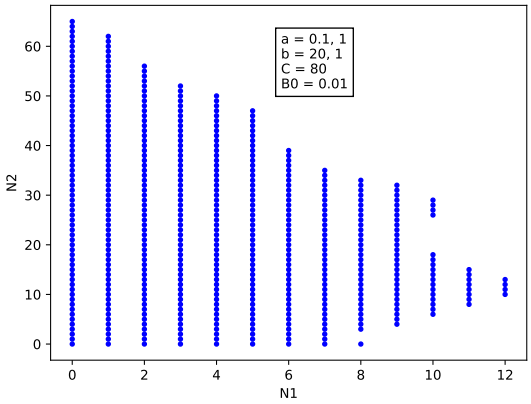

Figure 24: Actual admission region, bursty sources

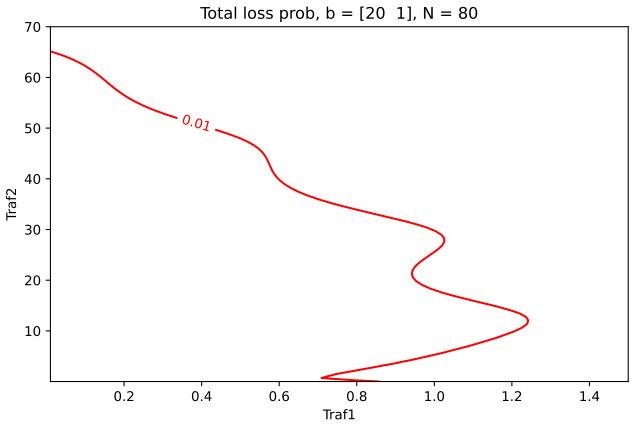

Figure 25: Contour of P for $\overline{P} = 0.01$

To explain this, consider again the N_1 axis. When $N_2 = 0$, the system can be replaced with an equivalent queue with 4 servers each representing a group of 20 actual servers. Each customer takes up one of these blocks so that the equivalent queue behaves like a pure Erlang B system. We can see from the figure that the QoS is not met whenever $N_1 > 8$ which is the maximum number of type 1 connections that can be admitted. Starting from the point $(8, 0)$, we now move in the N_2 direction. For small values of N_2, the queue still does not meet the QoS but if we keep adding small connections, the system becomes feasible again.

This may seem counter-intuitive but consider the same system when there are currently 60 busy servers. If $N_2 = 0$, a customer can only take a block of 20 servers, no less. Suppose now that small type 1 customers are offered to the queue and that one of these customers takes a server. This means that during a certain time, no large customer can be admitted but up to 9 small ones can be served. The trade-off is then between losing some large customers and accepting more small ones. In the present case, the arrival rate of large customers is relatively small so that we will be missing a small number of those while the arrival rate of the small customers is much larger, which means that we will be accepting more of these. Since the QoS is the aggregated call loss probability for both classes, losing one small call is the same as losing one large one. Given that a large call uses up many more servers than a small one, it is clear that for some values of the traffic, having small calls block larger ones will be beneficial, hence the improved QoS. We can see that this is indeed the case from Figure 26

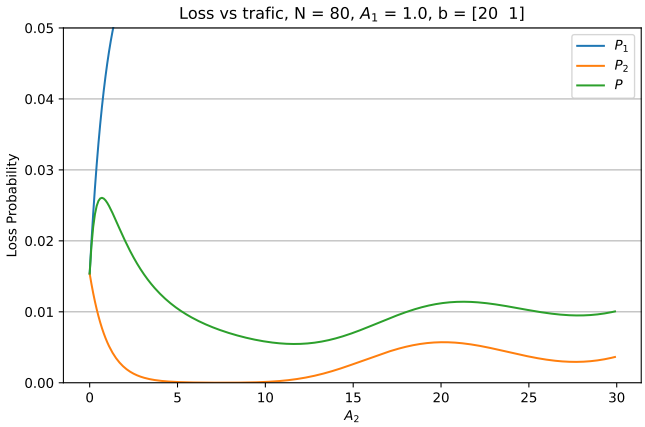

Figure 26: Loss probability for $N_1 = 10$

where we have plotted P, P_1 and P_2 as a function of the traffic generated by an increasing number of type 2 sources for the value $N_1 = 10$. We can readily see that the total loss probability P will fall below 0.01 over two disjoint regions of A_2.

We can see the same kind of region for delay systems. The admission region computed from the N_1 axis is shown on Figure 27. We see that the end-point effective bandwidth is at $N_1 = 16$ which leaves out a significant number of feasible states. The explanation is similar to the case of bursty traffic. We

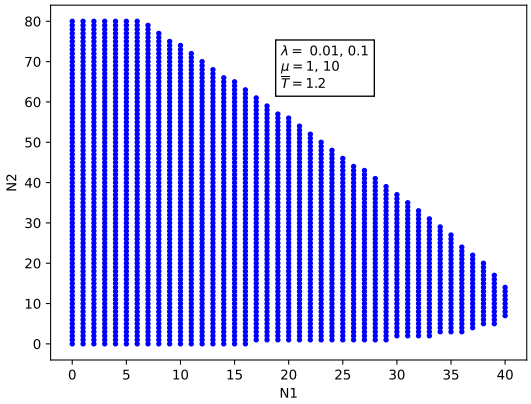

Figure 27: Actual admission region, M/H2/1 queue

have two different streams. Type 2 has a large μ which corresponds to a small packet size, but a large arrival rate. Type 1 packets are larger but with a smaller arrival rate. If we compute the average delay for large packets with no small packets present, all measured packets will have a relatively large holding time so that we cannot accept more than 16 sessions. If small packets are present, they will contribute many small values in the calculation of the average waiting time thus increasing the number of large packets that can come in and improve the overall QoS.

7 Further reading

There has been a large amount of work done on effective bandwidth in the context of ATM networks. A generalization of the bufferless system discussed in Sections 3 and 4 can be found in [12] where approach is to propose a definition of effective bandwidth that is applicable to systems with and without buffers. It is based on the amount of work produced by a source in the time interval $[0, t]$ and is denoted by the random variable $X(0, t)$. If the source has independent increments, the effective bandwidth is defined as

$$b(s, t) = \frac{1}{st} \log E\left\{\exp sX(0, t)\right\}.$$

The work contains many examples on the application of this definition to different sources. The references also give pointers to the computation of appropriate values of s and t for different situations. Note however that this is still related to the Chernoff bound and thus would be applicable to cases where the QoS is defined in terms of the tail of some distribution.

The concept of effective bandwidth has been extended to more complex cases in numerous work. An early example would be [5] which examines a number of possible definitions for the effective bandwidth for a number bursty sources and a queue with finite buffer.

8 Conclusion

The decision whether to accept a session or not in a multi-service network is based on the concept of the admission region. We have shown a few examples that can easily be computed for two mixes of services. In the first case, all streams are modeled by a multi-rate burst traffic with Poisson arrivals and exponential service times offered to a bufferless set of servers. The streams differ only in the values of the arrival and service rates and the burst bandwidth. The second is a mixture of flows where each one is modeled by a Poisson arrival process and exponential service time, all of which are offered to a single server with an infinite buffer. Here too the only differences between the streams are the actual values of the arrival and service parameter.

The reason why we can compute the admission exactly is because in both cases, the individual streams are homogeneous, i.e., they have the same distributions. In these cases, the aggregated stream is relatively simple and the performance function can be computed easily, either by the Kaufman-Roberts multi-rate model for the burst traffic or by the M/H/1 queue for the Poisson streams.

Things get much more complicated when the streams are not homogeneous. Mixtures of voice and data streams cannot be analyzed exactly and even approximations are complicated to evaluate. Other combinations are often intractable and the only way to check feasibility is by simulation.

For this reason, one needs some simple measures of the impact of a new session on the system performance. For bufferless systems, for instance, the effective bandwidth in effect transforms the lower-level QoS requirements into a bandwidth requirement which can be evaluated using the Kaufman-Roberts algorithm. Equivalently, this amounts to a linearization of the boundary of the admission region from which we can use a simple admission rule. As we have seen, this linearization may not be feasible everywhere if we use the end-point technique. Using the Chernoff bound guarantees feasibility but in some cases, the shape of the admission region is nowhere near a linear region so that an effective

bandwidth technique may leave out significant parts of the region. This would yield a conservative admission policy which may not be suitable for wireless systems where bandwidth must be used as efficiently as possible.

To summarize,

1. Effective bandwidth is used to replace by a linear approximation an admission region that is difficult to evaluate. There are two standard definitions that can always be used
2. End-point effective bandwidth
   (a) This may produce infeasible regions
   (b) It requires inverting the QoS function to get the maximum load as a function of the system parameters
   (c) The admission region may be very conservative
3. Chernoff-bound effective bandwidth
   (a) It is based on a quality of service defined in terms of the tail of the distribution of some random variable and is not directly usable for other measures such as an average
   (b) It always produces a feasible solution
   (c) It requires the the computation of the moment-generating function
   (d) The effective bandwidth depends on the value of one or more arbitrary parameters of the moment-generating function that have to be estimated depending on the particular traffic mix
   (e) This admission region may be very conservative
4. The boundary is not generally monotone decreasing and may not even be convex so that a linear approximation may very well be inaccurate.

In practice, a resonable procedure is

1. Define an effective bandwidth. This may be the end-point, Chernoff or some other definition more suitable for the problem at hand
2. Choose a suitable set of cases and check by simulation that the values produced by the effective bandwidth are in fact feasible

There does not seem to be any simple way to check how conservative the effective bandwidth might be except by doing a large number of simulations in the region outside the effective bandwidth boundary.

A Moment-generating functions

If X is a random variable with probability density function f_x, its moment-generating function G_X is defined as

$$G_X(s) = E\left\{e^{sX}\right\}$$

for all $s \geq 0$. If we define the linear transformation

$$Y = a + bX$$

then

$$\begin{aligned} G_Y(s) &= E\left\{e^{s(a+bX)}\right\} \\ &= E\left\{e^{sa}e^{sbX}\right\} \\ &= e^{sa}E\left\{e^{sbX}\right\} \\ &= e^{sa}G_X(bs). \end{aligned} \tag{28}$$

Consider now the set of *independent* random variables $X_i, i = 1 \dots N$. If we define

$$Y = \sum_{i=1}^{N} X_i$$

then

$$\begin{aligned} G_Y(s) &= E\left\{\exp\left(s\sum_{i=1}^{N} X_i\right)\right\} \\ &= E\left\{\exp\left(\sum_{i=1}^{N} sX_i\right)\right\} \\ &= E\left\{\prod_{i=1}^{N} \exp\left(sX_i\right)\right\} \\ &= \prod_{i=1}^{N} E\left\{\exp\left(sX_i\right)\right\} \\ &= \prod_{i=1}^{N} G_{X_i}(s) \end{aligned} \tag{29}$$

where we have used the independence of the X_i in (29). Similarly, if

$$Y = \sum_{i=1}^{N}\sum_{j=1}^{N_i} X_{i,j}$$

where the $X_{i,j}$ are iid for any given i, we have

$$G_Y(s) = \prod_{i=1}^{N} G_{X_i}(s)^{N_i}. \tag{30}$$

B The hyperexponential distribution

The hyperexponential distribution belongs to the class of phase type distributions. In the present case, at a given instant, the source can be in one of N phases during which the service rate of phase i is μ_i with fixed probabiliy p_i of being selected at that instant. The probability density function of X is given by

$$f_X(x) = \sum_{i=1}^{N} p_i f_{X_i}(x) \tag{31}$$

$$\sum_{i=1}^{N} p_i = 1$$

$$p_i \geq 0. \tag{32}$$

The average $\overline{X}$, the variance $\mathrm{var}\,(X)$ and the coefficient of variation C_b are given by [20]

$$\overline{X} = \sum_i \frac{p_i}{\mu_i} \tag{33}$$

$$\mathrm{var}\,(X) = \overline{X}^2 + \sum_{i,j} p_i p_j \left(\frac{1}{\mu_i} - \frac{1}{\mu_j}\right)^2 \tag{34}$$

$$C_b = \frac{\sqrt{\mathrm{var}\,(X)}}{\overline{X}}. \tag{35}$$

Using (28) and (29), the moment-generating function is given by

$$G_X(s) = \prod_{i=1}^{N} G_{X_i}(p_i s) = \prod_{i=1}^{N} \frac{\mu_i}{(\mu_i - p_i s)}. \tag{36}$$

C The M/G/1 queue

We summarize here some useful results for the $M/G/1/\infty$ queue. The Pollaczek-Khintchine formula [15] yields the average sojourn time $\overline{T}$

$$\overline{T} = \overline{S} + \frac{\rho \overline{S}(1 + C_b^2)}{2(1 - \rho)} \tag{37}$$

from which we get the average waiting time

$$\overline{W} = \overline{T} - \overline{S} \tag{38}$$

where

λ is the arrival rate
$\overline{S}$ is the average service time
ρ is the average utilization where $\rho = \lambda \overline{S}$
C_b is the coefficient of variation of S

The Laplace transform of $A^*(T)$ is given by [14, Eq. (5.100)]

$$A^*(t) = \frac{t(1 - \rho)}{t - \lambda + \lambda S^*(t)} S^*(t)$$

where λ is the arrival rate, $1/\mu$ is the average service time, $\rho = \lambda/\mu$ and $S^*(t)$ is the Laplace transform of the service time distribution.

References

[1] Diego Cruz Abrahão, Flávio Henrique Teles Vieira, and Marcus Vinícius Gonzaga Ferreira. Resource allocation algorithm for LTE networks using βMWM modeling and adaptive effective bandwidth estimation. In International Workshop on Telecommunications (IWT), pages 1–8, June 2015.

[2] M.H. Ahmed. Call admission control in wireless networks: A comprehensive survey. IEEE Communications Surveys & Tutorials, 7(1):49–68, 2005.

[3] Shiyao Chen and Lang Tong. Multiuser cognitive access of continuous time markov channels: Maximum throughput and effective bandwidth regions. In Information Theory and Applications Workshop (ITA), pages 1–10, January 2010.

[4] J.S. Evans and D. Everitt. Effective bandwidth-based admission control for multiservice CDMA cellular networks. IEEE Transactions on Vehicular Technology, 48(1):36–46, January 1999.

[5] R. Guérin, H. Ahmadi, and M. Naghshineh. Equivalent capacity and its applications to bandwidth allocation in high-speed networks. IEEE Journal on Selected Areas in Communications, 9(7):968–981, September 1991.

[6] Pengchao Han, Lei Guo, and Yejun Liu. VNE2: A virtual network embedding framework based on equivalent bandwidth in fiber-wireless enhanced 5G networks. In 22nd International Conference on Transparent Optical Networks (ICTON), pages 1–4, July 2020.

[7] Pengchao Han, Yejun Liu, Xu Zhang, and Lei Guo. Energy-efficient service placement based on equivalent bandwidth in cell zooming enabled mobile edge cloud networks. IEEE Transactions on Vehicular Technology, 71(11):12275–12290, November 2022.

[8] J.H. Hui. Resource allocation for broadband networks. IEEE Journal on Selected Areas in Communications, 6(9):1598–1608, December 1988.

[9] Ksenia Kalinina, Evsey Morozov, and Vladimir Rykov. Effective bandwidth estimation in highly reliable regenerative networks. In Second International Symposium on Stochastic Models in Reliability Engineering, Life Science and Operations Management (SMRLO), pages 323–327, February 2016.

[10] J.S. Kaufman. Blocking in a shared resource environment. IEEE Transactions on Communications, 29(10):1474–1481, October 1981.

[11] F. P. Kelly. Effective bandwidths at multi-class queues. Queuing Systems, 9:5–16, 1991.

[12] F.P. Kelly. Notes on effective bandwidths. In F. Kelly, S. Zachary, and I. Ziedins, editors, Stochastic Networks: Theory and Applications, volume 4 of Royal Statistical Society Lecture Notes Series, pages 141–168. Oxford University Press, 1996.

[13] Bosung Kim, Gyu-Min Lee, Geunkyung Choi, and Byeong-Hee Roh. Effective bandwidth based capacity request for real-time voice in GEO satellite system. In IEEE Military Communications Conference, pages 379–384, November 2016.

[14] L. Kleinrock. Communication Nets: Stochastic Message Flow and Delay. McGraw-Hill, 1964.

[15] L. Kleinrock. Queuing Systems. Wiley, 1975.

[16] Hyun-Jin Lee and Jae-Hyun Kim. An effective bandwidth based admission control for multimedia service in WLAN. In IEEE Military Communications Conference, pages 1003–1008, October 2014.

[17] Lintao Li, Wei Chen, and Khaled B. Letaief. Short packet communications with random arrivals: An effective bandwidth approach. In IEEE Global Communications Conference (GLOBECOM), pages 1–6, December 2021.

[18] Indra H. Mulyadi, N. M. A. Sumarang, Norlaili M. Safri, and Eko Supriyanto. Effective bandwidth allocation algorithm for medical device wireless network. In 2nd International Conference on Instrumentation, Communications, Information Technology, and Biomedical Engineering, pages 304–309, November 2011.

[19] Su Pan, Cheng Li, Sheng Zhang, and Danwei Chen. Cross-layer resource allocation based on equivalent bandwidth in OFDMA systems. Journal of Systems Engineering and Electronics, 27(4):754–762, August 2016.

[20] H.T. Papadopolous, C. Heavey, and J. Browne. Queueing Theory in Manufacturing Systems Analysis and Design. Springer, 1993.

[21] J.W. Roberts. A service system with heterogeneous user requirements: Application to multi-services telecommunications systems. In G. Pujolle, editor, Performance of Data Communication Systems and their Applications, pages 423–431. North-Holland Publishing Co., 1981.

[22] S. Ebrahim Safavi and K. P. Subbalakshmi. Effective bandwidth for delay tolerant secondary user traffic in multi-PU, multi-SU dynamic spectrum access networks. IEEE Transactions on Cognitive Communications and Networking, 1(2):175–184, June 2015.

[23] B.A. Sevastyanov. An ergodic theorem for Markov processes and its application to telephone systems with refusals. Theory Probab Appl, 2:104–113, 1957.

[24] Thiago Stahlschmidt, Walter Godoy, and Augusto Foronda. Effective bandwidth based connection admission control for stored variable bit rate video streams. In IEEE Latin-American Conference on Communications, pages 1–6, September 2010.

[25] Maciej Stasiak, Mariusz Glabowski, Arkadiusz Wisniewski, and Piotr Zwierzykowski. Equivalent Bandwidth, pages 175–190. Wiley, 2010.

[26] Flávio H. T. Vieira, Bruno H. P. Gonçalves, Flávio G. C. Rocha, and Luan L. Lee. Dynamic time slot allocation for multiuser OFDM/TDMA networks using effective bandwidth and βMWM network traffic modeling. In 10th International Conference on Network and Service Management (CNSM) and Workshop, pages 340–343, November 2014.

[27] Steven Wright. Admission control in multi-service IP networks: A tutorial. IEEE Communications Surveys & Tutorials, 9(2):72–87, 2007.

[28] DongXu Zhou, Hong Zhu, WenDi Wang, LiSha Gao, and LinQing Yang. 5G radio resource slicing based on equivalent bandwidth for IOT of the power grid. In 13th International Conference on Wireless Communications and Signal Processing (WCSP), pages 1–5, October 2021.

[29] Moshe Zukerman. Introduction to Queueing Theory and Stochastic Teletraffic Models. Number arXiv:1307.2968v16 in math.PR. arXiv, June 2017.