Centrum voor Wiskunde en Informatica
REPORT*RAPPORT*

On the effective bandwidth in buffer design for the multi-server channels

J.W. Cohen

Department of Operations Reasearch, Statistics, and System Theory

BS-R9406 1995

Report BS-R9406
ISSN 0924-0659

CWI
P.O. Box 94079
1090 GB Amsterdam
The Netherlands

On the Effective Bandwidth in Buffer Design for the Multi-Server Channels

J.W. Cohen
CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

January 1994

Abstract

GIBBENS and HUNT [3] derive the expression for the effective bandwidths for a uniform arrival and service model of traffic flow offered to a multi-service channel. Their model is based on negative exponentially distributed on-times of the traffic sources. The present study leads to the expression for the effective bandwidth without specifying these distributions explicitly. By the results so obtained the influence of the burstiness of the individual source traffic on its effective bandwidth can be analysed; it is shown that it can be quite substantial.

AMS Subject Classification (1991): 90B22, 60K25
Keywords & Phrases: effective bandwidth, multi-service channels, instantaneous input, gradual input, maximal buffer content, $M/G/1$ model, tail behaviour, burstiness, in-and outflow rate.

1. INTRODUCTION

In communication systems for data handling and transport the incoming traffic stream usually passes a buffer before it is admitted to the processing network. The buffer transforms the incoming irregular traffic stream into a fairly smooth one. This incoming traffic is usually generated by a number of sources. Generally, N types of sources can be distinguished. Denote by n_j the number of type j-sources, $j = 1, \ldots, N$, all j-sources have the same traffic characteristics.

The performance of the buffer is a crucial element for the processing of traffic. The performance measure (*p.m.*) of the buffer may be expressed in terms of its blocking probability, of its average content, of its probability of being empty, in general in terms of characteristics of the probability distribution of the buffer content.

The total traffic stream offered to the buffer is a mix of the N traffic streams, each stream generated by a class of j-sources; this mixed stream is characterised by the vector $(n_1, \ldots, n_N)$. A performance measure P of the buffer is obviously a function of this vector.

In his study KELLY [1] shows that for many, relevant p.m.'s P the components n_j of the mix $(n_1, \ldots, n_N)$ satisfy in first approximation a linear relation for given value of P i.e.

$$\sum_{j=1}^{N} c_j n_j \;=\; 1, \tag{1.1}$$

with of course c_j depending on the type of chosen p.m. P.

In (1.1) the coefficient c_j of n_j is called the *effective bandwidth* of a j-source (for the chosen P). Obviously c_j measures the contribution of the j-source to the p.m. P, and as such the knowledge of the $c_j, j = 1, \ldots, N$, is most useful in composing a well balanced mix when the traffic characteristics of the j-sources are known.

In many applications the workload process of the buffer can be modelled by that of an $M/G/1$ queueing model. In the standard $M/G/1$ model the workload process increases instantaneously at the arrival of a customer, the jump being the required service time of this customer.

In data transport systems the activities of a source consist of alternating idle ("off") and sending ("on") periods. Only during an on-period the source feeds the buffer. Here the workload process $\mathbf{v}_t$

of the buffer does not increase discontinuously but continuously, i.e. gradually; its rate of increase depends on the sending speeds of the active sources and on the outflow rate of the buffer. For this model of traffic streams, the so-called uniform arrival and service model (u.a.s.), GIBBENS and HUNT [3] show that an effective bandwidth approximation exists when for the p.m. P is taken the probability that the buffer content $\mathbf{v}_t$ exceeds a level v with v large (for the distribution of $\mathbf{v}_t$ they take the stationary distribution of the $\mathbf{v}_t$-process). In their analysis Gibbens and Hunt assume that the off- as well as the on-periods are negative exponentially distributed, the parameters of these distributions and the sending speed depend all on the type of the source.

In the present paper we investigate the concept of effective bandwidth for the $M/G/1$ model with instantaneous input and for the u.a.s. model, i.e. for gradual input. In the study of both models the distribution of the traffic supplied to the buffer by a single active source is not specified, it is an arbitrary distribution but source dependent. In both cases the p.m. P will be based on the distribution of the maximum workload of the buffer during a busy cycle of the buffer. For the determination of the capacity of the buffer this distribution is the appropriate one.

Although the distribution of the traffic fed into the buffer by a single source activity is not specified it turns out that an effective bandwidth approximation can be formulated for both models. The fact that the distribution of the supplied traffic by a j-source is not specified provides a good insight in the relation between the effective bandwidth and the higher moments of this distribution, so that the effect of burstiness of the traffic can be judged. Of further interest is the comparison of the effective bandwidths for both models.

We proceed with a short description of the various sections.

In Section 2 we start with the definitions of the various traffic characteristics of a single source for the case of instantaneous input. The input epochs of a single source form a Poisson process. N types of sources are considered, and the workload process $\mathbf{v}_t$ of the buffer is modelled by that of an $M/G/1$ queue. It is assumed that the traffic load of the buffer is less than one, so that its busy cycle has a finite first moment, and that *its output rate is equal to one* (when the buffer is not empty). For this model the distribution of $\mathbf{v}_{\max}$, the maximum workload of the buffer during a busy cycle, is a simple functional of the stationary waiting time distribution, and the tail behaviour of the distribution of $\mathbf{v}_{\max}$ can be readily formulated for the case that the Laplace-Stieltjes transform $\beta(\rho)$ of the distribution of the traffic generated by a single j-source activity has a negative abscissa of convergence for all $j = 1, \ldots, N$; this is assumed, see (2.7) and remark 2.1 below; in applications this assumption is hardly a restriction. With the knowledge of the tail behaviour of the distribution of $\mathbf{v}_{\max}$ we can proceed along the lines indicated by GIBBONS and HUNT [3] with the investigation of the effective bandwidth. The resulting expression for the effective bandwidth c_j of a j-source contains the L.S.-transform $\beta_j(\cdot)$, see (2.19) below. The Section 2 concludes with a first order approximation for the parameter which characterises the tail behaviour of the distribution of $\mathbf{v}_{\max}$, and for the effective bandwidth c_j. This first order approximation indicates the influence of the second moments of the traffic load on c_j.

In Section 3 the gradual input process. (the u.a.s.-model) is formulated. First, the traffic stream delivered by a single source is described. The outflow rate of the buffer is scaled to be equal to one and the sending rates of all sources are assumed to be at least one, cf. (3.2). This assumption is rather essential in our analysis, see with regard to this Section 5. and remark 4.6. Further an expression is derived for the traffic load which is generated by a single source during a time interval (t_1, t_2), where t_1 and t_2 are two epochs each of which is overlapped by an off-period of the source.

Section 4 starts with an analysis of the workload process, it is based on an earlier study of the author, cf. [2]. It is shown that this workload process contains an embedded discrete time parameter Markov process which can be described as the actual waiting time process of an $M/G/1$ queueing model. It is this observation which leads to the characterisation of the distribution of the maximum $\mathbf{v}_{\max}$ of the workload process of the buffer during a busy cycle. The tail behaviour of this distribution can be analyzed and along the same lines as in Section 2 an expression for the effective bandwidth is obtained. Again a first order approximation is derived.

In Section 5 our results holding for nonspecified distributions of the on-periods of the various sources are compared. with those obtained by GIBBENS and HUNT [3], where these distributions are negative exponential. Both results agree. The results in [3] are independent of our assumption (3.2) concerning the inflow rates of the sources. Section 5 contains a discussion of the motives which indicate that our expressions for the effective bandwidths may be assumed to be independent of the assumption (3.2). The appendices contain the derivations of some results needed in the other sections.

2. INTRODUCTION

In this section we shall discuss the concept of effective bandwidth for a buffer fed by various traffic sources for the case that the traffic load stemming from a single activity of a source is put instantaneously into the buffer. The buffer content is processed by a single server, whose service speed is taken to be equal to one.

We first describe the traffic characteristics of a single source. Such a source generates inputs according to a Poisson process with rate λ. Each of those inputs results in an instantaneous increase of the workload of the buffer, which is assumed to have an unlimited capacity. The successive additions of the buffer content are assumed to be independent, identically distributed, positive stochastic variables with distribution $B(\cdot)$, put

$$\beta := \int_0^\infty t\mathrm{d}B(t)\ , \quad \beta^{(2)} := \int_0^\infty t^2\mathrm{d}B(t), \quad B(0+) = 0. \tag{2.1}$$

So far for the description of the traffic stream generated by a single source.

The total traffic stream is composed of a number of streams generated by the single sources. We shall distinguish N types of sources. For a source of type $j, j = 1, \ldots, N$, the characteristics as introduced above will be indicated by a subscript j, e.g. λ_j is the rate of the Poisson process of inputs originated by a j-source. The number of type j-sources will be indicated by n_j, so the traffic stream fed into the buffer is generated by

$$M := \sum_{j=1}^N n_j \tag{2.2}$$

sources. It will henceforth be assumed that the M traffic streams generated by the M sources are independent stochastic processes. Put

$$\Lambda := \sum_{j=1}^N \lambda_j n_j \quad , \quad p_j := \frac{\lambda_j n_j}{\Lambda} \quad , \quad j = 1, \ldots, N, \tag{2.3}$$

$$B(\tau) := \sum_{j=1}^N p_j B_j(\tau), \quad \beta(\rho) := \sum_{j=1}^N p_j \beta_j(\rho), \ \mathrm{Re}\,\rho \geq 0.$$

Obviously, the epochs at which the successive inputs are generated by the M sources is again a Poisson process and has rate equal to Λ. Further p_j is the probability that an input stems from a j-source and $B(\cdot)$, cf. (2.3), is the distribution of the workload supplied to the buffer if the sending source is not specified.

With $\mathbf{v}_t$ the buffer content at time t, i.e. the workload of the server, it is readily seen from the definitions above that the $\mathbf{v}_t$-process is the workload process of an $M/G/1$ queue.

It will be assumed that

$$A := \Lambda \sum_{j=1}^{N} p_j \beta_j = \sum_{j=1}^{N} n_j \lambda_j \beta_j < 1. \tag{2.4}$$

This assumption implies that the busy cycles of the $\mathbf{v}_t$-process are finite with probability one and have finite first moment given by, cf. [2].

$$\frac{\Lambda^{-1}}{1-A}. \tag{2.5}$$

Denote by $\mathbf{v}_{\max}$ the supremum of the workload $\mathbf{v}_t$ during a busy cycle. The distribution of $\mathbf{v}_{\max}$ has been studied in [2], and we have, cf. [2], p. 619,

$$\mathrm{E}\{\mathbf{v}_{\max}\} = \Lambda^{-1} \log (1-A)^{-1}. \tag{2.6}$$

For the study of the concept of effective bandwidth we need information concerning the tail of the distribution of $\mathbf{v}_{\max}$.

Suppose that the abscissa $\hat{\rho}_j$ of convergence of $\beta_j(\rho)$ is negative and $\beta_j(\rho) \to \infty$ for $\rho \downarrow \hat{\rho}_j, j = 1, \ldots, N$, so that for $j = 1, \ldots, N$,

$$\hat{\rho}_j < 0 \quad \text{and} \quad |\beta_j(\rho)| = |\int_0^\infty \mathrm{e}^{-\rho t} \mathrm{d}B_j(t)| < \infty \quad \text{for Re } \rho > \hat{\rho}_j. \tag{2.7}$$

With the assumptions (2.4) and (2.7) we have, cf. [2] p. 624, that

$$\Pr\{\mathbf{v}_{\max} \geq v\} \sim b\mathrm{e}^{\epsilon v} \quad \text{for } v \to \infty, \tag{2.8}$$

where ϵ is the unique zero of, cf. (2.3),

$$\beta(\rho) + \Lambda^{-1}\rho - 1, \quad \hat{\rho} < \text{Re } \rho < 0,$$

which is nearest to the axis $\mathrm{Re}\,\rho = 0, \hat{\rho}$ and b are defined by

$$\hat{\rho} := \max\{\hat{\rho}_1, \hat{\rho}_2, \ldots, \hat{\rho}_N\},$$

$$b := \frac{1-A}{1+\Lambda\beta^{(1)}(\epsilon)} \Lambda^{-1}\epsilon, \quad \beta^{(1)}(\rho) = \frac{\mathrm{d}}{\mathrm{d}\rho}\beta(\rho); \tag{2.9}$$

actually ϵ is real and

$$\hat{\rho} < \epsilon < 0.$$

REMARK 2.1. From (2.3) it is seen that $\hat{\rho}$ is the abscissa of convergence of $\beta(\rho)$ and

$$\beta(\rho) \to \infty \quad \text{for} \quad \rho \downarrow \hat{\rho}, \quad \rho \geq \hat{\rho}.$$

Henceforth the labeling of the type of sources is chosen in such way that

$$\hat{\rho}_1 \geq \hat{\rho}_2 \geq \ldots \geq \hat{\rho}_N. \qquad \square$$

Our further analysis is based on an approach due to GIBBENS and HUNT, cf. [3].
Put

$$c_j(\delta) := \lambda_j \beta_j \frac{1-\beta_j(\delta)}{\delta \beta_j} \quad \text{for} \quad \delta \in (\hat{\rho}_j, 0)\,, \; j = 1, \ldots, N, \tag{2.10}$$

$$\mathcal{N} := \{(n_1, \ldots, n_N)\} \quad \text{with} \quad n_j \in \{0, 1, 2, \ldots\}, \; j = 1, \ldots, N, \tag{2.11}$$

$$\mathcal{A}(\delta) := \{\mathcal{N} : \quad \sum_{j=1}^{N} n_j c_j(\delta) < 1\},$$

$$\bar{\mathcal{A}}(\delta) := \{\mathcal{N} : \quad \sum_{j=1}^{N} n_j c_j(\delta) \le 1\},$$

$$A(v,p) := \{\mathcal{N} : \Pr\{\mathbf{v}_{\max} \ge v\} \le p\}, \tag{2.12}$$

with

$$p := \mathrm{e}^{\delta v} d(v), \quad d(v) > 0 \quad \text{for} \quad v > 0,$$

and

$$\frac{\log d(v)}{v} \to 0 \quad \text{for} \quad v \to \infty.$$

From (2.8) we have for v sufficiently large

$$\mathrm{e}^{\delta v} d(v) \ge \Pr\{\mathbf{v}_{\max} \ge v\} \sim b\mathrm{e}^{\epsilon v},$$

and so since $\epsilon < 0$ and $\delta < 0$ we have

$$\epsilon + \frac{\log b}{v} \ge \delta + \frac{\log d(v)}{v} \quad \text{for} \quad v >> 1. \tag{2.13}$$

It follows that for v sufficiently large

$$\Pr\{\mathbf{v}_{\max} \ge v\} \le p \quad \Leftrightarrow \quad \epsilon \ge \delta. \tag{2.14}$$

Next note that for $\hat{\rho}_j < \delta_1 < \delta < \delta_2$, $j = 1, \ldots, N$,

$$\frac{1-\beta_j(\delta_1)}{\delta_1 \beta_j} > \frac{1-\beta_j(\delta)}{\delta \beta_j} > \frac{1-\beta_j(\delta_2)}{\delta_2 \beta_j} \ge 1, \tag{2.15}$$

since

$$\frac{1-\beta_j(\rho)}{\rho \beta_j} \;=\; \frac{1}{\beta_j} \int_0^\infty \mathrm{e}^{-\rho t}\{1 - B_j(t)\}\mathrm{d}t \quad , \quad \mathrm{Re}\,\rho > \hat{\rho}_j. \tag{2.16}$$

From (2.10), (2.11), (2.14) and (2.15) it follows that

i. $\quad \delta = \epsilon \quad \Rightarrow \quad \mathcal{A}(\epsilon) \subset A(v,P) \subseteq \bar{\mathcal{A}}(\epsilon),$ (2.17)

because the definition of ϵ implies, cf. (2.8), (2.10),

$$\sum_{j=1}^{N} c_j(\epsilon) n_j \;=\; \sum_{j=1}^{N} \lambda_j \beta_j \frac{1-\beta_j(\epsilon)}{\epsilon \beta_j} n_j = 1;$$

ii. $\quad \hat{\rho} < \delta < \epsilon \quad \Rightarrow \quad A(v,p) \not\subseteq \bar{\mathcal{A}}(\delta),$

because for $\delta < \epsilon$,

$$\sum_{j=1}^{N} \lambda_j \beta_j \frac{1-\beta_j(\delta)}{\delta \beta_j} n_j > \sum_{j=1}^{N} \lambda_j \beta_j \frac{1-\beta_j(\epsilon)}{\epsilon \beta_j} = 1;$$

iii. $\delta > \epsilon \quad \Rightarrow \quad A(v,p) \subset A(\delta) \subseteq \bar{A}(\delta),$

because for $\delta > \epsilon$,

$$\sum_{j=1}^{N} \lambda_j \beta_j \frac{1-\beta_j(\delta)}{\delta \beta_j} n_j < \sum_{j=1}^{N} \lambda_j \beta_j \frac{1-\beta_j(\epsilon)}{\epsilon \beta_j} = 1.$$

From (2.7) it is seen that the condition

$$\sum_{j=1}^{N} c_j(\epsilon) n_j = 1, \tag{2.18}$$

with

$$c_j := c_j(\epsilon) = \lambda_j \beta_j \frac{1-\beta_j(\epsilon)}{\epsilon \beta_j}, \tag{2.19}$$

implies

$$\mathcal{A}(\epsilon) \subset A(v,p) \subset \bar{A}(\epsilon),$$

for v sufficiently large.
Consequently this condition (2.18) characterises the set of traffic mixes $(n, \ldots, n_N)$ for which holds

$$\Pr\{\mathbf{v}_{\max} \geq v\} \leq p = d(v)\mathrm{e}^{\epsilon v}, \tag{2.20}$$

for v sufficiently large.

The coefficient c_j in (2.18) is called the *effective bandwidth* of a j-source. It describes the influence of the number of j-sources in the traffic mix $(n_1, \ldots, n_N)$ on the probability (2.20).

Actually c_j is a first approximation because actually ϵ depends on $n_1, \ldots, n_N$, cf. (2.3) and (2.8). However, this dependence is generally rather weak even for more moderate values of v. The quality of this approximation dependes on the value of A with respect to one, cf. (2.4), as it may be seen from the following rather qualitative considerations, see also Remark 2.2.

From (2.7), (2.18) and Remark 2.1 it is readily seen that

$$\hat{\rho} = \hat{\rho}_1 \; < \; \epsilon < 0.$$

Denote by ϵ_1 the zero of

$$A \frac{1-\beta_1(\rho)}{\rho \beta_1} = 1, \quad \hat{\rho} \; < \mathrm{Re}\, \rho > 0,$$

which is nearest to $\mathrm{Re}\, \rho = 0$. It is then readily seen that $\epsilon_1 \downarrow \hat{\rho} = \hat{\rho}_1$ for $A \downarrow 0$.

The functions

$$\frac{1-\beta_j(\rho)}{\rho \beta_j}, \quad \hat{\rho} < \rho < 0,$$

are all monotonically decreasing, and uniformly bounded for all $j, j = 1$ excepted since $\hat{\rho} = \hat{\rho}_1$. Hence it is seen that the term with index $j = 1$ in

$$\sum_{j=1}^{N} n_j \lambda_j \beta_j \frac{1-\beta_j(\epsilon)}{\epsilon \beta_j} = 1,$$

is the dominant term of the sum if A is sufficiently small possibly with exception of the case that $n_1 \lambda_1 \beta_1 << \max(n_2 \lambda_2 \beta_2, \ldots, n_N \lambda_N \beta_N)$. Hence for A small $|\epsilon - \epsilon_1|$ varies hardly with $(n_1, \ldots, n_N)$.

Next we consider the case that $|\epsilon|$ is small. We may then write, because (2.9) implies that all moments of $B_j(\cdot)$ are finite,

$$\beta_j(\epsilon) = 1 - \epsilon\beta_j + \frac{1}{2}\epsilon^2\beta_j^{(2)} + o(\epsilon^2) \quad \text{for} \quad \epsilon \downarrow 0. \tag{2.21}$$

By using (2.21) the first order approximation of ϵ is given by, cf. (A.3.2),

$$-\epsilon \sim \frac{1 - \sum_{j=1}^{N} n_j\lambda_j\beta_j}{\sum_{j=1}^{N} n_j\lambda_j\beta_j^{(2)}/2}, \tag{2.22}$$

or

$$-\epsilon\beta \sim \frac{1-A}{A}\{\frac{\beta^{(2)}}{2\beta^2}\}^{-1},$$

with β and $\beta^{(2)}$ the first and second moment of $B(\cdot)$, cf. (2.3). From (A.3.3) it is seen that this first order approximation will in general be quite accurate if

$$\frac{1-A}{A}\{\frac{\beta^{(2)}}{2\beta^2}\}^{-1} \quad << 1. \tag{2.23}$$

From (2.21) and (2.22) we obtain the first order approximation of c_j, i.e.

$$c_j \sim \lambda_j\beta_j\{1 + \frac{\beta_j^{(2)}}{2\beta_j}\frac{1 - \sum_{j=1}^{N} n_j\lambda_j\beta_j}{\sum_{j=1}^{N} n_j\lambda_j\beta_j^{(2)}/2}\} = \lambda_j\beta_j\{1 + \frac{\beta_j}{\beta}\frac{\beta_j^{(2)}}{2\beta_j^2}\frac{1-A}{A}[\frac{\beta^{(2)}}{2\beta^2}]^{-1}\}. \tag{2.24}$$

Hence it is seen that if (2.23) applies that again c_j depends weakly on the traffic mix $(n_1, \ldots, n_N)$.

The condition (2.23) obviously applies if $1-A$ is small and/or $\beta^{(2)}/2\beta^2$ large. In the latter case the total traffic is quite bursty, and from (2.24) it is then seen how the burstiness of a j-source influences its effective bandwidth, see also Remark 2.2.

REMARK 2.1. From

$$\lim_{v\to\infty} \frac{1}{v} \log \Pr\{\mathbf{v}_{\max} \geq v\} = \epsilon,$$

it seen that if ϵ is known then for v sufficiently large $\Pr\{\mathbf{v}_{\max} \geq v\}$ is approximately known. Further if $\beta(\rho)$ is explicitly known then, by using wellknown asymptotical techniques, it is not difficult to obtain sharper approximations for $\Pr\{\mathbf{v}_{\max} \geq v\}$. □

REMARK 2.2. The relation (2.22) is illustrative to judge the influence of bursty traffic. Consider e.g. the case with two source types, with same rates λ_j and first moments β_j, but different second moments $\beta_j^{(2)}$, so

$$N = 2, \quad \tilde{\lambda} := \lambda_1 = \lambda_2 \quad , \quad \tilde{\beta} := \beta_1 = \beta_2, \quad n_1 = n_2 = \frac{1}{2}M.$$

For this case (2.22) becomes

$$-\epsilon\tilde{\beta} \sim \frac{1 - M\tilde{\lambda}\tilde{\beta}}{[\frac{\beta_1^{(2)}}{2\tilde{\beta}^2} + \frac{\beta_2^{(2)}}{2\tilde{\beta}^2}]\tilde{\beta}M/2},$$

and it is seen that $-\epsilon\tilde{\beta}$ strongly depends on the ratio of $\beta_1^{(2)}/\beta_2^{(2)}$, and hence so does $\Pr\{\mathbf{v}_{\max} \geq v\}$, cf. Remark 2.1. Note that a large value of $\beta_j^{(2)}/2\beta_j > 1$ characterises the j-source as a bursty source. It is not so difficult to verify that this conclusion not only holds for the first order approximations of ϵ but also for the exact value of ϵ as a function of the second moments $\beta_j^{(2)}$. Note that

$$\frac{\beta_j^{(2)}}{2\beta_j^{(2)}} = \int_0^\infty t\{1 - B_j(t)\}\mathrm{d}t,$$

and that

$$\int_0^\infty t\mathrm{e}^{-\rho t}\{1 - B_j(t)\}\mathrm{d}t$$

decreases from-∞ to $\beta_j^{(2)}/2\beta_j$ on $[\hat{\rho}_j, 0]$. □

3. Gradually input

In this section we shall derive an expression for the bandwidth for the case that the traffic load of an input is gradually fed into the buffer. The buffer content is processed by a single server, whose service speed is again taken equal to one.

Again we start with the description of the traffic characteristics of a single source.

fig. 1.

Let $\{\boldsymbol{\sigma}_n, n = 1, \ldots\}$ and, similarly $\{\boldsymbol{\tau}_n, n = 1, 2, \ldots\}$, be a sequence of independent, identically distributed, positive stochastic variables, and moreover both these families of stochastic variables are assumed to be independent families. The $\boldsymbol{\sigma}_n$ are negative exponentially distributed with parameter λ. Put

$$A(\sigma) := \Pr\{\boldsymbol{\sigma}_n < \sigma\} = 1 - \mathrm{e}^{-\lambda\sigma}, \quad \sigma > 0, \tag{3.1}$$

$$B(\tau) := \Pr\{\boldsymbol{\tau}_n < \tau\}, \quad \tau > 0, \quad B(0+) = 0,$$

$$\beta := \int_0^\infty \tau\mathrm{d}B(\tau), \quad \beta^{(2)} := \int_0^\infty \tau^2\mathrm{d}B(\tau),$$

$$\beta(\rho) := \int_0^\infty \mathrm{e}^{-\rho\tau}\mathrm{d}B(\tau), \quad \mathrm{Re}\,\rho \geq 0.$$

The time intervals $\boldsymbol{\sigma}_n, n = 1, 2, \ldots$, are the successive off-periods of the source, the intervals $\boldsymbol{\tau}_n, n = 1, 2, \ldots$, the successive on-periods of the source, i.e. at the start of a τ-interval the source generates an input, and during such a τ-interval it generates traffic which is fed into the buffer with a rate γ, so $\gamma\boldsymbol{\tau}_n$ is the traffic fed into the buffer by the nth input. It will always be assumed that

$$\gamma \geq 1, \tag{3.2}$$

this assumption is rather essential for the analysis in section 4, cf. further section 5 and remark 4.6. Define, see fig. 1,

$$\begin{aligned} \mathbf{z}_t = 1 \quad & if \quad epoch \quad t \quad is \quad overlapped \quad by \quad a \quad \sigma - interval, \\ = 0 \quad & ,, \quad ,, \quad t \quad ,, \quad ,, \quad ,, \quad ,, \quad \tau - interval, \end{aligned} \tag{3.3}$$

(and take $\mathbf{z}_t$ to be continuous from the right).

It follows that

$$\mathbf{h}_t := \gamma t - \int_0^t \mathbf{z}_\tau \, \mathrm{d}\tau, \tag{3.4}$$

is the traffic load supplied to the buffer by the source during the time interval $[0, t]$.

It is seen that : for $\mathrm{Re}\,\rho \geq 0$,

$$\mathrm{E}\{\mathrm{e}^{-\rho \mathbf{h}_t}(\mathbf{z}_t = 1) | \mathbf{z}_0 = 1\} = \tag{3.5}$$

$$\sum_{n=0}^{\infty} \int_{v=0-}^{t} \mathrm{e}^{-\rho\gamma v}\{A^{n*}(t-v) - A^{(n+1)*}(t-v)\}\mathrm{d}B^{n*}(v),$$

where $n*$ stands for the n-fold convolution and $A^{0*}(\cdot)$, $B^{0*}(\cdot)$ are probability distributions degenerated at zero.

From (3.1) and (3.5) it follows readily: for $\mathrm{Re}\,\rho \geq 0$, $\mathrm{Re}\,s > 0$,

i. $$\int_0^\infty \mathrm{e}^{-st}\mathrm{E}\{\mathrm{e}^{-\rho \mathbf{h}_t}(\mathbf{z}_t = 1) | \mathbf{z}_0 = 1\}\mathrm{d}t = \frac{1}{s + \lambda\{1 - \beta(\gamma\rho + s)\}}, \tag{3.6}$$

ii. for $\mathrm{Re}\,\rho \geq 0$, $t > 0$,

$$\mathrm{E}\{\mathrm{e}^{-\rho \mathbf{h}_t}(\mathbf{z}_t = 1) | \mathbf{z}_0 = 1\} = \frac{1}{2\pi\mathrm{i}} \int_{u=\omega-\mathrm{i}\infty}^{\omega+\mathrm{i}\infty} \frac{\mathrm{e}^{ut}\mathrm{d}u}{u + \lambda\{1 - \beta(\gamma\rho + u)\}}, \quad \omega > 0;$$

here the relation (3.6)ii follows from (3.6)i by using the wellknown inversion formula for the Laplace transform.

REMARK 3.1. Obviously, (3.6)i also holds for $s = 0$, $\mathrm{Re}\,\rho > 0$. □

So far for the description of the traffic stream generated by a single source.

The total input traffic stream of the buffer is generated by n_j sources of type $j, j = 1, \ldots, N$, here N is the number of different types. Each of these M sources with

$$M = \sum_{j=1}^{N} n_j, \tag{3.7}$$

generates a traffic stream with a structure as described above for the single source, For the characteristics of the traffic stream of a j-source the same symbols as above will be used but then *indexed* by the subscript j.

Put

$$\Lambda := \sum_{j=1}^{N} \lambda_j n_j, \quad p_j := \frac{\lambda_j n_j}{\Lambda}, \quad j = 1, \ldots, N, \tag{3.8}$$

$$B(\tau) := \sum_{j=1}^{N} p_j B_j(\tau), \quad \beta(\rho) := \sum_{j=1}^{N} p_j \beta_j(\rho), \quad \mathrm{Re}\,\rho \geq 0.$$

Next we derive an expression for the total traffic offered to the buffer by the M sources during the interval $[0, t]$. We label the M sources as follows:

$$1, \ldots, n_1,\ n_1 + 1, \ldots, n_1 + n_2,\ n_1 + n_2 + 1, \ldots, n_1 + n_2 + \cdots + n_N. \tag{3.9}$$

Define for $k = 1, \ldots, M$,

$$\begin{aligned} \mathbf{z}_t^{(k)} &= 0 \quad \textit{when} \quad \textit{source} \quad k \quad \textit{is} \quad \text{“on”}, \\ &= 1 \quad ,, \qquad ,, \qquad k \quad ,, \quad \text{“off”}, \end{aligned} \tag{3.10}$$

and put

$$\mathbf{Z}_t^{(M)} := \prod_{k=1}^{M} \mathbf{z}_t^{(k)}. \tag{3.11}$$

Obviously, $\mathbf{Z}_t^{(M)} = 1$ is the event that all sources are idle at time t.

From now on it is *assumed* that the M families of stochastic variables which characterise the traffic process of a single source are independent families.

From this assumption it is readily seen that the alternating intervals for which $\mathbf{Z}_t^{(M)} = 1$ and $\mathbf{Z}_t^{(M)} = 0$ are independent, and that zero intervals are identically distributed. Similarly, the 1-intervals are identically distributed; moreover these latter intervals are negative exponentially distributed with first moment Λ^{-1}.

Denote by $\mathbf{h}_t^{(M)}$ the total traffic load generated by the M traffic sources during the interval $[0, t]$. Because all the individual traffic streams are independent stochastic processes it follows from (3.6)ii that for $\mathrm{Re}\,\rho \geq 0$, $t > 0$,

$$\mathrm{E}\{\mathrm{e}^{-\rho \mathbf{h}_t^{(M)}}\ (\mathbf{Z}_t^{(M)} = 1) | \mathbf{Z}_0^{(M)} = 1\} = \tag{3.12}$$

$$\prod_{j=1}^{N} \Big[\frac{1}{2\pi \mathrm{i}} \int_{-\mathrm{i}\infty - \omega}^{\mathrm{i}\infty + \omega} \frac{\mathrm{e}^{ut}\mathrm{d}t}{u + \lambda_j\{1 - \beta_j(\rho\gamma_j + u)\}}\Big]^{n_j}, \quad \omega > 0.$$

4. The workload process of the buffer

By $\mathbf{v}_t$ we shall indicate the workload of the buffer at time t. In figure 2 a realisation of this $\mathbf{v}_t$-process

is shown.

fig. 2.

In the lower part of the figure the realisation of the $\mathbf{Z}_t^{(M)}$-process is shown, cf. (3.11). In the middle part of the figure the off- and on- periods of the sources are plotted, it shows for this realisation that three times two sources are simultaneously busy, Initially, at time $t = 0$ all sources are idle, at time $\mathbf{s}_1$ a first request arrives and the buffer content starts to increase with rate $\gamma_{j_1} - 1$, denoting by j_1 the type of the source from which the first input originates. Immediately after the arrival of the second input, say of j_2-type, the buffer content increases with rate $\gamma_{j_1} + \gamma_{j_2} - 1$, and so on. At time $\mathbf{s}_i + \mathbf{t}_1$ the buffer content is equal to $\mathbf{b}_1$. During $(\mathbf{s}_1 + \mathbf{t}_1, \mathbf{s}_1 + \mathbf{t}_1 + \mathbf{s}_2)$ there is no inflow and the buffer content decreases with unit rate; $\mathbf{w}_2 = \mathbf{b}_1 - \mathbf{s}_2$ is the buffer content at the start of the next inflow originating from a j_4-type source. At time

$$\mathbf{s}_1 + \mathbf{t}_1 + \mathbf{s}_2 + \mathbf{t}_2 + \mathbf{s}_3 + \mathbf{t}_3 + \mathbf{w}_3 + \mathbf{b}_3$$

the buffer becomes empty for the first time after $t = \mathbf{s}_1$, and so

$$\mathbf{p} = \mathbf{t}_1 + \mathbf{s}_2 + \mathbf{t}_2 + \mathbf{s}_3 + \mathbf{t}_3 + \mathbf{w}_3 + \mathbf{b}_3$$

is the first busy period of the buffer for the realisation drawn in figure 2. At time

$$\mathbf{s}_1 + \mathbf{t}_1 + \mathbf{s}_2 + \mathbf{t}_2 + \mathbf{s}_3 + \mathbf{t}_3 + \mathbf{s}_4$$

a new busy period of the buffer starts and so

$$\mathbf{c} = \mathbf{t}_1 + \mathbf{s}_2 + \mathbf{t}_2 + \mathbf{s}_3 + \mathbf{t}_3 + \mathbf{s}_4$$

is the first busy cycle of the buffer.

Obviously, we have for $n = 1, 2, \ldots$,

$$\mathbf{w}_{n+1} = [\mathbf{w}_n + \mathbf{b}_n - \mathbf{s}_{n+1}]^+, \quad n = 1, 2, \ldots, \tag{4.1}$$

$$\mathbf{w}_1 \quad = 0.$$

Apparently, $\mathbf{b}_n$, $n = 1, 2, \ldots$, are i.i.d. stochastic variables, also $\mathbf{s}_n, n = 1, 2, \ldots$, are i.i.d. stochastic variables, moreover these families of stochastic variables are independent, since the stochastic traffic flows of the individual sources are assumed to be independent, see the preceding section. Because $\mathbf{s}_n$ is negative exponentially distributed with first moment Λ^{-1}, it follows that the sequence $\mathbf{w}_n$, $n = 1, 2, \ldots$, has the same stochastic properties as the sequence of the actual waiting times of an $M/G/1$ queueing model with arrival rate Λ and with service time distribution the distribution of $\mathbf{b}_n$.

For the analysis of the workload process $\mathbf{v}_t$ of the present model we need the distribution of $\mathbf{b}_n$ which is the net workload offered to the buffer during the inflow period $\mathbf{t}_n$, here $\mathbf{b}_n$ is equal to the workload produced by the M sources during $\mathbf{t}_n$ reduced by $\mathbf{t}_n$, since the service rate is equal to one.

The distribution of $\mathbf{b}_n$ is calculated as follows. The traffic stream into the buffer is characterised by the idle periods $\mathbf{s}_n$ and the inflow periods $\mathbf{t}_n$ and during $\mathbf{t}_n$ a traffic load $\mathbf{b}_n$ is produced, note that $\mathbf{b}_n$ and $\mathbf{t}_n$ are not independent.

This traffic input process has a structure quite similar to that of a single source. Denote by $\mathbf{k}_t$ the total workload offered to the buffer by the M sources during $[0, t]$. Put

$$F(s) := \Pr\{\mathbf{s}_n < s\} = 1 - \mathrm{e}^{-\Lambda s}, \quad s > 0, \tag{4.2}$$

and let $(\mathbf{b}, \mathbf{t})$ be a pair of stochastic variables with the same joint distribution as that of $(\mathbf{b}_n, \mathbf{t}_n)$. It follows: for $\mathrm{Re}\,\rho \geq 0$, $\mathrm{Re}\,s > 0$,

$$\int_0^\infty \mathrm{e}^{-st} \mathrm{E}\{\mathrm{e}^{-\rho \mathbf{k}_t} (\mathbf{Z}_t^{(M)} = 1) | \mathbf{Z}_0^{(M)} = 1\} \mathrm{d}t = \tag{4.3}$$

$$\int_{v=0-}^\infty \mathrm{e}^{-st} \mathrm{d}t \int_{v=0-}^t \sum_{n=0}^\infty \{F^{n*}(t-v) - F^{(n+1)*}(t-v)\} \mathrm{d}_v \, \mathrm{E}\{\mathrm{e}^{-\rho \sum_{i=1}^n (\mathbf{b}_i + \mathbf{t}_i)} (\mathbf{t}_1 + \cdots + \mathbf{t}_n < v)\} =$$

$$= \frac{1}{s + \Lambda\{1 - \mathrm{E}\{\mathrm{e}^{-\rho(\mathbf{b}+\mathbf{t}) - s\mathbf{t}}\}\}}.$$

It is seen from (3.12) and (4.3) that: for $\mathrm{Re}\,\rho \geq 0$, $\mathrm{Re}\,s > 0$,

$$\frac{1}{s + \Lambda\{1 - \mathrm{E}\{\mathrm{e}^{-\rho(\mathbf{b}+\mathbf{t}) - s\mathbf{t}}\}\}} = \tag{4.4}$$

$$\int_0^\infty \mathrm{e}^{-st} \mathrm{d}t \prod_{j=1}^N [\frac{1}{2\pi \mathrm{i}} \int_{-\mathrm{i}\infty+\omega}^{\mathrm{i}\infty+\omega} \frac{\mathrm{e}^{ut} \mathrm{d}t}{u + \lambda_j\{1 - \beta_j(\rho\gamma_j + u)\}}]^{n_j}, \quad \omega > 0.$$

The relation (4.4) is a functional relation for $\mathrm{E}\{\mathrm{e}^{-\rho \mathbf{b}}\}$, i.e. for the distribution of $\mathbf{b}$. Obviously $\mathrm{E}\{\mathrm{e}^{-\rho \mathbf{b}}\}$, $\mathrm{Re}\,\rho \geq 0$, can be obtained immediately from (4.4), by taking $s = -\rho$, if the validity of (4.4) can be extended to the domain $\mathrm{Re}\,\rho \geq 0$, $\mathrm{Re}\,s > -a - \mathrm{Re}\,\rho$, for some $a > 0$, see [4] where a similar functional relation has been studied.

REMARK 4.1. Since λ_j^{-1} is the average idle time of a j-source and β_j is the average inflow period, it is seen that $\lambda_j/(1+\lambda_j\beta_j)$ is the arrival rate per unit of idle time of the j-source. Hence the traffic A generated by the M sources is given by

$$A := \sum_{j=1}^{N} \frac{\lambda_j}{1+\lambda_j\beta_j} \gamma_j \beta_j n_j. \tag{4.5}$$

Obviously A is the traffic load offered to the buffer by the M sources, cf. also (4.20) below. □

In the following analysis it will be assumed, cf. [2.7], that for the abscissa of convergence $\hat{\rho}_j$ of $\beta_j(\rho)$ holds: for $j = 1, \ldots, N$,

$$\hat{\rho}_j < 0 \text{ and } \beta_j(\rho) \to \infty \quad for \quad \rho \downarrow \hat{\rho}_j. \tag{4.6}$$

It is readily shown that: for fixed real ρ, the function

$$f_j(r,\rho) := r + \lambda_j\{1 - \beta_j(\rho\gamma_j + r)\}, \quad \rho\gamma_j + \mathrm{Re}\,r > \hat{\rho}_j, \tag{4.7}$$

possesses a unique zero $r_j(\rho)$ such that

$$f_j(r,\rho) \neq 0 \quad for \quad \mathrm{Re}\,r > \mathrm{Re}\,r_j(\rho); \tag{4.8}$$

this zero $r_j(q)$ is real and further:

$$f_j(r,\rho) \neq 0 \quad for \quad \mathrm{Re}\,r = r_j(\rho),\ r \neq r_j(\rho), \tag{4.9}$$

$$\begin{aligned} -\rho\gamma_j < r_j(\rho) &< 0 && for \quad \rho > 0, \\ r_j(\rho) &= 0 && ,, \quad = 0, \\ -\rho\gamma_j > r_j(\rho) &> \max(0, \hat{\rho} - \gamma_j\rho) && ,, \quad < 0, \end{aligned}$$

and the zero $r_j(\rho)$ has multiplicity one.

For a proof of analogous statements see [4], lemma 2.1, where a similar zero is studied. Note that $\beta_j(\rho)$ is monotonic on $(\hat{\rho}_j, \infty)$. Because

$$\beta_j(\rho\gamma_j + u) \text{ is regular for } \mathrm{Re}\,(\hat{\rho}_j\gamma_j + u) > 0, \tag{4.10}$$

$$\text{and } r_j(\rho) \text{ is a simple zero of } f_j(r,\rho),$$

it follows by using Cauchy's theorem on contour integration that: for $\rho \geq 0$, $\omega > 0$,

$$\frac{1}{2\pi\mathrm{i}} \int_{-\mathrm{i}\infty+\omega}^{\mathrm{i}\infty+\omega} \frac{\mathrm{e}^{ut}\mathrm{d}u}{u + \lambda_j\{1 - \beta_j(\rho\gamma_j + u)\}} = \tag{4.11}$$

$$\mathrm{e}^{r_j(\rho)t}R_j(\rho) + \frac{1}{2\pi\mathrm{i}} \int_{-\mathrm{i}\infty+r_j(\rho)-}^{\mathrm{i}\infty+r_j(\rho)-} \frac{\mathrm{e}^{ut}\mathrm{d}u}{u+\lambda_j\{1-\beta_j(\rho\gamma_j+u)\}},$$

with

$$\begin{aligned} R_j(\rho) &:= \lim_{u\to r_j(\rho)} \frac{u-r_j(\rho)}{u+\lambda_j\{1-\beta_j(\rho\gamma_j+u)\}} \\ &= [1-\lambda_j\beta_j^{(1)}(\rho\gamma_j+r_j(\rho))]^{-1}. \end{aligned} \tag{4.12}$$

From (4.4) and (4.11) it follows that the abscissa of convergence $s_0(\rho)$ of the Laplace transform in the righthand side of (4.4) is given by:

$$\begin{aligned} s_0(\rho) = \sum_{j=1}^{N} n_j r_j(\rho) &< 0 \quad for \quad \rho > 0, \\ &= 0 \quad ,, \qquad = 0, \\ &> 0 \quad ,, \qquad < 0. \end{aligned} \tag{4.13}$$

By using (4.9) and noting that $\gamma_j \geq 1$, cf. (3.2), it follows that

$$s_0(\rho) < -\rho\sum_{j=1}^{N} n_j\gamma_j < -\rho, \quad \rho > 0. \tag{4.14}$$

Consequently, the relation (4.4) holds by analytic continuation for $\mathrm{Re}\, s > s_0(\rho)$, and hence we may take in (4.4) $s = -\rho$. From (4.4) and (4.11) we have for $\rho > 0$ and $\mathrm{Re}\, s > s_0(\rho)$,

$$\frac{1}{s+\Lambda\{1-\mathrm{E}\{\mathrm{e}^{-\rho(\mathbf{b}+\mathbf{t})-s\mathbf{t}}\}\}} = \tag{4.15}$$

$$[\prod_{j=1}^{N}\{R_j(\rho)\}^{n_j}] \int_0^{\infty} \mathrm{e}^{-(s-s_0(\rho))t}\mathrm{d}t \prod_{j=1}^{N}\{1+\mathrm{e}^{-r_j(\rho)t}F_j(\rho,t)\}^{n_j},$$

where for $\rho > 0$,

$$F_j(\rho,t) := \frac{1}{R_j(\rho)}\frac{1}{2\pi\mathrm{i}} \int_{-\mathrm{i}\infty+r_j(\rho)-}^{\mathrm{i}\infty+r_j(\rho)-} \frac{\mathrm{e}^{ut}\mathrm{d}u}{u+\lambda_j\{1-\beta_j(\rho\gamma_j+u)\}}. \tag{4.16}$$

Note that for $\mathrm{Re}\, s > s_0(\rho)$, $\rho > 0$,

$$\int_0^{\infty} \mathrm{e}^{-(s-s_0(\rho))t}\mathrm{d}t = \frac{1}{s-s_0(\rho)} = \frac{1}{s-\sum_{j=1}^{N} n_j r_j(\rho)}.$$

By taking $s = -\rho > s_0(\rho)$ with $\rho > 0$, it is seen that for $\rho > 0$.

$$[\prod_{j=1}^{N}\{R_j(\rho)\}^{n_j}]^{-1}\frac{\rho+s_0(\rho)}{\rho-\Lambda\{1-\mathrm{E}\{\mathrm{e}^{-\rho\mathbf{b}}\}\}} = 1+G(\rho), \tag{4.17}$$

where for $\rho > 0$,

$$G(\rho) := -1 - \{\rho + s_0(\rho)\} \int_0^\infty \mathrm{e}^{(\rho+s_0(\rho))t}\mathrm{d}t \prod_{j=1}^N \{1 + \mathrm{e}^{-r_j(\rho)t} F_j(\rho,t)\}^{n_j}. \tag{4.18}$$

Obviously (4.17) determines $\mathrm{E}\{\mathrm{e}^{-\rho\mathbf{b}}\}, \rho > 0$, so that the Laplace Stieltjes transform of the distribution of $\mathbf{b}$, has been obtained.

In the appendices A.1 and A.2 it is shown that, cf.(4.5),

i. $$\Lambda\mathrm{E}\{\mathbf{t}\} = -1 + \prod_{j=1}^N (1 + \lambda_j\beta_j)^{n_j}, \tag{4.19}$$

ii. $$1 - \Lambda\mathrm{E}\{\mathbf{b}\} = (1 - A)\prod_{j=1}^N (1 + \lambda_j\beta_j)^{n_j};$$

a direct proof of (4.19)i can be obtained by using simple properties of alternating renewal processes.

It is noted that the relations (4.19) hold independently of the assumption introduced in (4.6).

From (4.19) it is seen that

$$0 < (1 - A)\prod_{j=1}^N (1 + \lambda_j\beta_j)^{n_j} < 1 \;\Rightarrow\; 0 < \Lambda\mathrm{E}\{\mathbf{b}\} < 1. \tag{4.20}$$

From now on it will be assumed that

$$\Lambda\mathrm{E}\{\mathbf{b}\} < 1. \tag{4.21}$$

so that the stored, i.e. the delayed, traffic $\Lambda\mathrm{E}\{\mathbf{b}\}$ of the traffic offered to the buffer by the M traffic sources, is less than one. Hence the M/G/1 queueing model introduced above, see below (4.1), has busy periods with a finite first moment.

For the study of the effective bandwidth of our model we are interested in $\mathbf{v}_{\max}$, the supremum of the $\mathbf{v}_t$-process, i.e. the supremum of the buffer content during a busy cycle $\mathbf{c}$, see figure 2. A realisation of $\mathbf{v}_{\max}$ during a busy cycle occurs at an epoch which belongs to the set of endpoints of the $\mathbf{t}_n$ intervals, cf. fig. 2.

Hence it follows from (4.1) that the distribution of $\mathbf{v}_{\max}$ is identical with the distribution of the supremum of the virtual waiting time of an $M/G/1$ queue with arrival rate Λ and service time distribution that of $\mathbf{b}$. The tail behaviour of this distribution is actually described by the zero ϵ of

$$\rho - \Lambda\{1 - \mathrm{E}\{\mathrm{e}^{-\rho\mathbf{b}}\}\} \tag{4.22}$$

in $\rho < 0$, which is nearest to the axis $\mathrm{Re}\,\rho = 0$, cf. section 2. By using assumption (4.6) and with

$$\hat{\rho} := \max_{j=1,\ldots,N} \hat{\rho}_j < 0, \tag{4.23}$$

and by noting that the righthand side of (4.17) is finite for $\rho > \hat{\rho}$, it is seen that ϵ is that zero of

$$\rho + s_0(\rho) \equiv \rho + \sum_{j=1}^N n_j r_j(\rho) \; in \; \rho < 0, \tag{4.24}$$

which is nearest to the axis $\mathrm{Re}\,\rho = 0$, or equivalently of, cf. (4.7) and (4.13),

$$\rho - \sum_{j=1}^{N} n_j \lambda_j \{1 - \beta_j(\rho\gamma_j + r_j(\rho))\} \; in \; \rho < 0. \tag{4.25}$$

REMARK 4.2. The existence of ϵ follows immediately from (4.23) and the fact that $\Pr\{\mathbf{b} > 0\} > 0$. However, it is of interest to prove i by starting from (4.25). By using the substitution $u = \gamma\rho + r$ it follows readily from (4.7) that $r_j(\rho)$ is a monotonically decreasing function of ρ on $(-\infty, \infty)$ and that, cf. (3.2),

$$\lim_{\rho \to -\infty} \frac{r_j(\rho)}{-\rho} = -\gamma_j \leq -1. \tag{4.26}$$

Since, cf. (A.2.1),

$$-1 < \left. \frac{\mathrm{d}r_j(\rho)}{\mathrm{d}\rho} \right|_{\rho=0} < 0,$$

it is seen that (4.24) has a zero. □

From the considerations above it is seen that the study of the bandwidth $c_j(\epsilon)$ of a j-source for the case of gradual inflow into the buffer has been reduced to that for a buffer with instantaneous input, cf. section 2, there is only a slight difference between the equations which determine ϵ, cf. (4.25) and section 2, in particular below (2.8).

To derive for the present case of gradual inflow the expression for the effective bandwidth c_j of a j-source, put

$$\delta_j(\rho) := 1 + \frac{r_j(\rho)}{\rho\gamma_j}, \quad -\infty < \rho < \infty, \tag{4.27}$$

with $r_j(\rho)$ as defined above, cf. (4.7), it then follows from (4.25),

$$\sum_{j=1}^{N} n_j \lambda_j \beta_j \gamma_j \delta_j(\epsilon) \frac{1 - \beta_j(\epsilon\gamma_j\delta_j(\epsilon))}{\epsilon\beta_j\gamma_j\delta_j(\epsilon)} = 1. \tag{4.28}$$

The effective bandwidth c_j of a j-source is now defined by, cf. (2.19),

$$c_j := \lambda_j \beta_j \gamma_j \; \delta_j(\epsilon) \frac{1 - \beta_j(\epsilon\gamma_j\delta_j(\epsilon))}{\epsilon\beta_j\gamma_j\delta_j(\epsilon)}, \tag{4.29}$$

with ϵ the zero of (4.25) nearest to $\mathrm{Re}\,\rho = 0$.
It follows from (4.28) that

$$\sum_{j=1}^{N} n_j c_j = 1. \tag{4.30}$$

It is of interest to compare (4.29) with (2.19). It is quite obvious that β_j in (2.19) is replaced by $\beta_j\gamma_j$ in (4.29). Next note that, cf. (4.9) and (4.27),

$$-1 < \frac{r_j(\epsilon)}{\epsilon\gamma_j} < 0. \tag{4.31}$$

Further if we approximate $r_j(\rho)/\rho\gamma_j$ by its value at $\rho = 0$, cf. (A.2.1), i.e.

$$1 + \frac{r_j(\epsilon)}{\epsilon\gamma_j} \sim \frac{1}{1+\lambda_j\beta_j},$$

then (4.29) becomes:

$$c_j \sim \lambda_j \frac{\beta_j\gamma_j}{1+\lambda_j\beta_j} \frac{1-\beta_j(\frac{\epsilon\gamma_j}{1+\lambda_j\beta_j})}{\epsilon\frac{\beta_j\gamma_j}{1+\lambda_j\beta_j}}. \tag{4.32}$$

Here the righthand side is the effective bandwidth of a j-source for the case of instantaneous input with λ_j the arrival rate and

$$\frac{\gamma_j\boldsymbol{\tau}_j}{1+\lambda_j\beta_j} \quad \textit{instead of } \boldsymbol{\tau}_j,$$

the load generated by an input of a j-source.

A first order approximation of ϵ for the case with gradual input has been derived in appendix A.4, see (A.4.5). By using this approximation in (4.32) we obtain as a first order approximation of c_j:

$$c_j \sim \frac{\lambda_j\beta_j\gamma_j}{1+\lambda_j\beta_j}\,[1+\gamma_j\frac{\beta_j^{(2)}}{2\beta_j}\,\frac{1-\sum_{j=1}^{N}\frac{n_j\lambda_j\beta_j\gamma_j}{1+\lambda_j\beta_j}}{\sum_{j=1}^{N}\frac{n_j\lambda_j\gamma_j^2}{1+\lambda_j\beta_j}\beta_j^{(2)}/2}], \tag{4.33}$$

with, (cf. A.4.5),

$$-\epsilon \sim \frac{1-\sum_{j=1}^{N}\frac{n_j\lambda_j\beta_j\gamma_j}{1+\lambda_j\beta_j}}{\sum_{j=1}^{N}\frac{n_j\lambda_j\gamma_j^2}{(1+\lambda_j\beta_j)^3}\,\frac{\beta_j^{(2)}}{2}}. \tag{4.34}$$

REMARK 4.3. It is readily seen that if we let $\beta_j \to 0$, $\beta_j^{(2)} \to 0$, $\gamma_j \to \infty$ with $\gamma_j\beta_j \to \tilde{\beta}_j$, and $\gamma_j^2\beta_j^{(2)} \to \tilde{\beta}_j^{(2)}$ then the relation (4.33) becomes identical to (2.24) as it could be expected, because this limiting procedure transforms the case with gradual input into that of instantaneous input. □

Above the expression for the effective bandwidth of a j-source, cf. (4.29), for the case of gradual input with $\gamma_j \geq 1$, cf. (3.2), has been derived, see also remark 4.6 below. As in the case with instantaneous input, c_j also here depends on the traffix mix $(n_1, \ldots, n_N)$, and along the same lines as in section 2 it may be shown that also for the case with gradual input this dependence of c_j on $(n_1, \ldots, n_N)$ is rather weak, of course A as defined in (2.4) should then be replaced by A according to (4.5).

REMARK 4.4. In [3] the characteristics of a traffic stream by a single group of sources, i.e. $N = 1$ has been studied also for the case that $n_1 \to \infty$, $\lambda_1 \to 0$ such that $n_1\lambda_1 \to \hat{\lambda}$. Therefore we could also consider the case that a number $\hat{M}$ of such traffic streams ($n_j \to \infty, \lambda \to 0$, $n_j\lambda_j \to \hat{\lambda}_j$) are fed into the buffer. It is readily verified that for this case the effective bandwidth c_j is expressed by (4.29) with λ_j replace by $\hat{\lambda}_j$, and ϵ being determined from (4.25) with $r_j(\rho) \equiv 0$ and $n_j\lambda_j$ replaced by $\hat{n}_j\hat{\lambda}_j$, $j = 1, \ldots, \hat{N}$; $\hat{M} = \Sigma_{j=1}^{\hat{N}}\hat{n}_j$ where $\hat{n}_j$ is the number of j-streams. □

REMARK 4.5. In our analysis the assumption, cf. (4.6), that the abscissa of convergence ρ_j of $\beta_j(\rho)$ is a pole singularity is essential, but the case $\rho_j = -\infty$ is not excluded. However, if ρ_j is a more

complicated singularity, e.g. a branch point then the assymptotic analysis of $\Pr\{\mathbf{v}_{\max} \geq v\}$ for v large becomes more complicated. □

REMARK 4.6. It is readily seen that the analysis of the tail behaviour of the distribution of $\mathbf{b}$ remains valid if assumption (3.2) is replaced by

$$\sum_{j=1}^{N} n_j \gamma_j > 1,$$

because if this applies then we can still take $s = -\rho$ in (4.4). Note that (4.35) is a weaker condition than (3.2), and if it does not hold no traffic will be stored in the buffer. □

5. On the condition $\gamma_j \geq 1$.

The assumption, cf. (3.2),

$$\gamma_j \geq 1, \quad j = 1, 2, \ldots, N, \tag{5.1}$$

is rather essential in the analysis of section 4, cf. (4.1). Only if (5.1) holds $\mathbf{b}_n$ is the buffer content at the end of the on-period $\mathbf{t}_n$, see fig. 2. If (5.1) does not apply then $\mathbf{b}_n + \mathbf{t}_n$ is still the total traffic inflow during $\mathbf{t}_n$ but the buffer content at the end of the interval $\mathbf{t}_n$ may be some what larger then $\mathbf{b}_n$. Actually it may happen that during one or more subintervals of $\mathbf{t}_n$ the total input rate of the sending sources is less than one, so that then the output rate of the buffer, which is one, cannot be fully used whenever the buffer does not contain stored traffic. A situation which occurs if $\mathbf{w}_n = 0$ and the sending source which starts the inflow interval t_n has a sending rate $\gamma < 1$.

Hence the question arises whether the results obtained in the preceding section are still valid if (5.1) does not hold. Before discussing this question we consider the case studied in [5].

Kosten [5] considers the same model as we do; the assumption (5.1) is not made, but it is assumed that all distributions $B_j(\cdot)$ are negative exponential. For this case he obtains an expression for the tail behaviour of the stationary distribution of the buffer content. The expressions for the bandwidths based on this asymptotic result are derived in [3]. It is readily seen that they are identical to those derived in section 4 with all the $B_j(\cdot)$ negative exponential. Compare the expression $c = \Sigma_{i=1}^{N} \alpha_i(\eta_0) N_i$ in the proof of theorem 1 of [3] with our expression (4.25), note that the inflow rate γ_j in [3] should be then properly scaled since in (4.25) the output rate c of the buffer has been chosen so that $c = 1$.

REMARK 5.1. In [3] the results are based on the tail of the stationary distribution of the buffer content $\mathbf{v}_t$, our results are based on the tail of the $\mathbf{v}_{\max}$ distribution. But note that for an $M/G/1$ queueing model the tail behaviour of the stationary distribution of the $\mathbf{v}_t$-process is the same as that of the $\mathbf{v}_{\max}$-distribution. Hence this difference in approach is not very essential in the agreement of our results with those of [3]. □

The agreement of our results with those of [3], if (5.1) does not apply, may be explained as follows. Consider the successive busy periods of the buffer, a busy period being the longest time interval during which the output of the buffer is positive. We distinguish two types of busy periods, the *weak* and the *strong* ones. A strong busy period is characterised by the fact that the output rate of the buffer is equal to one during (nearly) all inflow periods of the busy period (possibly excepted the first and last one). A weak busy period is a busy period which is not a strong one.

Our derivation of the expression for the effective bandwidth is based on the distribution of $\mathbf{v}_{\max}$, the maximum buffer content during a busy cycle under the condition (5.1). Clearly the higher peaks of the buffer content process $\mathbf{v}_t$ occur mainly during the strong busy periods, and within a strong busy periods they occur mainly at the ends of those inflow periods of the buffer during which the inflow rate into the buffer is nearly always larger than one. This leads to the conjecture that the

tail statistics of the $\mathbf{v}_t$-process, in particular that of $\mathbf{v}_{\max}$, is controled by the distribution of $\mathbf{b}$, and determined by the zero ϵ of (4.22), with (5.1) replaced by (4.35), cf. remark 4.6. This explains the agreement of our results with those obtained in [5]. An analogous result occurs for an $M/G/1$ buffer model with outflow rate depending on the buffer content $\mathbf{v}$, viz. on $\mathbf{v} \leq K$ and on $\mathbf{v} > K$, cf. [7]. From the results obtained in [7] it is readily deduced that the tail behaviour of the distribution of $\mathbf{v}$, i.e. $\log \Pr\{\mathbf{v} \geq v\}/v$ for v large depends only on the outflow rate above level K, i.e. on the peaks of the content process which exceed level K.

Finally a remark on the model with all the $B_j(\cdot)$ negative exponential, cf. [5], [6], [7]. It is the model studied by Kosten, who investigate the stochastic process with state variables the buffer content and the number of sending j-souces, $j = 1, \ldots, N$, at time t. This process is a continuous type Markov process and from the Kolmogorov equations for the stationary distribution of this process the tail behaviour of the distribution of the buffer content is obtained in [5]. Such an approach cannot be worked out if one or more of the $B_j(\cdot)$ are not negative exponential. The difficulties which occur then are of a similar type as in the analysis of an $M/G/m$ queue with $m > 1$.

ACKNOWLEDGEMENTS

The author is indebted to Mr. S.C. Borst for several helpful comments.

APPENDIX A.1.

In this appendix we derive the expression for the first moment of $\mathbf{t}$.
From (4.4) with $\rho = 0$ we have for $\rho > 0, \omega > 0$,

$$\frac{s}{s + \Lambda\{1 - \mathrm{E}\{\mathrm{e}^{-s\mathbf{t}}\}\}} = \tag{A.1.1}$$

$$s \int_0^\infty \mathrm{e}^{-st}\mathrm{d}t \prod_{j=1}^N [\frac{1}{2\pi\mathrm{i}} \int_{-\mathrm{i}\infty+\omega}^{\mathrm{i}\infty+\omega} \frac{\mathrm{e}^{ut}}{u} \frac{\mathrm{d}u}{1 + \lambda_j\beta_j \frac{1-\beta_j(u)}{\beta_j u}}]^{n_j} =$$

$$s \int_0^\infty \mathrm{e}^{-st}\mathrm{d}t \prod_{j=1}^N [\sum_{k=0}^\infty (-\lambda_j\beta_j)^k \{H(t)\}^{k*}]^{n_j},$$

if $\lambda_j\beta_j < 1$;
here

$$\begin{aligned} H(t) &= \tfrac{1}{\beta_j} \textstyle\int_0^t \{1 - B_j(\tau)\}\mathrm{d}\tau, & t > 0, \\ &= 0 & t \leq 0, \end{aligned}$$

and $\{H(t)\}^{k*}$ is the k-fold convolution of $H(\cdot)$.

It is readily seen that the sum in the righthand side of (A.1.1) has a finite limit for $t \to \infty$ and so by using an Abel-theorem for the Laplace transform we have

$$\frac{1}{1 + \Lambda\mathrm{E}\{\mathbf{t}\}} = \lim_{s \downarrow 0} \frac{s}{s + \Lambda\{-\mathrm{E}\{\mathrm{e}^{-s\mathbf{t}}\}\}} =$$

$$\prod_{j=1}^N \{\sum_{k=0}^\infty (-\lambda_j\beta_j)^k\}^{n_j} = \prod_{j=1}^N \frac{1}{(1 + \lambda_j\beta_j)^{n_j}},$$

Hence

$$\Lambda\mathrm{E}\{\mathbf{t}\} = -1 + \prod_{j=1}^N (1 + \lambda_j\beta_j)^{n_j}. \tag{A.1.2}$$

REMARK A.1.1. The condition $\lambda_j\beta_j < 1$ used in the derivation of A.1.2 is superfluous, actually (A.1.1) can be derived immediately by using simple properties of the alternating renewal process.

APPENDIX A.2.

In this appendix we derive the expression for the first moment of $\mathbf{b}$.
From the definition of $r_j(\rho)$ it is readily seen that it has a derivative and that

$$\frac{\mathrm{d}r_j(\rho)}{\mathrm{d}\rho}|_{\rho=0} = -\frac{\lambda_j\beta_j\gamma_j}{1+\lambda_j\beta_j}, \quad r_j(0) = 0. \tag{A.2.1}$$

Because $s_0(\rho) \to 0$ for $\rho \to 0$ it follows since the integral in (4.18) is finite for $\rho > \hat{\rho}$, that

$$G(\rho) \to 0 \quad \text{for} \quad \rho \to 0. \tag{A.2.2}$$

Further from (4.12)

$$\lim_{\rho \downarrow 0} R_j(\rho) = [1+\lambda_j\beta_j\gamma_j]^{-1}, \tag{A.2.3}$$

and from (4.5), (4.13) and (A.2.1),

$$\frac{\mathrm{d}}{\mathrm{d}\rho}s_0(\rho) = -\sum_{j=1}^{N}\frac{\lambda_j\beta_j n_j}{1+\lambda_j\beta_j} = -A. \tag{A.2.4}$$

Hence from (4.17) we obtain for $\rho \to 0$,

$$1 - \Lambda \mathrm{E}\{\mathbf{b}\} = (1-A)\sum_{j=1}^{N}\{1+\lambda_j\beta_j\}^{n_j}. \tag{A.2.5}$$

APPENDIX A.3.

In this appendix we derive a first and a second approximation for the root ϵ, cf. (2.8), of

$$\sum_{j=1}^{N} n_j\lambda_j\beta_j\frac{1-\beta_j(\epsilon)}{\epsilon\beta_j} = 1. \tag{A.3.1}$$

For $|\epsilon| \to 0$ we have

$$\beta_j(\epsilon) = 1 - \epsilon\beta_j + \frac{\epsilon^2}{2}\beta_j^{(2)} - \frac{\epsilon^3}{6}\beta_j^{(3)} + o(\epsilon^3),$$

with $\beta_j^{(m)}$ the mth moment of $\boldsymbol{\tau}_j$; because of (4.6) this moment is finite for every $m = 1, 2, \ldots$. From (A.3.1) we obtain

$$1 = \sum_{j=1}^{N} n_j\lambda_j\beta_j[1 - \epsilon\beta_j\frac{\beta_j^{(2)}}{2\beta_j^2} + \epsilon^2\beta_j^2\frac{\beta_j^{(3)}}{6\beta_j^3} + o(\epsilon^2)]. \tag{A.3.2}$$

From which we obtain the first order approximation

$$-\epsilon \sim \frac{1-\sum_{j=1}^{N} n_j\lambda_j\beta_j}{\sum_{j=1}^{N} n_j\lambda_j\beta_j^{(2)}/2}, \tag{A.3.2}$$

and the second order approximation reads

$$\epsilon \sim -\frac{1-\sum_{j=1}^{N} n_j\lambda_j\beta_j}{\sum_{j=1}^{N} n_j\lambda_j\beta_j^{(2)}/2}[1-\frac{1}{2}\{1-\sum_{j=1}^{N} n_j\lambda_j\beta_j\}\frac{\sum_{j=1}^{N} n_j\lambda_j\beta_j^{(3)}/6}{\{\sum_{j=1}^{N} n_j\lambda_j\beta_j^{(2)}/2\}^2}]. \tag{A.3.3}$$

APPENDIX A.4

In this appendix we derive a first order approximation for the root ϵ, cf. (4.28), of

$$\sum_{j=1}^{N} n_j\lambda_j\beta_j\gamma_j\delta_j(\epsilon)\frac{1-\beta_j(\epsilon\gamma_j\delta_j(\epsilon))}{\epsilon\beta_j\gamma_j\delta_j(\epsilon)} = 1, \tag{A.4.1}$$

with $\delta_j(\rho)$ defined by (4.27).

We start from, cf. (4.7),

$$r + \lambda_j\{1-\beta_j(\rho\gamma_j + r)\} = 0 \quad \text{for} \quad \rho \to 0. \tag{A.4.2}$$

Hence

$$\frac{\mathrm{d}r}{\mathrm{d}\rho} = \frac{\lambda_j\gamma_j\beta_j^{(1)}(\rho\gamma_j + r)}{1-\lambda-\gamma_j\{\beta_j^{(1)}(\rho\gamma_j + r)\}},$$

$$\frac{\mathrm{d}^2r}{\mathrm{d}\rho^2} = \frac{1}{[1-\lambda_j\{\beta_j^{(1)}(\rho\gamma_j + r)\}]}\lambda_j\gamma_j[\gamma_j + \frac{\mathrm{d}r}{\mathrm{d}\rho}]\beta_j^{(2)}(\rho\gamma_j + r)$$

$$+\frac{\lambda_j\gamma_j\beta_j^{(1)}(\rho\gamma_j + r))}{[1-\lambda_j\{\beta_j^{(1)}(\rho\gamma_j + r)\}]^2}\lambda_j[\gamma_j + \frac{\mathrm{d}r}{\mathrm{d}\rho}]\beta_j^{(2)}(\rho\gamma_j + r)),$$

and so

$$r(0) = 0 \quad , \quad \frac{\mathrm{d}r}{\mathrm{d}\rho}|_{\rho=0} = \frac{-\lambda_j\beta_j\gamma_j}{1+\lambda_j\beta_j},$$

$$\frac{\mathrm{d}^2r}{\mathrm{d}\rho^2}|_{\rho=0} = \frac{\lambda_j\gamma_j^2\beta_j^{(2)}}{[1+\lambda_j\beta_j]^3},$$

so that

$$r(\rho) = -\frac{\lambda_j\beta_j\gamma_j}{1+\lambda_j\beta_j}\rho + \frac{\lambda_j\gamma_j^2\beta_j^{(2)}/2}{[1+\lambda_j\beta_j]^3}\rho^2 + o(\rho^3)), \tag{A.4.3}$$

$$\delta_j(\rho) = 1 + \frac{r(\rho)}{\rho\gamma_j} = \frac{1}{1+\lambda_j\beta_j} + \frac{\lambda_j\gamma_j\beta_j^{(2)}/2}{(1+\lambda_j\beta_j)^3}\rho + o(\rho).$$

Further

$$\frac{1-\beta_j(\epsilon\gamma_j\delta_j(\epsilon))}{\epsilon\beta_j\gamma_j\delta_j(\epsilon)} = 1-\epsilon\gamma_j\delta_j(\epsilon)\frac{\beta_j^{(2)}}{2\beta_j} + \epsilon^2\gamma_j^2\delta_j^2(\epsilon)\frac{\beta_j^{(3)}}{6\beta_j} + o(\epsilon^2), \tag{A.4.4}$$

$$\delta_j(\epsilon)\frac{1-\beta_j(\epsilon\gamma_j\delta_j(\epsilon))}{\epsilon\beta_j\gamma_j\delta_j(\epsilon)} = \delta_j(\epsilon) - \epsilon\gamma_j\delta_j^2(\epsilon)\frac{\beta_j^{(2)}}{2\beta_j} + \epsilon^2\gamma_j^2\delta_j^3(\epsilon)\frac{\beta_j^{(3)}}{6\beta_j} + o(\epsilon^2)$$

$$= \frac{1}{1+\lambda_j\beta_j} - \frac{\epsilon\gamma_j}{[1+\lambda_j\beta_j]^3}\frac{\beta_j^{(2)}}{2\beta_j} + o(\epsilon^2).$$

Hence the first order approximation of ϵ reads:

$$-\epsilon = \frac{1-\sum_{j=1}^{N}\frac{n_j\lambda_j\beta_j\gamma_j}{1+\lambda_j\beta_j}}{\sum_{j=1}^{N}\frac{n_j\lambda_j\gamma_j^2}{(1+\lambda_j\beta_j)^3}\frac{\beta_j^{(2)}}{2}}. \tag{A.4.5}$$

References

1. Kelly, F. (1991). Effective bandwidth at multi-class queues, Queueing Systems **9**, p. 5-16.
2. Cohen, J.W., The Single Server Queue, North-Holland Publ. Co, Amsterdam 1982, revised edition.
3. Gibbens, R.J. and Hunt, P.J. (1991). Effective bandwidths for the multi-type UAS channel, Queueing Systems **9**, p. 17-28.
4. Cohen, J.W. (1974). Superimposed renewal processes and storage with gradual input, Stoch. Proc. Appl. **2**, p. 31-58.
5. Kosten, L., Stochastic theory of data handling systems with groups of multiple sources, Proc. 2nd Int. Symp. Performance Comp. Comm. Systems, eds. H. Rudin, W. Bux, North-Holland, Amsterdam, 1984, p. 321-331.
6. Anick, D., Mitra, D., Sondhi, M.M. (1982). Stochastic theory of a data-handling system with multiple sources, Bell Syst. Techn. J. **61**, 1871-1894.
7. Cohen, J.W. (1976). On the optimum switching level for an $M/G/1$ queueing system, Stoch. Proc. Appl. **4**, p. 297-316.
8. Kino, I. & Miyazawa, M. (1993). The stationary work in system of $GI/G/1$ gradual input queue, J. Appl. Prob. **30**, p. 207-222.