Towards Efficient Decision Rules for Admission Control Based on the Many Sources Asymptotics

Gergely Seres[1], Árpád Szlávik[1], János Zátonyi[2], and József Bíró[2]

[1] Traffic Lab, Ericsson Research Hungary, P.O. Box 107, H-1300 Budapest, Hungary, Gergely.Seres@eth.ericsson.se

[2] HSN Lab, DTT, Budapest University of Technology and Economics, P.O. Box 91, H-1521 Budapest, Hungary

Abstract. This paper introduces new admission criteria that enable the use of algorithms based on the many sources asymptotics in real-life applications. This is achieved by a significant reduction in the computational requirements and by moving the computationally intensive tasks away from the timing-sensitive decision instant. It is shown that the traditional overflow-probability type admission control method can be reformulated into a bandwidth-requirement type and a buffer-requirement type methods and that these methods are equivalent when used for admission control. The original and the two proposed methods are compared through the example of fractional Brownian motion traffic.

1 Introduction

Bandwidth requirement estimation is a key function in networks intending to provide quality of service (QoS) to their users. Network devices in QoS-capable networks must be able to control the amount of traffic they handle. This is generally performed by using some form of admission control. There are two commonly used methods for determining whether a new connection can be allowed to enter the system: in the first one an estimate of the buffer overflow probability is computed based on the properties of the new and the already active flows in the system, while the second method computes the bandwidth requirement of the existing traffic flows. When using the first method for admission control decisions, the devices check the computed overflow probability against the target overflow probability. If the second method is used, the bandwidth requirement of the existing flows is increased by the predicted bandwidth usage of the new flow and the result is compared to the capacity of the system.

Often, the second method is preferred over the first, mainly because it results in a quantity – the bandwidth requirement – that is more tractable and more useful than the estimate of the overflow probability. The on-line estimation of the bandwidth requirement of the traffic enables the network operator to track the amount of allocated (and free) capacity in the network. Furthermore, the impact of network management actions (e.g. directing more traffic on the link) on the resource status of the network can be more easily assessed. The overflow probability on the other hand is a less straightforward quantity that depends on the

E. Gregori et al. (Eds.): NETWORKING 2002, LNCS 2345, pp. 166–177, 2002.

parameters of the queueing system in a more complex way, thus changes in them imply a less tractable and computationally more complex update procedure.

Accordingly, most of the work to date has focused on algorithms that quantify the bandwidth requirement of traffic flows. The most widespread approaches are based on the notion of the effective bandwidth, a comprehensive review of which is given in [5]. A group of algorithms use the Chernoff bound or the Hoeffding bound to derive simplified and directly applicable formulae for the effective bandwidth in case of bufferless statistical multiplexing [9]. For buffered resources, the theory of large deviations was shown to be a very capable method for calculating the bandwidth requirement of traffic flows. There are two asymptotics that can be used for this purpose: the large buffer asymptotics and the many sources asymptotics. The large buffer asymptotics provide a rate function describing the decay rate of the tail of the probability of buffer overflow when the size of the buffer gets very large. The many sources asymptotics also offer a rate function but with the assumption that the number of traffic flows in the system gets very large, while the traffic mix, per-source buffer space and system per-source capacity are held constant. Both asymptotics discussed so far provide an overflow-probability type quantity.

Using the large buffer asymptotics it is easy to switch from the overflow probability representation to the bandwidth requirement representation. However, algorithms relying on this asymptotics [6] do not account for the gain arising from the statistical multiplexing of many traffic flows. In recent years, the second asymptotic regime, the many sources asymptotics (and its Bahadur-Rao improvement) have been described and investigated in [3], [2], [1] and [7]. In the native form, the many sources asymptotics provide a rate function that can be used to estimate the probability of overflow. The computation of this rate function involves two optimisations in two variables. Yet, if it is the bandwidth requirement that is of interest, another optimisation has to be performed that requires the recomputation of the two original optimisations in each step. Despite of its complexity, this bandwidth requirement estimate is appealing because it incorporates the statistical properties of the traffic along with its QoS requirements and it also embraces the statistical multiplexing gain that occurs on the multiplexing link. However, the use of this estimator in real-time applications is not feasible because of its computational complexity.

This paper introduces a new method for computing the bandwidth requirement of traffic flows that is based on the many sources asymptotics as well. Instead of the three embedded optimisations that previous approaches required, it comprises only of two optimisations that directly result in an estimate of the bandwidth requirement. The method is favourable to on-line measurement-based application, since the admission decision step is simplified and the more involving computations can be done in the background. It is shown that the new and the old methods for obtaining the bandwidth requirement are equivalent.

The rest of this paper is organised as follows. Section 2 presents a brief overview of the many sources asymptotics and describes three admission decision methods based directly on this asymptotics. The proposed computationally more favourable method for computing the bandwidth requirement is introduced in

Sect. 3, and the equivalence is proven. The operation of the novel method is demonstrated with the example of fractional Brownian motion traffic in Sect. 4. Conclusions are given in Sect. 5. The Appendix shows that similar results can be achieved for computing the buffer requirement of traffic flows.

2 Overflow-Probability Based Admission Criteria

This section presents an overview on the many sources asymptotics. Next, a collection of admission control methods are reviewed, all of which build on the asymptotic property of the overflow probability.

2.1 Many Sources Asymptotics

The asymptotic regime described by the many sources asymptotics can be used to form an estimate of the probability of buffer overflow in the system as follows. Let us consider a buffered communication link with transmission capacity C, buffer size B, which carries N independent flows multiplexed in the system. N is viewed as a scaling factor, i.e. we can identify a per-source transmission capacity $c = C/N$ and a per-source buffer size $b = B/N$. Further, let the stochastic process $X[0,t)$ denote the total amount of work arriving at the system during the time interval $[0,t)$. Let us assume that $X[0,t)$ has stationary increments.

Conclusions on the behaviour of this system can be derived by investigating a queueing system of infinite buffer size that is served by a finite capacity server with service rate $C = cN$. In order to account for the finite buffer size $B = bN$ of the real system, the probability of buffer overflow in the original system can be deduced from the proportion of time over which the queue length, $Q(C,N)$, is above the finite level B. In this system, where the system parameters (cN, bN) and the workload $(X[0,t))$ are scaled by the number of sources, an asymptotic equality can be obtained in N for the probability of overflow:

$$\lim_{N\to\infty} \frac{1}{N} \log P\{Q(cN,N) > bN\} = \sup_{t>0} \inf_{s>0} \left\{ st\, \frac{\alpha(s,t)}{N} - s(b+ct) \right\} \stackrel{\text{def}}{=} -I \quad . \tag{1}$$

Here $\alpha(s,t)$ (the so-called effective bandwidth [5]) is defined as

$$\alpha(s,t) \stackrel{\text{def}}{=} \frac{1}{st} \log E\left[e^{sX[0,t)} \right] \tag{2}$$

and I is called the asymptotic rate function, which depends on the per-source system parameters and on the scaled workload process. This result was proven for discrete time in [2] and for continuous time in [3]. Equation (2) practically means that for N large, the probability of overflow can be approximated as $P\{Q(C,N) > B\} \approx e^{-NI}$, where $-NI$ can be computed from (1) as

$$-NI = \sup_{t>0} \inf_{s>0} \{ st\, \alpha(s,t) - s(B+Ct) \} \quad . \tag{3}$$

The approximation above can also be reasoned in a less rigorous, but brief and intuitive manner as follows [7]. The Chernoff bound can be used to approximate the probability that the workload $X[0,t)$ exceeds Ct, the offered service in $[0,t)$ and in addition it fills up the buffer space B: $P\{X[0,t) > B + Ct\} \approx \inf_{s>0} \exp\{st\,\alpha(s,t) - s(B+Ct)\}$. The steady state queue length distribution can be described by $Q = \sup_{t>0}\{X[0,t) - Ct\}$ provided that the $X[0,t)$ process has stationary increments. This way, the probability of the queue length exceeding the buffer level B is $P\{Q > B\} \approx P\{\sup_{t>0}\{X[0,t) - Ct\} > B\} \approx \sup_{t>0} P\{X[0,t) > B + Ct\} \approx e^{-NI}$.

For the sake of simplifying further discussions, let us define the function $J(s,t) \stackrel{\text{def}}{=} st\,\alpha(s,t) - s(B+Ct)$. In (3), the evaluation of $\sup_{t>0}\inf_{s>0} J(s,t)$ is computationally complex as a double optimisation has to be performed after the computation or estimation of the effective bandwidth of $X[0,t)$. Since the optimisations are embedded, first the optimal (minimal) s has to be found which still depends on t. Placing this optimal s into $J(s,t)$, the task is its maximisation with respect to t. For a more formal and concise discussion the following notation is introduced:

$$s^*(t) \stackrel{\text{def}}{=} \arg\inf_{s>0} J(s,t), \quad t^* \stackrel{\text{def}}{=} \arg\sup_{t>0} J(s^*(t),t) \quad \text{and} \quad s^* \stackrel{\text{def}}{=} s^*(t^*)\ . \tag{4}$$

Now, the extremising pair of $J(s,t)$ is (s^*,t^*) and thus $-NI = J(s^*,t^*)$. The extremising values t^* and s^* are commonly termed as the critical time and space scales, respectively. The intuitive explanation of the critical time scale is that it is the most probable time interval after which overflows occur in the multiplexing system (i.e. the most likely length of the busy period prior to overflow). Although many other busy periods may contribute to the total overflow, large deviation theory takes into account only the most probable one, which is the most dominant in the asymptotic sense. The rationale behind the critical space parameter is that it captures the statistical behaviour of the workload process, that is the amount of achievable statistical multiplexing gain and the burstiness. Critical space values close to 0 describe a source (or an aggregate) that can benefit from statistical multiplexing, while larger values infer a higher bandwidth requirement. Finally, it is also worth noting that s^* and t^* always depend on the system parameters C, B and the statistical properties of $X[0,t)$.

In practical applications there is a QoS requirement, which is often specified as a constraint for the probability of buffer overflow ($e^{-\gamma}$). In order to admit a source the following criterion has to be satisfied:

$$P\{Q(C,N) > B\} \approx e^{-NI} \le e^{-\gamma} \quad \text{or} \quad \sup_{t>0}\inf_{s>0} J(s,t) \le -\gamma\ . \tag{5}$$

2.2 Equivalent Admission Criteria

The inequalities in (5) define an admission rule that uses the method of the many sources asymptotics in the native form. In this original form, the probability of buffer overflow is estimated using $X[0,t)$, B and C as the input quantities, whilst the target overflow probability is used as the performance criterion.

It is possible to set up two other criteria that can be used for admission control decisions. As it was mentioned in the introduction, it is often preferable to express the bandwidth requirement of the traffic and compare this quantity to the server capacity. In order to form an estimate of the bandwidth requirement of the traffic, another optimisation has to be performed. For this, the server capacity has to be treated as a free variable and given the workload process, the buffer size and the QoS requirement, the smallest server capacity has to be identified for which the system still satisfies the performance criterion put forward in (5). The resulting quantity

$$C_{\mathrm{equ}} \stackrel{\mathrm{def}}{=} \inf\left\{C : \sup_{t>0}\inf_{s>0} J(s,t) \le -\gamma\right\} \tag{6}$$

is termed in the rest of the paper as the equivalent capacity.[1] Then the admission criterion can be written as

$$C_{\mathrm{equ}} \le C \ . \tag{7}$$

A similar, but less frequently used criterion can be defined that allows admission decisions to be made based on the available buffer space. In this case, the buffer requirement of the traffic is determined using a similar triple optimisation as in (6), but this time taking $X[0,t)$, C and the QoS requirement as the input quantities and B as the performance constraint:

$$B_{\mathrm{req}} \stackrel{\mathrm{def}}{=} \inf\left\{B : \sup_{t>0}\inf_{s>0} J(s,t) \le -\gamma\right\} \quad \text{and} \quad B_{\mathrm{req}} \le B \ . \tag{8}$$

Figure 1 presents a summary of the three methods with respect to their input parameters and the quantity they use as a constraint in the decision criterion. The methods are equivalent in the sense that in a given context they arrive at the same decision. When it comes to numerical evaluation, the first (original) method with the double optimisation is, however, significantly less demanding than the others involving three embedded optimisations.

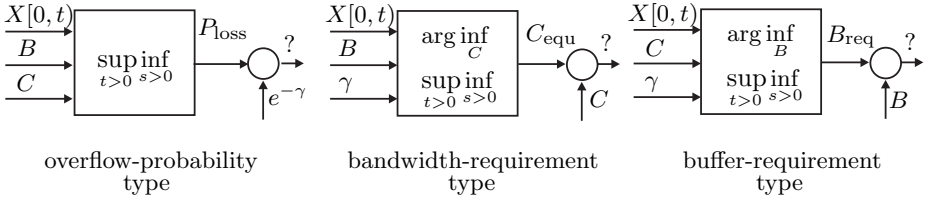

Fig. 1. Admission decision methods

[1] Following the terminology of previous works, the term effective bandwidth is reserved for $\alpha(s,t)$, which is not directly associated with the minimal service rate required to meet the QoS target.

3 The Improved Bandwidth Requirement Estimator

This section introduces an alternative method for computing the equivalent capacity. The advantage of this new method is that its computational complexity is reduced to a double optimisation, resulting in a similar formula to the one used in the rate-function based estimation of the buffer overflow probability. It is shown that the estimation of the equivalent capacity using the proposed method arrives at the same decision as the method in (6) and (7). The proposed method infer a new optimisation function resulting in an alternative set of space and time scales. The equivalence of the respective methods for estimating the buffer requirement can be proven in an identical manner (see the Appendix).

3.1 Alternative Definition of the Equivalent Capacity

Let us introduce $K(s,t)$ as

$$K(s,t) \stackrel{\text{def}}{=} \alpha(s,t) + \frac{\gamma}{st} - \frac{B}{t} \ , \tag{9}$$

which is obtained from the isolation of C from $J(s,t) = -\gamma$. Namely, $K(s,t) = C$ holds after the rearrangement.[2] By defining a new double optimisation

$$\widetilde{C}_{\text{equ}} \stackrel{\text{def}}{=} \sup_{t>0} \inf_{s>0} K(s,t) \ , \tag{10}$$

similarly to (3), the extremisers are attained in the form of

$$s^\dagger(t) \stackrel{\text{def}}{=} \arg\inf_{s>0} K(s,t), \quad t^\dagger \stackrel{\text{def}}{=} \arg\sup_{t>0} K(s^\dagger(t),t) \quad \text{and} \quad s^\dagger \stackrel{\text{def}}{=} s^\dagger(t^\dagger) \ . \tag{11}$$

The extremising pair of the double optimisation in (10) is then $(s^\dagger, t^\dagger)$ and these are the alternative space and time scales, respectively.

It can be proven that $\widetilde{C}_{\text{equ}} = C_{\text{equ}}$ holds. In other words, we need only two optimisations instead of three to arrive at the equivalent capacity C_{equ}. This is shown in the next subsection using the subsequent theorem.

Theorem 1. *The following two strict inequalities are equivalent:*

$$J(s^*,t^*) < -\gamma \iff K(s^\dagger,t^\dagger) < C \ , \tag{12}$$

furthermore the equations

$$J(s^*,t^*) = -\gamma \ , \tag{13}$$

$$K(s^\dagger,t^\dagger) = C \tag{14}$$

[2] The so-called Bahadur-Rao improvement, as described in [1], introduces a prefactor to the estimate of the overflow probability in (5). For the improvement of the equivalent capacity (and the buffer requirement) this manifests in a modified QoS constraint for (5) in the following form: $\gamma' = \gamma - \frac{\frac{1}{2}\log(4\pi\gamma)}{1+\frac{1}{2\gamma}}$.

are equivalent as well and consequently the two strict inequalities below also imply each other:

$$J(s^*, t^*) > -\gamma \ , \tag{15}$$
$$K(s^\dagger, t^\dagger) > C \ . \tag{16}$$

Proof. First of all, note that any two of the three equivalences, (12), (13)⇔(14) and (15)⇔(16), imply the third one, therefore it is enough to prove the second and the third assertion only. The proof of the statements of Theorem 1 uses three lemmas.

Lemma 1. *For all $t > 0$, (17) ⇔ (18):*

$$J(s^*(t), t) < -\gamma \ , \tag{17}$$
$$K(s^\dagger(t), t) < C \ . \tag{18}$$

Proof. Suppose (17) holds. The isolation of C gives $K(s^*(t), t) \overset{\text{by rearr. (17)}}{<} C$. Then, the definition of $s^\dagger(t)$ can be used to obtain $K(s^\dagger(t), t) \overset{\text{by def. of } s^\dagger(t)}{\leq} K(s^*(t), t)$. These two inequalities together entail that (18) holds as well. In the other direction, if (18) holds, $J(s^*(t), t) \overset{\text{by def. of } s^*(t)}{\leq} J(s^\dagger(t), t) \overset{\text{by rearr. (18)}}{<} -\gamma$, consequently (18)⇒(17) as well and thus Lemma 1 is proven. □

Lemma 2. *For all $t > 0$, (19) ⇔ (20):*

$$J(s^*(t), t) = -\gamma \ , \tag{19}$$
$$K(s^\dagger(t), t) = C \ . \tag{20}$$

Proof. If (19) holds, then $K(s^\dagger(t), t) \geq C$ because otherwise (17) should hold by the assertion of Lemma 1 contradicting (19). This means that $C \overset{\text{by L1 and (19)}}{\leq} K(s^\dagger(t), t) \overset{\text{by def. of } s^\dagger(t)}{\leq} K(s^*(t), t) \overset{\text{by rearr. (19)}}{=} C$, therefore equality has to hold throughout this chain of inequalities, consequently (20) holds. Now suppose that (20) holds. Then $-\gamma \overset{\text{by L1 and (20)}}{\leq} J(s^*(t), t) \overset{\text{by def. of } s^*(t)}{\leq} J(s^\dagger(t), t) \overset{\text{by rearr. (20)}}{=} -\gamma$, so (20)⇒(19) as well and hence Lemma 2 is proven. □

Lemma 3. *For all $t > 0$*

$$J(s^*(t), t) > -\gamma \Longleftrightarrow K(s^\dagger(t), t) > C \ . \tag{21}$$

Proof. Lemma 3 is a straightforward consequence of Lemma 1 and Lemma 2. □

Continuing with the proof of Theorem 1, let us suppose that (13) holds. Substituting t^* in Lemma 2 implies that $K(s^\dagger(t^*), t^*) = C$. On the other hand, $K(s^\dagger, t^\dagger) = K(s^\dagger(t^\dagger), t^\dagger) \geq K(s^\dagger(t^*), t^*)$ by the definition of $t^\dagger$ (note that in general $s^\dagger$ can not be used instead of $s^\dagger(t^*)$). Thus $K(s^\dagger(t^\dagger), t^\dagger) \overset{\text{by def. of } t^\dagger}{\geq}$

$K(s^\dagger(t^*),t^*) \overset{\text{by L2 with } t=t^*}{=} C$. By Lemma 2 and Lemma 3 with $t = t^\dagger$ this inequality implies that $J(s^*(t^\dagger),t^\dagger) \geq -\gamma$. Hence $-\gamma \overset{(13)}{=} J(s^*(t^*),t^*) \overset{\text{by def. of } t^*}{\geq} J(s^*(t^\dagger),t^\dagger) \overset{\text{prev., L2 and L3 with } t=t^\dagger}{\geq} -\gamma$, from which it can be concluded that equality holds along the chain of inequalities, accordingly $J(s^*(t^\dagger),t^\dagger) = -\gamma$. Lemma 2 with $t = t^\dagger$ then implies that (14) holds. Vice versa, if (14) holds, then $J(s^*,t^*) = J(s^*(t^*),t^*) \overset{\text{by def. of } t^*}{\geq} J(s^*(t^\dagger),t^\dagger) \overset{\text{by L2 with } t=t^\dagger}{=} -\gamma$, thereupon $C \overset{(14)}{=} K(s^\dagger(t^\dagger),t^\dagger) \overset{\text{by def. of } t^\dagger}{\geq} K(s^\dagger(t^*),t^*) \overset{\text{prev., L2 and L3 with } t=t^*}{\geq} C$. The consequence is that $K(s^\dagger(t^*),t^*) = C$ holds and by Lemma 2 with $t = t^*$ it turns out that (13) holds as well. Thus the proof of equivalence (13)⇔(14) is done. For the proof of the third statement of Theorem 1, first suppose (15) holds. Then $K(s^\dagger,t^\dagger) = K(s^\dagger(t^\dagger),t^\dagger) \overset{\text{by def. of } t^\dagger}{\geq} K(s^\dagger(t^*),t^*) \overset{\text{by L3 with } t=t^*}{>} C$, consequently (15)⇒(16). Finally, supposing (16) holds, $J(s^*,t^*) = J(s^*(t^*),t^*) \overset{\text{by def. of } t^*}{\geq} J(s^*(t^\dagger),t^\dagger) \overset{\text{by L3 with } t=t^\dagger}{>} -\gamma$, subsequently (16)⇒(15) as well. Therefore the equivalence (15)⇔(16) is proven as well. Equivalence (12) follows from the other two equivalences, hence the proof of Theorem 1 is completed. □

3.2 Equivalence of the Two Definitions of the Equivalent Capacity

Theorem 1 can now be used to prove that the equivalent capacities defined by (6) and (10) are equal.

Corollary 1. *The equivalent capacity defined by the double optimisation in (10) equals the one defined by the triple optimisation in (6):* $K(s^\dagger,t^\dagger) = \widetilde{C}_{\text{equ}} = C_{\text{equ}}$.

Proof. Observe that $J(s,t)$ is (strictly) monotonously decreasing in the variable C for fixed B and γ. It is easy to see that $C_{\text{equ}} = \inf\{C : \sup_{t>0} \inf_{s>0} J(s,t) \leq -\gamma\} \overset{\text{str. mon. decr.}}{=} \inf\{C : \sup_{t>0} \inf_{s>0} J(s,t) = -\gamma\}$. The consequence of this is that $J((s^*(C_{\text{equ}}),t^*(C_{\text{equ}})) = -\gamma$ (note that the extremisers depend on the variables C and B in this case, but B is fixed here, therefore only the dependence on variable C is indicated). By Theorem 1 it follows that $K(s^\dagger,t^\dagger) = C_{\text{equ}}$ as well (here the optimising parameters depend on variables B and γ, but those are fixed). However, this is exactly the definition of $\widetilde{C}_{\text{equ}}$ and that corresponds to the assertion. □

The respective optimiser pairs $(s^*(B,C),t^*(B,C))$ and $(s^\dagger(B,\gamma),t^\dagger(B,\gamma))$ do not coincide in general, they are not even comparable as such, since they depend on a different set of variables. Nevertheless, on the boundary of the acceptance region $(J(s^*,t^*) = -\gamma \Leftrightarrow (C_{\text{equ}} =)K(s^\dagger,t^\dagger) = C)$ the same parameter values are the optimisers of the two problems:

Proposition 1. *If one of the double optimisations (3) and (10) has a unique extremising pair and* $J(s^*,t^*) = -\gamma$ *(or* $K(s^\dagger,t^\dagger) = C$*), then the two extremiser pairs coincide,* $t^* = t^\dagger$ *and* $s^* = s^\dagger$.

Proof. Assume $(s^\dagger, t^\dagger)$ is unique. By supposing any of the two (equivalent) equalities as the second assumption, $J(s^*(t^*), t^*) = -\gamma$ and $K(s^\dagger(t^\dagger), t^\dagger) = C$ hold. By Lemma 2 with $t = t^*$ this means that $K(s^\dagger(t^*), t^*) = C$ holds as well. Since $K(s^\dagger(t^\dagger), t^\dagger) = C$, then $t^* = t^\dagger$ by the uniqueness property. On the other hand, by rearranging $K(s^\dagger(t^*), t^*) = C$, $J(s^\dagger(t^*), t^*) \overset{\text{rearr.}}{=} -\gamma = J(s^*(t^*), t^*)$ is obtained. This proves $s^* = s^*(t^*) \overset{\text{uniq.}}{=} s^\dagger(t^*) \overset{t^*=t^\dagger}{=} s^\dagger$. The proof of the statement is almost the same when assuming the uniqueness of (s^*, t^*). □

4 Comparison of the Methods for fBm Traffic

This section presents a comparison of the three admission control methods discussed in Sect. 2.2 using the new formulae developed in Sect. 3.1 and in the Appendix. The traffic case used is fractional Brownian motion (fBm), which involves closed-form formulae due to its Gaussian nature.

4.1 Key Formulae for Fractional Brownian Motion Traffic

The stochastic process $\{Z_t, t \in \mathbb{R}\}$ is called normalized fractional Brownian motion with self-similarity (Hurst-) parameter $H \in (0, 1)$ if it has stationary increments and continuous paths, $Z_0 = 0$, $E[Z_t] = 0$, $Var[Z_t] = |t|^{2H}$ and if Z_t is a Gaussian process. Let us define the process $X[0, t) \overset{\text{def}}{=} mt + Z_t$, for $t > 0$. It is known as fractional Brownian traffic and can be interpreted as the amount of traffic offered to the multiplexer in the time interval $[0, t)$. This is a so-called self-similar model, which has been suggested for the description of Internet traffic aggregates [8], [4].

Using this model the effective bandwidth (2) can be written as $\alpha(s, t) = m + \frac{s\sigma^2 t^{2H-1}}{2}$ and accordingly $J(s, t) = st\, m + \frac{s^2\sigma^2 t^{2H}}{2} - s(B + Ct)$. The extremisers for $J(s, t)$ and $-NI$ can be found in Table 1[3], where $\kappa(H) \overset{\text{def}}{=} H^H(1 - H)^{1-H}$.

The equivalent capacity can be evaluated in two ways, either using the definition in (6) or the method proposed in this paper (10). $\widetilde{C}_{\text{equ}}$ requires the direct evaluation of $K(s, t)$ (9) at the alternative critical space and time scales $(s^\dagger, t^\dagger)$ (11), i.e. "only" a double optimisation is necessary. For fBm traffic $K(s, t) = m + \frac{1}{2}\sigma^2 s t^{2H-1} + \frac{\gamma}{st} - \frac{B}{t}$, its extremisers and $\widetilde{C}_{\text{equ}}$ are listed in Table 1. If C_{equ} is calculated in the conventional way (6), the third optimisation (with respect to C) can be exchanged for solving $-NI = -\gamma$ for $C = C_{\text{equ}}$ (as seen in the proof of Corollary 1).[4] It can be checked that $C_{\text{equ}} = \widetilde{C}_{\text{equ}}$ as expected (using the definition of $\kappa(H)$). In a similar way, s', t' and $\widetilde{B}_{\text{req}}$ (see the Appendix) can be computed (see Table 1) and it also turns out that $\widetilde{B}_{\text{req}} = B_{\text{req}}$.

[3] An identical expression for the approximation of the overflow probability was obtained in [8] with a different approach.

[4] This simplification can be done only because $-NI$ is an explicit function of C in the fBm case (C can be isolated from the equation). In most other cases the third optimisation must be done through several double optimisations of $J(s, t)$ for different values of C in order to locate $C = C_{\text{equ}}$ for which $-NI = -\gamma$.

Table 1. Comparison of the three admission control methods for fBm traffic

	$f(s,t)$		
	$J(s,t)$	$K(s,t)$	$L(s,t)$
$s^{\text{opt}}(t) = \arg\inf_{s>0} f(s,t)$	$s^*(t) = \frac{t^{-2H}(B+(C-m)t)}{\sigma^2}$	$s^\dagger(t) = \frac{\sqrt{2\gamma}t^{-H}}{\sigma}$	$s'(t) = \frac{\sqrt{2\gamma}t^{-H}}{\sigma}$
$t^{\text{opt}} = \arg\sup_{t>0} f(s^{\text{opt}}(t),t)$	$t^* = \frac{H}{1-H}\frac{B}{C-m}$	$t^\dagger = 2^{-\frac{1}{2H}}\left(\frac{B}{(1-H)\sqrt{\gamma}\sigma}\right)^{\frac{1}{H}}$	$t' = \left(\frac{H\sqrt{2\gamma}\sigma}{C-m}\right)^{\frac{1}{1-H}}$
$s^{\text{opt}} = s^{\text{opt}}(t^{\text{opt}})$	$s^* = \frac{1-H}{\kappa(H)^2}\cdot\frac{(C-m)^{2H}B^{1-2H}}{\sigma^2}$	$s^\dagger = \frac{2(1-H)\gamma}{B}$	$s' = \left(\frac{C-m}{H}\right)^{\frac{H}{1-H}}\cdot\left(\sqrt{2\gamma}\sigma\right)^{\frac{1}{1-H}}2\gamma$
$\sup_{t>0}\inf_{s>0} f(s,t)$	$-NI = -\frac{(C-m)^{2H}B^{2-2H}}{2\kappa(H)^2\sigma^2}$	$\widetilde{C}_{\text{equ}} = m + H\cdot\left(2\gamma\sigma^2\right)^{\frac{1}{2H}}\left(\frac{1-H}{B}\right)^{\frac{1-H}{H}}$	$\widetilde{B}_{\text{req}} = \left(\frac{H}{C-m}\right)^{\frac{H}{1-H}}\cdot(1-H)\left(\sqrt{2\gamma}\sigma\right)^{\frac{1}{1-H}}$

Confirming the statements in the previous section, it is apparent from Table 1 that the critical space scales s^*, $s^\dagger$ and s' are usually different and depend on different parameter sets. For given B, C, γ, m, H and σ, the corresponding scales match only when equalities (13) and (14) hold. An interesting consequence of this fact is that the solution of $t^*(B,C,m,H) = t^\dagger(B,\gamma,\sigma,H)$ for C results in the equivalent capacity C_{equ} and its solution for γ is NI. Similar statements are valid for the space scales as well.

4.2 A Numerical Example

In this subsection a numerical example is presented to demonstrate the results of the previous subsection. Let us take the fBm model of one of the Bellcore Ethernet data traces [4]: $m_1 = 138\,135\,\text{byte/s}$, $\sigma_1 = 89\,668\,\text{byte/s}^H$, $H = 0.81$. Assume that $N = 100$ of such sources are multiplexed into a buffer. Hence, the model parameters of the fBm model for the aggregate traffic workload become: $m = 13.8135\,\text{Mbyte/s}$, $\sigma = 0.89668\,\text{Mbyte/s}^H$, $H = 0.81$. The buffer size is chosen to be $B = 5.3\,\text{Mbyte}$, the service rate is $C = 16\,\text{Mbyte/s}$ and let the constraint for the overflow be $e^{-16} \approx 10^{-7}$ ($\gamma = 16$). For these system parameters, the extremiser pair is $(s^*, t^*) = (1.453, 7.091)$ and therefore $-NI = -20.26$. Clearly, $-NI < -\gamma$, i.e. the QoS requirement is fulfilled. The alternative critical scales and the equivalent capacity are obtained as $(s^\dagger, t^\dagger) = (1.147, 8.203) \neq (s^*, t^*)$ and $C_{\text{equ}} = \widetilde{C}_{\text{equ}} = 15.568\,\text{Mbyte/s}$, thus $C_{\text{equ}} < C$ holds and there is $0.432\,\text{Mbyte/s}$ of free service capacity.

5 Conclusion

This paper has introduced a new method for the computation of the equivalent capacity (and the buffer requirement) of traffic flows that is based on the many

sources asymptotics. In contrast to the method directly building on the asymptotic rate function, the new method involves only two embedded optimisations instead of three, thus it significantly reduces the computational complexity of the task. It has been shown that the two methods are equivalent.

The presented method of deriving the equivalent capacity leads to an alternative domain of time and space scales. In a given system the optimisation defininig the equivalent capacity estimate ($C_{\text{equ}} =) \widetilde{C}_{\text{equ}}$ (10) yields different optimal parameter values than that defining the estimate of the overflow probability e^{-NI} (3). Consequently, the substitution of the extremisers of $J(s,t)$ into $K(s,t)$ (9) does not lead to a correct estimate of the equivalent capacity. The only exception is the boundary of the admission region, where the two extremising pairs coincide.

In terms of applicability, it can be shown that the method of the equivalent capacity computation is more appropriate for real-time operation than those based on the asymptotic rate function, especially if the workload process is measured on-line (measurement-based admission control). Recall the admission methods defined by (5) and (6), (7). In practice these admission rules are performed at the arrival of a new flow. The effective bandwidth estimate has to be adjusted in order to take the new flow into account. For example, let us assume that the new flow is described by its peak rate only. Then $\alpha^+(s,t) = \alpha(s,t) + p$ is a conservative adjustment. With the rate-function based admission method, the double optimisation has to be re-evaluated in order to update the estimate of the overflow probability: $-NI^+ = \sup_{t>0} \inf_{s>0} \{st\,\alpha^+(s,t) - s(B+Ct)\}$. The decision criterion remains the same in this case.

Using the equivalent-capacity based admission criterion is more convenient. Here, the estimation of the equivalent capacity of the existing flows can be maintained in the background, i.e. the estimate of C_{equ} can be recomputed based on periodic measurements. At the arrival of a new flow, the $C_{\text{equ}}+p \leq C$ criterion has to be checked, which differs from (7) only in a correction term that is the peak rate of the new flow. Hence, the timing-sensitive operation (the admission decision) involves only a simple addition and a comparison, while the time-consuming double optimisation can be performed in the background, with more relaxed timing requirements.

The proposed method thus enables the deployment of the many sources asymptotics in practice not only through the reduction of its complexity, but through shifting the computations away from the critical decision instant.

References

[1] C. Courcoubetis, V. A. Siris, and G. D. Stamoulis. Application of the many sources asymptotic and effective bandwidths to traffic engineering. *Telecommunication Systems*, 12:167–191, 1999.

[2] C. Courcoubetis and R. Weber. Buffer over ow asymptotics for a buffer handling many traffic sources. *Journal of Applied Probability*, 33:886–903, 1996.

[3] N. G. Duffield. Economies of scale for long-range dependent traffic in short buffers. *Telecommunication Systems*, 7:267–280, 1997.

[4] R. J. Gibbens and Y. C. Teh. Critical time and space scales for statistical multiplexing in multiservice networks. In *Proceedings of International Teletraffic Congress (ITC)*, pages 87–96, Edinburgh, Scottland, 1999. ITC'16.
[5] F. P. Kelly. Notes on effective bandwidths. *Stochastic Networks: Theory and Applications*, 4:141–168, Oxford University Press, 1996.
[6] J. T. Lewis, R. Russell, F. Toomey, S. Crosby, I. Leslie, and B. McGurk. Statistical properties of a near-optimal measurement-based CAC algorithm. In *Proceedings of IEEE ATM*, pages 103–112, Lisbon, Portugal, June 1997.
[7] M. Montgomery and G. de Veciana. On the relevance of time scales in performance oriented traffic characterizations. In *Proceedings of the Conference on Computer Communications (IEEE INFOCOM)*, volume 2, pages 513–520, San Francisco, USA, March 1996.
[8] I. Norros. A storage model with self-similar input. *Queueing Systems*, 16(3/4):387–396, 1994.
[9] P. Tran-Gia and N. Vicari, editors. *Impacts of New Services on the Architecture and Performance of Broadband Networks - COST257 Final Report*. compuTEAM 2000, 2000.

Appendix: The Improved Buffer Requirement Estimator

Let us introduce $L(s,t) \stackrel{\text{def}}{=} t\,\alpha(s,t) + \frac{\gamma}{s} - Ct$, resulting from the isolation of the variable B from $J(s,t) = -\gamma$. That is, $L(s,t) = B$ holds after the rearrangement. Similarly to (4) and (11) the critical space and time scales of

$$\widetilde{B}_{\text{req}} \stackrel{\text{def}}{=} \sup_{t>0} \inf_{s>0} L(s,t) \tag{22}$$

are $s'(t) \stackrel{\text{def}}{=} \arg\inf_{s>0} L(s,t)$, $t' \stackrel{\text{def}}{=} \arg\sup_{t>0} K(s'(t),t)$ and $s' \stackrel{\text{def}}{=} s'(t')$. The extremiser pair of (22) is then (s',t'), like in the previous cases.

Analogously to the equivalent capacity, it is also true that the buffer requirement defined by the triple optimisation in (8) $\widetilde{B}_{\text{req}} = B_{\text{req}}$ holds. Consequently, only two optimisations are needed instead of three to determine the buffer requirement B_{req}, matching the case of the equivalent capacity.

The statements of Sect. 3.1 and Sect. 3.2 can now be reformulated.

Theorem 2. *(12) $\Leftrightarrow$ $L(s',t') < B$, (13) $\Leftrightarrow$ (14) $\Leftrightarrow$ $L(s',t') = B$ and (15) $\Leftrightarrow$ (16) $\Leftrightarrow$ $L(s',t') > B$.*

Lemma 4. *(17) $\Leftrightarrow$ (18) $\Leftrightarrow$ $L(s'(t),t) < B$, (19) $\Leftrightarrow$ (20) $\Leftrightarrow$ $L(s'(t),t) = B$, (21) $\Leftrightarrow$ $L(s'(t),t) > B$.*

Corollary 2. *The buffer requirement defined by the double optimisation in (22) equals the one defined by the triple optimisation in (8): $L(s',t') = \widetilde{B}_{\text{req}} = B_{\text{req}}$.*

Proposition 2. *If one of the double optimisations (3), (10) and (22) has a unique extremising pair and $J(s^*,t^*) = -\gamma$ (or $K(s^\dagger,t^\dagger) = C$ or $L(s',t') = B$), then the three extremiser pairs coincide, $t^* = t^\dagger = t'$ and $s^* = s^\dagger = s'$.*

Proof. The proofs follow the structure of those in Sect. 3.1 and Sect. 3.2. □