Centrum voor Wiskunde en Informatica

REPORTRAPPORT

PNA

Probability, Networks and Algorithms

PNA

Probability, Networks and Algorithms

Generalized processor sharing: characterization of the admissible region and selection of optimal weights

P.M.D. Lieshout, M.R.H. Mandjes

REPORT PNA-R0605 MARCH 2006

Centrum voor Wiskunde en Informatica (CWI) is the national research institute for Mathematics and Computer Science. It is sponsored by the Netherlands Organisation for Scientific Research (NWO).
CWI is a founding member of ERCIM, the European Research Consortium for Informatics and Mathematics.

CWI's research has a theme-oriented structure and is grouped into four clusters. Listed below are the names of the clusters and in parentheses their acronyms.

Probability, Networks and Algorithms (PNA)

Software Engineering (SEN)

Modelling, Analysis and Simulation (MAS)

Information Systems (INS)

P.O. Box 94079, 1090 GB Amsterdam (NL)
Kruislaan 413, 1098 SJ Amsterdam (NL)
Telephone +31 20 592 9333
Telefax +31 20 592 4199

ISSN 1386-3711

Generalized processor sharing: characterization of the admissible region and selection of optimal weights

ABSTRACT

We consider a two-class Generalized Processor Sharing (GPS) queueing system, in which each class has its specific traffic characteristics and Quality-of-Service (QoS) requirements. Traffic of both classes is assumed to be Gaussian (a versatile family of models that covers both long-range dependent and short-range dependent traffic). In this paper we address the question how to select the GPS weight values. To do so, we first characterize the admissible region of the system for fixed weights. Then we obtain the realizable region by taking the union of the admissible regions over all possible weight values. The results indicate that, under a broad variety of traffic characteristics and QoS requirements, nearly the entire realizable region can be obtained by strict priority scheduling disciplines. In addition, we indicate how the buffer thresholds, QoS requirements and the traffic characteristics of the two classes determine which class should get high priority.

2000 Mathematics Subject Classification: 60G15;60K25;68M20;90B18

Keywords and Phrases: Generalized processor sharing; Priority scheduling; Gaussian traffic; Loss probabilities; Admissible region; Weight setting

Note: This research has been funded by the Dutch BSIK/BRICKS (Basic Research in Informatics for Creating the Knowledge Society) project.

Generalized Processor Sharing: Characterization of the Admissible Region and Selection of Optimal Weights

P. Lieshout and M. Mandjes

CWI
P.O. Box 94079, 1090 GB Amsterdam, the Netherlands
Email: {lieshout, michel}@cwi.nl

3rd March 2006

Abstract

We consider a two-class Generalized Processor Sharing (GPS) queueing system, in which each class has its specific traffic characteristics and Quality-of-Service (QoS) requirements. Traffic of both classes is assumed to be Gaussian (a versatile family of models that covers both long-range dependent and short-range dependent traffic). In this paper we address the question how to select the GPS weight values. To do so, we first characterize the admissible region of the system for fixed weights. Then we obtain the realizable region by taking the union of the admissible regions over all possible weight values. The results indicate that, under a broad variety of traffic characteristics and QoS requirements, nearly the entire realizable region can be obtained by strict priority scheduling disciplines. In addition, we indicate how the buffer thresholds, QoS requirements and the traffic characteristics of the two classes determine which class should get high priority. [1]

[1] This research has been funded by the Dutch BSIK/BRICKS (Basic Research in Informatics for Creating the Knowledge Society) project. M. Mandjes is also affiliated with the Korteweg-de Vries Institute, University of Amsterdam, the Netherlands, and EURANDOM, Eindhoven, the Netherlands.

1 Introduction

Future communication networks are expected to support a wide range of heterogeneous services, including the 'traditional' data, video, and voice-applications, but in addition also more demanding multimedia applications, such as gaming, remote surgery, etc. Services may not only have different traffic characteristics, but may also have different Quality-of-Service (QoS) requirements, where QoS is usually expressed in terms of constraints on loss probabilities (buffer overflow) or delay. Thus, the integration of heterogeneous applications with different QoS requirements in the network raises the need for service differentiation. An obvious alternative to service differentiation could be to strive for the most stringent QoS requirement for all classes and to serve all traffic in a FIFO manner. This policy, however, inevitably leads to a waste of resources, as some classes get a better QoS than desired, and is therefore unattractive.

Generalized Processor Sharing (GPS) is a queueing discipline that is capable of supporting heterogeneous QoS-levels. The GPS discipline assigns weights to the traffic classes, and the link capacity is shared according to the weights of the backlogged classes. Hence, GPS provides some sort of isolation among competing classes, by guaranteeing a certain minimum rate to each backlogged class. Assigning all weight to a single class, implies that other classes can only be served if there is no traffic of this single class queued; i.e., priority queueing can be viewed as a special case of GPS.

Although the selection of GPS weights is, at least from an operational point of view, a key problem, most of the work on GPS describes the queueing performance of a GPS system for fixed weights. Parekh & Gallager [22, 23] derived deterministic worst-case delay guarantees for leaky-bucket controlled traffic. Subsequent papers focused on statistical performance guarantees, often based on asymptotic approximations. We briefly mention some results.

Yaron & Sidi [28] derived bounds for GPS queues fed by so-called exponentially-bounded burstiness traffic. Bertsimas *et al.* [3], Massoulié [20], and Zhang [29] established large-deviations results for light-tailed traffic (i.e., short-range dependent) sources. Large-buffer asymptotics for heavy-tailed traffic (i.e., long-range dependent) processes were obtained in Borst *et al.* [4, 5] and Kotopoulos *et al.* [13]. Van Uitert & Borst [26, 27] extended these results to networks of GPS queues. Borst *et al.* [6, 7] analyzed the buffer asymptotics in a two-class GPS system with a mixture of heavy-tailed and light-tailed traffic.

In practice most (real-time) applications do not tolerate large delays, hence the large buffer asymptotics are not always appropriate. It can be argued that in many situations the so-called *many-sources* asymptotic regime is more justified. Mannersalo & Norros [19] developed accurate approximations for the overflow probabilities in this regime. They considered a GPS system shared by two heterogeneous classes of Gaussian sources, with a relatively large number of sources in both classes. The obtained approximations were validated by extensive simulations. Mandjes & Van Uitert [17] further justified and refined these approximations, and established an interesting connection with tandem queues fed by Gaussian traffic, see also [18]. For the special case of Brownian inputs, Mandjes [16] showed the exactness of the resulting decay rates.

As mentioned, the inverse problem of mapping the QoS requirements on suitable GPS weights has received considerably less attention in literature. Dukkipati *et al.* [9] and Panagakis *et al.* [21] developed algorithms to allocate optimal weights to leaky-bucket constrained traffic with deterministic service guarantees, in the presence of best effort traffic, i.e., weights are chosen such that the throughput of the best effort class is maximized. Again for leaky-bucket regulated traffic, Elwalid & Mitra [10] first derived the admissible region for a two-class GPS system for fixed weights (i.e., all combinations of flows that satisfy the QoS for both classes), and then show that nearly the entire realizable region (i.e., the union of the admissible regions over all possible weight values) is obtained by selecting either one or two specific weights. Further results along these lines may be found in Kumaran *et al.* [14].

The results of Elwalid & Mitra [10] on the weight setting problem rely on the restrictive assumption of leaky-bucket controlled traffic. The contribution of this paper is that we extend the results on the weight setting to an extremely general and versatile class of input processes, covering a broad range of correlations, viz. the class of *Gaussian* inputs. Importantly, Gaussian models include both short-range and long-range dependent traffic. They arise as limiting processes of the superposition of a large number independent traffic sources, and are thus appropriate if the aggregation level is sufficiently large. Fraleigh *et al.* [11] empirically showed that a relatively low aggregation level is already sufficient for Gaussianity (average rates in the order of 50 Mbps suffice, and in many cases even considerably lower rates). A complicating issue is the fact that elastic traffic is controlled through feedback loops like TCP. Kilpi & Norros [12] however argued that (non-feedback) Gaussian traffic models are still justified as long as the level of aggregation is sufficiently large (both in time and number of flows).

In this paper we consider a two-class GPS system with Gaussian traffic sources. The QoS criterion is that the loss probability should be kept below some class-specific value. We focus on a two-class system, as the majority of the traffic can broadly be categorized into *streaming* and *elastic* traffic (see e.g. [24]), each one having its own QoS requirements, thus justifying our choice. The large-deviations approximations of Mannersalo & Norros [19] on GPS for Gaussian inputs are the key tool in our analysis. As a first step, we use these approximations to find the admissible region for class 1 for fixed weights, i.e., all numbers of sources n_1, n_2 of class 1 and class 2 such that the QoS requirement of class 1 is met. By taking the intersection of the admissible region of both classes, we then obtain the admissible region (of the system), i.e., all combinations of flows that satisfy the QoS for both classes. In the special case of Brownian inputs, we explicitly determine the boundary of the admissible region.

We also explicitly derive the realizable region as the union of the admissible regions over all possible weights values, in case of Brownian inputs. A remarkable finding is that nearly the entire realizable region is achieved by strict priority scheduling disciplines. A further key observation is that the QoS requirements and the buffer thresholds fully determine which class should have high priority if such a strict priority policy would be imposed. Importantly, the above two remarkable findings also hold for general Gaussian inputs. As we lack here an explicit description of the boundary of the realizable region, we have relied on extensive numerical experimentation.

An important purpose of GPS is to run the system at maximum efficiency, i.e., to realize an admissible region that is as large as possible. Each application can be guaranteed its required QoS. As a result GPS outperforms FIFO, in which each class is guaranteed the QoS of the most stringent class, leading to inefficient use of resources. The results above indicate that from an efficiency point of view, GPS does not outperform a simple priority discipline. In other words, it suggests that there is hardly any efficient enhancement due to implementing GPS (compared to priority), in that the admissible region corresponding to some GPS weight, is contained in the admissible region corresponding to one of the priority cases. A second purpose of GPS, however, could be still accomplished: the protection against starvation effects. Under priority scheduling, low priority traffic may be excluded from service over substantial time intervals, which can be prevented under GPS.

The remainder of this paper is organized as follows. In Section 2 we describe our two-class GPS model with Gaussian inputs, and we review the Mannersalo-Norros approximations [19] for loss probabilities, which consists of three regimes. In Section 3 the stable region is partitioned into three subsets, each subset corresponding to one of the three regimes. Using the partitioning of the stable region and the Mannersalo-Norros approximations, we derive the admissible region in Section 4. In Section 5 we consider Brownian inputs, and we explicitly derive the boundary of the admissible region and the boundary of the realizable region. In Section 6 we perform numerical analysis. In particular, we consider systems shared by two types of applications with heterogeneous QoS requirements, and we numerically derive the realizable regions. In Section 7 we make some concluding remarks.

2 Preliminaries

In this section we introduce the notation of the two-class GPS model and discuss Gaussian sources. Then we present approximations for the overflow probabilities.

2.1 GPS model

We consider a model with two queues and that share a server of rate C. Traffic of class i is buffered in queue i, $i = 1, 2$. The scheduling discipline is GPS, with weight $\phi_i \geq 0$ assigned to class i, $i = 1, 2$. Without loss of generality we assume that $\phi_1 + \phi_2 = 1$. The weight ϕ_i determines the guaranteed minimum rate for class i. If a class does not fully use the minimum rate, then the excess capacity becomes available to the other class.

A formal description of GPS is given by Parekh & Gallager [22]. Let $B_i(s,t)$ denote the amount of traffic of class i served in the time interval $[s,t]$. If the queue of class i is non-empty (backlogged) in the corresponding interval, then GPS satisfies the following property:

$$\frac{B_i(s,t)}{B_j(s,t)} \geq \frac{\phi_i}{\phi_j}, \quad i = 1, 2, \quad j \neq i. \tag{1}$$

Obviously, there is equality in (1) if class j is also continuously backlogged in the interval $[s,t]$.

Note that GPS is a work-conserving scheduling discipline. That is, the server always works at maximum speed if at least one of the queues is non-empty. GPS assumes that a server can serve different classes simultaneously and that the traffic is infinitely divisible, which is obviously not true in practice. However, the difference between a real-life implementation of packetized traffic (e.g., 'packet-by-packet GPS') and the 'theoretical GPS' is usually negligible (see [22] and [23]).

2.2 Gaussian input traffic, overflow probabilities

As our first goal in Sections 3-4 is to characterize the admissible region (for a given weight), we first present the Mannersalo-Norros approximations [19] for the overflow probabilities for given numbers of sources of both classes.

Let class 1 (class 2) consist of a superposition of n_1 (n_2) i.i.d. flows (or: sources), modeled as Gaussian processes with stationary increments. Clearly $n_1, n_2 \in \mathbb{N}_0$, but for convenience we let $n_1, n_2 \in \mathbb{R}_+$; as n_1, n_2 are relatively large the resulting error is typically small. We denote the mean traffic rate and variance function of a single class-i flow by $\mu_i > 0$ and $v_i(\cdot) : \mathbb{R}_+ \to \mathbb{R}_+$, respectively, for $i = 1, 2$; this mean rate and variance curve fully characterize the probabilistic behavior of the flow. Hence, if $A_i(s,t)$ denotes the amount of traffic generated by a single flow of type i in the interval $[s,t]$, then $\mathbb{E}A_i(s,t) = \mu_i \cdot (t-s)$ and $\mathbb{V}\mathrm{ar}A_i(s,t) = v_i(t-s)$; note that the assumption of stationary increments entails that the law of $A_i(s,t)$ only depends on the *length* of the interval $[s,t]$. To guarantee stability we assume that $n_1\mu_1 + n_2\mu_2 \leq C$ (which we refer to as the 'capacity constraint'). We impose the following (weak) assumptions on $v_i(\cdot)$.

Assumption 2.1 *For $i = 1, 2$,*
A1 $v_i(\cdot) \in C_2([0,\infty))$.
A2 $v_i(\cdot)$ is strictly increasing.
A3 For some $\alpha < 2$ it holds that $v_i(t)t^{-\alpha} \to 0$, as $t \to \infty$.

We need assumptions A1 and A3 in order for the results on the overflow probabilities to be valid. Assumption A2 is needed in the proofs of some lemmas.

As mentioned earlier, the derivation of the admissible regions relies on the Mannersalo-Norros approximations [19] for the overflow probabilities; these require assumptions A1 and A3. On the basis of extensive simulation experiments, Mannersalo & Norros [19] showed the accuracy of their approximations.

Let Q_i denote the stationary buffer content in the GPS model of class i, and $\triangle_i(n_1,n_2)$ the Mannersalo-Norros approximation of $-\log \mathbb{P}(Q_i > B_i)$. Define t^F as a minimizer of

$$\psi(t) := \frac{1}{2} \inf_{t\geq 0} \frac{(B_1 + (C - n_1\mu_1 - n_2\mu_2)t)^2}{n_1 v_1(t) + n_2 v_2(t)},$$

and

$$\phi_2^F := \frac{n_2\mu_2}{C} + \left(\frac{n_2 v_2(t^F)\left(B_1 + (C - n_1\mu_1 - n_2\mu_2)t^F\right)}{Ct^F\left(n_1 v_1(t^F) + n_2 v_2(t^F)\right)} \right).$$

Then

$$\triangle_1(n_1,n_2) = \begin{cases} (i) & \frac{1}{2}\inf_{t\geq 0} \frac{(B_1+(\phi_1 C - n_1\mu_1)t)^2}{n_1 v_1(t)} & \text{for } \phi_2 \in [0, \frac{n_2\mu_2}{C}]; \\ (ii) & \frac{1}{2}\inf_{t\geq 0} \frac{(B_1+(\phi_1 C - n_1\mu_1)t)^2}{n_1 v_1(t)} + \frac{(\phi_2 C - n_2\mu_2)^2 t^2}{n_2 v_2(t)} & \text{for } \phi_2 \in (\frac{n_2\mu_2}{C}, \phi_2^F); \\ (iii) & \frac{1}{2}\inf_{t\geq 0} \frac{(B_1+(C - n_1\mu_1 - n_2\mu_2)t)^2}{n_1 v_1(t) + n_2 v_2(t)} & \text{for } \phi_2 \in [\phi_2^F, 1]. \end{cases}$$

The approximations $\triangle_2(n_1,n_2)$ are analogous; evidently, we can now approximate $\mathbb{P}(Q_i > B_i)$ by $\exp(-\triangle_i(n_1,n_2))$. We now heuristically explain the three regimes (i), (ii), (iii). As the first and the third have the easiest explanation we start there, before proceeding to the second regime.

In regime (i) we have that $\phi_2 C \leq n_2\mu_2$. That is, the mean traffic rate generated by class 2 exceeds the guaranteed rate of service to class 2 (we call this: class 2 in overload). Therefore, it is very likely that type-2 sources claim their guaranteed service rate $\phi_2 C$ essentially all the time. Hence, overflow in queue 1 resembles overflow in a FIFO queue with service rate $\phi_1 C$. The approximation $\triangle_1(n_1, n_2)$ of regime (i) is based on this principle, cf. [1]. The minimizing t represents the (most likely) length of the interval between the epoch queue 1 starts to build up, until it reaches buffer content B_1.

Regime (iii) requires ϕ_2 to be at least as large as ϕ_2^F. It can be verified (by using the explicit formulae for conditional means of Normal random variables) that ϕ_2^F is equal to the value of ϕ_2 for which

$$\mathbb{E}\left(A_2(-t^F, 0) | A_1(-t^F, 0) + A_2(-t^F, 0) = B_1 + Ct^F\right)$$

equals $\phi_2 Ct^F$. Hence, if $\phi_2 \geq \phi_2^F$, conditioned on the total queue building up B_1 in t^F time units, then all this traffic is in queue 1, and queue 2 is essentially empty.

Regime (ii) applies if class 2 is underloaded, but $\phi_2 \leq \phi_2^F$. When the total queue reaches level B_1, it is now very likely that the queue of class 2 is non-empty. Hence, an additional constraint must be imposed to keep the buffer content of queue 2 small. The approximation is such that the flows of class 1 generate $B_1 + \phi_1 Ct$, while the class-2 sources generate $\phi_2 Ct$ (i.e., the class-2 sources claim their guaranteed rate). Note that in the approximation it is used that the interval in which the class-2 sources claim rate $\phi_2 C$ coincides with the interval in which queue 1 builds up. For a refinement of this approximation we refer to [17], which allows scenarios in which the first queue starts to build up before the queue reaches traffic rate $\phi_2 C$.

3 Partitioning of the stable region

In order to derive the admissible region (for given weights) of the two-class GPS system, we have to determine the admissible region of each class separately and then take the intersection of these two. In Sections 3-4, without loss of generality, we focus on the admissible region of the first class (i.e., the set of sources (n_1, n_2) for which the class-1 sources receive the desired QoS), as the second one can be treated in the same fashion. Before the admissible region of the first class can be obtained, which we will do in Section 4, we first determine all (n_1, n_2) for which (i) $\phi_2 \in [0, n_2\mu_2/C]$, (ii) $\phi_2 \in (n_2\mu_2/C, \phi_2^F)$ and (iii) $[\phi_2^F, 1]$, thus partitioning the stable region $T := \{(n_1, n_2) : n_1\mu_1 + n_2\mu_2 \leq C\}$ into three sets. In these three sets we can use the approximation of $\triangle_1(n_1, n_2)$ presented in Section 2.2.

Lemma 3.1 *Let* $\phi_1 \in (0, 1)$. *Then* $T = T_1^i(\phi_1) \cup T_1^{ii}(\phi_1) \cup T_1^{iii}(\phi_1)$ *for disjoint non-empty* $T_1^i(\phi_1)$, $T_1^{ii}(\phi_1)$ *and* $T_1^{iii}(\phi_1)$, *where*

$$\begin{aligned}
T_1^i(\phi_1) &= \left\{(n_1, n_2) \in T : n_2 \geq \frac{\phi_2 C}{\mu_2}\right\}; \\
T_1^{ii}(\phi_1) &= \left\{(n_1, n_2) \in T : \frac{B_1 + (\phi_1 C - n_1\mu_1)t^F}{n_1 v_1(t^F)} > \frac{(\phi_2 C - n_2\mu_2)t^F}{n_2 v_2(t^F)}\right\}; \\
T_1^{iii}(\phi_1) &= \left\{(n_1, n_2) \in T : n_2 < \frac{\phi_2 C}{\mu_2}, \quad \frac{B_1 + (\phi_1 C - n_1\mu_1)t^F}{n_1 v_1(t^F)} \leq \frac{(\phi_2 C - n_2\mu_2)t^F}{n_2 v_2(t^F)}\right\},
\end{aligned}$$

such that regime (j) applies in $T_1^j(\phi_1)$, for $j \in \{i, ii, iii\}$.

Proof: $T_1^i(\phi_1)$ follows from the fact that we must have $\phi_2 \in [0, n_2\mu_2/C]$. In order to be in $T_1^{ii}(\phi_1)$ we must have that $\phi_2 \in (n_2\mu_2/C, \phi_2^F)$, or equivalently $n_2 < \phi_2 C/\mu_2$ and $\phi_2 < \phi_2^F$. The latter inequality can be rewritten as

$$\phi_2 < \frac{n_2\mu_2}{C} + \left(\frac{n_2 v_2(t^F)\left(B_1 + (C - n_1\mu_1 - n_2\mu_2)t^F\right)}{Ct^F\left(n_1 v_1(t^F) + n_2 v_2(t^F)\right)}\right).$$

Multiply both sides with Ct^F, and rearrange the right-hand side to obtain

$$\phi_2 Ct^F < \left(\frac{n_2 v_2(t^F)\left(B_1 + Ct^F - n_1\mu_1 t^F\right)}{n_1 v_1(t^F) + n_2 v_2(t^F)}\right) + \frac{n_1 v_1(t^F) n_2\mu_2 t^F}{n_1 v_1(t^F) + n_2 v_2(t^F)}.$$

Multiplying both sides with $n_1 v_1(t^F) + n_2 v_2(t^F)$ and collecting ‘equivalent terms’ leads to

$$n_1 v_1(t^F)\left(\phi_2 Ct^F - n_2\mu_2 t^F\right) < n_2 v_2(t^F)\left(B_1 + \phi_1 Ct^F - n_1\mu_1 t^F\right).$$

Dividing both sides by $n_1 v_1(t^F)$ and $n_2 v_2(t^F)$ respectively gives

$$\frac{B_1 + (\phi_1 C - n_1\mu_1)t^F}{n_1 v_1(t^F)} > \frac{(\phi_2 C - n_2\mu_2)t^F}{n_2 v_2(t^F)}. \tag{2}$$

Note that the constraint $n_2 < \phi_2 C/\mu_2$ is redundant, as it is automatically satisfied if $\phi_2 < \phi_2^F$ (given that $(n_1, n_2) \in T$). The characterization of $T_1^{iii}(\phi_1)$ follows similarly.

In case $\phi_1 \in (0, 1)$, all three sets are non-empty, and this proves the stated. Note that $T = T_1^{iii}(0)$ for $\phi_1 = 0$ and $T = T_1^i(1)$ for $\phi_1 = 1$. □

Remark. Note that t^F implicitly depends on n_1 and n_2 (see Section 2.2), i.e., $t^F \equiv t^F(n_1, n_2)$. Due to A3, $\lim_{t\to 0}\psi(t) = \lim_{t\to\infty}\psi(t) = \infty$, and thus a minimizer t^F of $\psi(\cdot)$ clearly exists, but it is not necessarily unique. Dębicki & Mandjes [8] find a sufficient condition for the minimizer to be unique, but this condition is not necessarily fulfilled under A1-A3. In virtually all cases we considered t^F was unique; in fact, it turned out to be a non-trivial exercise to find a situation with multiple minimizers. Our example, depicted in Figure 1, has a correlation structure with strong negative correlations on short time scales (due to $v_1(\cdot)$), and strong positive correlations on long time scales (due to $v_2(\cdot)$). Since t^F is non-unique, we have that t^F is not continuous in n_1 and n_2. This is also illustrated in Figure 1 (right). By slightly increasing (n_1, n_2), the minimizing t jumps from 0.2773 to 32.3586. For a related example, see Section 5 of [15].

Now consider the boundary between $T_1^{ii}(\phi_1)$ and $T_1^{iii}(\phi_1)$, i.e., combinations of (n_1, n_2) such that (2) holds with equality. For most of the $v_i(\cdot)$ curves we considered, this boundary could not be explicitly expressed in terms of a function $f_1(n_2) = n_1$; to compute the boundary, one needs to resort to numerical methods. However, the following characteristics of $f_1(\cdot)$ can easily be derived: $f_1(0) = 0$; $f_1(\phi_2 C/\mu_2) = \phi_1 C/\mu_1$; $f_1(\cdot)$ only intersects the n_1-axis and n_2-axis at $(n_1, n_2) = (0, 0)$; $f_1(\cdot)$ only intersects the capacity constraint at $(n_1, n_2) = (\phi_1 C/\mu_1, \phi_2 C/\mu_2)$; $f_1(\cdot)$ only intersects the line $n_2 = \phi_2 C/\mu_2$ at $(n_1, n_2) = (\phi_1 C/\mu_1, \phi_2 C/\mu_2)$. In our numerical experiments with ‘popular’ variance functions $v_i(\cdot)$ (as the ones presented in [1]), $f_1(\cdot)$ is strictly increasing, see Figure 2. By choosing rather ‘extreme’ $v_i(\cdot)$, however, we have been able to construct fairly ‘exotic’ shapes for $f_1(\cdot)$.

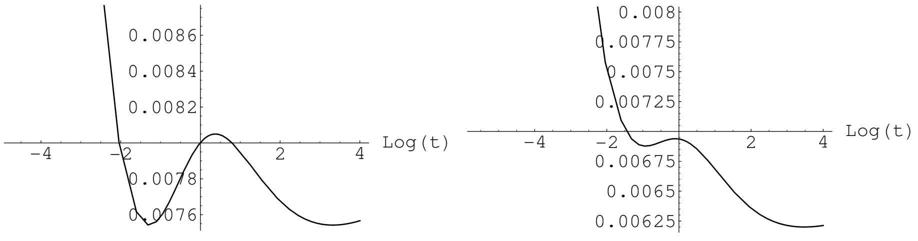

Figure 1: Left: The function $\psi(\cdot)$ with parameters $n_1 = n_2 = 1$, $C = 1$, $B_1 = 0.04$, $\mu_1 = 0.4$, $\mu_2 = 0.5$, $v_1(t) = t^{0.55}$ and $v_2(t) = 1.4955t^{1.98}$. The minimizers are $t_1^F = 0.2773$ and $t_2^F = 27.6691$, with $\psi(t_1^F) = \psi(t_2^F) = 0.00754$. Right: The same setting but now with $n_1 = n_2 = 1.01$. The minimizer is $t^F = 32.3586$.

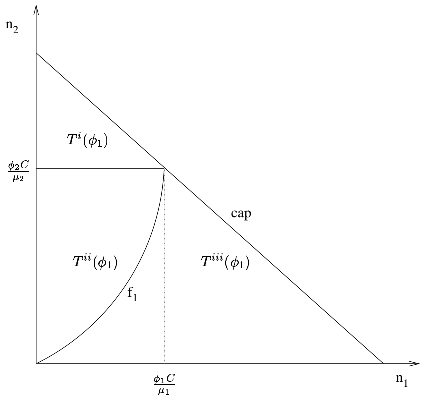

Figure 2: The partitioning of the stable region T

4 Analysis of the admissible region

In this section we analyze the admissible region of the first class (for given weights), i.e., all combinations of (n_1, n_2) that satisfy $\triangle_1(n_1, n_2) \geq \delta_1$, for some $\delta_1 > 0$. We show that this set consists of three disjoint subsets: $S_1(\phi_1) = S_1^i(\phi_1) \cup S_1^{ii}(\phi_1) \cup S_1^{iii}(\phi_1)$, with $S_1^j(\phi_1) \subset T_1^j$, $j \in \{i, ii, iii\}$, which we derive below. Finally, we present our main result that characterizes the boundary of $S_1(\phi_1)$. Again we concentrate on $S_1(\phi_1)$, but of course $S_2(\phi_1)$ can be treated analogously, thus determining the admissible region $S(\phi_1) := S_1(\phi_1) \cap S_2(\phi_1)$.

4.1 Region $S_1^i(\phi_1)$

We define $S_1^i(\phi_1)$ as the subset of $T_1^i(\phi_1)$ (see Section 3), for which $\triangle_1(n_1, n_2) \geq \delta_1$. That is,

$$\triangle_1(n_1, n_2) = \frac{1}{2} \inf_{t \geq 0} \frac{(B_1 + (\phi_1 C - n_1 \mu_1)t)^2}{n_1 v_1(t)} \geq \delta_1.$$

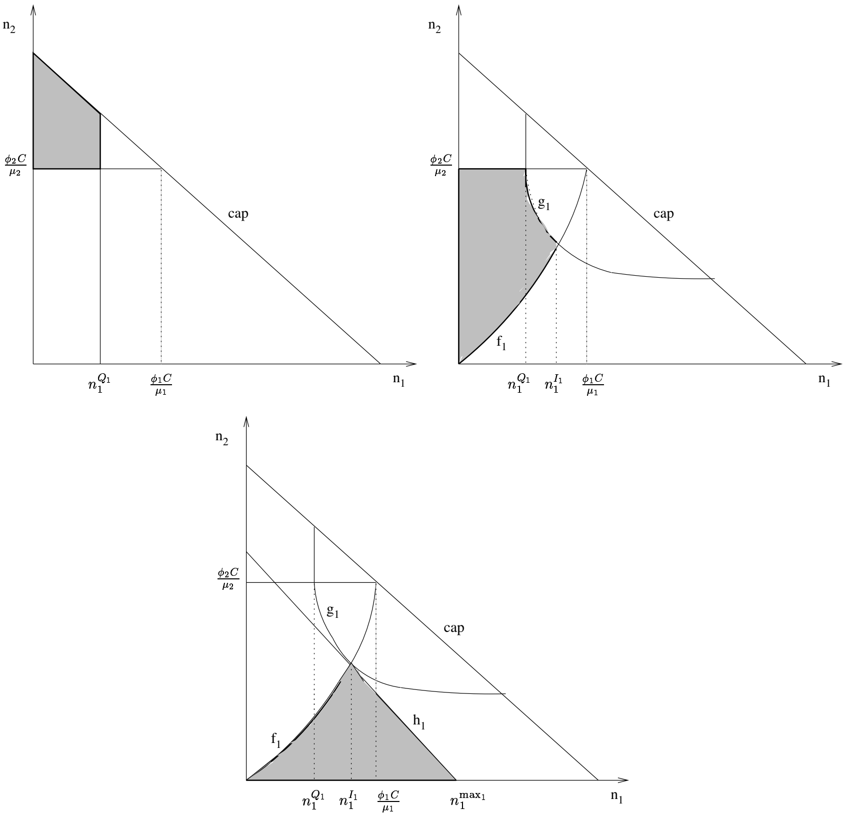

Figure 3: The partitioning of the admissible region of the first queue $S_1(\phi_1)$. Top, left: $S_1^i(\phi_1)$. Top, right: $S_1^{ii}(\phi_1)$. Bottom: $S_1^{iii}(\phi_1)$.

Rearranging and collecting terms yields

$$n_1 \leq \max\left\{ n_1 : \quad \forall t \geq 0 : X_t n_1^2 + Y_t n_1 + Z_t \geq 0 \right\},$$

where

$$\begin{aligned} X_t &:= \mu_1^2 t^2; \\ Y_t &:= -2B_1\mu_1 t - 2\phi_1 C \mu_1 t^2 - 2\delta_1 v_1(t); \\ Z_t &:= B_1^2 + \phi_1^2 C^2 t^2 + 2B_1\phi_1 C t. \end{aligned}$$

This eventually leads to

$$n_1 \leq n_1^{Q_1} := \inf_{t \geq 0} \frac{-Y_t - \sqrt{Y_t^2 - 4X_t Z_t}}{2X_t}. \tag{3}$$

Note that $\triangle_1(\phi_1 C/\mu_1, n_2)$ of regime (i) equals 0, as it is minimized for $t = \infty$ by A2. Since we assumed that $\triangle_1(n_1, n_2) \geq \delta_1 > 0$, this implies that $n_1^{Q_1} < \phi_1 C/\mu_1$. An example of a set $S_1^i(\phi_1)$ is depicted in Figure 3 (top, left).

4.2 Region $S_1^{ii}(\phi_1)$

In this regime $S_1^{ii}(\phi_1)$ consists of all combinations (n_1, n_2) in $T_1^{ii}(\phi_1)$ such that $\triangle_1(n_1, n_2) =$

$$\frac{1}{2}\inf_{t\geq 0}\frac{(B_1+(\phi_1 C-n_1\mu_1)t)^2}{n_1 v_1(t)}+\frac{(\phi_2 C-n_2\mu_2)^2 t^2}{n_2 v_2(t)}\geq \delta_1.$$

Proceeding in the same manner as above, this reduces to

$$n_2 \leq g_1(n_1) := \inf_{t\geq 0}\frac{-Y_t-\sqrt{Y_t^2-4X_tZ_t}}{2X_t}, \tag{4}$$

where

$$\begin{aligned} X_t &:= \mu_2^2 t^2/v_2(t); \\ Y_t &:= \frac{(B_1+(\phi_1 C-n_1\mu_1)t)^2}{n_1 v_1(t)}-\frac{2\phi_2 C\mu_2 t^2}{v_2(t)}-2\delta_1; \\ Z_t &:= \phi_2^2 C^2 t^2/v_2(t). \end{aligned}$$

As $g_1(\cdot)$ plays an important role in describing the boundary of $S_1(\phi_1)$, the remainder of this subsection is devoted to some structural properties of $g_1(\cdot)$. First notice that

$$\frac{1}{2}\inf_{t\geq 0}\frac{(B_1+(\phi_1 C-n_1\mu_1)t)^2}{n_1 v_1(t)}+\frac{(\phi_2 C-n_2\mu_2)^2 t^2}{n_2 v_2(t)}\geq$$
$$\frac{1}{2}\inf_{t\geq 0}\frac{(B_1+(\phi_1 C-n_1\mu_1)t)^2}{n_1 v_1(t)}+\frac{1}{2}\inf_{t\geq 0}\frac{(\phi_2 C-n_2\mu_2)^2 t^2}{n_2 v_2(t)};$$

the first part of the right-hand side of the last equation coincides with the loss probability of regime (i). By definition all $n_1 \leq n_1^{Q_1}$ satisfy the loss constraint of regime (i). Hence, all $n_1 \leq n_1^{Q_1}$ $(\forall n_2)$ satisfy the loss constraint of regime (ii) as well. One can easily see that if $n_2 = \phi_2 C/\mu_2$, then the loss probability of the middle regime reduces to that of the first regime. Thus, this implies that $g_1(n_1^{Q_1}) = \phi_2 C/\mu_2$ and that $g_1(\cdot)$ is only defined on the interval $[n_1^{Q_1}, \infty)$.

Lemma 4.1 *$g_1(\cdot)$ is continuous on the interval $[n_1^{Q_1}, \infty)$.*

Proof: $g_1(n_1)$ can be expressed as $\inf_{t\geq 0} k(t, n_1)$, see (4). Now, since $k(t, n_1)$ is continuous in n_1, $\inf_{t\geq 0} k(t, n_1)$ is continuous as well. □

It is tempting to believe that the differentiability assumption imposed on $v_i(\cdot)$ (A1), $i = 1, 2$, would imply differentiability of $g_1(\cdot)$ as well, but this turns out to be false, and can be seen as follows. Let $t^o \equiv t^o(n_1)$ be a minimizer of $\inf_{t\geq 0} k(t, n_1)$ (see previous lemma), so that $\inf_{t\geq 0} k(t, n_1) = k(t^o, n_1)$. Now,

$$\frac{\mathrm{d}g_1(n_1)}{\mathrm{d}n_1} = \frac{\partial k(t,n)}{\partial t}\Bigg|_{\substack{t=t^o\\ n=n_1}} + \frac{\partial k(t,n)}{\partial n}\Bigg|_{\substack{t=t^o\\ n=n_1}};$$

note that the first of the two partial derivatives in the right-hand side is 0 (since t^o optimizes $k(t, n_1)$). As k is continuous in both arguments, we see that $g_1(\cdot)$ is differentiable when t^o is continuous in n_1. As before, counterexamples for the latter property can be constructed, cf. the remark made in Section 3.

Lemma 4.2 *$g_1(\cdot)$ is strictly decreasing on the interval $[n_1^{Q_1}, \phi_1 C/\mu_1]$.*

Proof: Consider $(n_1, n_2) = (a, b)$, with $g_1(a) = b$, where $n_1^{Q_1} < a < \phi_1 C/\mu_1$ and $b < \phi_2 C/\mu_2$, or equivalently

$$\inf_{t\geq 0} \frac{(B_1 + (\phi_1 C - a\mu_1)t)^2}{2av_1(t)} + \frac{(\phi_2 C - b\mu_2)^2 t^2}{2bv_2(t)} = \delta_1.$$

Let an optimizer be denoted by t^o. Now, consider the point $(n_1, n_2) = (a + \epsilon_a, b + \epsilon_b)$, with $\epsilon_a \in (0, \phi_1 C/\mu_1 - a)$ and $\epsilon_b \in (0, \phi_2 C/\mu_2 - b)$. Clearly,

$$\begin{aligned}
&\inf_{t\geq 0} \frac{(B_1 + (\phi_1 C - (a+\epsilon_a)\mu_1)t)^2}{2(a+\epsilon_a)v_1(t)} + \frac{(\phi_2 C - (b+\epsilon_b)\mu_2)^2 t^2}{2(b+\epsilon_b)v_2(t)} \\
&\leq \frac{(B_1 + (\phi_1 C - (a+\epsilon_a)\mu_1)t^o)^2}{2(a+\epsilon_a)v_1(t^o)} + \frac{(\phi_2 C - (b+\epsilon_b)\mu_2)^2 (t^o)^2}{2(b+\epsilon_b)v_2(t^o)} \\
&\qquad\qquad < \delta_1,
\end{aligned}$$

Thus, $(n_1, n_2) = (a+\epsilon_a, b+\epsilon_b)$ cannot satisfy the loss constraint. The same holds for $(n_1, n_2) = (a+\epsilon_a, b)$ and $(n_1, n_2) = (a, b+\epsilon_b)$. In the same manner we can prove that $\triangle(a-\epsilon_a, b-\epsilon_b) > \delta_1$, with $\epsilon_a \in [0, a - n_1^{Q_1})$ and $\epsilon_b \in [0, b)$, but not both 0. Recall that $g_1(\cdot)$ corresponds to all combinations (n_1, n_2) for which $\triangle_1(n_1, n_2) = \delta_1$. This proves that $g_1(\cdot)$ must be a strictly decreasing function of n_1 on the interval $[n_1^{Q_1}, \phi_1 C/\mu_1]$. □

In Section 3 we remarked that under quite general circumstances the function $f_1(\cdot)$, which separates regime (ii) from regime (iii), is increasing on the interval $[0, \phi_2 C/\mu_2]$, with $f(\phi_2 C/\mu_2) = \phi_1 C/\mu_1$. As $g_1(\cdot)$ is continuous and strictly decreasing on the interval $[n_1^{Q_1}, \phi_1 C/\mu_1]$, with $g_1(n_1^{Q_1}) = \phi_2 C/\mu_2$, we expect that $f_1(\cdot)$ and $g_1(\cdot)$ intersect at a unique point $(n_1, n_2) = (n_1^{I_1}, n_2^{I_1})$, with $n_1^{Q_1} < n_1^{I_1} < \phi_1 C/\mu_1$ and $n_2^{I_1} < \phi_2 C/\mu_2$; in Section 5 we will show that for Brownian motion inputs this claim is true. Then a typical shape of the region $S_1^{ii}(\phi_1)$ would be like Figure 3 (top, right).

4.3 Region $S_1^{iii}(\phi_1)$

$S_1^{iii}(\phi_1)$ consists of all combinations of (n_1, n_2) in $T_1^{iii}(\phi_1)$ such that

$$\triangle_1(n_1, n_2) = \frac{1}{2} \inf_{t\geq 0} \frac{(B_1 + (C - n_1\mu_1 - n_2\mu_2)t)^2}{n_1 v_1(t) + n_2 v_2(t)} \geq \delta_1.$$

Once again, standard rewriting yields

$$n_2 \leq h_1(n_1) = \inf_{t\geq 0} \frac{-Y_t - \sqrt{Y_t^2 - 4X_t Z_t}}{2X_t}, \tag{5}$$

where

$$\begin{aligned}
X_t &:= \mu_2^2 t^2; \\
Y_t &:= 2n_1\mu_1\mu_2 t^2 - 2\delta_1 v_2(t) - 2B_1\mu_2 t - 2C\mu_2 t^2; \\
Z_t &:= B_1^2 + 2B_1 Ct + C^2 t^2 + n_1^2\mu_1^2 t^2 - 2B_1 n_1\mu_1 t - 2Cn_1\mu_1 t^2 - 2\delta_1 n_1 v_1(t).
\end{aligned}$$

Let $n_1^{\max_1}$ denote the value of n_1 that solves $h_1(n_1) = 0$. The following lemma states some properties of $h_1(\cdot)$.

Lemma 4.3 *$h_1(\cdot)$ is continuous, strictly decreasing on the interval $[0, n_1^{\max_1}]$ and tighter than the capacity constraint. Furthermore, $g_1(n_1) \geq h_1(n_1)$ for all $n_1 \in [n_1^{Q_1}, n_1^{\max_1}]$.*

Proof: The proof of the first statement is similar to Lemma 4.1 and the proof of the second statement is similar to Lemma 4.2. If $n_1\mu_1 + n_2\mu_2 = C$, then the optimizing t in the approximation $\triangle_1(n_1, n_2)$ of regime (iii) equals ∞ (due to A2). Subsequently, we obtain the inequality $0 \geq \delta_1$, thus contradicting $\delta_1 > 0$. It follows that $n_1\mu_1 + n_2\mu_2 < C$. Note that this also implies that $n_1^{\max_1} < C/\mu_1$.

We now show that all combinations of (n_1, n_2) that meet the loss constraint for regime (iii), will also meet that of regime (ii) for $n_1 \in [n_1^{Q_1}, n_1^{\max_1}]$. Let $a_1 := B_1 + (\phi_1 C - n_1\mu_1)t$, $a_2 := (\phi_2 C - n_2\mu_2)t$, $v_1 := n_1 v_1(t)$ and $v_2 := n_2 v_2(t)$. It can be seen that it suffices to prove that for all $t \geq 0$,

$$\frac{a_1^2}{v_1} + \frac{a_2^2}{v_2} \geq \frac{(a_1 + a_2)^2}{v_1 + v_2}. \tag{6}$$

Rearranging (6) yields $a_1^2 v_2^2 + a_2^2 v_1^2 - 2a_1 a_2 v_1 v_2 \geq 0$, which is equivalent to $(a_1 v_2 - a_2 v_1)^2 \geq 0$, thus proving the last statement. Note that there is equality if $a_1 v_2 = a_2 v_1$, so in that case $\triangle_1(n_1, n_2)$ of regime (ii) and (iii) coincide and they have the same optimizer t^F. Recall from Section 3 that $a_1 v_2 = a_2 v_1$, with $t = t^F$, corresponds to the line $f_1(\cdot)$. □

As in the case of $g_1(\cdot)$, $h_1(\cdot)$ is not necessarily differentiable, because its optimizing t might not be unique (see Section 4.2). By definition, for $(n_1, n_2) = (f_1(n_2), n_2)$ the approximations of $\triangle_1(n_1, n_2)$ are equal for regimes (ii) and (iii) (see previous lemma). Hence, if $f_1(\cdot)$ and $g_1(\cdot)$ intersect at $(n_1^{I_1}, n_2^{I_1})$ (see Section 4.2), then this is also the point where $f_1(\cdot)$ and $h_1(\cdot)$ intersect. Figure 3 (bottom) illustrates the region $S_1^{iii}(\phi_1)$.

4.4 Region $S_1(\phi_1)$

$S_1(\phi_1)$ can be obtained by taking the union of the three described regions, i.e., $S_1(\phi_1) = S_1^{i}(\phi_1) \cup S_1^{ii}(\phi_1) \cup S_1^{iii}(\phi_1)$. We now state our main result, which follows from Sections 4.1, 4.2, and 4.3.

Theorem 4.4 *The boundary of the admissible region of the first queue, $S_1(\phi_1)$, is defined as follows:*

$$\begin{array}{ll} 0 \leq n_1 \leq n_1^{Q_1}: & n_2 = (C - n_1\mu_1)/\mu_2; \\ n_1^{Q_1} < n_1 < n_1^{I_1}: & n_2 = g_1(n_1); \\ n_1^{I_1} \leq n_1 \leq n_1^{\max_1}: & n_2 = h_1(n_1). \end{array}$$

In addition, the boundary is continuous.

5 Brownian inputs

For most Gaussian inputs that satisfy A1-A3 the boundary of $S(\phi_1)$ cannot be explicitly computed; consequently, in those cases one has to rely on numerical techniques (as will be done in the numerical examples in Section 6). For the 'canonical model' with Brownian inputs though, we have succeeded in finding closed-form expressions for the boundary. As indicated in [16], Brownian motions can be used to approximate weakly-dependent traffic streams, cf. also the celebrated 'Central Limit Theorem in functional form'. We let the variance functions be characterized through $v_i(t) = \lambda_i t$, with $\lambda_i > 0$, $i = 1, 2$.

5.1 Region $S_1(\phi_1)$

It is a matter of straightforward calculus to show that $t^F = B_1/(C - n_1\mu_1 - n_2\mu_2)$. Now, the Mannersalo-Norros approximation reduces to the following. The critical weight ϕ_2^F equals

$$1 - \frac{n_1\lambda_1 - n_2\lambda_2}{n_1\lambda_1 + n_2\lambda_2}\left(1 - \frac{n_1\mu_1 + n_2\mu_2}{C}\right) - \frac{n_1\mu_1}{C}.$$

Then we get the approximations

$$\triangle_1(n_1, n_2) = \begin{cases} (i) & 2B_1\frac{\phi_1 C - n_1\mu_1}{n_1\lambda_1} & \text{for } \phi_2 \in [0, \frac{n_2\mu_2}{C}]; \\ (ii) & \frac{1}{2}\left(\frac{(B_1 + (\phi_1 C - n_1\mu_1)t^*)^2}{n_1\lambda_1 t^*} + \frac{(\phi_2 C - n_2\mu_2)^2}{n_2\lambda_2}t^*\right) & \text{for } \phi_2 \in (\frac{n_2\mu_2}{C}, \phi_2^F); \\ (iii) & 2B_1\frac{C - n_1\mu_1 - n_2\mu_2}{n_1\lambda_1 + n_2\lambda_2} & \text{for } \phi_2 \in [\phi_2^F, 1], \end{cases}$$

with the 'critical time scale' t^* given by

$$\frac{B_1}{\sqrt{(\phi_1 C - n_1\mu_1)^2 + (\phi_2 C - n_2\mu_2)^2\frac{n_1\lambda_1}{n_2\lambda_2}}}.$$

Mandjes [16] shows that the resulting expressions are 'asymptotically exact' in the many-sources regime.

Let us first derive the function $f_1(\cdot)$. Recall from Section 3 that $f_1(\cdot)$ is equivalent to all pairs of (n_1, n_2) that satisfy (2) with equality. Plugging in the expression for t^F and some rearranging yields

$$n_1 = \frac{C\lambda_2(1+\phi_1) - n_2\lambda_2\mu_2}{\frac{\phi_2 C\lambda_1}{n_2} + 2\lambda_2\mu_1 - \lambda_1\mu_2} =: f_1(n_2). \tag{7}$$

It can easily be verified that $f_1(0) = 0$ and $f_1(\phi_2 C/\mu_2) = \phi_1 C/\mu_1$. The following lemma states some properties of $f_1(\cdot)$; define

$$\xi := \frac{(1+\phi_1)\mu_1}{\phi_1\mu_2}.$$

Lemma 5.1 *$f_1(\cdot)$ is continuous and has a continuous derivative on the interval $[0, \phi_2 C/\mu_2]$. Furthermore, $f_1(\cdot)$ is concave on $[0, \phi_2 C/\mu_2]$ if $\lambda_1 > \xi\lambda_2$; $f_1(\cdot)$ is convex on $[0, \phi_2 C/\mu_2]$ if $\lambda_1 < \xi\lambda_2$; $f_1(\cdot)$ has a constant positive derivative on $[0, \phi_2 C/\mu_2]$ if $\lambda_1 = \xi\lambda_2$ and this derivative has the value $\phi_1\mu_2/(\phi_2\mu_1)$.*

Proof: Note that

$$\frac{\mathrm{d}^2}{\mathrm{d}x^2}\frac{1-\alpha x}{\beta/x+\gamma}=\frac{-2\beta(\alpha\beta+\gamma)}{(\beta+\gamma x)^3}. \tag{8}$$

In other words, due to (7), $f_1''(n_2)$ changes sign only at

$$n_2=\frac{\phi_2 C\lambda_1}{\lambda_1\mu_2-2\lambda_2\mu_1} \tag{9}$$

(corresponding to $x=-\beta/\gamma$). Note that expression (9) does not lie in $[0,\phi_2 C/\mu_2]$, so $f_1(\cdot)$ is either convex or concave on this interval. From (8) we conclude that there is concavity when $\lambda_1>\xi\lambda_2$ (corresponding to $\alpha\beta=-\gamma$), and convexity otherwise. □

Subsequently, in order to fully characterize the areas $S_1^i(\phi_1)$, $S_1^{ii}(\phi_1)$, $S_1^{iii}(\phi_1)$, we now derive $n_1^{Q_1}$, $g_1(\cdot)$ and $h_1(\cdot)$. We do this by relying on (3), (4) and (5), respectively. This yields

$$\begin{aligned}&\frac{2\phi_1 CB_1}{\delta_1\lambda_1+2B_1\mu_1}=:n_1^{Q_1};\\&\frac{(\phi_2 C-n_2\mu_2)^2B_1^2}{n_2\lambda_2(\delta_1^2\lambda_1+2B_1\delta_1\mu_1)}+\frac{2\phi_1 CB_1\delta_1}{\delta_1^2\lambda_1+2B_1\delta_1\mu_1}=:g_1^{-1}(n_2);\\&\frac{2CB_1}{2B_1\mu_2+\delta_1\lambda_2}-n_1\frac{2B_1\mu_1+\delta_1\lambda_1}{2B_1\mu_2+\delta_1\lambda_2}=:h_1(n_1).\end{aligned}$$

Note that $h_1(\cdot)$ is linear in n_1 and that

$$h_1(n_1^{\max_1})=h_1\left(\frac{2CB_1}{2B_1\mu_1+\delta_1\lambda_1}\right)=0.$$

Due to Lemma 5.1, $f_1(\cdot)$, $g_1(\cdot)$ and $h_1(\cdot)$ have a unique intersection point (n_1,n_2) given by

$$(n_1^{I_1},n_2^{I_1})=\left(\frac{CB_1(\delta_1\lambda_2(1+\phi_1)+2B_1\mu_2\phi_1)}{(\delta_1\lambda_1+2B_1\mu_1)(\delta_1\lambda_2+B_1\mu_2)},\frac{\phi_2 CB_1}{\delta_1\lambda_2+B_1\mu_2}\right).$$

Now we have all the ingredients to describe the boundary of $S_1(\phi_1)$ explicitly. The admissible region of the second queue can be treated analogously. Both are depicted in Figure 4.

5.2 Region $S(\phi_1)$

A combination (n_1,n_2) is contained in $S(\phi_1)$ if it satisfies the QoS requirements for both classes. That is, if it is contained in $S_1(\phi_1)\cap S_2(\phi_1)$. In this subsection we characterize the boundary of $S(\phi_1)$. In the analysis below the ratios B_1/B_2 and δ_1/δ_2 turn out to be crucial. We therefore introduce $b:=B_1/B_2$ and $d:=\delta_1/\delta_2$. Let us first mention some useful facts.

Lemma 5.2 *If $b<d$ then $h_2^{-1}(n_1)>h_1(n_1)$ for all n_1 that satisfy $h_2^{-1}(n_1)\geq 0$ and $h_1(n_1)\geq 0$.*

Proof: This can be seen as follows. We know that

$$h_1(n_1)=\frac{2CB_1}{2B_1\mu_2+\delta_1\lambda_2}-n_1\frac{2B_1\mu_1+\delta_1\lambda_1}{2B_1\mu_2+\delta_1\lambda_2};\quad h_2^{-1}(n_1)=\frac{2CB_2}{2B_2\mu_2+\delta_2\lambda_2}-n_1\frac{2B_2\mu_1+\delta_2\lambda_1}{2B_2\mu_2+\delta_2\lambda_2}.$$

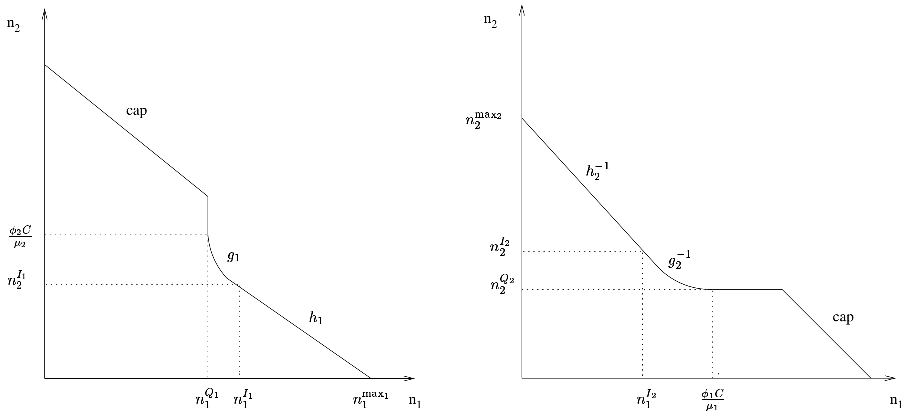

Figure 4: Left: $S_1(\phi_1)$. Right: $S_2(\phi_1)$.

Now, $h_2^{-1}(0) > h_1(0)$ implies that

$$\frac{2CB_2}{2B_2\mu_2 + \delta_2\lambda_2} > \frac{2CB_1}{2B_1\mu_2 + \delta_1\lambda_2} \text{ or } b < d,$$

but we also have that $h_2(0) > h_1^{-1}(0)$ implies that

$$\frac{2CB_2}{2B_2\mu_1 + \delta_2\lambda_1} > \frac{2CB_1}{2B_1\mu_1 + \delta_1\lambda_1} \text{ or } b < d.$$

Since $h_1(\cdot)$ and $h_2^{-1}(\cdot)$ are linear, this proves the stated. In the remaining this is denoted by $h_2^{-1}(\cdot) > h_1(\cdot)$. Likewise, if $b > d$ then $h_1(\cdot) > h_2^{-1}(\cdot)$. Note that $h_1(\cdot)$ and $h_2^{-1}(\cdot)$ are identical if $b = d$. □

Lemma 5.3 *If $b < d/2$ then $n_1^{Q_1} < n_1^{I_2}$ and $n_2^{Q_2} > n_2^{I_1}$.*

Proof: Use the explicit expressions for $n_1^{Q_1}$ and $n_1^{I_2}$. Thus

$$\frac{2\phi_1 CB_1}{\delta_1\lambda_1 + 2B_1\mu_1} < \frac{\phi_1 CB_2}{\delta_2\lambda_1 + B_2\mu_1} \quad \text{or} \quad 2\delta_2 B_1\lambda_1 + 2B_1B_2\mu_1 < \delta_1 B_2\lambda_1 + 2B_1B_2\mu_1.$$

Omitting common terms and some rearranging directly yields $b < d/2$.

Likewise, it holds that $n_2^{Q_2} > n_2^{I_1}$, since

$$\frac{2\phi_2 CB_2}{\delta_2\lambda_2 + 2B_2\mu_2} > \frac{\phi_2 CB_1}{\delta_1\lambda_2 + B_1\mu_2} \quad \text{or} \quad 2\delta_1 B_2\lambda_2 + 2B_1B_2\mu_2 > \delta_2 B_1\lambda_2 + 2B_1B_2\mu_2,$$

reduces to $b < 2d$. Analogously, if $b > 2d$ then $n_1^{Q_1} > n_1^{I_2}$ and $n_2^{Q_2} < n_2^{I_1}$. If $d/2 \leq b \leq 2d$ then $n_1^{Q_1} \geq n_1^{I_2}$ and $n_2^{Q_2} \geq n_2^{I_1}$. □

Combining the two previous lemmas leads to the conclusion that we have to distinguish between three cases: (a) $b < d/2$, (b) $d/2 \leq b \leq 2d$ and (c) $b > 2d$. Note that in the middle case it is not clear whether $h_2^{-1}(\cdot) < h_1(\cdot)$ or $h_2^{-1}(\cdot) > h_1(\cdot)$. Below we show that the shape of the boundary of $S(\phi_1)$ depends on (a), (b) or (c) if $\phi_1 \in (0,1)$. First we characterize the boundary of $S(\phi_1)$ for $\phi_1 = 0$ and $\phi_1 = 1$. The boundary of $S(0)$ is given by

$$0 \leq n_1 \leq n_1^O : \quad n_2 = n_2^{Q_2};$$

$$n_1^O < n_1 < n_1^{\max_1} : \quad n_2 = h_1(n_1),$$

where $n_2^{Q_2}$ is evaluated at $\phi_1 = 0$, and $n_1^O := h_1^{-1}(n_2^{Q_2})$. The boundary of $S(1)$ is

$$0 \leq n_1 \leq n_1^{Q_1} : \quad n_2 = h_2^{-1}(n_1),$$

where $n_2^{Q_1}$ is evaluated at $\phi_1 = 1$.

Remark. One can easily show that $S(0) \subset S(1)$ if $b < d$, $S(1) \subset S(0)$ if $b > d$ and $S(0) = S(1)$ if $b = d$.

In the following we show that there are different generic shapes of the boundary of $S(\phi_1)$, $\phi_1 \in (0,1)$, within each of the three cases.

5.2.1 Case $b < d/2$

It can easily be seen that the boundary of $S_1(\phi_1)$ has four possible shapes in this case (see Figure 5). The shape of the boundary ((a_1), (a_2), (a_3) or (a_4)) depends on the value of ϕ_1, but each shape occurs as will be shown in the following lemmas.

Lemma 5.4 *The boundary of $S(\phi_1)$ has shape (a_1) if $\phi_1 \in [X_3, 1)$, where*

$$X_3 := \frac{\delta_2\lambda_2(\delta_1\lambda_1 + 2B_1\mu_1)}{\delta_2\lambda_2(\delta_1\lambda_1 + 2B_1\mu_1) + 2\lambda_1\mu_2(\delta_1 B_2 - \delta_2 B_1)}.$$

Proof: In order to have shape (a_1) we must have that $h_2^{-1}(n_1^{Q_1}) \geq \phi_2 C/\mu_2$ for some value of $\phi_1 \in (0,1)$. That is,

$$\frac{2CB_2}{\delta_2\lambda_2 + 2B_2\mu_2} - \frac{2\phi_1 CB_1(\delta_2\lambda_1 + 2B_2\mu_1)}{(\delta_2\lambda_2 + 2B_2\mu_2)(\delta_1\lambda_1 + 2B_1\mu_1)} \geq \frac{(1-\phi_1)C}{\mu_2}. \tag{10}$$

One can easily show that this reduces to a constraint of the form $-A + B\phi_1 \geq 0$, with $A, B > 0$. For $\phi_1 = 0$ the left hand side of the constraint (10) is equivalent to

$$\frac{2CB_2}{\delta_2\lambda_2 + 2B_2\mu_2} - \frac{C}{\mu_2},$$

which is smaller than 0 (assuming that $\delta_2, \lambda_2 > 0$). Hence, the constraint is not satisfied. For $\phi_1 = 1$ the left hand side of (10) equals

$$\frac{2C\lambda_1(\delta_1 B_2 - \delta_2 B_1)}{(\delta_2\lambda_2 + 2B_2\mu_2)(\delta_1\lambda_1 + 2B_1\mu_1)},$$

which is larger than 0 if $b/d < 1$, which is true, as we required $b < d/2$. Thus, since the constraint is a linear function of ϕ_1, there must be value of $\phi_1 \in (0,1)$ for which $h_2^{-1}(n_1^{Q_1}) = \phi_2 C/\mu_2$. Straightforward calculus shows there is equality for $\phi_1 = X_3$. □

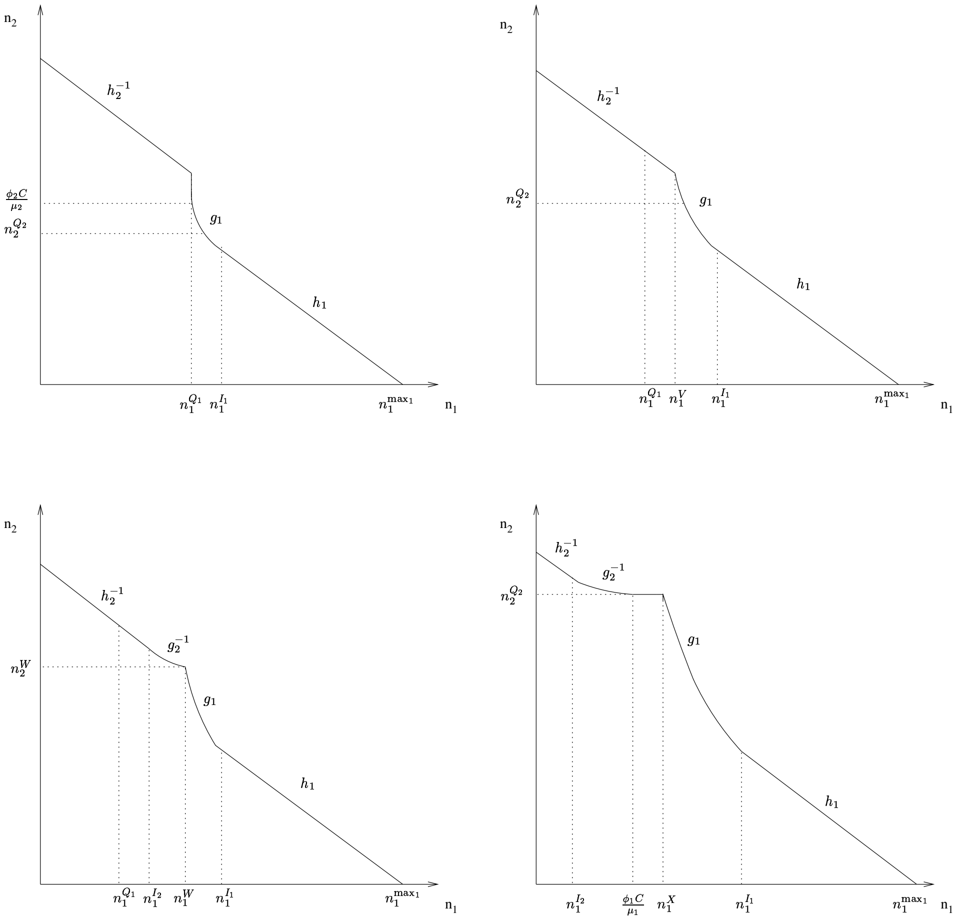

Figure 5: Top, from left to right: shape (a_1) and (a_2). Bottom, from left to right: shape (a_3) and (a_4).

Lemma 5.5 *The boundary of $S(\phi_1)$ has shape (a_4) if $\phi_1 \in (0, X_1]$, where*

$$X_1 := \frac{\delta_2^2 B_1^2 \lambda_2 \mu_1}{\delta_2^2 B_1^2 \lambda_2 \mu_1 + 2\delta_1^2 B_2 \lambda_1 (\delta_2 \lambda_2 + 2B_2 \mu_2)}.$$

Proof: The proof is analogous to that of Lemma 5.4. Shape (a_4) occurs if there exists a value of $\phi_1 \in (0,1)$ for which $g_1^{-1}(n_2^{Q_2}) \geq \phi_1 C/\mu_1$. This constraint can be rewritten as $A - B\phi_1 \geq 0$, with $A, B > 0$. Since, it is satisfied for $\phi_1 = 0$, but not for $\phi_1 = 1$, there exists a unique value of $\phi_1 \in (0,1)$, X_1, such that there is equality, i.e., $g_1^{-1}(n_2^{Q_2}) = \phi_1 C/\mu_1$. □

Lemma 5.6 *The boundary of $S(\phi_1)$ has shape (a_3) if $\phi_1 \in (X_1, X_2)$, where X_2 is the value of ϕ_1 such that $n_2^{I_2} = h_2^{-1}(n_1^{I_2}) = g_1(n_1^{I_2})$.*

Proof: The shape of the boundary is like (a_3) if $h_2^{-1}(n_1^{I_2}) < g_1(n_1^{I_2})$ and if $g_1^{-1}(n_2^{Q_2}) < \phi_1 C/\mu_1$. The latter constraint is satisfied if $\phi_1 > X_1$ (Lemma 5.5). Unfortunately, the former does not reduce to a constraint that is a linear function of ϕ_1. It can be shown that there exists a unique value of ϕ_1, X_2, such that $h_2^{-1}(n_1^{I_2}) = g_1(n_1^{I_2})$. Now, the constraint is satisfied for all $\phi_1 \in [0, X_2)$.

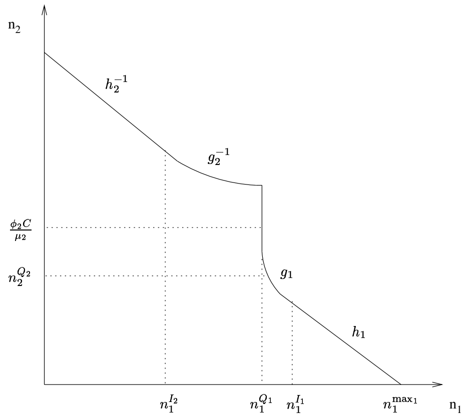

Figure 6: Shape (b_1)

We now show that $X_2 \in (X_1, X_3)$. First recall that $g_1(n_1)$ is defined on the interval $(n_1^{Q_1}, n_1^{I_1})$ in $S_1(\phi_1)$, whereas $h_2^{-1}(n_1)$ is defined on the interval $[0, n_1^{I_2}]$ in $S_2(\phi_1)$. Therefore, if $g_1(\cdot)$ and $h_2^{-1}(\cdot)$ are part of the boundary of $S(\phi_1)$, then they are defined on (parts of) the mentioned intervals. If $\phi_1 \in (0, X_1]$, then $g_1(\cdot)$ is defined on the interval $(n_1^X, n_1^{I_1})$, with $n_1^X \geq \phi_1 C/\mu_1$ (see shape (a_4)). By definition $n_1^{I_2} < \phi_1 C/\mu_1$, so this implies that $g_1(\cdot)$ and $h_1^{-1}(\cdot)$ cannot intersect if $\phi_1 \in (0, X_1]$. Furthermore, if $\phi_1 \in [X_3, 1)$, then $h_2^{-1}(n_1) > g_1(n_1)$ for all $n_1 \in (n_1^{Q_1}, \min\{n_1^{I_1}, n_1^{I_2}\})$ (see shape (a_1)), so $X_2 \notin [X_3, 1)$. Hence, we conclude $0 < X_1 < X_2 < X_3 < 1$. The expression of X_2 is not presented here (as it is quite intransparent); it depends on the parameters $\delta_1, \delta_2, B_1, B_2, \lambda_1, \lambda_2, \mu_1$ and μ_2. □

Lemma 5.7 *The boundary of $S(\phi_1)$ has shape (a_2) if $\phi_1 \in [X_2, X_3)$.*

Proof: One observes shape (a_2) if $h_2^{-1}(n_1^{Q_1}) < \phi_2 C/\mu_2$ and $h_2^{-1}(n_1^{I_2}) \geq g_1(n_1^{I_2})$. From Lemmas 5.4 and 5.6 we know that this coincide with $\phi_1 < X_3$ and $\phi_1 \geq X_2$ respectively. □

We now state our main result. The proof follows directly from Lemmas 5.4-5.7.

Proposition 5.8 *If $b < d/2$, then the boundary of $S(\phi_1)$ has*
shape (a_4) for $0 < \phi_1 \leq X_1$;
shape (a_3) for $X_1 < \phi_1 < X_2$;
shape (a_2) for $X_2 \leq \phi_1 < X_3$;
and shape (a_1) for $X_3 \leq \phi_1 < 1$.
Here X_1 is the value of ϕ_1 such that $g_1^{-1}(n_2^{Q_2}) = \phi_1 C/\mu_1$, X_2 is the value of ϕ_1 such that $n_2^{I_2} = h_2^{-1}(n_1^{I_2}) = g_1(n_1^{I_2})$, and X_3 is the value of ϕ_1 that solves $h_2^{-1}(n_1^{Q_1}) = \phi_2 C/\mu_2$.

5.2.2 Case $d/2 \leq b \leq 2d$

As proved in Lemma 5.3, this criterion leads to $n_1^{Q_1} \geq n_1^{I_2}$ and $n_2^{Q_2} \geq n_2^{I_1}$. Now, the boundary of $S(\phi_1)$ can have three shapes ((b_1), (b_2) and (b_3)). Shape (b_1) is depicted in Figure 6. Shape (b_2) corresponds to (a_3), and (b_3) to (a_4) (see Figure 5).

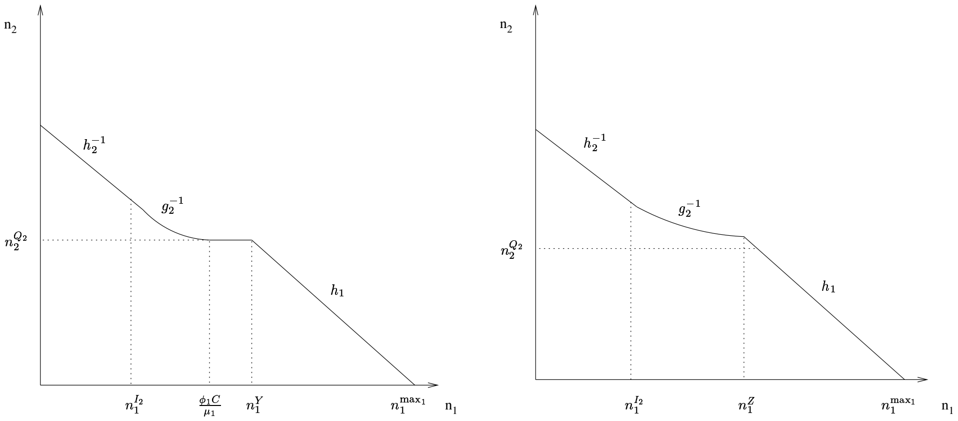

Figure 7: Left: Shape (c_1). Right: Shape (c_2).

As in the case of $b < d/2$, one can easily prove that each shape is observed. The proofs are omitted as they are similar to the proofs of Lemmas 5.4 and 5.5. We directly state the following proposition.

Proposition 5.9 *If $d/2 \leq b \leq 2d$, then the boundary of $S(\phi_1)$ has*
shape (b_3) for $0 < \phi_1 \leq Y_1$;
shape (b_2) for $Y_1 < \phi_1 < Y_2$;
and shape (b_3) for $Y_2 \leq \phi_1 < 1$.
Here Y_1 is the value of ϕ_1 such that $g_1^{-1}(n_2^{Q_2}) = \phi_1 C/\mu_1$ and Y_2 coincides with the value of ϕ_1 such that $g_2^{-1}(n_1^{Q_1}) = \phi_2 C/\mu_2$.

5.2.3 Case $b > 2d$

The last case is the counterpart of the first case. Therefore, the proofs are also omitted in the following. Now, $n_1^{Q_1} > n_1^{I_2}$ and $n_2^{Q_2} < n_2^{I_1}$. There are four possible shapes of $S(\phi_1)$, $\phi_1 \in (0,1)$. Shapes (c_1) and (c_2) are depicted in Figure 7. Shape (c_3) corresponds to (a_3), and (c_4) to (b_1) (see Figures 5 and 6 respectively).

Proposition 5.10 *If $b > 2d$, then the boundary of $S(\phi_1)$ has*
shape (c_1) for $0 < \phi_1 \leq Z_1$;
shape (c_2) for $Z_1 < \phi_1 \leq Z_2$;
shape (c_3) for $Z_2 < \phi_1 < Z_3$;
and shape (c_4) for $Z_3 \leq \phi_1 < 1$.
Here Z_1 corresponds to the value of ϕ_1 such that $h_1^{-1}(n_2^{Q_2}) = \phi_1 C/\mu_1$, Z_2 is the value of ϕ_1 that solves $n_1^{I_1} = h_1^{-1}(n_2^{I_1}) = g_2(n_2^{I_1})$ and Z_3 is the value of ϕ_1 such that $g_2^{-1}(n_1^{Q_1}) = \phi_2 C/\mu_2$.

5.3 The realizable region

Let R denote the realizable region, i.e., the admissible region if one would be allowed to adjust the weights at any time:

$$R := \bigcup_{\phi_1 \in [0,1]} S(\phi_1).$$

In the following we show that we do not always need all values of $\phi_1 \in [0,1]$ to compose R. We now state our main result.

Theorem 5.11 *The realizable region R can be obtained as follows:*

$$b < d/2: \quad R = \bigcup_{\phi_1 \in (0,X_2) \cup \{1\}} S(\phi_1);$$

$$d/2 \leq b \leq d \qquad R = \bigcup_{\phi_1 \in (0,1]} S(\phi_1);$$

$$d < b \leq 2d: \quad R = \bigcup_{\phi_1 \in [0,1)} S(\phi_1);$$

$$b > 2d: \quad R = \bigcup_{\phi_1 \in \{0\} \cup (Z_2,1)} S(\phi_1).$$

Proof: Let us consider the case $b < d/2$. Recall that this implies that $S(0) \subset S(1)$. Furthermore, $S(\phi_1) \subset S(1)$ for all $\phi_1 \in [X_2, 1)$. To see this, compare boundaries (a_1) and (a_2) with the boundary of $S(1)$, and recall that $h_2^{-1}(\cdot) > h_1(\cdot)$ if $b < d/2$. However, this does not hold for all $\phi_1 \in (0, X_2)$, because otherwise R would be equal to $S(1)$. One can see this as follows. If $\phi_1 \in (X_1, X_2)$ then we must have shape (a_3). But then $S(\phi_1)$ contains (n_1^W, n_2^W), with $h_2^{-1}(n_1^W) < n_2^W$, which cannot be part of $S(1)$. From Lemma 5.6 it follows that for all $\phi_1 \in (X_1, X_2)$, n_1^W (n_2^W) increases (decreases) as ϕ_1 increases (but not linearly). That is, we need all values of $\phi_1 \in (X_1, X_2)$ to compose R. Likewise, shape (a_4) arises if $\phi_1 \in (0, X_1]$. The point $(n_1^X, n_2^{Q_2})$ will then be contained in $S(\phi_1)$, which cannot be contained in $S(1)$ either. From Lemma 5.5 it follows that as ϕ_1 increases in the corresponding interval, n_1^X ($n_2^{Q_2}$) linearly increases (decreases), i.e., we also need all values of $\phi_1 \in (0, X_1]$.

The other statements follow in a similar fashion. Recall that for $b > d$ we have $S(1) \subset S(0)$. □

The boundary of R can now also be determined using Theorem 5.11. Below we discuss each of the four cases of Theorem 5.11. Let us first introduce some notation. From now on, we write $z(\phi_1)$ if z depends on ϕ_1. Note that $\phi_2 = 1 - \phi_1$, thus if an expression contains ϕ_2, we can also easily rewrite it as function of ϕ_1.

5.3.1 Case $b < d/2$

We need all values $\phi_1 \in (0, X_2)$ and $\phi_1 = 1$ to compose R. As we will see, $S(1)$ contributes a large part to the boundary of R. All values of $\phi_1 \in (0, X_2)$ contribute to the boundary in the following way (straightforward calculus):

$$\phi_1 \in (0, X_1]: \quad (n_1, n_2) = (g_1^{-1}(n_2^{Q_2}(\phi_1)), n_2^{Q_2}(\phi_1)); \tag{11}$$

$$\phi_1 \in (X_1, X_2): \quad (n_1, n_2) = (n_1^W(\phi_1), n_2^W(\phi_1)), \tag{12}$$

with

$$n_1^H := g_1^{-1}(n_2^{Q_2}(0)) > 0; \quad n_2^{Q_2}(0) = h_2^{-1}(0);$$

$$g_1^{-1}(n_2^{Q_2}(X_1)) = n_1^W(X_1); \quad n_2^{Q_2}(X_1) = n_2^W(X_1);$$

$$n_2^W(X_2) = h_2^{-1}(n_1^W(X_2)).$$

As mentioned, it can be shown that (11) and (12) correspond to two lines that decrease in n_2 as n_1 increases (the former linearly, but the latter non-linearly). Let us denote the former by $k_1(n_1)$ and the latter by $k_2(n_1)$. Moreover, $k_1(\cdot)$, $k_2(\cdot)$ and $h_2^{-1}(\cdot)$ 'perfectly connect' (see Figure 8 (top, left)), as one can show that

$$\left.\frac{\frac{\partial n_2^{Q_2}(\phi_1)}{\partial \phi_1}}{\frac{\partial g_1^{-1}(n_2^{Q_2}(\phi_1))}{\partial \phi_1}}\right|_{\phi_1 = X_1} = \left.\frac{\frac{\partial n_2^W(\phi_1)}{\partial \phi_1}}{\frac{\partial n_1^W(\phi_1)}{\partial \phi_1}}\right|_{\phi_1 = X_1}; \quad \left.\frac{\partial \frac{n_2^W(\phi_1)}{\partial \phi_1}}{\frac{\partial n_1^W(\phi_1)}{\partial \phi_1}}\right|_{\phi_1 = X_2} = \frac{\partial h_2^{-1}(n_1)}{\partial n_1}.$$

We are now able to describe to boundary of R, which follows from above.

Proposition 5.12 *If $b < d/2$, then the boundary of R, denoted by r_1 (see Figure 8), is continuous.*

5.3.2 Case $d/2 \leq b \leq d$

The approach is very similar to that in the previous case. We introduce line $k_3(n_1)$ and $k_4(n_1)$, that correspond to the following equations respectively:

$$\phi_1 \in (Y_1, Y_2): \quad (n_1, n_2) = (n_1^W(\phi_1), n_2^W(\phi_1));$$

$$\phi_1 \in [Y_2, 1): \quad (n_1, n_2) = (n_1^{Q_1}(\phi_1), g_2^{-1}(n_1^{Q_1}(\phi_1))).$$

Recall that we obtain line $k_1(\cdot)$ from $\phi_1 \in (0, Y_1)$, as $Y_1 = X_1$. It can be shown that $k_3(\cdot)$ is a non-linearly decreasing function, whereas $k_4(\cdot)$ is a linearly decreasing function. Furthermore, it can be shown that $k_1(\cdot)$, $k_3(\cdot)$ and $k_4(\cdot)$ 'connect perfectly' (see Figure 8 (top, right)). Now we have all the ingredients to describe the boundary.

Proposition 5.13 *If $d/2 \leq b \leq d$, then the boundary of R, denoted by r_2 (see Figure 8), is continuous.*

5.3.3 Case $d < b \leq 2d$

We directly state our result on the boundary of R, since it is very similar to the previous case.

Proposition 5.14 *If $d < b \leq 2d$, then the boundary of R, denoted by r_3 (see Figure 8), is continuous.*

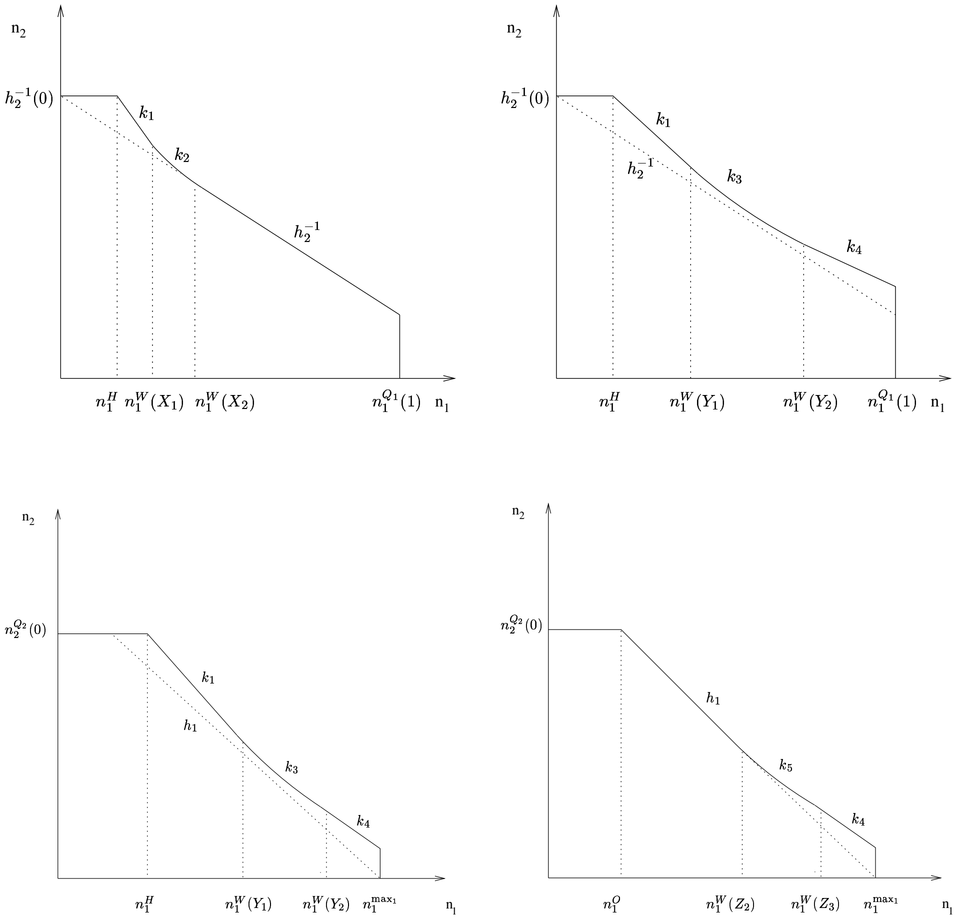
n2
$h_2^{-1}(0)$
k_1
k_2
h_2^{-1}
n_1^H $n_1^W(X_1)$ $n_1^W(X_2)$
$n_1^{Q_1}(1)$ n1
n2
$h_2^{-1}(0)$
k_1
h_2^{-1}
k_3
k_4
n_1^H $n_1^W(Y_1)$ $n_1^W(Y_2)$ $n_1^{Q_1}(1)$ n1
n2
$n_2^{Q_2}(0)$
k_1
h_1
k_3
k_4
n_1^H $n_1^W(Y_1)$ $n_1^W(Y_2)$ $n_1^{\max_1}$ n1
n2
$n_2^{Q_2}(0)$
h_1
k_5
k_4
n_1^O $n_1^W(Z_2)$ $n_1^W(Z_3)$ $n_1^{\max_1}$ n1

Figure 8: Top, from left to right: shape (r_1) and (r_2) (area below dotted line represents $S(1)$). Bottom, from left to right: shape (r_3) and (r_4) (area below dotted line represents $S(0)$).

5.3.4 Case $b > 2d$

We introduce line $k_5(n_1)$, that corresponds to the following:

$$\phi_1 \in (Z_2, Z_3): \quad (n_1, n_2) = (n_1^W(\phi_1), n_2^W(\phi_1)).$$

Note that $Z_3 = Y_2$. Once again, we directly state the boundary of R.

Proposition 5.15 *If $b > 2d$, then the boundary of R, denoted by r_4 (see Figure 8), is continuous.*

Although we need a range of weights to obtain R, the results suggest that almost all of R is obtained by the priority scheduling discipline, e.g., $\phi_1 = 0$ or $\phi_1 = 1$. In case $b \leq d$, the admissible region $S(1)$ covers most of R, whereas in case $b > d$ the region $S(0)$ approximates R. We further explore this issue in the next section.

6 Numerical analysis

In this section we numerically compute the boundary of the realizable region for two realistic examples of Gaussian inputs. As the inputs are non-Brownian, the boundary of the admissible region (and thus the realizable region) has to be obtained numerically. We compare the realizable region with the admissible region corresponding to the priority cases. The following examples illustrate that either $S(0)$ or $S(1)$ (or both) covers most of the realizable region (as was the case for Brownian inputs, see Section 5).

We remark that, in addition to the examples presented here, we have considered many other parameter settings. The result that priority strategies cover nearly the entire realizable region appears to remain valid under quite general circumstances.

6.1 Example 1

Consider two traffic classes sharing a total capacity (C) of 100 Mbps. The first class consists of data traffic, whereas the second class corresponds to voice traffic. Traffic of the first class is modeled as fractional Brownian motion, i.e., $v_1(t) = \alpha t^{2H}$, with $H \in (0, 1)$. The mean traffic rate μ_1 is 0.2 Mbps and its variance function is given by $v_1(t) = 0.0025t^{1.6}$ (such that at time scale $t = 1$ s the standard deviation is 0.05 Mbps). The value of $H = 0.8$ is in line with several measurement studies (one commonly finds a value between, say, 0.7 and 0.85).

Traffic of the second class corresponds to the Gaussian counterpart of the Anick-Mitra-Sondhi (AMS) [2] model, see Section 3.5 of [1]. In the AMS model work arrives from sources in bursts which have peak rate h and Exp(β) distributed lengths. After each burst, the source switches off for an Exp(λ) distributed period. The variance curve of a single source is given by

$$v_2(t) = \frac{2\lambda\beta h^2}{(\lambda+\beta)^3}\left(t - \frac{1}{\lambda+\beta}(1 - \exp(-(\lambda+\beta)t))\right). \tag{13}$$

We choose $h = 0.032$, $\lambda = 1/0.65$ and $\beta = 1/0.352$ in (13), in line with the parameters for coded voice given in Sriram & Whitt [25]. Hence, the mean traffic rate of a source of class 2 (μ_2) is 0.021 Mbps. Note that traffic of class 1 is long-range dependent (i.e., the autocorrelations are

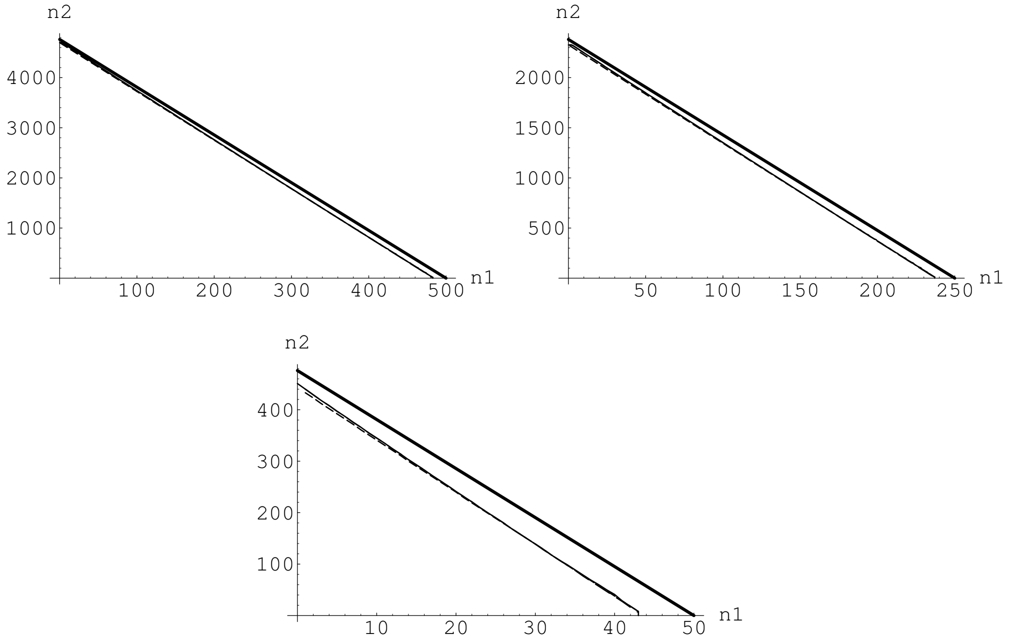

Figure 9: Realizable regions of Example 1. The boundary of $S(1)$ is given by the small dashed line. The long dashed line represents the boundary of $S(0)$. The solid line above the boundary of $S(0)$ and $S(1)$ is the boundary of R. The thick line on top is the boundary of T. Top, left: $C = 100$ Mbps. Top, right: $C = 50$ Mbps. Bottom: $C = 10$ Mbps.

non-summable), whereas the traffic of class 2 is short-range dependent. We allow an overflow probability of 10^{-6} for the first class and 10^{-3} for the second class (corresponding to $\delta_1 \approx 13.8$ and $\delta_2 \approx 6.9$). We choose B_1 such that $B_1/C = 0.05$ (i.e., 50 ms) and B_2 such that $B_2/C = 0.01$ (i.e., 10 ms). Hence we allow a (relatively) large delay but small loss for the data traffic, and a small delay but (relatively) large loss for the voice traffic.

Figure 9 (top, left) depicts the admissible region for the priority cases ($S(0)$ and $S(1)$), the realizable region (R) and the stable region (T). Obviously, $R \subseteq T$, but they almost coincide. Furthermore, the boundaries of $S(0)$, $S(1)$ and R almost match (the boundaries of $S(0)$ and $S(1)$ are hardly visible). That is, most of R can be obtained by giving priority to class 1 or 2. In fact, any weight of $\phi_1 \in [0, 1]$ yields an admissible region $S(\phi_1)$ that closely resembles R.

We have also experimented with other values for C as depicted in Figure 9 (top, right and bottom). As the value of C becomes smaller (with still $B_1/C = 0.05$, $B_2/C = 0.01$, and all other parameters left unchanged), the difference between the boundary of R and T becomes clear. Note that R still closely resembles $S(0)$ and $S(1)$. This indicates that GPS scheduling is only marginally more effective than a strict priority discipline ($\phi_1 = 0$ or $\phi_1 = 1$).

As the values of δ_1 and δ_2 increase, the QoS requirements become more stringent and therefore the difference between the regions R and T becomes more substantial. For large values of C this is hardly visible, and therefore we show this for the case that $C = 10$ Mbps (with the parameter values corresponding to Figure 9 (bottom)).

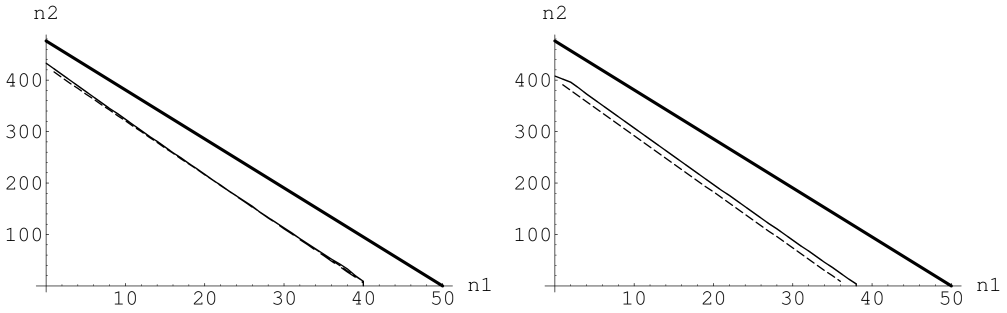

Figure 10: Realizable regions of Example 1. The boundary of $S(1)$ is given by the small dashed line. The long dashed line represents the boundary of $S(0)$. The solid line above the boundary of $S(0)$ and $S(1)$ is the boundary of R. The thick line on top is the boundary of T. Left: $\delta_1 = 27.6$ and $\delta_2 = 13.8$. Right: $\delta_1 = 41.4$ and $\delta_2 = 27.6$.

Figures 9 (bottom) and 10 show the expected impact of the 'QoS requirements' δ_i, $i = 1, 2$. Although the difference between R and T becomes larger as δ_1 and δ_2 increase, R continues to be closely approximated by $S(0)$ or $S(1)$. Compare Figure 9 (bottom) with Figure 10 (left) and observe that if the δ_is are doubled, then R decreases by less than 15 percent.

6.2 Example 2

In this example we let the two traffic classes share a total capacity of 1 Gbps. The traffic of the first class is data traffic with a higher access rate, and traffic of the second class with a lower access rate. The mean traffic rate of a source of the first (second) class is 3 Mbps (0.2 Mbps). The variance functions are given by $0.5625t^{1.6}$ and $0.0025t^{1.6}$, respectively, such that at time scale $t = 1$ s the standard deviations are 0.75 Mbps and 0.05 Mbps, respectively. We allow an overflow probability of 10^{-8} (10^{-3}) for the first (second) class ($\delta_1 \approx 18.4$ and $\delta_2 \approx 6.9$). The buffer thresholds are such that $B_1/C = 0.04$ and $B_2/C = 0.01$.

Figure 11 (top, left) shows the resulting realizable region. Once again, most of R is covered by the admissible region of a priority strategy. Furthermore, also the influence of C is as before, as can be seen in Figure 11 (top, right and bottom). For large values of C, the boundaries of $S(0)$, $S(1)$, R and T almost coincide. As C decreases, the difference between the boundaries of R and T becomes significant. Note also that the difference between the boundaries of $S(0)$ and $S(1)$ becomes visible for small values of C. In all the experiments as depicted in Figure 11, R nearly coincides with $S(1)$.

For $C = 100$ Mbps (setting of Figure 11 (bottom)), Figure 12 depicts the sensitivity with respect to the δ_is. Again, R becomes considerably smaller when the QoS requirements become more stringent (i.e., increasing δ_i, $i = 1.2$). Furthermore, the boundary of R still seems to closely match the boundary of $S(1)$.

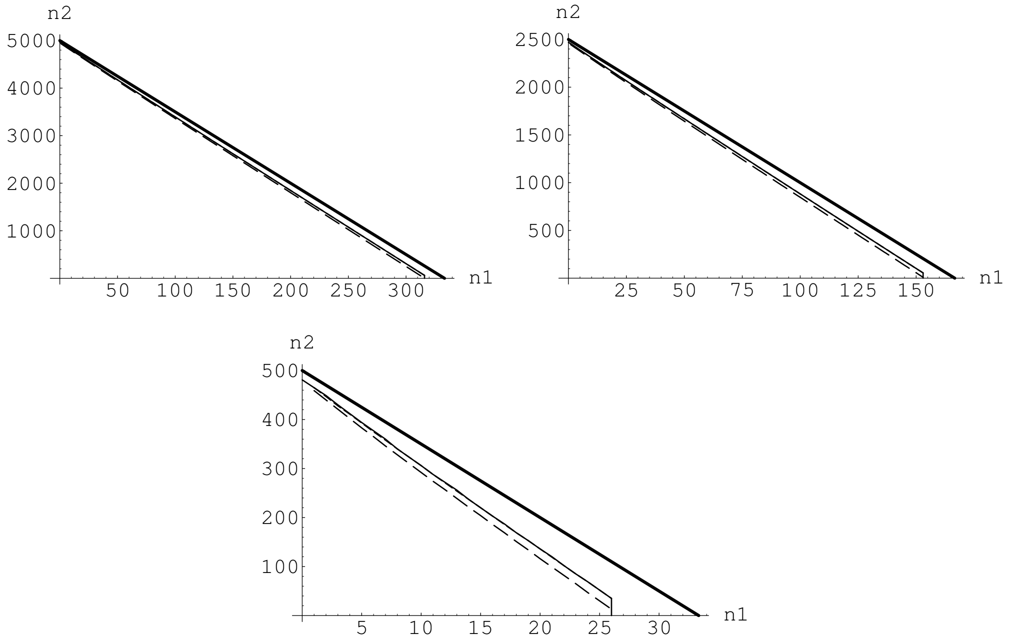

Figure 11: Realizable regions of Example 2. The boundary of $S(1)$ is given by the small dashed line. The long dashed line represents the boundary of $S(0)$. The solid line above the boundary of $S(0)$ and $S(1)$ is the boundary of R. The thick line on top is the boundary of T. Top, left: $C = 1000$ Mbps. Top, right: $C = 500$ Mbps. Bottom: $C = 100$ Mbps.

6.3 Discussion

For larger values of C, the boundaries of T, R, $S(0)$ and $S(1)$ match even better than the ones depicted in Figure 9 (top, left) and Figure 11 (top, left); then the realizable region R nearly coincides with the stability region T. For smaller values of C there is some discrepancy between R and T, particularly when the required overflow probabilities are small. However, it seems that still either $\phi_1 = 0$ or $\phi_1 = 1$ suffices to (nearly) cover R.

Note that the results of the two examples suggest that the boundary of the admissible (realizable) region is approximately linear, which corroborates with the results of Elwalid & Mitra [10].

In case of Brownian inputs, we saw in Section 5 that R was accurately approximated by $S(1)$ if $b \leq d$, and by $S(0)$ otherwise. Therefore, if the ratio of the buffer thresholds is less than the ratio of the (exponential) decay rates of the overflow probabilities, then one should select $(\phi_1, \phi_2) = (1, 0)$, otherwise $(\phi_1, \phi_2) = (0, 1)$. Interestingly, this criterion does not involve the characteristics of the sources. The numerical analysis presented in this section (as well as the additional numerical experiments that we performed) suggest that for other Gaussian sources there is a similar criterion. However, it is in general not given by $b \leq d$ versus $b > d$; it seems that the traffic characteristics of the two classes should be taken into account as well.

We now give some arguments that may informally explain why nearly the entire realizable region is achievable by strict priority scheduling strategies. First consider the scenario that

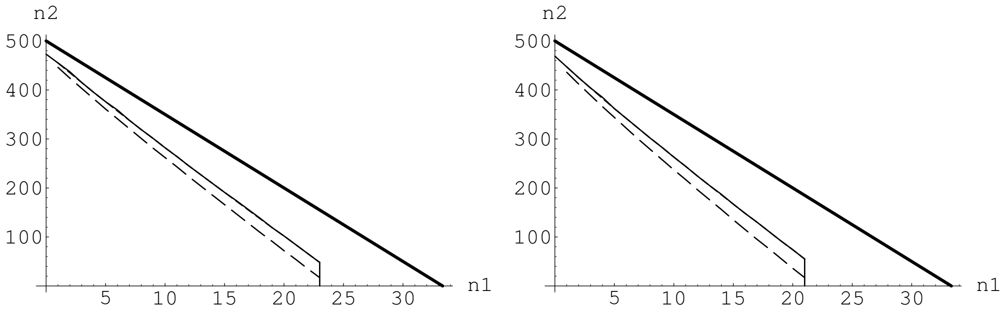

Figure 12: Realizable regions of Example 2. The boundary of $S(1)$ is given by the small dashed line. The long dashed line represents the boundary of $S(0)$. The solid line above the boundary of $S(0)$ and $S(1)$ is the boundary of R. The thick line on top is the boundary of T. Left: $\delta_1 = 36.8$ and $\delta_2 = 13.8$. Right: $\delta_1 = 55.2$ and $\delta_2 = 20.7$.

both classes have similar traffic characteristics. In that scenario the buffer asymptotics of each individual class will resemble that of the aggregate traffic stream, implying that each work-conserving discipline will give similar performance. Hence, since GPS is a work-conserving discipline, the performance of the system is insensitive to the weights in this scenario. Now, consider the scenario that one class has heavy traffic and loose QoS requirements, whereas for the other class it is the reverse (smooth traffic and stringent QoS requirements). Then the buffer asymptotics of the bursty traffic class will not be affected by the weights (may be even completely insensitive), as long as the traffic intensity of the smooth traffic class does not exceed its weight. The latter will necessarily hold, as otherwise the heavy traffic class would be negatively influenced by the smooth traffic class. This insensitivity implies that there is little lost by simply giving strict priority to the smooth traffic class. The only scenario that remains is where the bursty traffic class has tighter QoS requirements than the smooth traffic class, but that scenario appears to be atypical.

Our choice to focus in this paper on two-class GPS is motivated by the fact that most traffic can be categorized into streaming and elastic traffic. In general, in order to keep the complexity as low as possible, one should attempt to minimize the number of classes. However, the weight setting problem in the case with more than two classes is not fundamentally different from the two-class case; we expect our conclusions to carry over.

7 Conclusions

In this paper we determined the admissible region for a two-class GPS system with Gaussian traffic sources. The analysis relied on the powerful large-deviations approximations of [19, 17]. These are particularly useful, as they cover general correlation structures, thus allowing both short-range dependent and long-range dependent traffic processes, and avoid the rather restrictive traffic assumptions in previous work.

We showed that the admissible region for each class may be partitioned into three subsets, which facilitated the derivation of the joint admissible region for both classes by taking the intersection for given weight values. We then obtained the realizable region as the union of the admissible regions over all possible weight values.

In the case of Brownian inputs, the boundary of the admissible region can be explicitly derived, and it can be shown that nearly the entire realizable region is achievable by simple priority strategies. A further key observation is that the choice of which class to prioritize is entirely determined by the Quality-of-Service requirements, particularly the ratio of the buffer thresholds compared to the ratio of the exponential decay rates of the violation probabilities. Thus, the proper priority ordering is not influenced by the traffic characteristics or even the number of sources, but of course how many sources of the two classes actually can be supported *does* strongly depend on the statistical traffic properties. Extensive numerical experiments indicated that these remarkable findings also hold for general Gaussian traffic sources. The results suggest that the precise selection of scheduling weights is not that critical, and that simple priority strategies may suffice for practical purposes.

Acknowledgment. The authors thank Sem Borst for valuable discussions.

References

[1] R. Addie, P. Mannersalo, I. Norros (2002). Most probable paths and performance formulae for buffers with Gaussian input traffic. *European Trans. Telecommun.*, 13: 183–196.

[2] D. Anick, D. Mitra, M. Sondhi (1982). Stochastic theory of a data handling system with multiple resources. *Bell Syst. Techn. J.*, 61: 1871–1894.

[3] D. Bertsimas, I. Paschalidis, J. Tsitsiklis (1999). Large deviations analysis of the generalized processor sharing policy. *Queueing Systems*, 32: 319–349.

[4] S. Borst, O. Boxma, P. Jelenković (2000). Asymptotic behavior of Generalized Processor Sharing with long-tailed traffic sources. In: *Proc. IEEE Infocom 2000*, Tel-Aviv, Israel, 912–921.

[5] S. Borst, O. Boxma, P. Jelenković. (2003). Reduced-load equivalence and induced burstiness in GPS queues with long-tailed traffic flows. *Queueing Systems*, 43: 273–306.

[6] S. Borst, M. Mandjes, M. van Uitert. (2002). GPS queues with heterogeneous traffic classes. In: *Proc. IEEE Infocom 2002*, New York, USA, 74–83.

[7] S. Borst, M. Mandjes, M. van Uitert (2003). Generalized Processor Sharing queues with light-tailed and heavy-tailed input. *IEEE/ACM Trans. Netw.*, 11: 821–834.

[8] K. Dȩbicki, M. Mandjes (2003). Exact overflow asymptotics for queues with many Gaussian inputs. *J. Appl. Prob.*, 40: 704–720.

[9] N. Dukkipati, J. Kuri, H. Jamadagni (2001). Optimal call admission control for generalized processor sharing (GPS) schedulers. In: *Proc. IEEE Infocom 2001*, Anchorage, USA, 468–477.

[10] A. Elwalid, D. Mitra (1999). Design of generalized processor sharing schedulers which statistically multiplex heterogeneous QoS classes. In: *Proc. IEEE Infocom 1999*, New York, USA, 1220–1230.

[11] C. Fraleigh, F. Tobagi, C. Diot (2003). Provisioning IP backbone networks to support latency sensitive traffic. In: *Proc. IEEE Infocom 2003*, San Francisco, USA.

[12] J. Kilpi, I. Norros (2002). Testing the Gaussian approximation of aggregate traffic. In: *Proc. Internet Measurement Workshop*, Marseille, France.

[13] C. Kotopoulos, N. Likhanov, R. Mazumdar (2001). Asymptotic analysis of the GPS system fed by heterogeneous long-tailed sources. In: *Proc. IEEE Infocom 2001*, Anchorage, Alaska, USA, 299–308.

[14] K. Kumaran, G. Margrave, D. Mitra, K. Stanley (2000). Novel techniques for the design and control of generalized processor sharing schedulers for multiple QoS classes. In: *Proc. IEEE Infocom 2000*, Tel-Aviv, Israel.

[15] M. Mandjes (2004). A note on the benefits of buffering. *Stochastic Models*, 20: 43–54.

[16] M. Mandjes (2005). Large deviations for complex buffer architectures: the short-range dependent case. To appear in: *Stochastic Models.*

[17] M. Mandjes, M. van Uitert (2005). Sample-path large deviations for generalized processor sharing queues with Gaussian inputs. *Perf. Eval.*, 61: 225–256.

[18] M. Mandjes, M. van Uitert (2005). Sample-path large deviations for tandem and priority queues with Gaussian inputs. *Ann. Appl. Prob.*, 15: 1193–1226.

[19] P. Mannersalo, I. Norros (2002). GPS schedulers and Gaussian traffic. In: *Proc. IEEE Infocom 2002*, New York, USA, 1660–1667.

[20] L. Massoulié (1999). Large deviations estimates for polling and weighted fair queueing service systems. *Adv. Perf. Anal.*, 2: 103–128.

[21] A. Panagakis, N. Dukkipati, I. Stavrakakis, J. Kuri (2004). Optimal call admission control on a single link with a GPS scheduler. *IEEE/ACM Trans. Netw.*, 12: 865–878.

[22] A. Parekh, R. Gallager (1993). A generalized processor sharing approach to flow control in integrated services networks: The single node case. *IEEE/ACM Trans. Netw.*, 1: 344–357.

[23] A. Parekh, R. Gallager (1994). A generalized processor sharing approach to flow control in integrated services networks: The multiple node case. *IEEE/ACM Trans. Netw.*, 2: 137–150.

[24] J. Roberts (2001). Traffic theory and the Internet. *IEEE Communications Magazine*, 39: 94–99 (Issue 1).

[25] K. Sriram, W. Whitt (1986). Characterizing superposition arrival processes in packet multiplexers for voice and data. *IEEE J. Sel. Areas Commun.*, 4: 833–846.

[26] M. van Uitert, S. Borst (2001). Generalised Processor Sharing networks fed by heavy-tailed traffic flows. In: *Proc. IEEE Infocom 2001*, Tel-Aviv, Israel, 269–278.

[27] M. van Uitert, S. Borst (2002). A reduced-load equivalence for Generalised Processor Sharing networks with long-tailed traffic flows. *Queueing Systems*, 41: 123–163.

[28] O. Yaron, M. Sidi (1994). Generalized Processor Sharing networks with exponentially bounded burstiness arrivals. In: *Proc. IEEE Infocom 1994*, 628–634.

[29] Z.-L. Zhang. (1998). Large deviations and the generalized processor sharing scheduling for a multiple-queue system. *Queueing Systems*, 28: 349–376.