Fair rate sharing models in a CDMA link with multiple classes of elastic traffic

Ioannis Koukoutsidis, Eitan Altman, Jean Marc Kelif

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Fair rate sharing models in a CDMA link with multiple classes of elastic traffic

Ioannis Koukoutsidis — Eitan Altman — Jean Marc Kelif

N° 5596

June 2005

Thème COM

ISSN 0249-6399 ISRN INRIA/RR--5596--FR+ENG

INRIA
SOPHIA ANTIPOLIS

Fair rate sharing models in a CDMA link with multiple classes of elastic traffic

Ioannis Koukoutsidis*, Eitan Altman*, Jean Marc Kelif †

Thème COM — Systèmes communicants
Projet MAESTRO

Rapport de recherche n° 5596 — June 2005 — 35 pages

Abstract: In this paper we describe a modeling approach for studying fair rate sharing on a CDMA link. Capacity models derived for CDMA indicate that fair rate sharing belongs to the class of *generalized processor sharing* (GPS) schemes, as these were defined and studied by J.W. Cohen. From this starting point, we examine the steady-state characteristics of flows on a CDMA link considering multiple classes of non-real-time, or elastic, traffic. These permit us to evaluate the expected transfer times of flows and their blocking probabilities, in loss systems. We study traffic models with Poisson arrivals as well as arrivals from a finite-source population, in an Engset-like manner. Along the way we unveil some interesting properties of the GPS model –partially hidden in Cohen's work– and extend some of its results. In this context, we revisit insensitivity and truncation properties of the stationary distributions encountered in GPS models, and extend to different access control policies.

Key-words: CDMA, elastic traffic, fairness, generalized processor sharing, insensitivity, access control

This work was supported by a CRE research contract with France Telecom R&D and by the EuroNGI network of excellence.

* {Giannis.Koukoutsidis}{Eitan.Altman}@sophia.inria.fr
† {JeanMarc.Kelif}@francetelecom.com, France Telecom R&D, Rue du Général Leclerc, 92794 Issy-les-Moulineaux Cedex 9

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles, BP 93, 06902 Sophia Antipolis Cedex (France)
Téléphone : +33 4 92 38 77 77 —Télécopie : +33 4 92 38 77 65

Modèles de partage équitable du débit sur un lien CDMA avec plusieurs classes de traffic élastique

Résumé : Dans cet article nous décrivons la modélisation du partage équitable du débit entre plusieurs sources sur un lien CDMA. L'étude de la capacité dans une cellule CDMA indique que le modèle du partage équitable du débit appartient à la classe de modèles suivant la discipline *processeur partagé généralisé* (GPS), comme celle-ci a été définie et étudiée par J.W. Cohen. De ce point de départ, nous examinons le comportement des flux de données dans l'état stationnaire, en considérant plusieurs classes de traffic non temps-réel, également appelé de type élastique. Les mesures de performance sont principalement l'espérance du temps de séjour des flux de données, aussi que leurs probabilités de blocage. Nous analysons des modèles de traffic avec des arrivées de type Poisson, ainsi que des arrivées par une population limitée, d'une façon similaire au modèle d'Engset. Notre travail permet également de démontrer quelques propriétés intéressantes du modèle GPS et de produire quelques résultats supplémentaires. Dans ce contexte, nous revisitons les propriétés d'insensibilité et de troncature des distributions stationnaires rencontrées dans les modèles GPS, ce qui nous permet d'étendre les résultats aux différentes politiques d'accès.

Mots-clés : CDMA, traffic élastique, partage équitable, processeur partagé généralisé, insensibilité, contrôle d'accès

Contents

1 Introduction

A wireless link in a CDMA network is characterized by its capacity and total throughput, under specific channel environment conditions. At the level of traffic sources transmitting in the link, the interest is shifted to the kind of resource-sharing performed between the sources and the throughput that results for each source.

The special case of *fair-rate sharing* is particularly important, since it may be desired that mobiles in a cell that belong to the same service class transmit or receive at the same rate, regardless of their position in the cell. Then, given a number of active or present mobiles in a cell, we would like to find out what can be the maximum throughput for each mobile, and how a stochastically varying number of these affects QoS parameters of the system.

Previous theoretical analyses of capacity and throughput in the uplink and downlink of a CDMA system ([14],[20]) have adressed our first inquiry and provided a basis of how to master the second. Most importantly, they have indicated that fair rate sharing appropriately fits the class of *generalized processor sharing* (GPS) models defined and studied by J.W.Cohen [10]. This permits us to employ known results regarding the steady-state characteristics of the system, in order to derive performance measures of interest for our study of traffic on the link.

Modern wireless communication networks aim to carry various types of applications on the same medium and define several classes of service. For example, the 3GPP standards for UMTS envisage four service classes, driven by specific types of applications [1]: conversational (VoIP, audio and video conferencing), streaming (broadcast services), interactive (web browsing), and background (e-mail, file transfers). For modeling purposes, we can more generally distinguish traffic in two categories: *real time* or *streaming*, and *non-real-time* or *elastic*. We remark that traffic of the latter type is more apt for a processor sharing setting, as there exists no guaranteed bit rate. Real-time and (although to a lesser extent) streaming traffic are characterized by an intrinsic rate and/or duration, which require much more stringent QoS guarantees and thus more complex multiplexing policies[1].

In this paper, we focus only on the transmission of non-real-time or elastic traffic, whose rate can be "more freely" adjusted. We will use and interchange in a not so strict manner the characterizations "non-real-time" and "elastic", although they may be distinguishable. The main attribute of interest is the "rate elasticity" of this kind of traffic. The major performance metrics we examine are then the transfer times of flows over the link and their blocking probabilitites, in loss systems. We are also interested in the interaction of multiple classes of traffic and different access control policies that can be applied for a certain class, such as *common* and *dedicated* access.

We analyze two major arrival models; the case of Poisson arrivals and that of finite-source arrivals in a closed system with think times, a variant of the widely known Engset model. Under generalized processor sharing, these are called here as the *GPS-Poisson* and *GPS-Engset* model, respectively. Apart from the specific CDMA problem analysis, we manage to extend the wealth of very general results of Cohen in various ways. First, the well-known result regarding the proportionality of the expected sojourn time of a deterministic service job to its service volume is shown to hold for all blocking and non-blocking cases, in the Poisson arrival model. We also present as a conjecture a more general formula regarding the expected sojourn time of jobs in the case of a multiple class, GPS-Engset, system, with or without blocking. Based on that, we also derive the analogous result regarding deterministic service job times in the GPS-Engset system with blocking. In the same system, we extend the formula that yields blocking probabilities for different class jobs. Finally, we manage to demonstrate more general facts regarding the *insensitivity* properties, ubiquitous throughout Cohen's results, and the ensuing truncation principles that can be applied in a blocking system.

[1] Even if we assume the existence of adaptive real-time compression algorithms, which can compensate for some rate reduction.

2 Modeling capacity and throughput in the CDMA link

Let us first start by giving a more precise characterization of the non-real-time or elastic traffic that passes through the link. Modern networks employ packet transmission, however we need to go a level of abstraction above packets for our modeling purposes. In so doing, we use the term 'flow' to refer to distinct 'jobs' in the system. Similarly to [13], a flow represents a stream of packets that have some criteria in common. Since we are considering a single link here, these criteria are mostly related to an application purpose (i.e. the transfer of a file or document, browsing of web pages, etc.). We will also associate a 'flow' with a 'user' or a 'mobile device' from which a transfer is originated. Further, we make the simplifying assumption that there is no need for packet retransmissions (e.g. by the existence of an appropriate forward error correction scheme) and the transfer of a flow is completed in the time suggested by its initial volume.

The starting point of the capacity and associated throughput analysis in a CDMA link is the formula relating the carrier to interference threshold, $(C/I)_s$, that should be satisfied at the receiver side for transmissions of a connection s, the energy per bit to noise density requirement, $(E_b/N_0)_s$, and the processing gain, which is the chip rate W divided by the rate of transmission R_s:

$$(C/I)_s = \left(\frac{E_b}{N_0}\right)_s \cdot \frac{R}{W}. \tag{1}$$

The subsequent analysis is based on defining the link's capacity as a function of the number of users that the system can theoretically sustain without the total power going to infinity; this being subject to the constraint that C/I threshold requirements for all users (in the downlink) and the cell base station (in the uplink) are satisfied.

Next we recall essential parts of this analysis presented in [14],[20] for the uplink and downlink of a CDMA cell, respectively. We confine ourselves to the case of homogeneous service for all mobiles, and emphasize on the maximum throughput that can be obtained for each.

2.1 Uplink analysis

Consider an arbitrary cell in the CDMA network, with M mobiles transmitting towards the base station, and assume perfect power control. For each mobile j ($j = 1 \dots M$), in order for its signal to be received properly, the ratio of its minimum received power to the sum of background noise and interference must be equal to some constant, $\tilde{\Delta}$. This condition writes:

$$\frac{P_j}{N + I_{own} + I_{other} - P_j} = \tilde{\Delta}, \tag{2}$$

where N is the background noise, I_{own} is the total power received from mobiles in the same cell, and I_{other} is the total power received from mobiles in other cells. From (1), the constant $\tilde{\Delta}$ depends on the type of service and rate of transmission, and thus is taken to be the same for all mobiles here. We have

$$I_{own} = \sum_{k=1}^{M} P_j.$$

To model the intercell interference, we make the standard simplifying assumption (cf. [17]) that

$$I_{other} = f_u \cdot I_{own},$$

for some given constant f_u obtained from measurements.

We find it more useful to rewrite (2) as

$$\frac{P_j}{N + I_{own} + I_{other}} = \Delta, \tag{3}$$

where

$$\Delta = \frac{\tilde{\Delta}}{1+\tilde{\Delta}} \;\Leftrightarrow\; \tilde{\Delta} = \frac{\Delta}{1-\Delta}. \tag{4}$$

Solving the system of M equations (3) yields

$$P_j = \frac{N\Delta}{1-[(1+f_u)\Delta]M}. \tag{5}$$

Given a certain service, the 'pole' or 'integer' capacity of the system can be defined as the number of mobiles M that makes the denominator of (5) vanish. Conversely, given a certain number of mobiles, the upper bound on Δ is $\frac{1}{M(1+f_u)}$. From this, with the help of (1),(4), we may compute the theoretical maximum throughput of each connection, when all mobiles transmit *simultaneously*[2]:

$$R_{max}^{(UL)} = \frac{1}{M(1+f_u)-1}\left(\frac{N_0}{E_b}\right)_u W. \tag{6}$$

This is the theoretical upper bound on throughput under a fair-rate sharing scheme in the limiting case where the received power from a mobile goes to infinity. In practice, due to physical limitations in the power of a mobile device the throughput will be much smaller. One can model this by taking a value $\Theta_u < 1$, such that $M(1+f_u)\Delta = \Theta_u$. We define this as the total *resource capacity* of the link. We then have for the throughput of a single mobile:

$$R^{(UL)} = \frac{\Theta_u}{M(1+f_u)-\Theta_u}\left(\frac{N_0}{E_b}\right)_u W. \tag{7}$$

Thus, the allocated throuhgput depends on the transmission environment and the total number of simultaneously transmitting mobiles, as well as the available bandwidth and E_b/N_0 requirements. The above analysis also shows that the maximization of throughput depends on maximum power constraints of a mobile device. In this sense, mobiles which are more distant from the base station must consume more power, and thus will determine the fair-rate throughput.

2.2 Downlink analysis

The analysis in the downlink follows the same lines, but is more involved. Let there be S base stations in a CDMA network with perfect power control. The minimum power that should be received at a mobile from a base station is again determined by condition (1), concerning the carrier to interference ratio for a certain service and transmission rate. Let now $P_{k,\ell}$ be the power transmitted to mobile k from the BS ℓ, $\ell = \{1,\ldots,S\}$. Assume that there are M mobiles in cell ℓ; the BS of that cell transmits at a total power P_{tot}^{ℓ}, given by

$$P_{tot}^{\ell} = \sum_{k=1}^{M} P_{k,\ell} + P_{SCH} + P_{CCH}. \tag{8}$$

P_{SCH}, P_{CCH} is the power transmitted for the synchronization and the control information common channels, respectively[3]. Note that these are not power controlled, and so they can be modeled by adding a constant power (this should further be calculated for the worst case user at the cell edge).

[2]In this paper, we consider the standard case where mobiles share the total throughput of the link by transmitting or receiving simultaneously. There exist also high data rate schemes, like the HSDPA scheme in WCDMA ([17]) and the HDR scheme in CDMA2000 ([4]), where the whole link capacity is allocated to one user for a very short-time, when conditions are favorable. These schemes implement a complex scheduler which evaluates channel conditions and pending transmissions for each connection to decide on which user should transmit. It is reasonably anticipated that even the fair-rate implementation of such schemes exhibits improved performance. Preliminary results on this matter were presented in [23].

[3]We include here all pilot, paging and timing signals.

Despite the use of orthogonal signaling in the downlink, due to the multipath propagation a fraction α_k of the received own cell power is experienced as intracell interference by the mobile k (non-orthogonality factor)[4]. Denoting by I^k_{intra} and I^k_{inter} the intracell and intercell interferences perceived by mobile k, respectively, and by $g_{k,\ell}$ the attenuation between base station ℓ and mobile k, we have

$$I^k_{intra} = \alpha_k \cdot (P_{SCH} + P_{CCH} + \sum_{j=1, j\neq k}^{M} P_{j,\ell})/g_{k,\ell}, \tag{9}$$

$$I^k_{inter} = \sum_{j=1, j\neq \ell}^{S} P^j_{tot}/g_{k,j}. \tag{10}$$

We then consider a homogeneous service and transmission rate for all mobiles in the cell and assume minimum transmitted powers such that for all $k = \{1, \ldots, M\}$:

$$\frac{P_{k,\ell}/g_{k,\ell}}{I^k_{inter} + I^k_{intra} + N} = \tilde{\Delta}, \tag{11}$$

where N is the background noise[5] and $\tilde{\Delta}$ is the assumed target ratio in the downlink. We define

$$F_{k,\ell} = \frac{\sum_{j=1, j\neq \ell}^{S} P^j_{tot}/g_{k,j}}{P^\ell_{tot}/g_{k,\ell}}, \tag{12}$$

i.e. the ratio between the received intercell and intracell power. Then, we find it again more useful to write

$$\frac{P_{k,\ell}/g_{k,\ell}}{(F_{k,\ell} + \alpha_k)P^\ell_{tot}/g_{k,\ell} + N} = \Delta_k, \tag{13}$$

where $\Delta_k = \frac{\tilde{\Delta}}{1+\alpha_k\tilde{\Delta}}$.

At this point we make a crucial remark: in CDMA dimensioning, it is important to estimate the total amount of base station power required. This should be based on the *average* transmission power for a user, not the *maximum* transmission power for the cell edge. In order to calculate the total base station power in our problem, we should normally solve for the individual transmission powers in (13) and substitute them in (8). However, even if we know the parameters of each individual connection this does not result in any useful dimensioning on the link. So the approach we follow is to calculate the minimum transmitted power for a mobile at an "average location" within the cell. Formally, this means that we take the sample average for $g_{k,\ell}$, $F_{k,\ell}$, α_k over all $k = 1, \ldots, M$.

The average approximation is used in all downlink dimensioning models of CDMA (see e.g. [17],[16], [29]), as it provides an easy way to estimate the pole capacity. Additionally, as was performed in [16], the accuracy of the approximative parameters can be improved by curve fitting, based on measurements for the total output power of the base station.

Replacing $g_{k,\ell}$, $F_{k,\ell}$, α_k by single parameters G, f_d, α and denoting $\Delta = \frac{\tilde{\Delta}}{1+\alpha\tilde{\Delta}}$, we finally get for the total output power of base station ℓ (we omit the index ℓ):

$$P_{tot} = \frac{P_{SCH} + P_{CCH} + N \cdot G \cdot \Delta \cdot M}{1 - (\alpha + f_d)\Delta \cdot M}.$$

Further assuming that the power in the synchronization and control common channels is a fraction ψ of the total output power, i.e. $P_{SCH} + P_{CCH} = \psi P_{tot}$, we get the simple expression:

$$P_{tot} = \frac{N \cdot G \cdot \Delta \cdot M}{1 - \psi - (\alpha + f_d)\Delta \cdot M}. \tag{14}$$

[4]For mathematical convenience, α_k weights on the total received power.

[5]The background noise may or may not depend on the mobile's position. For simplicity, we assume it does not.

Then, in the same way as in the uplink, we get the upper bound on Δ to be $\frac{1-\psi}{(\alpha+f_d)M}$, from which the theoretical maximum throughput of each connection when the base station transmits simultaneously to all mobiles, is found as:

$$R_{max}^{(DL)} = \frac{1-\psi}{M(\alpha+f_d)-\alpha(1-\psi)} \left(\frac{N_0}{E_b}\right)_d W. \tag{15}$$

We impose a limit on the total output power of a base station by taking a value of total capacity $\Theta_d < 1-\psi$, such that $M(\alpha+f_d)\Delta = \Theta_d$. Then the throughput of a single mobile is

$$R^{(DL)} = \frac{\Theta_d}{M(\alpha+f_d)-\alpha\Theta_d} \left(\frac{N_0}{E_b}\right)_d W. \tag{16}$$

Similarly to the uplink, the allocated fair throughput depends on transmission environment conditions, the total number of active mobiles in the downlink, the available bandwidth and E_b/N_0 requirements. Clearly, the base station must transmit stronger signals at mobiles with less favorable channel conditions (notably, mobiles near the cell edge) in order to maintain the same rate.

We end the whole section with some necessary remarks.

Remark 2.1. Both in the uplink and downlink, we assume that there exists for the system a stable state, where the C/I target is satisfied for all mobiles in the cell by an ideal version of closed-loop power control, in which case the transmit and receive powers are deterministic. In practice, these powers are constantly fluctuating for all users, especially in the uplink, because of random fading effects. In [14], the authors showed how shadow fading effects (log-normal fading channels) can be incorporated in the analysis, by multiplying the target C/I ratios by some constant that depends on log-normal fading characteristics.

Remark 2.2. Another difficulty in implementing closed-loop power control arises for data transmissions. More specifically, the time to transmit a packet may be too short for feedback control to converge. We will disregard this deficiency and assume that in our model, flows are large enough to allow power control to converge. In addition, it can be argued that longer transmission times induced by processor-sharing make closed-loop power control more applicable.

Remark 2.3. It is worth noting that one may also include in the analysis the *channel activity factors*. These indicate the portion of time, in a communication, that mobiles actually trasmit. Inactivity periods are very useful in CDMA, not only for energy conservation at the mobile devices, but mainly because with less interference the capacity of the system is increased. In this paper, we associate a mobile with a flow; thus we consider flows transmitted between sleep periods of a mobile as *separate*. More generally however, one may include channel activity factors in the above analysis by introducing an indicator random variable taking values 0 and 1 depending on whether the mobile is active or not, see e.g. [30].

Remark 2.4. Despite our constant use of the word 'mobiles', mobility will not be considered in the context of this paper. It can be argued that we take a 'snapshot' of a system with mobility, with certain E_b/N_0 targets and interference, and the above analysis is still valid in that case. However, mobility along with fading effects (especially fast fading) cause estimation problems for the power control function and make the practical implementation of fair-rate sharing almost impossible. Hence, it should be considered that mobile devices are fixed terminals, or that they move slightly around their positions so that transmitted signals and interferences in the cell do not additionally change because of mobility. Besides, the case of static or quasi-static users is usually the default in the majority of data transfers.

3 The processor-sharing model

Based on the above capacity and throughput analysis, we now examine more carefully the processor-sharing model. It is repeated that we consider the case where a number of mobile terminals transmit or receive flows of packets simultaneously on the link, under fair-rate sharing. In so doing, we must assume very fast closed-loop power control, as well as completely *fluid* traffic subject to an ideal transmission rate control, with negligible feedback delay between the receiver and the source.

Remark 3.1. Regarding further the implementation aspects of such a system, we reason that it is very difficult to assign transmission rates beforehand with the best possible fair utilization of the link's capacity, due to throughput fluctuations in a CDMA link. It is more probable that a self-adjusting control protocol, with combined rate and power control, can be applied; the base station should constantly monitor the signal power and transmission rate to/from each mobile, and send feedback regarding the gradual modification of both values, in an attempt to reach the best feasible state. Provided that users are almost static and transmissions are long enough compared to the time of convergence of the joint power and rate control scheme, fair-rate sharing could be of practical importance.

Remark 3.2. The fairness of such a processor-sharing policy is an issue that demands more discussion. Generally speaking, the objective of a transmission protocol or scheduler is to achieve a link utilization as high as possible, while at the same time ensuring a fair treatment for all flows. In our assumed setting, the link is utilized at full and all the flows receive the same rate, but we still risk unfairness. For example, if we consider a discrimination to classes of mobiles, it is easy for a single class of flows to dominate the link and thus leave little space for others. Therefore, we must insist on applying access control policies on the number of different class flows in the system, which can in part amend the problem. More discussion on these aspects is included in § 6.

For a number of n flows, emitted from/to an equal number of mobiles, summing up the corresponding individual throughputs in (7),(16), we write down the total throughput as a function of the number of flows present:

$$\textit{Uplink:} \quad R_{tot}(n) = \frac{n\Theta_u}{n(1+f_u)-\Theta_u}\left(\frac{N_0}{E_b}\right)_u W. \tag{17}$$

$$\textit{Downlink:} \quad R_{tot}(n) = \frac{n\Theta_d}{n(\alpha+f_d)-\alpha\Theta_d}\left(\frac{N_0}{E_b}\right)_d W. \tag{18}$$

Based on the above formulae, one can construct a general model as follows: if n is the number of requests for transmission on a link, the service rate for each of these is $f(n)$, where $f(\cdot)$ is an arbitrary positive function, and the total service rate is $n \cdot f(n) = R_{tot}(n)$. We have the following constraints:

$$\begin{aligned} 0 \leq \ & f(n) < \infty, \\ & n \cdot f(n) < \infty. \end{aligned}$$

The preceding formulation corresponds to a processor sharing problem with *equal* but *time-varying* service allocation, the service rates varying with time as flows randomly enter and leave the system. The total service rate is likewise time-varying. In our problem, it is a decreasing function of n and attains a limit $R_{min} = \lim_{n\to\infty} R_{tot}(n)$ as the number of flows tends to infinity. This manifests the impact of interference, which restrains the total throughput on the link.

It can readily be seen or verified that this model belongs to the class of *generalized processor sharing* models, as these were defined and studied by Cohen[6]. This permits us to incorporate results regarding the stationary state distributions of a system with Poisson arrivals of flows, as well as Engset-like arrivals from a finite source population, directly from [10]. Based on these we evaluate critical

[6]Nowadays it has prevailed that the term GPS be used for another class of queueing models, namely *Weighted Fair Queueing* [24]. However we prefer to stick to the original denomination in this work.

performance measures in our system. In the description of the steady-state characteristics that follows we present the general case of multiple classes of flows. In all cases, the ramifications of the analysis in our system are discussed and extensions to Cohen's work, where needed, are provided. It is worth noting that the stationary distribution for the GPS models can also be derived from the BCMP model analysis [3], both for the Poisson and Engset systems. However, the derivation is awkward since we have to 'tailor' the solution. It is the analysis of Cohen that sheds light into the characteristics of the problem and allows various performance measures to be derived.

4 Steady-state characteristics

We consider that *classes* of flows, with different arrival and required service characteristics are accommodated in the link. A class here may represent a different non-real-time or elastic application (e.g. web browsing, e-mail, file transfer), or the same application but with a different set of users that employ it (e.g. with different behavioral characteristics). However, flows are assumed to be served with the same discipline: irrespective of the class, when a flow enters the system it receives service at the same rate as a flow of any other class. Thus, schemes with different service or more complex capacity-sharing schemes, such as reservation or preemptive priority of service for one class and allocation and de-allocation of the unused capacity to other classes, are not covered here. Although analytical approaches are sometimes possible [23], such schemes are notoriously difficult to solve, see for example [11] and the references therein.

Before getting into the queueing theoretical context we deploy, it is noted that having in mind non real-time applications –whose transfer time depends on the service rate received– we replace the notion of a service time, for a flow entering the link, by that of a *service requirement*[7]. The service requirement is translated here to the size, or traffic volume of the flow to be transmitted over the link.

4.1 The GPS-Poisson model

Here we consider K different classes of flows that arrive into the system according to independent Poisson processes of rate λ_k, $k = 1, 2, \ldots, K$. Flow sizes of each Poisson stream are independent, identically distributed random variables. In addition, flow sizes are also independent between different streams. We denote these service requirement distributions by $F_\sigma^k(\sigma)$ with first moments $E[\sigma_k]$. We also define a *load* parameter of each Poisson stream, as $\rho_k := \lambda_k \cdot E[\sigma_k]$.

For this system with GPS service, Cohen derives the joint stationary distribution of the number of customers of each class, N_k, to assume a value x_k ($k = 1, \ldots, K$) and their attained services, denoted by the vector $\overline{\sigma}_k = (\sigma_k(1), \ldots, \sigma_k(x_k))$, by finding the solution to a system of integro-differential equations. We have for the defined probability $p(x_k, \sigma_k(h);\ h = 1, \ldots, x_k,\ k = 1, \ldots, K) d\overline{\sigma}_k := \Pr\{N_k = x_k, \sigma_k(h) \leq \boldsymbol{\sigma_k}(h) \leq \sigma_k(h) + d\sigma_k(h);\ h = 1, \ldots, x_k,\ k = 1, \ldots, K\}$[8] the density (Theorem 7.2 of [10]):

$$\begin{aligned} p(x_k, \sigma_k(h); h = 1, \ldots, x_k;\ k = 1, \ldots, K) = \\ = p_{\overline{0}} \cdot \phi(x) \prod_{k=1}^{K} \frac{(\rho_k)^{x_k}}{x_k!} \prod_{h=1}^{x_k} \frac{1 - F_\sigma^k(\sigma_k(h))}{E[\sigma_k]}, \end{aligned} \tag{19}$$

for a system with no limit on the number of flows, under the assumption that the service requirement distributions are *absolutely continuous* and have a rational Laplace-Stieltjes transform[9]. Here we denote

[7] Obviously, if the service requirement were worked off at a constant rate in the link these two concepts would be equivalent; but this is not the case here.

[8] The bold notation in $\boldsymbol{\sigma_k}(h)$ is just used to distinguish the random variable.

[9] Cohen imposes this condition in the infinite case because it is difficult to show the uniqueness of the solution of the system of integro-differential equations, and he resorts to the method of stages. However, the results hold for more

$x = x_1 + x_2 + \cdots + x_K$ and

$$p_{\overline{0}} = \left(\sum_{z=0}^{\infty} \frac{\rho^z}{z!} \cdot \phi(z) \right)^{-1}$$

is the probability that the system is empty. This is also termed as the *normalization constant* of the system. By $\phi(n) = (\prod_{i=1}^{n} f(i))^{-1}$ we denote a very useful function, encountered everywhere in the course of our study. By definition, $\phi(0) = 1$. Above any of the $\boldsymbol{\sigma_k}$ may be taken to mean the residual service time, since the density function is the same. It is worth noting that Eq. (19) says that *given* the number of flows in the system, the attained (and also residual) services of these flows are independent. This is a common attribute to all the models studied here.

By integrating (19) over all σ_k, we get the stationary distribution of the number of flows in the system from each class, given that the total number of flows is x. We denote this as:

$$p_{x_1,x_2,\ldots,x_K}(x) = \frac{\prod_{k=1}^{K} \frac{(\rho_k)^{x_k}}{x_k!} \cdot \phi(x)}{\sum_{z=0}^{\infty} \frac{\rho^z}{z!} \cdot \phi(z)}. \tag{20}$$

This expression depicts the famous insensitivity property, as it depends on the service time distributions only through their means.

Note that this is also the same system studied by Kelly, in his context of 'symmetric queues' [19], if we consider the total service effort $i \cdot f(i)$ for a total number of flows i and the corresponding parameter $g(n) := \prod_{i=1}^{n} i \cdot f(i)$. Following this analysis, we have that the distribution of the total number of flows in the system, N, is given by the following expression:

$$\Pr\{N = n\} = G^{-1} \frac{\rho^n}{g(n)}, \tag{21}$$

where the normalization constant

$$G := \sum_{z=0}^{\infty} \frac{\rho^z}{g(z)}.$$

We will study the condition under which this stationary distribution exists in our system. From the last expression, it is obvious that the following must be satisfied:

$$\sum_{z=0}^{\infty} \frac{\rho^z}{g(z)} < \infty. \tag{22}$$

We can derive a more specific condition by considering the special structure that appears in our problem. Specifically, since $i \cdot f(i) > R_{min} \; \forall \, i$, we have

$$\frac{1}{g(z)} < \frac{1}{(R_{min})^z} \quad \forall \, z.$$

Then (22) becomes:

$$\sum_{z=0}^{\infty} \frac{\rho^z}{g(z)} < \sum_{z=0}^{\infty} \left(\frac{\rho}{R_{min}} \right)^z.$$

arbitrary distributions. In general, the different assumptions made about the general distributions in the original works cited here is something worth noting, and reflects the different solution approaches: Cohen assumes, as a common basis, that all general distributions are absolutely continuous. If we follow the insensitivity approach of Burman [6] to show the same results, this condition is reduced to having distribution functions which possess a density which is bounded and continuous (although, strictly speaking, by the use of a Lebesgue integral implied in Cohen we may have discontinuities of measure zero). Furthermore, Kelly in his analysis [19] initially assumes that service requirements have a gamma distribution; he then shows that this assumtpion may be dropped, leaving the only requirement that we have distribution functions that are continuous.

Therefore, a sufficient condition for a stationary distribution to exist is $\rho < R_{min}$. In the Proposition that follows, we further proceed to prove that this is also a *necessary* condition. We rely on the a priori knowledge that the stochastic process of the number of flows is insensitive. Then we are able to apply known statements for the Markovian process.

Proposition 4.1. *The stochastic process of the number of flows in the GPS-Poisson system has a stationary distribution if and only if* $\rho < R_{min}$.

Proof. It is best to prove the Proposition first for the single-class case. Then the extension to multiple classes is easy. Considering a single class of flows with arrival rate λ and exponential service requirement with mean $E[\sigma]$, it is obvious that the stochastic process corresponds to a birth-death process with state-dependent service (death) rates. The service rate approaches the limit $R_{min}/E[\sigma]$ as the number of flows tends to infinity. For both convenience and generality we may consider at first positive birth rates $\{\lambda_n,\, n \geq 0\}$ and death rates $\{\mu_n,\, n \geq 1\}$, with the limits $\lambda = \lim(\lambda_n)$, $\mu = \lim(\mu_n)$ and $d = \lambda/\mu$. Define the sums

$$S_1 = \sum_{n=1}^{\infty} \frac{\mu_1 \cdots \mu_n}{\lambda_1 \cdots \lambda_n}, \qquad S_2 = 1 + \sum_{n=1}^{\infty} \frac{\lambda_0 \cdots \lambda_{n-1}}{\mu_1 \cdots \mu_n}.$$

According to the conditions given in Proposition 2.1 and Corollary 2.5 of [2], the process is recurrent iff $S_1 = \infty$ and ergodic iff, additionally, $S_2 < \infty$. It is not hard to see that these conditions are satisfied when $d < 1$ (i.e. $\rho < R_{min}$), while converse relations hold for $d > 1$.

For $d = 1$ ($\rho = R_{min}$) we can show that the Markov process is recurrent, since

$$\sum_{n=1}^{\infty} \frac{\mu_1 \cdots \mu_n}{\lambda_1 \cdots \lambda_n} > \sum_{n=1}^{\infty} 1 = \infty.$$

Note that unless explicit values for the arrival and service rates are specified, we cannot further decide and the process may be either positive or null recurrent. However, we can prove that for a constant arrival rate and service rates derived by (17),(18), the Markovian process is *null-recurrent*, so that the general non-Markovian process cannot be ergodic. For this, it suffices to show that the series defined by

$$\sum_{n=1}^{\infty} \frac{\mu^n}{\mu_1 \cdots \mu_n}$$

is divergent. Substituting the service rates $\mu_n = n \cdot f(n)/E[\sigma]$, and after elimination of the terms $\frac{N_o W}{E_b}$, $E[\sigma]$ we get both in the uplink and downlink a positive series of the following general form:

$$\sum_{n=1}^{\infty} \frac{(\alpha-\beta)(2\alpha-\beta)\cdots(n\alpha-\beta)}{\alpha^n \cdot n!}, \tag{23}$$

where $\alpha, \beta \in \mathbb{R}^+$, with $a > b$. We compare this against the *harmonic* series $\sum_{n=1}^{\infty} \frac{1}{n}$, which we know is divergent. Taking the so-called 'comparison test of the 2nd kind' in [21], we have

$$\frac{(n+1)\alpha - \beta}{n\alpha} \geq \frac{n}{n+1},$$

which is equivalent to

$$n\alpha + (n+1)(\alpha - \beta) \geq 0.$$

This holds $\forall\, n$, so the series diverges and $S_2 = \infty$. This concludes the proof for the single-class case.

In the multiple-class case, we can likewise decsribe the evolution of the system by considering a single state for the number of flows in the system. We have the aggregate arrival rate $\lambda = \sum_{k=1}^{K} \lambda_k$. The aggregate service rate is calculated based on the mean service size of a flow in the system:

$$E[\sigma] = \sum_{k=1}^{K} \frac{\lambda_k}{\sum_{j=1}^{K} \lambda_j} E[\sigma_k]. \tag{24}$$

Then, for a total number of n flows in the system the aggregate service rate again has the form $\mu_n = \frac{n \cdot f(n)}{E[\sigma]}$. For $n \to \infty$, the service rate is accordingly $\mu = \frac{R_{min}}{E[\sigma]}$, where again $\mu_1 > \mu_2 > \cdots > \mu$. One should see now that by simply redefining the values λ, μ_n, we have exactly the same situation as in the single-class case and the same conclusions apply. □

Remark 4.1. In the Markovian case, we have a necessary and sufficient ergodicity condition. However, despite the fact that the Markovian and non-Markovian have the same stationary distribution, we can only conjecture that the non-Markovian process is ergodic when the corresponding Markov process is. This is the reason we have avoided the use of the term 'ergodic' in the announcement of the Proposition.

Remark 4.2. In the Markovian case, this Proposition can also be viewed as a special case of the ergodicity theorem presented in [23] for a non-homogeneous quasi-birth-death (QBD) process. One can derive it by considering the absence of phases in the system, so that the QBD reduces to a standard birth-death process. However, a much simpler proof is amenable here, and we are able to prove the 'if and only if' clause for the ergodicity condition. Moreover, note that except for the equality case in this condition, the Proposition holds no matter what the initial transition rates in the sequences are (viz. they don't have to be monotone sequences).

Provided that an equilibrium distribution exists, we can easily derive the mean total number of flows in the system as

$$E[N] = G^{-1} \sum_{n=1}^{\infty} \frac{n \cdot \rho^n}{g(n)}. \tag{25}$$

From [19] (Theorem 3.8) we also have that, given a certain total number of flows, the probability that a flow belongs to a certain class equals the fraction $\frac{\rho_k}{\rho}$ of its load in the system. Therefore we conclude that:

$$E[N_k] = \frac{\rho_k}{\rho} E[N].$$

The expected sojourn time of a class-k flow in the system is then easily derived by applying Little's law, $E[T_k] = E[N_k]/\lambda_k$. For all models discussed here, this also equals the mean transfer time, since flows immediately receive service and leave the system upon completion. We arrive at:

$$E[T_k] = E[\sigma_k] \frac{\sum_{j=0}^{\infty} \frac{\rho^j (j+1)}{g(j+1)}}{\sum_{j=0}^{\infty} \frac{\rho^j}{g(j)}}. \tag{26}$$

Consider now the case where an upper bound is set on the number of allowed flows in the system, say $N_{max} = M$. Blocked flows are cleared. This may be necessary in order to constraint the transfer time, and thus ensure a minimal quality of service to accepted flows.

For this system, the same formula (21) applies with $G(M) := \sum_{z=0}^{M} \frac{\rho^z}{g(z)}$. The call blocking probability is then $P_B = \Pr\{N = M\}$, irrespective of the class. This equals the fraction of time the system is full, since we have Poisson arrivals (i.e. the well-known PASTA property, see [31]). It is readily seen that the blocking probability, better denoted here as $B(M)$, can be expressed directly in terms of the normalization constant as:

$$B(M) = 1 - \frac{G(M-1)}{G(M)}. \tag{27}$$

Finally by Little's law, we also have for the expected transfer delay of a class-k flow that

$$E[T_k] = \frac{E[N_k]}{\lambda_k \cdot (1 - P_B)}. \tag{28}$$

We arrive after some tedious calculations at:

$$E[T_k] = E[\sigma_k] \frac{\sum_{j=0}^{M-1} \frac{\rho^j (j+1)}{g(j+1)}}{\sum_{j=0}^{M-1} \frac{\rho^j}{g(j)}}. \tag{29}$$

Finally, an important result concerns the conditional transfer time of a flow whose service requirement is known deterministically. The result is stated in [10] only for the non-blocking system. However, it also applies to a system with blocking. We formulate it in the following theorem:

Theorem 4.1. *For a multiple-class GPS-Poisson system with a total load ρ and maximum finite or infinite number of admitted flows, the mean sojourn time of a flow or class of flows whose service requirement is deterministic, $c > 0$, is given by*

$$E[T(c)] = c \frac{E[T]}{E[\sigma]},$$

where $E[T]$ is the mean sojourn time in a corresponding single class system with the same ensemble characteristics, i.e. the same total load and maximum number of admitted flows and with mean service requirement $E[\sigma]$.

Proof. The theorem is shown for a system with blocking, whereas in the non-blocking case the same formulae apply with $M = \infty$ and $P_B = 0$ (remember that in Kelly's analysis, we may have an arbitrary service requirement distribution[10] (Lemma 3.9 of [19])). Consider flows of class-k whose service requirement is deterministic, c. For their mean number in system, we have that:

$$E[N_k] = \frac{\rho_k}{\rho} \cdot \frac{\sum_{n=1}^{M} \frac{n\rho^n}{g(n)}}{\sum_{z=0}^{M} \frac{\rho^z}{g(z)}}.$$

By applying $E[T_k] = \frac{E[N_k]}{\lambda_k (1-P_B)}$, the mean sojourn time writes

$$\frac{E[T_k]}{c} = \frac{1}{\rho(1 - P_B)} \cdot \frac{\sum_{n=1}^{M} \frac{n\rho^n}{g(n)}}{\sum_{z=0}^{M} \frac{\rho^z}{g(z)}}.$$

Now going back to the single-class case, it can be shown that the right part of the equation above is equal to $\frac{E[T]}{E[\sigma]}$ in Eq. (29) (we omit the subscripts k). By substituting this term, the theorem follows immediately. □

This theorem can be interpreted as yielding either the sojourn time of a class of flows with deterministic service requirement or, more appropriately, the sojourn time of a flow *conditioned* on its service volume. Then, it reveals that the mean sojourn time of a transfer request is proportional to the volume of the request; so flows requiring more service experience larger delays and vice-versa. This embodies the fairness principle which originates from equal resource sharing of the different flows.

[10] This helps us to rid the uncomfortable assumption about rational Laplace-Stieltjes transmforms made for the infinite system in Cohen.

4.2 The GPS-Engset model

Here we examine the GPS service regime of a CDMA link under a finite source model of arrivals. We have a fixed population of mobile terminals sending flows in the link and a maximum number of simultaneously served flows, M. We consider a situation similar to that of an Engset model, in that the transmission of a flow from a source is followed by a random think period of that source. This can better model the sending procedure of non real-time traffic, i.e. the succession of file transfers and think periods corresponding to the activity of a given user. Blocking may occur when the number of sources exceeds the number of allowed flows M. In this case, we make the standard assumption that the blocked source goes back to its 'thinking phase'.

We consider K different classes of flows. Each class consists of a finite population of S_k sources $(k = 1, \ldots, K)$. We assume that think times of sources in each class, as well as their successive flows' service requirements form independent families of independent, identically distributed random variables. The distribution functions are denoted by F_i^k, F_σ^k, with values referring to time and size, respectively, and corresponding first moments $E[\tau_k], E[\sigma_k]$. We assume that the distribution functions are absolutely continuous in $[0, \infty)$.

Similarly to [15], we will also say that the above description corresponds to a *modified* Engset model[11]. We define analogous load parameters here as $\rho_k := \frac{E[\sigma_k]}{E[\tau_k]}$. For this system, we may follow the analysis in [10] to derive the joint density of the number of flows from each class that are receiving service, $N_k^{(s)}$ $(k = 1, \ldots, K)$ (and thus the remaining idle flows, $N_k^{(i)}$), their attained or residual service denoted by the vector $\overline{\sigma}_k = (\sigma_k(1), \ldots, \sigma_k(x_k))$, as well as the time spent or time to go in the idle phase of the remaining flows from each class, denoted by $\overline{\tau}_k = (\tau_k(1), \ldots, \tau_k(S_k - x_k))$.

Maintaining an analogous notation as in § 4.1, we define the infinitesimal probability:
$p(x_k,\, \sigma_k(h),\, \tau_k(m);\, h = 1, \ldots, x_k,\, m = 1, \ldots, S_k - x_k,\, k = 1, \ldots, K) d\overline{\sigma}_k d\overline{\tau}_k :=$
$\Pr\{N_k^{(s)} = x_k, N_k^{(i)} = S_k - x_k, \sigma_k(h) \leq \boldsymbol{\sigma}_k(h) \leq \sigma_k(h) + d\sigma_k(h),\, \tau_k(m) \leq \boldsymbol{\tau}_k(m) \leq \tau_k(m) + d\tau_k(m);$
$h = 1, \ldots, x_k,\, m = 1, \ldots, S_k - x_k,\, k = 1 \ldots, K\}$.

For shortness, we denote the density by $p(x_k,\, \overline{\sigma}_k,\, \overline{\tau}_k; k = 1, \ldots, K)$. We have[12]:

$$p(x_k,\, \overline{\sigma}_k,\, \overline{\tau}_k; k = 1, \ldots, K) = p(x_1, x_2, \ldots, x_K) \cdot \prod_{k=1}^{K} \left\{ \prod_{h=1}^{x_k} \frac{1 - F_\sigma^k(\sigma_k(h))}{E[\sigma_k]} \right\} \left\{ \prod_{h=1}^{S_k - x_k} \frac{1 - F_i^k(\tau_k(h))}{E[\tau_k]} \right\} \quad (30)$$

with

$$p(x_1, x_2, \ldots, x_K) = \frac{\prod_{k=1}^{K} \binom{S_k}{x_k} \rho_k^{x_k} \cdot \phi(x_1 + x_2 + \cdots + x_K)}{\sum_{z_1=0}^{S_1} \sum_{z_2=0}^{S_2} \cdots \sum_{z_K=0}^{S_K} \prod_{k=1}^{K} \binom{S_k}{z_k} \rho_k^{z_k} \cdot \phi(z_1 + z_2 + \cdots + z_K)} \quad (31)$$
$$z_1 + z_2 + \cdots + z_K \leq M$$

for $0 \leq x1 + x_2 + \cdots + x_K \leq M$. By integrating over the size and time values, it can readily be seen that $p(x_1, x_2, \ldots, x_K)$ is the joint probability, in steady state, that x_k sources of class k are busy, for $k = 1, \ldots, K$. We again remark the insensitivity properties of this distribution, as it depends on think times and service requirements only through their means.

Remark 4.3. If we consider Markovian processes, we note that we can derive the stationary distribution of the GPS-Engset model from that of the GPS-Poisson model, by considering a closed network of quasi-reversible queues. In a similar way, this was done for loss systems in [22]. We include this derivation in Appendix A.

[11] The authors in [15] also considered a modified Engset model. Differences here are the total varying service rate, as well as the fact that we generally consider both service rate reduction and blocking.

[12] This expression is explicitly mentioned in [10] only for the 2-class case. It's general form can be derived from Eqs. (5.9),(5.10), by applying the notation of § 4.2 therein.

Of particular interest is the probability that a source belonging to a given class will be blocked upon a request for service, given that the system has reached its maximum number of admitted flows, M. Remember that we don't have Poisson arrivals here, therefore the call blocking probability is different from the time blocking one. We present the following theorem, generalized for an arbitrary number K of service classes. The derivation is based on but extends Cohen's approach for finding the blocking probability in the case of a single class traffic.

Theorem 4.2. *For the case of K service classes in the GPS-Engset system with a total maximum number of flows M, the blocking probability of a class-m source, $m = 1, \ldots, K$, is given by*

$$P_B^k = \frac{\sum_{x_1+x_2+\cdots+x_K=M} \prod_{\substack{k=1 \\ k\neq m}}^{K} \binom{S_k}{x_k}\binom{S_m-1}{x_m} \rho_k^{x_k} \rho_m^{x_m} \cdot \phi(x_1+x_2+\cdots+x_K)}{\sum_{z_1=0}^{S_1} \cdots \sum_{z_m=0}^{S_m-1} \cdots \sum_{z_K=0}^{S_K} \prod_{\substack{k=1 \\ k\neq m}}^{K} \binom{S_k}{z_k}\binom{S_m-1}{z_m} \rho_k^{z_k} \rho_m^{z_m} \cdot \phi(z_1+z_2+\cdots+z_K)}, \qquad (32)$$
$$z_1+z_2+\cdots+z_K \leq M$$

where the sum $\sum_{x_1+x_2+\cdots+x_K=M}$ extends over the set $\mathcal{B} = \left\{x_k \geq 0 : \sum x_k = M;\ x_k \leq S_k,\ k = 1, \ldots, m-1, m+1, \ldots, K,\ x_m \leq S_m - 1\right\}$.

Proof. To find the blocking probability in the K-class case, we consider the extended system with one more class, i.e. $K' = K+1$, the last class having an initial population of $S_{K+1} = 1$ (the initial population of other classes is indifferent, as long as it is greater than zero). Take a number of x_k customers in the first k classes, $k = 1, \ldots, K$ and assume the last class consists of only one customer in the idle phase, whose time spent in this phase is slightly greater than zero, denoted by $\tau_{K+1} = 0+$. Then from (30), (31) we can arrive at the conditional probability:

$$p(x_1+\cdots+x_K = M \mid x_{K+1} = 0, \tau_{K+1} = 0+) = \frac{\sum_{x_1+\cdots+x_K=M} \binom{S_1}{x_1}\cdots\binom{S_K}{x_K} \rho_1^{x_1}\cdots\rho_K^{x_K} \phi(x_1+\cdots+x_K)}{\sum_{\substack{z_1=0}}^{S_1}\cdots\sum_{z_K=0}^{S_K} \binom{S_1}{z_1}\cdots\binom{S_K}{z_K} \rho_1^{z_1}\cdots\rho_K^{z_K} \phi(z_1+\cdots+z_K)},$$
$$z_1+\cdots+z_K \leq M$$

where the sum $\sum_{x_1+\cdots+x_K=M}$ extends over the set

$$\mathcal{B} = \left\{x_k \geq 0 : \sum x_k = M;\ x_k \leq S_k,\ k = 1, \ldots, K\right\}.$$

The conditional expression above represents the probability that source $K+1$ has made a request which could not receive service because there are already M flows present. We may consider that this source belongs to either of the groups $m = 1, \ldots, K$, whose $S_m > 1$. Then the system is equivalent with that of $S_m - 1$ sources present, which gives us (32). □

Remark 4.4. This theorem says that the probability a new arrival of a certain class is blocked is equal to the time blocking probability, in a system with initial population of that class reduced by one. More generally, the distribution of the number of flows in service seen by an arriving flow of a certain class is the time average distribution that would be observed if the number of flows of that class were reduced by one. This is reminiscent of a known situation for finite source Engset arrivals. Similar results also hold for the generalized Engset model ([8]), as well as the generalized Engset loss station ([22]).

Remark 4.5. Similarly to the other Engset models mentioned above, the time and call blocking probabilities can be expressed simply from the normalization constant. We denote the latter, for given values of loads, by $G(\mathbf{S}, M)$, where $\mathbf{S}$ is the vector of the number of sources of each class. Then one can verify that the time blocking probability, $B(\mathbf{S}, M)$, and the call blocking probability of class-k, $L_k(\mathbf{S}, M)$ are given by:

$$B(\mathbf{S}, M) = 1 - \frac{G(\mathbf{S}, M-1)}{G(\mathbf{S}, M)} \qquad (33)$$

$$L_k(\mathbf{S}, M) = B(\mathbf{S} - \mathbf{1}_k, M), \tag{34}$$

where $\mathbf{1}_k$ is the unit vector of dimension K whose k-th component is unity. In general, formulae for the time and call congestion involving only the normalization constant can be very useful in numerical calculations involving large values of $\mathbf{S}$ and M. We refer the reader to Appendix B for more details regarding the efficient computation of this constant.

The next important result concerns the expected transfer time of a class-k flow in the system. We present it in the form of a conjecture (the reason for which is revealed at the end of the proof):

Conjecture 4.1. *For the case of K service classes in the GPS-Engset system with a total maximum number of flows M, the expected sojourn time of a class-k flow ($k = 1, \ldots, K$), is given by*

$$E[T_k] = \frac{E[\sigma_k]}{\frac{\sum_{(\mathbf{x}\in\mathcal{F}(\mathbf{S},M))} x_k f(x_1 + \cdots + x_K) p(x_1, \ldots, x_K)}{\sum_{(\mathbf{x}\in\mathcal{F}(\mathbf{S},M))} x_k p(x_1, \ldots, x_K)}}, \tag{35}$$

where for $\mathbf{x} = (x_1, \ldots, x_k)$, the sums extend over the whole feasible set defined by $\mathcal{F}(\mathbf{S}, M) = \{x_k \geq 0 : \sum x_k \leq M;\ x_k \leq S_k,\ k = 1, \ldots, K\}$.

Proof. The proof follows Cohen's approach by the use of regenerative process theory. Consider the process $\boldsymbol{x}_t = (x_t^1, x_t^2, \ldots, x_t^K)$ of the number of active flows of each class at time t. It is assumed that the distribution of the idle time of each class-k flow is a convolution of a negative exponential distribution with first moment ε and an arbitrary distribution with support $[0, \infty)$ and finite first moment $E[\tau_k]$. That is, every idle time is considered to be the sum of a negative exponential and an arbitrary nonnegative variable. The family of all these variables and of the service requirements are again assumed to be mutually independent.

Then, the end point of a time interval during which the idle times of all sources are in the 'negative exponential stage' is obviously a regeneration point of the process $\boldsymbol{x}_t$. Clearly, such a regeneration point exists since the probability $p(0, \ldots, 0)$ from (31) is always greater than zero. Denote by c the duration of a regeneration cycle, and by n_k the number of requests of class-k that originated during this cycle. The service requirement and sojourn time of requests of class-k during this cycle are denoted by $u_i^{(k)}, v_i^{(k)}$ respectively, where $i = 1, \ldots, n_k$.

Now

$$a_k := \int_0^c x_t^k f(x_t^1 + \cdots + x_t^K) dt$$

is the total amount of work that class-k 'brings into' the system during the regeneration cycle. This equals $\sum_{i=1}^{n_k} u_i^{(k)}$ by definition. Then from the theory of regenerative process ([31],[9]), the expectation of $x_k \cdot f(x_1 + \cdots + x_K)$, if it exists, equals the expected cumulative work over the regeneration cycle, divided by the expected cycle length. Hence

$$\begin{aligned}
\sum_{\mathbf{x}\in\mathcal{F}(\mathbf{S},M)} x_k\, f(x_1 + \cdots + x_K)\, p(x_1, \ldots, x_K) &= \frac{1}{E[c]} \cdot E\left\{ \int_0^c x_t^k f(x_t^1 + \cdots + x_t^k) dt \right\} = \\
&= \frac{E[n_k]}{E[c]} \cdot \frac{1}{E[n_k]} \cdot E\left\{ \sum_{i=1}^{n_k} u_i^{(k)} \right\} = \\
&= \frac{E[n_k]}{E[c]} \cdot E[\sigma_k] \qquad \text{by Wald's Theorem [31].}
\end{aligned} \tag{36}$$

Now the total sojourn time of class-k flows in the system is

$$\int_0^c x_t^k dt = \sum_{i=1}^{n_k} v_i^{(k)}.$$

We have in the same way as before,

$$\sum_{\mathbf{x}\in\mathcal{F}(\mathbf{S},M)} x_k\, p(x_1,\ldots,x_K) = \frac{1}{E[c]}\cdot E\left\{\int_0^c x_t^k dt\right\} =$$
$$=\frac{E[n_k]}{E[c]}\cdot\frac{1}{E[n_k]}\cdot E\left\{\sum_{i=1}^{n_k} v_i^{(k)}\right\} =$$
$$=\frac{E[n_k]}{E[c]}\cdot E[v_k]. \tag{37}$$

Note in the last equality that the random variables $v_i^{(k)}$, $i=1,\ldots,n_k$ are correlated and hence not independent. However, in Wald's theorem, every variable need only be independent *of the event that it is included in the sum* (see e.g. the proof in [31], p. 98). This holds for v_i^k in our case, so the theorem can be applied.

Combining (36),(37), we finally obtain the formula for $E[T_k^{(\varepsilon)}] := E[v_k]$, for an arbitrarily small value of ε. Note that the value of $E[T_k^{(\varepsilon)}]$ does not depend on ε; however, $E[c]$ is finite for ε arbitrarily small but not zero, since we cannot claim that a regenerative process with finite cycle time exists as $\varepsilon\to 0$. In order to formally complete the theorem, we need to show the continuity of the expected sojourn time at $\varepsilon=0$. We haven't been able to prove this so far, and the proof seems to be even more difficult in the case of blocking (what's more, the same point has neither been shown in the non-blocking case in [10] !). In view of that, this result is not rigorous, but has to be taken only as a *conjecture*. □

This conjecture also has an intuitive explanation. In the denominator of the complex fraction in (35), the numerator represents the total mean service rate offered to class-k flows. The value that occurs when this is divided by the mean number of class-k flows may then be 'tagged' as the mean service rate of a single flow. Therefore, it may seem intuitive that dividing the mean service requirement by the mean service rate yields the mean sojourn time. However, this argumentation by mean values has no theoretical basis.

An equivalent method that leads to the same end-point in the proof is by using time averages and arguments from regenerative process theory. Consider generally the countable state space of the system, $\mathcal{F}(\mathbf{S},M)$. First of all, we observe that in a processor-sharing system that has a stationary distribution, the arrival rate must equal the departure rate, since the system is in equilibrium. Then by manipulating Eq. (35) and using Little's law, it suffices to show that

$$\sum_{\mathbf{x}\in\mathcal{F}(\mathbf{S},M)} \frac{x_k f(x_1+\cdots+x_K)}{E[\sigma_k]} p(x_1,\ldots,x_K)$$

is the departure rate of class-k flows in the system, defined as

$$\lim_{t\to\infty}\frac{1}{t}\int_0^t \frac{x_k(v) f(x_1(v)+\cdots+x_k(v))}{E[\sigma_k]}dv.$$

If we consider the regenerative process structure described above, this is immediately shown from the relative theory, viz. that the time average equals the mean of the limiting distribution[13] (see e.g. [31]). In the absence of that structure, we would have to consider much more general arguments from the theory of continuous time ergodic processes.

Remark 4.6. The formula for the sojourn time in the single-class case can be derived immediately from the above analysis. In that case, one can readily see Little's law behind Eq. (37). Further, in the single

[13] The same of course holds for a function of a regenerative process.

class case with a population of S sources, mean service requirement of flows $E[\sigma]$ and load ρ, with a total maximum number of allowed flows M, one obtains the useful expression:

$$E[T] = E[\sigma]\frac{\sum_{n=0}^{M-1}\binom{S-1}{n}\rho^n\,\phi(n+1)}{\sum_{n=0}^{M-1}\binom{S-1}{n}\rho^n\,\phi(n)}. \tag{38}$$

In a manner totally analogous to the Poisson arrival system, the conditional sojourn time depending on the service requirement c of a flow is given by $E[T(c)] = c\frac{E[T]}{E[\sigma]}$. Again, this result is stated in [10] only for the non-blocking case; based on our previous conjecture, we proceed to show it for a system with blocking.

Theorem 4.3. *Consider a GPS-Engset system with a total population of S sources, of which $S-1$ sources belong to a single class (i.e. have the same service requirement distribution, with mean $E[\sigma]$ and the same idle time distribution). The S-th source has deterministic service requirement $c > 0$. Its idle time distribution can be arbitrary, but with positive finite mean. Then the sojourn time of this flow in the system equals*

$$E[T(c)] = c\frac{E[T]}{E[\sigma]},$$

where $E[T]$ is the mean sojourn time in a corresponding single-class, GPS-Engset system with an initial population of S sources with the same mean service requirement and idle time.

Proof. The result has been proved for the non-blocking case in [10]. We present here the proof in the case of a system with blocking, based on the preceding Conjecture 4.1. The actual idle time distribution of the deterministic service flow does not appear in the final equations and therefore is indifferent, however we insist that it has finite positive mean so that the previous result can apply. In the following, we will also use the notation $x_1 + x_2 \leq M$ in a sum to denote that the summation extends over the whole feasible set.

Let the deterministic service flow belong to class 1. Then from Conjecture 4.1 we have

$$E[T_1] = c \cdot \frac{\sum_{(x_1+x_2\leq M)} x_1 p(x_1,x_2)}{\sum_{(x_1+x_2\leq M)} x_1 f(x_1+x_2) p(x_1,x_2)}.$$

In the above fraction, x_1 is either 0 or 1. Therefore, considering the summands with $x_1 = 1$ we have

$$E[T_1] = c \cdot \frac{\sum_{x_2=0}^{M-1} p(1,x_2)}{\sum_{x_2=0}^{M-1} f(x_2+1) p(1,x_2)}.$$

The stationary distribution $p(1,x_2)$ is given by (since $S_1 = 1$):

$$p(1,x_2) = \frac{\binom{S_2}{x_2}\rho_1\rho_2^{x_2}\phi(x_2+1)}{\sum_{\substack{z_1=0\\ z_1+z_2\leq M}}^{S_1}\sum_{z_2=0}^{S_2}\binom{S_1}{z_1}\binom{S_2}{z_2}\rho_1^{z_1}\rho_2^{z_2}\phi(z_1+z_2)}.$$

Substituting in the the mean sojourn time above, we obtain

$$E[T_1] = c \cdot \frac{\rho_1 \cdot \sum_{x_2=0}^{M-1}\binom{S_2}{x_2}\rho_2^{x_2}\phi(x_2+1)}{\rho_1 \cdot \sum_{x_2=0}^{M-1}\binom{S_2}{x_2}\rho_2^{x_2}\phi(x_2)}.$$

Then, after the terms ρ_1 cancel out, we observe from Eq. (38) that the fraction in the right hand of the above expression equals exactly $\frac{E[T]}{E[\sigma]}$ in a system with S_2+1 identical flows. By considering $S := S_2 - 1$, the theorem is proved. □

Remark 4.7. Note that there are rudimentary differences with the analogous theorem in the Poisson case. Here, the deterministic service flow is taken to be a *single one* from a population of identical flows, and we don't allow for *any* number of classes as we may in the Poisson system. However, by considering the deterministic flow 'mingled' in the set of all flows, and no matter what the exact setting is, both theorems express the same proportionality principle of the sojourn time in the system, relative to the size of the service requirement.

5 Insensitivity and truncation properties

We draw attention to two important realizations made in the course of this work:

1. Insensitivity properties apply to all examined models.

2. In blocking systems, the stationary distribution can be derived from the corresponding infinite system (where that exists), or a corresponding superset system[14] by applying the well-known *truncation principle* [19].

This section is devoted to establishing these properties by an easier and more general method. This will then permit us to extend results regarding the steady-state characteristics of GPS service systems to other access or admission control policies. Different such policies may be necessary in order to coordinate access between various classes of flows, while maintaining the same rate of transmission.

In general, there exist various methods for establishing insensitivity properties, all of which depend on the specific problem at hand. Here the setting of a *generalized semi-Markov process* (GSMP) is most appropriate: we may either follow Schassberger's method of clocks ([27]) or the method of Burman ([6]) to get to the desired result. The two methods are very similar in their approach; we choose to follow the latter, since it is generally simpler[15].

Let us paraphrase the general structure of the processor sharing models in the previous section so as to match a GSMP framework. Consider the set G to be the countable state space of the system. We say that each transition from a state g to another g' is caused by the occurence of an *event* active in the first state. 'Events' here correspond to the completion of ongoing services of flows or to the arrival of a new flow in the system. The set of all events in every state forms a countable event space E, and we denote by $e(g)$ a subset of this space associated with a single state g. The event $e \in \mathsf{E}$ requires X_e processing units to be completed, which can be construed to refer to either time or service units.

Let X_e be drawn from an arbitrary distribution $H_e(x) = \Pr\{X_e \leq x\}$, continuously differentiable in $[0, \infty)$ and with finite mean $1/\mu_e$. More generally, *some* of the X_e may be exponential, which we need to discern by specifying $\mathsf{E}' \subseteq \mathsf{E}$ as the subset of events $e \in \mathsf{E}$ for which they are *not*. Upon the completion of an event $e \in e(g)$ we go to a state g' with a probability $p(g, e, g')$. A set of outgoing states $\Gamma_o(g)$ is composed by all states g' such that $p(g, e, g') > 0$ for some $e \in e(g)$. Similarly, $\Gamma_i(g)$ is the set of all states g'' such that $p(g'', e, g) > 0$ for some event $e \in e(g'')$.

We denote by $N(t)$ the generalized semi-Markov process of the number of flows in the above setting. For each active event e, let also $c_e(t)$ be the amount of processing attained by e at time t. The joint process $(N(t), c_e(t);\ e \in e(N(t)))$, is a *supplementary generalized semi-Markov process* (SGSMP), which we shall also call the *associated life process* of $N(t)$. Moreover, the Markov process that comes out if all X_e where exponentially distributed with means solely[16] dependent on events (here, on arrivals and service completions) is called the *corresponding* or *associated* Markov process of $N(t)$.

[14] We use the term 'superset system' here to allude to the underlying stochastic process whose state space is a proper superset of the state space of the blocking system. The latter may well be referred to as the 'truncated system'.

[15] Burman's method does not have to deal with problems regarding clock selection from an infinite population that Schassberger's more detailed model poses (cf. [28]).

[16] But not dependent on former or next states associated with those events or the event that causes a transition.

In the SGSMP setting, we denote by r_{eg} the rate at which processing of event $e \in e(g)$ occurs at state g. For this we may also define the rate $\mu_e(g,g') := r_{eg}\mu_e p(g,e,g')$ as the *transition flow*[17] (expected number of transitions per unit time) from state g to state g', caused by the completion of the active event e.

In the purely Markov case, if the process $N(t)$ is positive recurrent, one can easily construct (cf. [6]) a set of equations, referring to events, that is equivalent to the set of global balance equations and can be solved, with the help of the normalization condition, to find the steady state probabilities $p(g)$ of states $g \in \mathsf{G}$. The decisive theorem in [6] relates insensitivity to the satisfaction of another set of equations, called 'restricted flow equations' for *uniquely* identified transition flows in the system. To uniquely identify a transition flow, we associate a label $\ell_e(g,g') := (e(g) - \{e\}) \cap \mathsf{E}'$, depicting the actual transition between states and the precise event which causes the transition. The restricted flow equations then write:

$$p(g) \cdot \sum\nolimits' \mu_e(g,g') = \sum\nolimits'' \mu_\zeta(g'',g) \cdot p(g''), \tag{39}$$

where the first summation is over all $g' \in \Gamma_o(g)$, such that $\ell_e(g,g')$ has a fixed label ℓ, and the second is over all $g'' \in \Gamma_i(g)$ with $\ell_\zeta(g'',g) = \ell$ for some event $\zeta \in e(g'')$. For transitions to the same state g' caused by a number of different events n but which are *probabilistically indistinguishable*, i.e. for which the probability $\Pr\{N(s) = g' |\, N(0) = g,\ c(0) \geq t\}$, where $c(0) = (c_{e_1}(0), c_{e_2}(0), \ldots, c_{e_n}(0))$, is unchanged by any permutation of the events, the sum of all rates must be considered in the above summations. We refer to these as 'identical events' here.

If the restricted flow equations are satisfied, we have for the stationary distribution of the life process $(N(t), c_e(t);\ e \in e(N(t)))$ that ([6]):

$$p(g,t) = p(g) \cdot \prod_{e \in e(g) \cap \mathsf{E}'} \mu_e(1 - H_e(x)), \tag{40}$$

where $p(g)$ is the stationary distribution of the corresponding Markov process. This immediately reveals the insensitivity property, since by integrating over x we get the distribution $p(g)$, which cannot depend on anything else but the first moments of the distributions of events in the system.

We proceed to discuss exactly how these restricted flow equations apply in the case of our processor sharing systems. We choose, as a case-study, the Engset-like system with 2 classes of traffic. Suppose one knows *a priori* (e.g., by solving the system of global balance equations) that the equilibrium distribution of the number of flows of each class in the associated Markov process is given by:

$$p(x_1,x_2) = \frac{\binom{S_1}{x_1}\binom{S_2}{x_2}\rho_1^{x_1}\rho_2^{x_2}\phi(x_1+x_2)}{\sum_{\substack{z_1=0}}^{S_1}\sum_{\substack{z_2=0 \\ z_1+z_2 \leq M}}^{S_2}\binom{S_1}{z_1}\binom{S_2}{z_2}\rho_1^{z_1}\rho_2^{z_2}\phi(z_1+z_2)}. \tag{41}$$

Consider the state $g = (x_1, x_2)$. Then the complete set of events active in that state is

$$(\{e_1^{1,s}, e_2^{1,s}, \ldots, e_{x_1}^{1,s}\}, \{e_1^{1,i}, e_2^{1,i}, \ldots, e_{S_1-x_1}^{1,i}\}, \{e_1^{2,s}, e_2^{2,s}, \ldots, e_{x_2}^{2,s}\}, \{e_1^{2,i}, e_2^{2,i}, \ldots, e_{S_2-x_2}^{2,i}\}),$$

where $e_i^{k,j}$ is an active event corresponding to the flow i of class k, that is either 'receiving service' or 'thinking', if $j = s$ or i, respectively. Clearly, the completion of an event –or of identical events– in our system leads to a unique state, so that all transition probabilities equal to unity. Denoting by μ_k^{-1} the service requirement of class-k jobs, and by λ_k the arrival rate of these, the restricted flow equations write:

$$\begin{aligned} p(x_1,x_2) \cdot \mu_1 f(x_1+x_2) \cdot x_1 &= p(x_1-1,x_2) \cdot \lambda_1 \cdot (S_1 - (x_1-1)) \\ p(x_1,x_2) \cdot \lambda_1 \cdot (S_1 - x_1) &= p(x_1+1,x_2) \cdot \mu_1 \cdot f(x_1+x_2) \cdot (x_1+1), \end{aligned}$$

[17] By coincidence, the term 'flow' Burman uses is the same as the jobs requesting service in our setting. So we use 'transition flow' to distinguish between the two.

for a class-1 flow, and

$$p(x_1, x_2) \cdot \mu_2 f(x_1 + x_2) \cdot x_2 = p(x_1, x_2 - 1) \cdot \lambda_2 \cdot (S_2 - (x_2 - 1))$$
$$p(x_1, x_2) \cdot \lambda_2 \cdot (S_2 - x_2) = p(x_1, x_2 + 1) \cdot \mu_2 \cdot f(x_1 + x_2) \cdot (x_2 + 1),$$

for a class-2 flow, for all states on which the steady state probabilities are defined. But one can readily see that these are exactly the *detailed balance equations*, which are satisfied for this system since the corresponding Markovian process is reversible. Therefore, we can apply Eq. (40) from which insensitivity properties of the GSMP readily derive.

Remark 5.1. The more general approach in [12, Chapter 6] also includes the insensitivity results for GPS service queueing systems. More importantly, the point process analysis shows that for a certain class, successive service requirements, and successive think times in the GPS-Engset system do not have to be independent; it suffices that the random marked point processes defined by the lifetimes of service and think times for a certain class jobs have a stationary distribution with finite intensity (i.e. define stationary ergodic sequences). Note however, that lifetimes between different classes, or between service and think times of the same class are still assumed to be independent. Relaxing partially the independence assumptions has a great significance; for example, we may define correlation dependencies between the sizes of flows following the sending of an initial flow, since that initial flow usually specifies the purpose of sending data over the link. Similarly, successive think times may be dependent. In all cases, the stationary distribution of the number of flows in the system remains the same.

The fact that we are able to establish the insensitivity properties by first looking at the corresponding Markov process also answers our second inquiry regarding truncation properties of these processes. Specifically, it can be verified that the associated Markov process of the class of systems examined here is *reversible*, since it is (or assumed to be) ergodic and its *communication graph* (i.e. the graph linking all states for which transition probabilities are not zero) is a *tree*. Therefore (see e.g. [19]), truncation properties apply for the Markov process, and hence for the corresponding non-Markovian process.

6 Access control policies

The nature of a processor-sharing scheme is such that in case of a large number of simultaneously transmitted flows, the throughput that incurs for a single flow becomes very small, and thus, depending on its size, the transfer time of a flow may increase inordinately. Despite the elasticity of data traffic, very large transfer times are in principle unacceptable and may also lead to unwanted reneging phenomena, as a result of user impatience. Therefore, some form of access or admission control should be considered.

In the case of multiple classes of traffic on a CDMA link, the issue of how to coordinate access between different classes is particularly important, and pertains to problems of fairness and good utilization of the access space (it is convenient to use the term 'access space' to refer to the total number of allowed flows on the link). We note that fairness endorses several definitions in all contexts where it is used. Here, apart from equal rate transmission, one may consider that a class with greater load has more importance and hence should be allocated more space in the system, nonetheless without dominating the link. This leads to the application of some idea of 'proportional fairness'. In order to apply that while maintaining an equal transmission rate, the only way is access control.

We may distinguish the following general families of access or admission control policies: *common* access, *dedicated* access and *mixed* access policies. These are defined by the feasible sets of states for each family. Access limits for each class of traffic should be seen as control parameters for the policy at hand. The similarity of access control models with capacity-sharing models such as those considered in [18] is obvious.

Common access:

$$\mathcal{F}(M) = \{x_k \geq 0 : \sum x_k \leq M;\ k = 1, \ldots, K\} \qquad \text{(GPS-Poisson)}$$
$$\mathcal{F}(\mathbf{S}, M) = \{x_k \geq 0 : \sum x_k \leq M;\ x_k \leq S_k,\ k = 1, \ldots, K\} \quad \text{(GPS-Engset)}$$

This is the standard access model that has been considered in the models so far. Clearly it is the easiest to implement, but as we emphasized before risks unfairness, since a class with higher relative load dominates the link's resources.

Dedicated access:

$$\mathcal{F}(\mathbf{M}) = \{\mathbf{x} : 0 \leq x_k \leq M_k;\ k = 1, \ldots, K\}, \qquad \text{(GPS-Poisson, Engset)}$$

where for the Engset model we make the logical assumption that $S_k \geq M_k$. In this case, each class of flows has an individual maximum number of allowed flows, and its blocking behavior is only affected by its own load. Thus objectively it is a more fair policy. Its disadvantage is that the link may have a smaller access utilization if the mean number of flows from a class is much smaller than the reserved number of flows for this class. The chosen fairness policy for each class, as well as the variance in the distribution of the number of flows play an important role in the proper selection of individual flow limits.

Mixed access: This is a family of access control policies that lie somewhere between the aforementioned common and dedicated access control policies. Thus, one may consider an access control policy with a reserved number of flows for each class, and a remaining 'space' for a number of flows which may be occupied by a flow of any class. This is described by the feasible set:

$$\mathcal{F}(\mathbf{M}) = \left\{\mathbf{x} : 0 \leq x_k \leq M_k + M_c,\ 0 \leq \sum x_k \leq \sum M_k + M_c;\ k = 1, \ldots, K\right\}, \text{(GPS-Poisson,Engset)}$$

where in the Engset model we assume that $S_k \geq M_k + M_c$. This can be described as a policy with *partially common access and guaranteed reservation.* A variant of this policy is the following:

$$\mathcal{F}(\mathbf{M}) = \left\{\mathbf{x} : 0 \leq x_k \leq M_k,\ 0 \leq \sum x_k \leq M < \sum M_k;\ k = 1, \ldots, K\right\}, \quad \text{(GPS-Poisson, Engset)}$$

where in the Engset model $S_k \geq M_k$. Here, a class of flows cannot surpass a predefined limit, however in general there is not a guarantee on a certain 'free' number of flows from each class. Thus we may call this as a policy with *dedicated access but no guaranteed reservation.*

It can be understood that mixed access control policies are more flexible, and hence they can easier meet a compromise between fairness and total utilization of the access space, based on the selection of the control parameters that define the feasible set. Finally, it is worth noting that all these policies are *coordinate convex* [18], since departures (or arrivals) are never blocked.

The question of how to analytically model different access control policies is not very difficult to answer. In fact, it can easily be shown that for the separate GPS-Poisson or GPS-Engset cases, all the different access models have the same stationary distribution, within a normalization constant. Further, in all the above models the insensitivity property holds.

This can easily be shown by following the line of thought introduced in § 5. First consider the associated reversible[18] Markov process, and apply the truncation principle from a corresponding superset system. Then, follow Burman's analysis shown in § 5 (or another equivalent analysis) to demonstrate insensitivity. The result of interest is that we can apply the truncation principle directly for a different blocking regime; the normalization constant is just the sum of the stationary probabilities (without this constant) in the set of all allowable states.

[18] It can easily be shown that all coordinate convex policies correspond to reversible stochastic processes.

As examples, one can easily show (by choosing appropriate superset systems)[19] that for the GPS-Engset system with 2 classes of traffic, source populations S_1, S_2, and for which we impose *individual* limits on the number of flows in the system, say M_1, M_2, the stationary distribution is given by:

$$p(x_1,x_2)=\frac{\binom{S_1}{x_1}\binom{S_2}{x_2}\rho_1^{x_1}\rho_2^{x_2}\phi(x_1+x_2)}{\sum_{z_1=0}^{M_1}\sum_{z_2=0}^{M_2}\binom{S_1}{z_1}\binom{S_2}{z_2}\rho_1^{z_1}\rho_2^{z_2}\phi(z_1+z_2)}. \tag{42}$$

Or, in the GPS-Poisson system, if we have different limits for the 2 classes of customers in the system, M_1, M_2, and a common limit M, such that $M < M_1+M_2$ (dedicated access control with no guaranteed reservation), it is easily shown that the steady-state distribution is:

$$p(x_1,x_2)=\frac{\frac{\rho_1^{x_1}}{x_1!}\frac{\rho_2^{x_2}}{x_2!}\phi(x_1+x_2)}{\sum_{\substack{z_1=0\\ z_1+z_2\le M}}^{M_1}\sum_{z_2=0}^{M_2}\frac{\rho_1^{z_1}}{z_1!}\frac{\rho_2^{z_2}}{z_2!}\phi(z_1+z_2)}. \tag{43}$$

Moreover, we can equally derive blocking probabilities in multiple class, Engset-like systems, under different access control policies. Specifically, by deriving the joint density (40) for an appropriate extended system by insensitivity and truncation arguments, one can then follow the proof of Theorem 4.2 to arrive at an expression for the blocking probability. For example, for the GPS-Engset system considered above we have for the blocking probability of a class-1 flow:

$$P_B^1=\frac{\sum_{x_2=0}^{M_2}\binom{S_1-1}{M_1}\binom{S_2}{x_2}\rho_1^{M_1}\rho_2^{x_2}\phi(M_1+x_2)}{\sum_{z_1=0}^{M_1}\sum_{z_2=0}^{M_2}\binom{S_1-1}{z_1}\binom{S_2}{z_2}\rho_1^{z_1}\rho_2^{z_2}\phi(z_1+z_2)}, \tag{44}$$

and similarly for a class-2 flow. Note also that we may follow exactly the same proof as in § 4.2 to conjecture the expected sojourn time of a class-k flow in this case, and also for any other blocking regime in our system.

More generally, the added flexibility of deriving steady-state characteristics by simple insensitivity and truncation arguments is primordial for such processor sharing systems, and it permits one to explore a wide range of situations or scenarios with respect to an admission control policy with a relative ease.

7 Numerical examples

In this section we present numerical evaluation results that aim at illustrating practical aspects of the fair-rate sharing models in the CDMA link. Selected test cases emphasize on user perceived performance metrics and exemplify the role of different parameters that affect the system. We examine blocking systems, where blocking occurs as a result of an access or admission control policy. Other possible causes of losses, such as limited buffer capacities, will not be considered here. Further, in multiple class scenarios, we only present results for the simplest common access control policy. For other policies the problem is largely diversified, since control parameters must be selected based on specific fairness, blocking and utilization criteria, which calls for an optimization approach. Optimization concepts in a close-by context are discussed and examined in [26].

We also stress that wherever results are presented both for the GPS-Poisson and GPS-Engset models, these are not directly comparable. On this aspect, it is worth noting the following. We know that the Poisson source model can be considered as the limiting case of a finite source model when in the latter the number of class-k sources $M_k \to \infty$ and the associated load $\rho_k^{Eng.} \to 0, \forall k = \{1,\dots,K\}$. This is so since $\binom{M_k}{x_k}(\rho_k^{Eng.})^{x_k} \to \frac{(\rho_k^{Poiss.})^{x_k}}{x_k!}$, for $M_k \cdot \rho_k^{Eng.} = \rho_k^{Poiss.}$. However, in the non-limiting case, one cannot choose values of traffic parameters such that the two models are comparable.

[19] By choosing appropriate values for S, $\mathbf{M}$, the systems examined in § 4 can be superset systems.

Uplink	*Downlink*
$W = 3.84$ Mcps	
$\Theta_u = 1 - 10^{-1}$	$\Theta_d = 0.8 - 10^{-2}$ ($\psi = 0.2$)
$(E_b/N_0)_u = 1$ dB	$(E_b/N_0)_d = 2.5$ dB
$f_u = 0.75$	$f_d = 0.55$
	$\alpha = 0.1$

Table 1: Numerical values

CDMA parameter values are taken from [17] and, unless specified elsewhere, are those described in Table 1. The total capacity values Θ_d, Θ_u are chosen according to Eqs. (5),(14) roughly as follows: in the downlink, for a background noise level of -100 dBm and a path loss exponent 4 in an urban environment, this yields a total BS output power of about 10 Watt for a mobile located at an average distance of 1 km. Likewise, in the uplink, for the same distance the chosen capacity corresponds to a power transmitted from the mobile of about 1 Watt. The E_b/N_0 targets are set for static users and 64 kbps data service. Considering more generally a static environment, a small value of the non-orthogonality factor α in the downlink has been assumed. Finally, we set intercell interference factors f_u, f_d to relatively high values, where a typically larger value is common in the uplink.

We begin by presenting a 'congestion diagram', showing the deterioration in total link throughput as the number of flows in the CDMA link increases. Link throughput deterioration is attributed to increased intracell interference (this is transparent in the derivation of the model in § 2). Further, intercell interference, as shown by the ratios f_u, f_d has a significant downgrading effect on performance. We also remark the very fast convergence of the throughput to asymptotic values. For a relatively small number of flows in the system, the throughput quickly approaches its minimum value R_{min}. For a number of 4 flows, the total throughput is nearly within 10% of its limit values for all cases studied. This is an important feature of the model that distinguishes our analysis of the CDMA link.

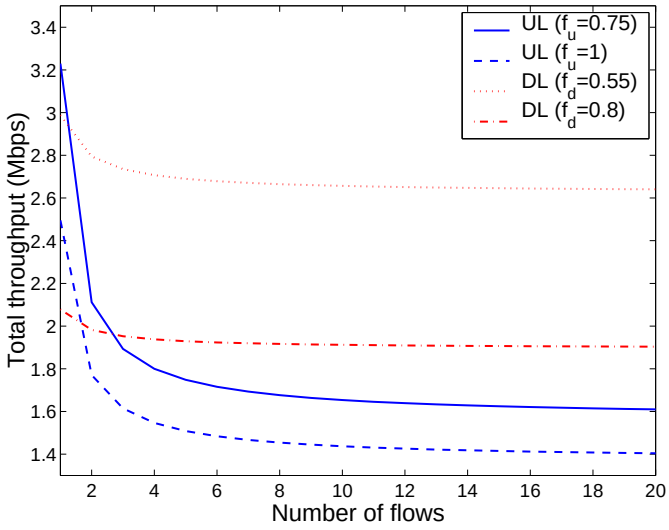


Figure 1: Link throughput deterioration in a CDMA link as the number of flows increases.

We also make the following observations regarding the bottleneck of the CDMA system. One the one hand, we have that the DL has higher E_b/N_0 requirements and smaller available resources, as a portion of the total capacity goes to SCH and CCH channels. Higher E_b/N_0 values in the DL are mainly due to smaller receiver sensitivity and antenna gain in the mobile units. Additionally,

antenna diversity which improves signal quality is not usually assumed in the DL. However, in a static configuration as presented here, these are largely eclipsed by the increased intracell interference in the UL, with a growing number of mobiles. In fact, as we see in the diagram, for almost all values of f_u, f_d, the UL has a higher throughput only for a single user in the cell (i.e. no intracell interference), while as the number of users increases the throughout rapidly drops.

Notwithstanding, further results can show that the bottleneck side is the opposite in the case of user mobility and increased intracell interference in the DL (cf. [23]). Practically, the CDMA bottleneck side is difficult to derive absolutely in the case of symmetric traffic on both sides (for data applications, there is usually much larger traffic carried on the DL). In reality, with time-varying channel and traffic conditions, both sides may be the bottleneck at one time or another.

On what concerns the processor-sharing model, the behavior is qualitatively the same both in the uplink and the downlink. Not to reiterate or confound the evaluation, we present hereafter results based on numerical values and formulae in the uplink which has been shown to be the bottleneck, while speaking more generally about CDMA and implicitly extending the conclusions to the downlink.

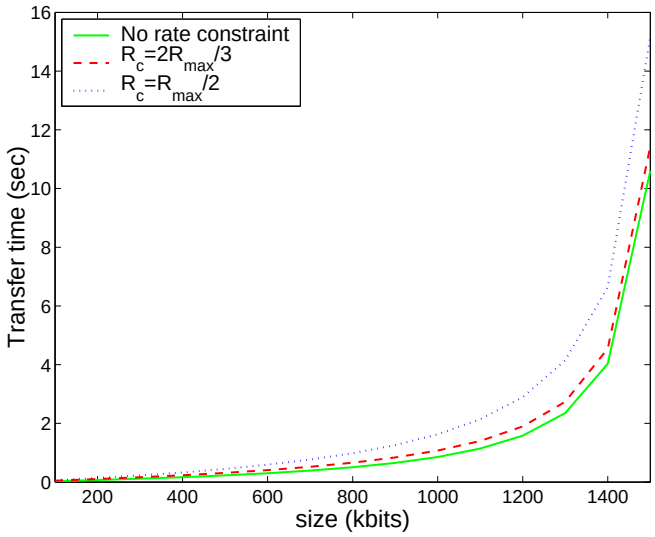


Figure 2: Single-class GPS-Poisson system with no blocking and maximum rate constraints.

For the case of non-real-time traffic the most important measure of performance is the transfer time of a flow. The behavior in the case of a class of identical traffic can be rather anticipated. Considering a single class Poisson arrival system without blocking and arrival rate $\lambda = 1$, it is shown in Fig. 2 that the transfer time increases with the average size of the flows. A more abrupt increase is observed when the size approaches values for which the system would be unstable.

We also study on this diagram a possible contraint that may exist on the transmission rate of flows on the link. More specifically, we have implicitly assumed so far that we are able to transmit on the link at the rate specified by the available capacity. The maximum throughput R_{max} is then attained when a single flow is transmitted on the link. However, the throughput of flows is often also limited by constraints other than interference, such as the modulation scheme, the handling of packets in limited-size buffers, the specific error correction/detection mechanisms, etc. Thus a total rate limit, say R_c, may exist in such cases. It is then easy to compute the relevant degradation in performance by considering a new service function

$$f'(n) = \begin{cases} R_c/n & \text{, if} \quad nf(n) > R_c \\ f(n) & \text{, if} \quad nf(n) \leq R_c \end{cases}$$

and applying the same analysis. As is shown in Fig. 2, the impact of a rate constraint increases for greater transfer times. It is worth noting that constraints may also be imposed on the individual flow

rates. However, apart from the –similar to the above– case where a common rate limit is imposed, the general case with different rate limits for each flow cannot be handled by the model in this paper.

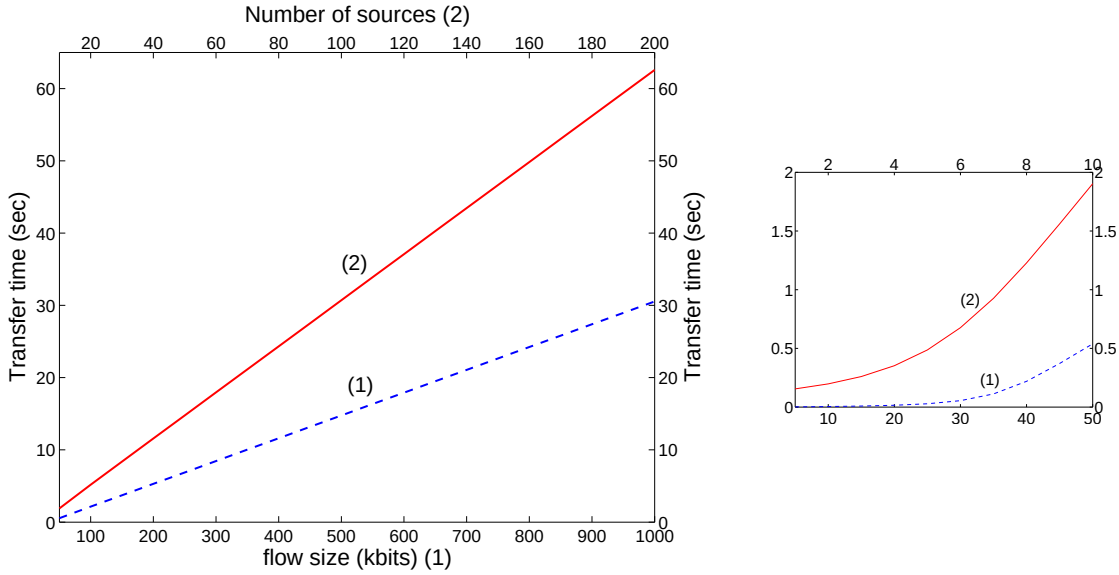
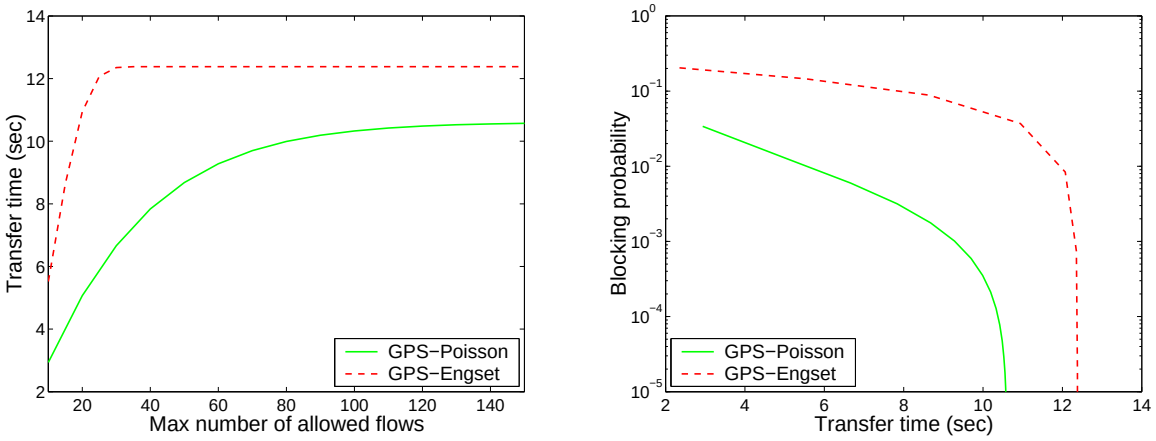


Figure 3: Expected transfer time in a GPS-Engset system without blocking as a function of mean flow size (1) and source population (2).

The elemental behavior of the GPS-Engset system is demonstrated in Fig. 3, where we depict the transfer time as a function of size (1) and the source population (2) in a system without blocking ($M \geq S$). The think time duration has been fixed to 1 sec, and we have taken a source population of $S = 50$ in (1) and a mean file size $E[\sigma] = 500$ kbits in (2). These curves appear to be linear, however in reality the slope of both changes very slowly (in the case of curve (1), it tends to 1 as $E[\sigma] \to \infty$, cf. Eq.(38)) and both curves are convex everywhere. This is shown better by taking a finer scale and smaller values of S, $E[\sigma]$ in the adjacent smaller graph. However, for practical purposes one may consider the evolution of the expected transfer time to be approximately linear for appropriately chosen small intervalls of the x-axis values. Approximately linear behavior of the mean transfer time with respect to mean file size occurs also in a blocking system, whereas for the GPS-Poisson model it only occurs for values far from saturation. These observations complement the useful results regarding the linearity of the conditional expectation presented in Theorems 4.1 and 4.3.

We illustrate the role of admission control on data traffic by showing the transfer time as a function of the maximum number of allowed flows, both in a GPS-Poisson and in a GPS-Engset model. Traffic parameters have been chosen such that the two models have a close behavior and are as follows. GPS-Poisson: $E[\sigma] = 1500$ kbits, $\lambda = 1$, GPS-Engset: $S = 50$, $E[\sigma] = 1500$ kbits, $E[\tau] = 33.3$ sec. It is shown in Fig. 4(a) that the transfer time can be restrained by limiting the maximum number of flows, in accordance with a certain blocking probability. In fact, since both the probability of blocking and transfer time are QoS parameters, an appropriate setting must be chosen from a block-delay diagram, shown in Fig. 4(b).

We begin to examine multiple-class cases. To avoid confusion from dimensionality, we restrict to the case of two classes. We then keep the total load on the link constant and define the *load coefficient* as ρ_1/ρ. This is a convenient way to evaluate the interaction between the two classes of flows. Other traffic parameters are chosen to be symmetric for both classes, since it is not our purpose to study specific test cases. However, the simple results presented here give a clear view of differences between the finite and infinite source models and can provide valuable intuition about the behavior of the system in more advanced cases.



(a) (b)

Figure 4: Admission control for data traffic flows in a single class GPS-Poisson and GPS-Engset system.

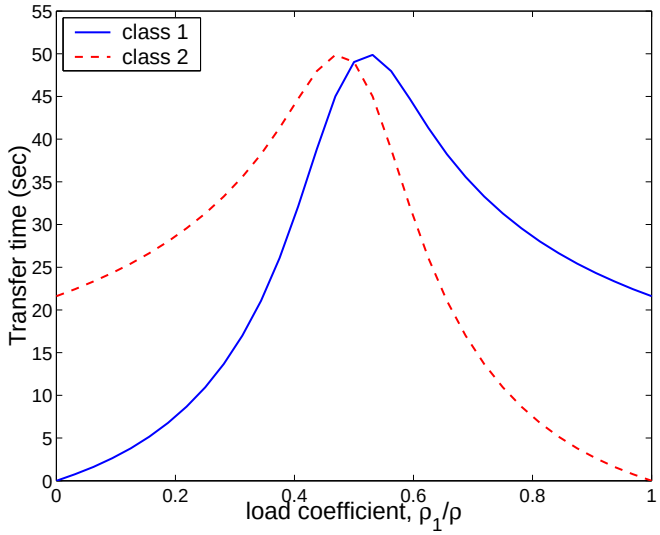
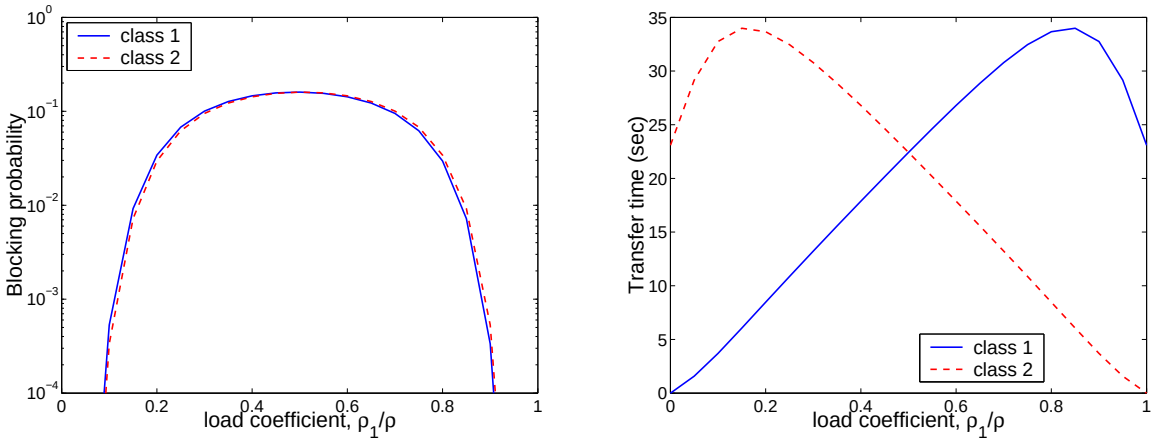


Figure 5: Expected transfer time of class 1 and class 2 flows in a GPS-Poisson model, for varying load coefficient ρ_1/ρ.

For Poisson arrivals, we depict in Fig. 5 the expected transfer time of the two classes of flows as the load coefficient changes. Arrival rates are taken constant, so that changes in the value of ρ_1/ρ correspond to increasing or decreasing mean flow sizes. More specifically, the setting is as follows: $M = 50$, $\rho = 1600$, $\lambda_1 = \lambda_2 = 1$. This corresponds to a blocking probability of 0.018. Clearly, as the relative load of one class increases, the transfer times of its flows in the link also increase.

In Fig. 6(a),(b) we depict the blocking probability and expected transfer time of classes 1, 2 with varying load coefficient, in a GPS-Engset model. The setting is as follows: $M = 50$, $S_1 = S_2 = 30$, and $E[\tau_1] = E[\tau_2] = 5$ sec. We observe that the blocking probability is nearly the same for the two sources. This is a result of the common constraint on the number of allowed flows. It can also be seen in Fig. 6(a) that for a relatively large range of values of the load coefficient, $0.2 \leq \rho_1/\rho \leq 0.8$, the blocking probability remains high, while it rapidly diminishes for smaller values (this is more evident here since the blocking probability of one class is zero in the absence of flows of the other class).



(a) (b)

Figure 6: Blocking probability (a) and expected transfer time (b) of class 1 and class 2 flows in the GPS-Engset system, for varying load coefficient ρ_1/ρ.

In the graph of the expected transfer time in Fig. 6(b), there is again an approximately linear increase in transfer time as the load of one class increases. With regard to the expected transfer time in both the GPS-Poisson and the GPS-Engset model, we also remark that the absence of, or very small load of one class permits the other to substantially reduce its transfer time. This can be deduced from the rapid fall of all curves as the load of one class diminishes. Thus in the processor-sharing system a class of flows takes advantage of low or intermittent traffic of other classes and obtains a better performance. Have in mind also that the discrimination of traffic into classes may well be an artificial one, so this observation carries over to any group of flows (or a single flow) with respect to the others.

8 Model extensions

We end by referring to a modeling extension related to this work. One can consider the transmission of flows on a link in a much more sophisticated and realistic way, by extending to a *session* model. A session is defined as the transmission of a specified number of flows, with possibly different service requirements, each flow being followed by an associated think period. The extension regards both the fact that we have control over the number of emitted flows as well as that stochastic dependencies between service and think periods may exist; these are common in the transfer of data in telecommuni-

cations networks. For example, it is most likely that think times are positively correlated with service requirements, a voluminus piece of data being usually followed by a longer idle period.

The idea is contained in [5],[13] (amongst other works of the same authors) in the context of TCP networks but can also be conveyed in the processor-sharing models of a CDMA link studied here. As it is explained in these works, it is possible to model such complex systems by taking advantage of the queueing network structure in [10],[19], for multiple-class systems. We expand a little on the relevant setting: we may consider a queueing network with a 'service' and 'think' station. A primitive, or basic class is used to dinstinguish flows with a given service requirement and think time distribution. To specify the number of flows in a session, a class may generate or terminate subclasses (with possibly different characteristics) by appending appropriate routing probabilities after the completion of a service cycle, defined by the exit from the think station.

Based on this main structure, it is then possible to consider any kind of class structures or correlations between flows, either for a finite source model or infinite Poisson arrivals. However, the derivation of sojourn times in the first is a complicated and unsolved problem. Besides this, the key issue here is what to model, so that we get an idea of behavior without having to specify so many classes that the system becomes untractable. Also it is desirable to investigate more into appropriate traffic models for a mobile environment. The nature of this traffic is difficult to determine, in view of the variety of data services to be offered in future wireless networks. Finally, another problem is the study of an access or admission control policy at a session level, since a user expects to maintain the same QoS throughout the whole session, and not just for the transmission of individual flows.

Appendix

A GPS-Engset from GPS-Poisson

The GPS-Engset station can be viewed as a closed network of two stations in tandem, a think station and the GPS station (see Fig. 7).

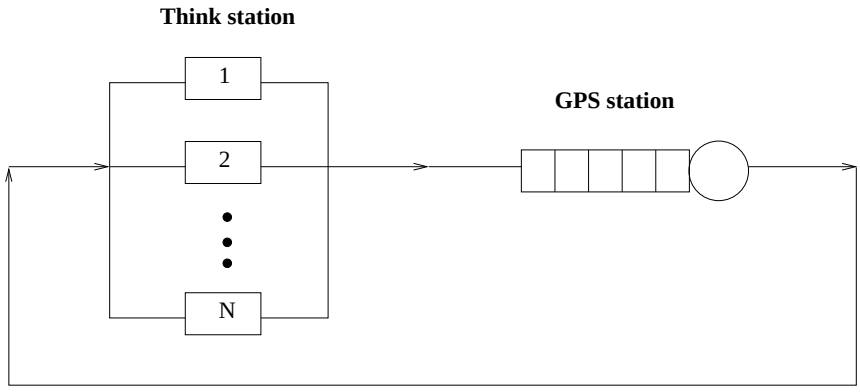


Figure 7: The GPS-Engset station

We will consider multiple classes of flows in both stations, from the set $\mathcal{K} = \{1, \ldots, K\}$. N in the figure is the total number of sources from all classes, i.e. $S_1 + S_2 + \cdots + S_K = N$. In the Markovian case, we adopt the more familiar notation that class-k sources in the think station have inter-generation times that are exponentially distributed with mean $1/\nu_k$. Additionally, in the GPS station flow sizes are exponentially distributed with mean $1/\mu_k$.

If the stations were in isolation and fed by Poisson arrivals, it is easily shown that they are both *quasi-reversible*. The first is a classic Erlang station for which this result is well-known (see e.g. [22]). It is easy to show this for the GPS station as well, given the known form of the stationary distribution

from (19). Consider the transition rates of the forward process $q(\boldsymbol{x}', \boldsymbol{x})$ between states $\boldsymbol{x}'$ and $\boldsymbol{x}$, denoted as vectors of the number of flows from each class. If the stationary distribution at state $\boldsymbol{x}$ is denoted by $\pi(\boldsymbol{x})$, the transition rates of the reversed process are defined from

$$q'(\boldsymbol{x}, \boldsymbol{x}') = \frac{\pi(\boldsymbol{x}')}{\pi(\boldsymbol{x})} \cdot q(\boldsymbol{x}', \boldsymbol{x}). \tag{45}$$

Consider now in the reversed process the arrival of a class-k flow at state $\boldsymbol{x} = (x_1, x_2, \ldots, x_K)$. The next state is, in vector notation, $\boldsymbol{x}' = \boldsymbol{x} + \mathbf{1}_k$, where $\mathbf{1}_k$ is the unity vector with 1 at the k-th position and 0s elsewhere.

From the form of the stationary distribution in (19) and from (45) we have that

$$\begin{aligned} q'(\boldsymbol{x}, \boldsymbol{x}') &= \frac{\rho_k}{f(x_1 + \cdots + x_k)(x_k + 1)} \cdot \mu_k f(x_1 + \cdots + x_k)(x_k + 1) = \\ &= \lambda_k. \end{aligned}$$

Hence the arival rate in the reverse process equals the arrival rate in the forward and is independent of the state of the process. This identifies quasi-reversibility.

Therefore, we have a closed network of quasi-reversible stations. Then from [19], the stationary distribution of the tandem connection has the form

$$\pi(\boldsymbol{S} - \boldsymbol{x}, \boldsymbol{x}) \propto \pi_1(\boldsymbol{S} - \boldsymbol{x}) \cdot \pi_2(\boldsymbol{x}),$$

where π_1, π_2 denote the marginal distributions of the two stations with Poisson arrivals of rate λ_k for each class. We thence have

$$\pi(\boldsymbol{S} - \boldsymbol{x}, \boldsymbol{x}) \propto \prod_{k \in \mathcal{K}} \frac{(\lambda_k/\nu_k)^{S_k - x_k}}{(S_k - x_k)!} \frac{(\lambda_k/\mu_k)^{x_k}}{x_k!} \phi(x_1 + \cdots + x_K).$$

This can then be normalized to yield the stationary distribution (31).

Finally, we note that a similar approach is used to derive the stationary distribution from the BCMP model [3].

B Numerical computations and recursive algorithms

The results regarding the steady state characteristics presented in § 4, mainly involving the computation of a normalization constant, have a deceptively simple form. In practice, the computation of this constant may be very difficult in cases where a large of classes or a large number of jobs in the system are involved. The risk is that it may lead to overflows, rounding errors, or, more generally numerical instabilities in the arithmetic operations. This situation may be aggravated considering decreased numerical capabilities in a mobile computing environment. Therefore, it is of interest to develop efficient computational algorithms for the GPS models.

In the following we develop recursive algorithms for this task, both for GPS-Poisson and GPS-Engset models. The main focus is on the calculation of the normalization constant, however in the Poisson case other measures can be computed recursively as well. It should be emphasized that these recursive algorithms apply only to models with common access. Other access control policies should be examined separately.

B.1 GPS-Poisson

In the Poisson arrivals case, efficient recursive algorithms can be constructed easily by considering the state description with the total number of jobs in the system. From $G(M) = \sum_{z=0}^{M} \frac{\rho^z}{g(z)}$, an easy first result is:

$$G(M) = G(M-1) + \frac{\rho^M}{g(M)}, \tag{46}$$

with $G(0) = 1$. This is a sufficient recursion for calculating all values of the normalization constant. Further, since $B(M) = 1 - \frac{G(M-1)}{G(M)}$ the blocking probabilities can be computed from the same recursion. However, we note that in practice the normalization constants may be quite large, which could lead to numerical instabilities in taking their ratio and then subtracting from unity. It is therefore of interest to derive a recursion solely for the blocking probabilities. We may apply the following manipulation:

$$B(M) = \frac{\frac{\rho^M}{g(M)}}{G(M)} = \frac{\frac{\rho^{M-1}}{g(M-1)} \cdot \frac{\rho}{Mf(M)}}{G(M-1) + \frac{\rho^M}{g(M)}} = \frac{\frac{\rho}{Mf(M)}}{\frac{G(M-1) + \frac{\rho^M}{g(M)}}{\frac{\rho^{M-1}}{g(M-1)}}}.$$

Therefore,

$$B(M) = \frac{\frac{\rho}{Mf(M)}}{\frac{1}{B(M-1)} + \frac{\rho}{Mf(M)}}, \tag{47}$$

where $M \geq 1$ and $B(0) = 1$.

More awkward recursions can be derived for the mean number of jobs in the system and the mean sojourn time of a class-k job. Define $\mathcal{E}(M) := \{E[N] \mid max.\ jobs = M\}$. We arrive at the following recursion:

$$\mathcal{E}(M) = \frac{\mathcal{E}(M-1) + \frac{\rho}{f(M)} \cdot B(M-1)}{1 + \frac{\rho}{Mf(M)} \cdot B(M-1)}, \tag{48}$$

in terms of $\mathcal{E}(M-1)$ and $B(M-1)$, where $\mathcal{E}(0) = 0$. Since $E[N_k] = \frac{\rho_k}{\rho} E[N]$, it is also straightforward to compute the mean number of class-k jobs recursively. By denoting also the mean sojourn time of a class-k job as

$$T_k(M) = \frac{\rho_k}{\rho} \cdot \frac{\mathcal{E}(M)}{\lambda_k(1 - B(M))}$$

we end up in:

$$T_k(M) = \frac{\rho_k}{\rho} \cdot \frac{\mathcal{E}(M-1) + \frac{\rho}{f(M)} \cdot B(M-1)}{\lambda_k[1 - B(M)][1 + \frac{\rho}{Mf(M)} \cdot B(M-1)]}. \tag{49}$$

Note that in all the above recursions at most one or two previous values need to be stored, which makes these relationships very attractive for computations in systems with limited memory space. Finally, we may express distributions in terms of normalization constants, so as to compute probability measures more efficiently. For example, we have:

$$\Pr\{N \geq a\} = \frac{\sum_{n=a}^{M} \frac{\rho^n}{n!}}{G(M)} = \frac{G(M) - G(a)}{G(M)} = 1 - \frac{G(a)}{G(M)}. \tag{50}$$

Additionally,

$$\begin{aligned} \Pr\{N = a\} &= \Pr\{N \geq a-1\} - \Pr\{N \geq a\} = \\ &= \frac{G(a) - G(a-1)}{G(M)}. \end{aligned} \tag{51}$$

Then probability measures for a certain class of jobs derive according to its load in the system.

B.2 GPS-Engset

In the case of a finite source model with Engset-like arrivals, the derivation of efficient recurrence relationships is much more involved. A first observation is that the cardinality of the feasible set is much greater, and that it grows more than linearly with the number of sources from each class, as well as with the common limit on the number of jobs. This will make the computation more lengthy and

increase storage requirements. However, the real difficulty comes from the existence of the ϕ-function in the steady-state characteristics.

For example, if it weren't for the ϕ-function, the folowing recurrence relation would apply for the modified Engset system (we maintain the notation introduced in § 4.2):

$$G(\mathbf{S}, M) = G(\mathbf{S} - \mathbf{1}_k, M) + \rho_k \cdot G(\mathbf{S} - \mathbf{1}_k, M - 1).$$

Further, the same recursion with ρ_k replaced by ρ_k/c would apply in case the service rate function remained constant, $f(n) = c\ \forall n$.

Unfortunately however, this doesn't work here. Furthermore, other Buzen-type [7] recursions do not lead anywhere. In addition, note that we cannot simplify things by considering one system state, as in the Poisson arrivals model. Hence, it seems highly unlikely that any kind of recurrence relationship exists for the normalization constant.

Nevertheless, we can consider an approach similar to the famous Kaufman–Roberts recursion ([18],[25]), and try to recursively compute the probability

$$Q(\mathbf{S}, n) = \sum_{\substack{\mathbf{x} \in \mathcal{F}(\mathbf{S}, M) \\ x_1 + \cdots + x_K = n}} p(x_1, \ldots, x_K).$$

In this way, we are only occupied with the sum of the variables which enters as an argument in the ϕ-function. Also, we can straightforwardly obtain the normalization constant since

$$G(\mathbf{S}, M)^{-1} = Q(\mathbf{S}, 0).$$

We develop a recurrence relation for $Q(\mathbf{S}, n)$ as follows:

$$\begin{aligned}
Q(\mathbf{S}, n) &= \sum_{x_1=0}^{S_1} \cdots \sum_{\substack{x_k=0 \\ x_1+\cdots+x_K=n}}^{S_k} \cdots \sum_{x_K=0}^{S_K} \prod_{j=1}^{K} \binom{S_j}{x_j} \rho_j^{x_j} \cdot \phi(n) = \\
&= \sum_{x_1=0}^{S_1} \cdots \sum_{\substack{x_k=0 \\ x_1+\cdots+x_K=n}}^{S_k} \cdots \sum_{x_K=0}^{S_K} \left[\binom{S_k - 1}{x_k} + \binom{S_k - 1}{x_k - 1} \right] \prod_{\substack{j=1, \\ j \neq k}}^{K} \binom{S_j}{x_j} \rho_j^{x_j} \rho_k^{x_k} \cdot \phi(n) = \\
&= \sum_{x_1=0}^{S_1} \cdots \sum_{\substack{x_k=0 \\ x_1+\cdots+x_K=n}}^{S_k - 1} \cdots \sum_{x_K=0}^{S_K} \binom{S_k - 1}{x_k} \prod_{\substack{j=1, \\ j \neq k}}^{K} \binom{S_j}{x_j} \rho_j^{x_j} \rho_k^{x_k} \cdot \phi(n) + \\
&\quad + \sum_{x_1=0}^{S_1} \cdots \sum_{\substack{x_k=1 \\ x_1+\cdots+x_K=n}}^{S_k} \cdots \sum_{x_K=0}^{S_K} \binom{S_k - 1}{x_k - 1} \prod_{\substack{j=1, \\ j \neq k}}^{K} \binom{S_j}{x_j} \rho_j^{x_j} \rho_k^{x_k} \cdot \phi(n)
\end{aligned}$$

By performing the change of variable $x'_k = x_k - 1$ in the second large summand above and noting that $\phi(n) = \frac{\phi(n-1)}{f(n)}$, it can readily be seen that

$$Q(\mathbf{S}, n) = Q(\mathbf{S} - \mathbf{1}_k, n) + \frac{\rho_k}{f(n)} Q(\mathbf{S} - \mathbf{1}_k, n - 1), \tag{52}$$

where we insist that $n \geq 1$. This is a recursion in two arguments which, together with the condition $\sum_{n=1}^{M} Q(\mathbf{S}, n) = 1\ \forall\, \mathbf{S}$ and 'appropriate' initial conditions, can eventually give the value of the normalization constant. By appropriate initial conditions we mean any non-zero value of $Q(\mathbf{S}, n)$ which can be easily computed.

We should note that this recursive formula may have again high computational and storage requirements, but perhaps it is the only plausible one. To a large extent, the efficient computation of the normalization constant in the GPS-Engset system remains an open problem. Another, more powerful approach one can take is the inversion of the probability generating function for the stationary distribution (see e.g. [22]). This can yield exact (by a direct numerical inversion) or approximate solutions for the normalization constant.

References

[1] 3GPP, "QoS concept and architecture", 3GPP Recommendation TS 23.107, v. 5.3.0, 2002.

[2] S. Asmussen, *Applied probability and queues*, 2nd Edition, Springer–Verlag, 2003.

[3] F. Baskett, K.M. Chandy, R.R. Muntz, F.G. Palacios, "Open, closed, and mixed networks of queues with different classes of customers", *Journal of the ACM*, 22(2), 248–260, April 1975.

[4] P. Bender, P. Black, M. Grob, R. Padovani, N. Sindhushayana, A. Viterbi, "CDMA/HDR: A bandwidth-efficient high-speed wireless data service for nomadic users", *IEEE Communications Magazine*, 70–77, July 2000.

[5] T. Bonald, A. Proutière, G. Régnié, J. Roberts, "Insensitivity results in statistical bandwidth sharing", *Proc. ITC 17*, Brazil, December 2001.

[6] D. Burman, "Insensitivity in queueing systems", *Advances in Applied Probability*, 13, 846–859, 1981.

[7] J.P. Buzen, "Computational algorithms for closed queueing networks with exponential servers", *Communications of the ACM*, 16(9), 527–531, September 1973.

[8] J.W. Cohen, "The generalized Engset formulae", *Philips Telecommunication Review*, 18(4), 158–170, November 1957.

[9] J.W. Cohen, *On regenerative processes in queueing theory*. Lecture Notes in Economics and Mathematical Statistics, Springer–Verlag, 1976.

[10] J.W. Cohen, "The multiple phase service network with generalized processor sharing", *Acta Informatica* 12, 245–284, Springer-Verlag, 1979.

[11] F. Delcoigne, A. Proutière, G. Régnié, "Modeling integration of streaming and data traffic", *Performance Evaluation*, 55(3-4), 185–209, Elsevier Science, 2004.

[12] P. Franken, D. König, U. Arndt, V. Schmidt, *Queues and Point Processes*, John Wiley & Sons, 1982.

[13] S. Ben Fredj, T. Bonald, A. Proutière, G. Régnié, J.W. Roberts, "Statistical bandwidth sharing: A study of congestion at flow level", *Proc. ACM Sigcomm '01*, San Diego, California, USA, August 2001.

[14] N. Hegde, E. Altman, "Capacity of multiservice WCDMA Networks with variable GoS", *Proc. of IEEE WCNC*, New Orleans, Louisiana, USA, March, 2003.

[15] D.P. Heyman, T.V. Lakshman, A.L. Neidhardt, "A new method for analysing feedback-based protocols with applications to engineering web traffic over the Internet", *Proc. ACM Sigmetrics 1997*, Seattle, USA.

[16] K. Hiltunen, R. De Bernardi, "WCDMA downlink capacity estimation", *Proc. VTC Spring 2000*, Tokyo, Japan, May 2000.

[17] H. Holma, A. Toskala (Eds.), *WCDMA for UMTS: Radio access for third generation mobile communications*, 2nd Edition, John Wiley & Sons, 2002.

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

[19] F.P. Kelly, *Reversibility and stochastic networks*, John Wiley & Sons, 1979.

[20] J.M. Kelif, E. Altman, "Admission and Gos control in multiservice WCDMA system", *Proc. ECUMN '04*, Porto, Portugal, October 2004.

[21] K. Knopp, *Theory and application of infinite series*, Dover Publications, 1990.

[22] H. Kobayashi, B.L. Mark, "Product-form loss networks", In J.H. Dshalalow (Ed.): *Frontiers in queueing: Models and Applications in Engineering and Science*, 147–195, CRC Press, 1997.

[23] I. Koukoutsidis, E. Altman, J.M. Kelif, "A non-homogeneous QBD approach for the admission and GoS control in a multiservice WCDMA system", INRIA Research Report No. RR–5358, November 2004.

[24] A.K. Parekh, R.G. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: The single-node case", *IEEE/ACM Transactions on Networking*, 1(3), 344–357, 1993.

[25] J.W. Roberts, "A service system with heterogeneous user requirements – Application to multiservices telecommunications systems", In G. Pujolle (Ed.): *Performance of Data Communication Systems and their Applications*, 423–431, North-Holland Publishing Company, 1981.

[26] K.W. Ross, *Multiservice loss models for broadband telecommunication networks*, Springer–Verlag, 1995.

[27] R. Schassberger, "Insensitivity of steady-state distributions of generalized semi-Markov processes with speeds", *Advances in Applied Probability*, 10, 836–851, 1978.

[28] R. Schassberger, "Two remarks on insensitive stochastic models", *Advances in Applied Probability*, 18, 791–814, 1986.

[29] K. Sipilä , Z.-C. Honkasalo, J. Laiho-Steffens, A. Wacker, "Estimation of capacity and required transmission power of WCDMA downlink based on a downlink pole equation", *Proc. VTC Spring 2000*, Tokyo, Japan, May 2000.

[30] A.M. Viterbi and A.J. Viterbi, "Erlang capacity of a power controlled CDMA system", *IEEE Journal on Selected Areas in Communications*, 11(6), 892–900, 1993.

[31] R.W. Wolff, *Stochastic modeling and the theory of queues*. Prentice–Hall, Inc., 1989.

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Unité de recherche INRIA Futurs : Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)
Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399