

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

Ioannis Koukoutsidis¹, Eitan Altman¹ and Jean-Marc Kelif²

¹ INRIA, B.P. 93, 06902 Sophia-Antipolis, France
{gkoukout, altman}@sophia.inria.fr

² France Telecom R&D, Rue du General Léclerc, 92794 Issy-les-Moulineaux, France
jeanmarc.kelif@francetelecom.com

Abstract. In this paper we describe a modeling approach for studying fair rate sharing in 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 in 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 addition, we revisit insensitivity and truncation properties of the stationary distributions encountered in GPS models, and extend to different access control policies.

Keywords: CDMA, elastic traffic, fairness, generalized processor sharing, insensitivity, access control.

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 on 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 that belong to the same service class transmit or receive at the same rate, regardless of their position in the cell, or their channel conditions. Then, given a number of active 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 ([7],[10]) have addressed 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 [5]. 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.

We focus on the transmission of non-real-time or elastic traffic. Traffic of this 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].

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

The major performance metrics we examine are then the transfer times of flows over the link and their blocking probabilities, 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 use the term ‘flow’ to refer to distinct ‘jobs’ in the system. Similarly to [6], a flow represents a stream of packets that have some criteria in common, 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.

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 sojourn time of jobs in the case of a multiple class GPS-Engset system, with or without blocking. Based on that, we derive an 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.

2 The processor-sharing model

The capacity and associated throughput analysis in a CDMA link is based on determining a set of feasible *states* of transmission that the system can theoretically sustain under specific signal power constraints; this being subject to the condition that SIR requirements for all transmissions are satisfied. Given a number n of simultaneously active mobiles, the total maximum link throughput that can be sustained under equal rate transmissions writes [7, 10]:

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

$$\text{Downlink: } R_{tot}(n) = \frac{n\Theta_d}{n(\alpha + f_d) - \alpha\Theta_d} \left(\frac{N_0}{E_b} \right)_d W. \quad (2)$$

Here $(E_b/N_0)_{u,d}$ are the energy per bit to noise density requirements and W is the chip rate. In the uplink, f_u is the intercell to intracell interference ratio; a similar parameter f_d is employed in the downlink for the average received intercell to intracell power ratio. Also in the downlink, parameter α depicts the non-orthogonality factor, i.e. the fraction of received own cell power experienced as intracell interference by a mobile, due to multipath propagation. The values of $\Theta_{u,d}$ are chosen according to physical limitations in the power of a mobile device or base station.

The underlying analysis in [7, 10] has shown that the maximization of throughput in the uplink 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. Likewise, in the downlink 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.

Remark 1. For the model analysis, we must assume very fast closed-loop power control, as well as completely *fluid* traffic subject to an ideal rate transmission control, with negligible feedback delay between the receiver and the source. Further, regarding 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 find a practical implementation.

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: $0 \leq f(n) < \infty$, $n \cdot f(n) < \infty$.

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 $R_{min} = \lim_{n \rightarrow \infty} R_{tot}(n)$ as the number of flows tends to infinity. This manifests the impact of interference, which restrains the total throughput of the link.

It can readily be seen that this model belongs to the class of *generalized processor sharing* models, as these were defined and studied by Cohen¹ [5]. This permits us to incorporate results regarding the steady-state characteristics of a system with Poisson arrivals of flows, as well as Engset-like arrivals from a finite source population, directly from [5]. In all cases, the ramifications of the analysis in our case are discussed and extensions to Cohen's work, where needed, are provided. For brevity, proofs to propositions and theorems presented here are omitted. Interested readers are referred to our research report [13].

3 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.

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*³. The service requirement is translated to the size, or traffic volume of the flow to be transmitted over the link.

3.1 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 , $k = 1, 2, \dots, K$. 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 ($k = 1, \dots, K$) and their attained services, denoted by the vector $\bar{\sigma}_k = (\sigma_k(1), \dots, \sigma_k(x_k))$, by finding the solution to a system of integro-differential equations. We have for the defined probability⁴ $p(x_k, \sigma_k(h); h = 1, \dots, x_k, k = 1, \dots, K) d\bar{\sigma}_k := \Pr\{N_k = x_k, \sigma_k(h) \leq \sigma_k(h) \leq \sigma_k(h) + d\sigma_k(h); h = 1, \dots, x_k, k = 1, \dots, K\}$ the density (Theorem 7.2 of [5]):

$$p(x_k, \sigma_k(h); h = 1, \dots, x_k; k = 1, \dots, K) = p_{\bar{0}} \cdot \phi(x) \prod_{k=1}^K \frac{(\rho_k)^{x_k}}{x_k!} \prod_{h=1}^{x_k} \frac{1 - F_{\sigma}^k(\sigma_k(h))}{E[\sigma_k]}, \quad (3)$$

¹ Nowadays it has prevailed that the term GPS be used for another class of queueing models, namely *Weighted Fair Queueing* [14]. However we prefer to stick to the original denomination in this work.

³ 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.

⁴ The bold notation in $\sigma_k(h)$ is just used to distinguish the random variable.

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 transform⁵. Here we denote $x = x_1 + x_2 + \dots + x_K$ and $p_{\bar{0}} = \left(\sum_{z=0}^{\infty} \frac{\rho^z}{z!} \cdot \phi(z) \right)^{-1}$ is the probability that the system is empty. This is also termed as the *normalization constant*⁶ of the system. By $\phi(n) = \left(\prod_{i=1}^n f(i) \right)^{-1}$ we denote a very useful function, encountered everywhere in the course of our study. By definition, $\phi(0) = 1$. It is worth noting that Eq. (3) 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 (3) 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 . 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’ [9], 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:

$$\Pr\{N = n\} = G^{-1} \frac{\rho^n}{g(n)}, \quad \text{where } G := \sum_{z=0}^{\infty} \frac{\rho^z}{g(z)}. \quad (4)$$

We will study the condition under which this stationary distribution exists in our system. We have the following:

Proposition 1. *The stochastic process of the number of flows in the GPS-Poisson system has a stationary distribution if and only if $\rho < R_{min}$.*

Remark 2. In the Markovian case, we have a necessary and sufficient *ergodicity* condition. Further in such a case, this Proposition can also be viewed as a special case of the ergodicity theorem presented in [12] 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.

Provided that an equilibrium distribution exists, we can easily derive the mean total number of flows in the system, $E[N]$. From [9] (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 deduce 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.

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 (4) 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 [15]). The expected transfer time of a class- k flow can also be derived by applying Little’s law.

Finally, an important result concerns the conditional transfer time of a flow whose service requirement is known deterministically. The result is stated in [5] only for the non-blocking system. However, it also applies to a system with blocking. By simply solving for the mean number of “deterministic flows” from (4) and applying Little’s law, we have that:

Theorem 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*

⁵ 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 arbitrary distributions.

⁶ Efficient methods for computing the normalization constant in both GPS-Poisson and GPS-Engset systems are discussed in [13].

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 $E[\sigma]$.

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.

3.2 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 on the link and a maximum number of simultaneously served flows, M . We consider a situation similar to 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 represent 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 ($k = 1, \dots, K$). 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 $[0, \infty)$. 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 [5] to derive the joint density of the number of flows from each class that are receiving service, $N_k^{(s)}$ ($k = 1, \dots, K$) (and thus the remaining idle flows, $N_k^{(i)}$), their attained service denoted by the vector $\bar{\sigma}_k = (\sigma_k(1), \dots, \sigma_k(x_k))$, as well as the time spent in the idle phase of the remaining flows from each class, denoted by $\bar{\tau}_k = (\tau_k(1), \dots, \tau_k(S_k - x_k))$.

Maintaining an analogous notation as in § 3.1, we define the infinitesimal probability:

$$p(x_k, \sigma_k(h), \tau_k(m); h = 1, \dots, x_k, m = 1, \dots, S_k - x_k, k = 1, \dots, K) d\bar{\sigma}_k d\bar{\tau}_k := \\ \Pr\{N_k^{(s)} = x_k, N_k^{(i)} = S_k - x_k, \sigma_k(h) \leq \sigma_k(h) \leq \sigma_k(h) + d\sigma_k(h), \tau_k(m) \leq \tau_k(m) \leq \tau_k(m) + d\tau_k(m); \\ h = 1, \dots, x_k, m = 1, \dots, S_k - x_k, k = 1, \dots, K\}.$$

For shortness, we denote the density by $p(x_k, \bar{\sigma}_k, \bar{\tau}_k; k = 1, \dots, K)$. We have⁷:

$$p(x_k, \bar{\sigma}_k, \bar{\tau}_k; k = 1, \dots, K) = p(x_1, x_2, \dots, x_K) \cdot \prod_{k=1}^K \left\{ \prod_{h=1}^{x_k} \frac{1 - F_\sigma^k(\sigma_k(h))}{E[\sigma_k]} \right\} \left\{ \prod_{h=1}^{S_k - x_k} \frac{1 - F_i^k(\tau_k(h))}{E[\tau_k]} \right\} \quad (5)$$

with

$$p(x_1, x_2, \dots, x_K) = \frac{\prod_{k=1}^K \binom{S_k}{x_k} \rho_k^{x_k} \cdot \phi(x_1 + x_2 + \dots + x_K)}{\sum_{z_1=0}^{S_1} \sum_{z_2=0}^{S_2} \dots \sum_{z_K=0}^{S_K} \prod_{k=1}^K \binom{S_k}{z_k} \rho_k^{z_k} \cdot \phi(z_1 + z_2 + \dots + z_K)} \quad (6)$$

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

Remark 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. This derivation is included in [13].

⁷ This expression is explicitly mentioned in [5] only for the 2-class case. Its general form can be derived from Eqs. (5.9),(5.10) in the original manuscript, 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, when 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 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, $m = 1, \dots, K$, is given by*

$$P_B^k = \frac{\sum_{x_1+x_2+\dots+x_K=M} \prod_{\substack{k=1 \\ k \neq m}}^K \binom{S_k}{x_k} \binom{S_m-1}{x_m} \rho_k^{x_k} \rho_m^{x_m} \cdot \phi(x_1 + x_2 + \dots + x_K)}{\sum_{z_1=0}^{S_1} \dots \sum_{z_m=0}^{S_m-1} \dots \sum_{z_K=0}^{S_K} \prod_{\substack{k=1 \\ k \neq m}}^K \binom{S_k}{z_k} \binom{S_m-1}{z_m} \rho_k^{z_k} \rho_m^{z_m} \cdot \phi(z_1 + z_2 + \dots + z_K)}, \quad (7)$$

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

Remark 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 [4], as well as the generalized Engset loss station [11].

The next important result concerns the derivation of the transfer time of a class- k flow in the system. We can only present it in the form of a conjecture:

Conjecture 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 ($k = 1, \dots, K$), is given by

$$E[T_k] = \frac{E[\sigma_k]}{\frac{\sum_{(\mathbf{x} \in \mathcal{F}(\mathbf{S}, M))} x_k f(x_1 + \dots + x_K) p(x_1, \dots, x_K)}{\sum_{(\mathbf{x} \in \mathcal{F}(\mathbf{S}, M))} x_k p(x_1, \dots, x_K)}}, \quad (8)$$

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

Remark 5. This result is shown by employing a regenerative process, considering the distribution of the idle time of each class- k flow to be the convolution of an arbitrary and a negative exponential one. It is extremely difficult to show the continuity of the expected sojourn time as the mean of the exponential tends to 0. Without this rigorousness, this has to be taken only as a conjecture (cf. [13]).

This conjecture also has an intuitive explanation. In the denominator of the complex fraction in (8), 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.

Moreover, in an analogous manner to the Poisson arrivals system, the conditional sojourn time of a deterministic size flow is proportional to its size. If Conjecture 1 is true, then the following result also applies to a system with blocking.

Theorem 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 think 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 and the same mean service requirement and idle time.*

4 Insensitivity properties and access control policies

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 system⁸ by applying the well-known *truncation principle* [9]. We are able to establish these properties by an easier and more general method. This can then permit 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.

We use the setting of a *generalized semi-Markov process* (GSMP) and follow the method of Burman ([3]) to establish insensitivity. Denote by $N(t)$ the GSMP of the number of flows in such a setting. We say that transitions from one state to another are caused by the occurrence 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 \mathbf{E} , and we denote by $e(g)$ a subset of this space associated with a single state g . The event $e \in \mathbf{E}$ requires X_e processing units to be completed. 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 $1/\mu_e$. 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), also called the *associated life process* of $N(t)$. Moreover, the Markov process that comes out if all X_e were exponentially distributed is called the *corresponding* Markov process of $N(t)$.

The major result in [3] states that if a set of equations called *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:

$$p(g, t) = p(g) \cdot \prod_{e \in e(g)} \mu_e (1 - H_e(x)), \quad (9)$$

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.

For both the GPS-Poisson and GPS-Engset models it can be shown that the restricted flow equations reduce to the *detailed balance equations*, which are satisfied since the corresponding Markovian processes are reversible (cf. [13]). Reversibility of the two processes also accounts for the ensuing truncation properties, and allows us to extend to other blocking regimes and access control policies.

We may distinguish the following general families of access 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, additionally because they give rise to issues of fairness and good utilization of the access space.

Common access:

$$\begin{aligned} \mathcal{F}(M) &= \{x_k \geq 0 : \sum x_k \leq M; k = 1, \dots, K\} && \text{(GPS-Poisson)} \\ \mathcal{F}(\mathbf{S}, M) &= \{x_k \geq 0 : \sum x_k \leq M; x_k \leq S_k, k = 1, \dots, K\} && \text{(GPS-Engset)} \end{aligned}$$

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

Dedicated access:

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

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. The disadvantage is a potential smaller utilization of the access space.

⁸ 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’.

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:

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

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:

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

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*.

All the above policies are *coordinate convex* [8], since departures (or arrivals) are never blocked. The question of how to analytically model different access control policies is not difficult to answer. In fact, under the GPS service discipline, 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 at the beginning of this section. First consider the associated reversible⁹ Markov process, and apply the truncation principle from a corresponding superset system. Then, follow Burman’s method to demonstrate insensitivity. The equivalent 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. Moreover, the same reasoning permits to obtain blocking probabilities and sojourn times in multiple class, GPS-Engset systems, under different access control policies, by considering the analogous results in § 3.2.

5 Numerical examples

Practical aspects of the modeling analysis are illustrated in this section, based on numerical evaluations. CDMA parameter values are taken for static users and 64 kbps data service¹⁰. Fig. 1 shows the deterioration in total uplink (UL) and downlink (DL) throughput as the number of flows increases, attributed to increasing intracell interference. Notice the fast convergence of the throughput to its asymptotic values: for a number of 4 flows, the total throughput is nearly within 10% of its limit values for all cases studied. For the case of Fig. 1, the UL shows a higher throughput only for a single user in the cell, and thus it restrains capacity of the system. However, the bottleneck side may be the opposite in case of user mobility and increased intracell interference in the DL (cf. [12]). In general, the CDMA bottleneck is difficult to determine and with time varying channel and traffic conditions both sides may restrain capacity at one time or another.

On what concerns the processor-sharing model, the behavior is qualitatively the same both in the UL and DL. Hereafter results refer to the UL, which has been shown to be the bottleneck. Considering a single class Poisson arrival system without blocking, 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 constraint 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 and coding, 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 consider a new service function $f(n)$ in the system and apply the GPS model analysis. Fig. 2 shows that the impact of a rate constraint increases

⁹ It can easily be shown that all coordinate convex policies correspond to reversible stochastic processes.

¹⁰ For details on set parameters, see [13].

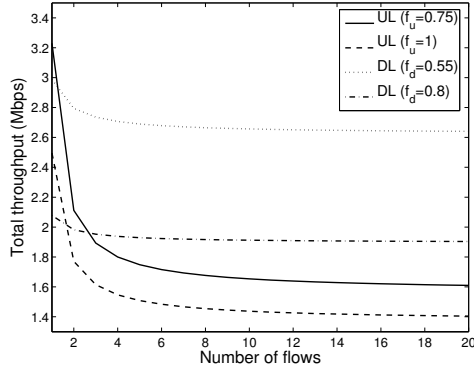


Fig. 1. Total CDMA link throughput as the number of flows increases.

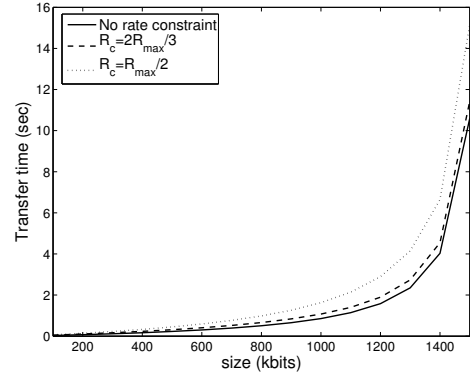


Fig. 2. Single-class GPS-Poisson system with no blocking and maximum rate constraints.

for greater transfer times. It is worth noting that constraints may also be imposed on the individual flow rates. However, apart from a common rate limit, the general case with different rate limits for each flow cannot be handled by the model in this paper.

The role of admission control on data traffic is illustrated by showing the blocking-delay trade-off in Fig. 3, for a GPS-Poisson and a GPS-Engset model. Traffic parameters have been chosen such that the two models have a close behavior. The transfer time can be restrained by limiting the maximum number of flows, in accordance with a certain blocking probability. Since both the probability of blocking and transfer time are QoS parameters, an appropriate setting must be chosen from this trade-off.

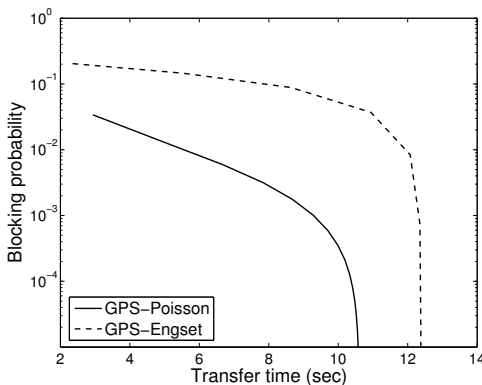


Fig. 3. Block-delay diagram in a single-class GPS-Poisson and GPS-Engset system.

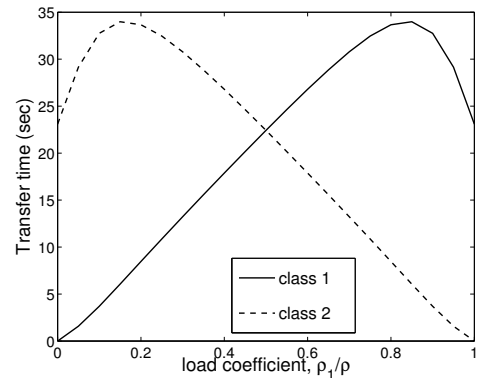


Fig. 4. Expected transfer time of class 1 and class 2 flows in a GPS-Engset system, for varying load coefficient ρ_1/ρ .

We give an example of transfer time behavior in case of two classes of traffic and common access control in Fig. 4. We keep the total load constant and define the load coefficient as ρ_1/ρ . This is a convenient way to evaluate the interaction between the two classes. Other traffic parameters are chosen to be the same for both classes. In a GPS-Engset model, there is an approximately linear increase in transfer time as the load of one class increases. Further, in both GPS-Poisson and GPS-Engset systems a major observation is that the absence of, or very small load of one class permits the other to substantially reduce its transfer time. 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.

6 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. We can then also consider correlations between successive service and think periods, which are common in the transfer of data in telecommunications networks. For example, it is most likely that think times are positively correlated with service requirements, a voluminous piece of data being usually followed by a longer idle period. Also, correlation dependencies may be defined for the sizes of flows following the sending of an initial flow, since this initial flow usually specifies the purpose of sending data over the link.¹¹

These ideas are contained in [2],[6] 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 [5],[9], 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 distinguish 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 possible to consider any kind of class structures or correlations between flows, either for finite sources or Poisson arrivals. However, the derivation of sojourn times in the first would be extremely complicated. Besides this, the key issue 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.

References

1. 3GPP. QoS concept and architecture. 3GPP Recommendation TS 23.107, v. 5.3.0, 2002.
2. T. Bonald, A. Proutière, G. Régnié, J. Roberts. Insensitivity results in statistical bandwidth sharing. *Proc. ITC 17*, Brazil, 2001.
3. D. Burman. Insensitivity in queueing systems. *Adv. Appl. Prob.*, 13: 846–859, 1981.
4. J.W. Cohen. The generalized Engset formulae. *Philips Telecomm. Review*, 18(4): 158–170, 1957.
5. J.W. Cohen. The multiple phase service network with generalized processor sharing. *Acta Informatica* 12: 245–284, 1979.
6. 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, USA, 2001.
7. N. Hegde, E. Altman. Capacity of multiservice WCDMA Networks with variable GoS. *Proc. IEEE WCNC*, New Orleans, USA, 2003.
8. J.S. Kaufman. Blocking in a shared resource environment. *IEEE Trans. Commun.*, 29(10): 1474–1481, 1981.
9. F.P. Kelly. *Reversibility and stochastic networks*. John Wiley & Sons, 1979.
10. J.M. Kelif, E. Altman. Admission and Gos control in multiservice WCDMA system. *Proc. ECUMN '04*, Porto, Portugal, 2004.
11. 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.
12. 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, 2004.
13. I. Koukoutsidis, E. Altman, J.M. Kelif. Fair rate sharing models in a CDMA link with multiple classes of elastic traffic. INRIA Research Report No. RR–5596, 2005.
14. A.K. Parekh, R.G. Gallager. A generalized processor sharing approach to flow control in integrated services networks: The single-node case. *IEEE/ACM Trans. Networking*, 1(3): 344–357, 1993.
15. R.W. Wolff. *Stochastic modeling and the theory of queues*. Prentice–Hall, Inc., 1989.

¹¹ Dependencies between successive service requirements, as well as between successive think times of the same class can already exist in the GPS-Engset model, since we can loosen the independence assumptions, see [13].