Beyond Processor Sharing

Samuli Aalto[1], Urtzi Ayesta[2], Sem Borst[3,4,5], Vishal Misra[6], Rudesindo Núñez-Queija[4,7]

[1]TKK Helsinki University of Technology, Networking Laboratory,
P.O. Box 3000, FI-02015 TKK, Finland

[2]LAAS-CNRS, 7 Avenue Colonel Roche, 31077 Toulouse Cedex, France

[3]Bell Laboratories, Alcatel-Lucent,
P.O. Box 636, Murray Hill, NJ 07974, USA

[4]CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

[5]Department of Mathematics & Computer Science, Eindhoven University of Technology,
P.O. Box 513, 5600 MB Eindhoven, The Netherlands

[6]Department of Computer Science, Columbia University, New York, NY 10027, USA

[7]TNO Information and Communication Technology, Delft, The Netherlands

samuli.aalto@tkk.fi, urtzi@laas.fr, sem@win.tue.nl, misra@cs.columbia.edu, sindo@cwi.nl

ABSTRACT

While the (Egalitarian) Processor-Sharing (PS) discipline offers crucial insights in the performance of fair resource allocation mechanisms, it is inherently limited in analyzing and designing differentiated scheduling algorithms such as Weighted Fair Queueing and Weighted Round-Robin. The Discriminatory Processor-Sharing (DPS) and Generalized Processor-Sharing (GPS) disciplines have emerged as natural generalizations for modeling the performance of such service differentiation mechanisms. A further extension of the ordinary PS policy is the Multilevel Processor-Sharing (MLPS) discipline, which has captured a pivotal role in the analysis, design and implementation of size-based scheduling strategies. We review various key results for DPS, GPS and MLPS models, highlighting to what extent these disciplines inherit desirable properties from ordinary PS or are capable of delivering service differentiation.

Keywords: Discriminatory Processor Sharing; Generalized Processor Sharing; Multilevel Processor Sharing; asymptotic analysis; insensitivity; queue length; size-based scheduling; slowdown; service differentiation; sojourn time; delay minimization; workload.

1. INTRODUCTION

Over the past few decades, the Processor-Sharing (PS) discipline has gained a prominent role in evaluating the performance of a variety of resource allocation mechanisms. Originally, the PS paradigm emerged as an idealization of Round-Robin scheduling algorithms in time-shared computer systems [46]. In recent years, the PS discipline has received renewed attention as a convenient abstraction for modeling the flow-level performance of bandwidth-sharing protocols in packet-switched networks, in particular TCP [13, 40, 60, 69, 70]. Although it is well-known that the queue length distribution in a PS queue with Poisson arrivals has a simple geometric distribution, regardless of the service requirement distribution, the sojourn time distribution is far less tractable, even for exponential service requirements. We refer to [21] for an extensive survey of results for the sojourn time distribution in PS queues.

While the PS model provides valuable insights, it critically relies on the assumption that the service capacity is shared in an egalitarian manner among competing users. The actual service shares may however show substantial variation among users with heterogeneous characteristics. For example, TCP flows that share a common bottleneck link but traverse distinct routes, may experience diverse packet loss rates and round-trip delays. Because of TCP mechanics, these differences result in a significant discrepancy in the bandwidth shares, see for instance [8].

Besides TCP-related effects, the heterogeneity in bandwidth shares may also be due to deliberate service differentiation among competing flows. As the Internet evolves to support an ever increasing range of services, there is a growing need for some form of service differentiation to accommodate the diverse Quality-of-Service requirements and traffic characteristics of heterogeneous applications. The ability to provide different bandwidth shares is arguably one of the most fundamental vehicles for service differentiation [16]. Discriminatory packet scheduling algorithms, such as Weighted Fair Queueing (WFQ) and Weighted Round-Robin (WRR), have been proposed as potential instruments to implement differentiated bandwidth sharing. Both WFQ and WRR constitute weighted versions of standard round-robin scheduling where the various users may receive different service quanta as specified by user-specific service weights.

The *Discriminatory* Processor-Sharing (DPS) discipline [34, 46, 61] provides a natural approach for modeling the flow-level performance of TCP with asymmetric bandwidth shares or differentiated scheduling mechanisms such as WFQ and WRR. DPS is a multi-class extension of the egalitarian PS policy, where the various classes are assigned positive weight factors. The service capacity is shared among all users present in proportion to the respective

class-dependent weights. In case all weight factors are equal, the DPS discipline reduces to the egalitarian PS policy.

The DPS discipline shows some resemblance with the *Generalized* Processor-Sharing (GPS) policy (or Generalized Head-Of-the-Line (HOL) PS), where the service capacity is also shared in accordance with class-dependent weight factors. In GPS however, the capacity is not divided among all users present, but distributed across (nonempty) classes irrespective of the actual number of users present. By guaranteeing a certain minimum rate to each back-logged class, the GPS discipline provides isolation among competing classes. In addition, it achieves statistical multiplexing gains by reallocating surplus bandwidth from non-backlogged classes.

While both the DPS and GPS disciplines arise as continuous idealizations of discriminatory packet scheduling algorithms, the difference in names reflects the traffic scenario and time scale considered in performance studies. The DPS discipline serves to evaluate flow-level performance, i.e., the interaction among a dynamic population of flows, where faster fluctuations at the packet level are assumed to average out. In contrast, the GPS discipline aims to describe the performance at the packet level, where the population of flows may be assumed (nearly) static due to the slower dynamics. Adding to the ambiguity in terminology, the term GPS was first used by Cohen [26] in the seventies to describe an extension of PS with state-dependent service rates, but here we adhere to the convention adopted after the seminal work in the nineties of Parekh & Gallager [64, 65].

A further extension of the ordinary PS policy is the family of Multilevel Processor-Sharing (MLPS) strategies as introduced in [45]. While DPS and GPS mainly serve to provide service differentiation, the goal of MLPS is to improve the overall performance by exploiting the variability in service demands, and in particular giving precedence to shorter requests over longer ones. The rationale for size-based scheduling has been amplified by findings that file sizes show extreme variability and commonly have heavy-tailed characteristics [9, 25, 28, 53].

An MLPS discipline is parameterized by a number of levels or thresholds, which are used to dynamically categorize jobs based on their current amounts of attained service. Lower classes (jobs with smaller amounts of attained service) receive priority over higher classes, and within each class the jobs may be served according to the ordinary PS policy or the Foreground-Background Processor-Sharing (FBPS, or just FB) discipline. The MLPS strategy may thus be interpreted as a coarse-grained approximation of the FBPS discipline. In fact, the family of MLPS strategies span a wide spectrum of policies, covering both the strict FBPS discipline and the ordinary PS policy as extreme (limiting) cases, and lend themselves better for actual implementation, especially at high line speeds. (We refer to [15] for an in-depth discussion of implementation issues involved in size-based scheduling.)

As described above, the DPS, GPS and MLPS strategies all three constitute extensions of the ordinary PS discipline. Given that strong commonality, we survey in the present paper various key results for DPS, GPS and MLPS models in a unified framework, emphasizing to what extent these disciplines inherit favorable properties from ordinary PS or are effective in providing service differentiation.

The remainder of the paper is organized as follows. In Section 2 we present a model description and introduce some notation. In Section 3 we review results for DPS models. Section 4 discusses GPS models. Since the results for GPS models center around workload characteristics rather than sojourn time distributions, we keep the section brief. We refer to [71] for a more elaborate overview of the GPS literature. Section 5 provides a survey of MLPS models. In Section 6 we make some concluding remarks.

2. MODEL DESCRIPTION AND NOTATION

We study a single-resource system of service capacity 1 with Poisson arrivals at rate λ and i.i.d. service requirements with distribution $F(x) = P\{B \leq x\}$ for all $x \geq 0$, where the generic random variable B denotes a single service requirement, with mean $E[B]$ and load $\rho = \lambda E[B]$. We assume that the system works at full rate 1 whenever it is nonempty, so that this M/G/1 system is stable (it empties in finite time) when $\rho < 1$. For convenience, we will impose additional technical assumptions on the service requirement distribution when needed.

In this paper we study several performance measures. With the random variable N we generically denote the number of jobs in steady state, while T denotes the sojourn time of an arbitrary job. Whenever appropriate, we supplement these variables with subscripts (for example N_j) when we wish to distinguish among different types of jobs in the system, or with an additional argument (for example $T(x)$) if the random variable is conditioned on the size of the particular job under consideration (i.e. x). As for the service requirements, we denote the mean of any random variable Y by $E[Y]$.

3. DISCRIMINATORY PROCESSOR SHARING

We will restrict our literature overview of DPS to some papers that are key to the results presented here. For an extensive overview of the literature on DPS we refer to the recent survey [7] and to [21] for results on tail distributions.

DPS is a class-based differentiation mechanism. Jobs are categorized into K different classes. Here we will assume that each new job is of class k with probability p_k, independent of the classes all previous jobs belong to. Therefore, the arrival processes of all individual classes are (independent) Poisson processes as well, with arrival rates $\lambda_k = p_k \lambda$. For class k we use B_k, $\rho_k = \lambda_k[B_k]$ and N_k to denote the service requirement, the class load and the number of jobs in system (in steady state). Naturally, $\rho = \sum_{k=1}^{K} \rho_k$.
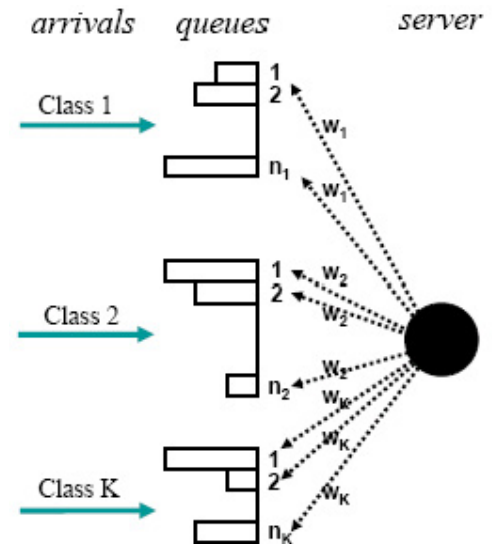

Figure 1: A queue operating under DPS

DPS operates as illustrated in Figure 1. Jobs of class k are assigned weights $w_k(> 0)$. If there are $N_l = n_l$ jobs of class $l =$

$1, 2, \ldots, K$, then *each job* of class k receives a fraction $w_k / \sum_{l=1}^{K} n_l w_l$ of the server's capacity.

The analysis of DPS is extremely difficult compared to that of egalitarian PS, which arises as a special case when all w_k are equal. Most notably, the simple geometric queue length distribution for the standard PS discipline does not have any counterpart for DPS. For PS, this distribution is fully determined by the traffic load ρ and, thus, not affected by higher moments of the service requirement distributions, a property which is commonly referred to as *insensitivity*. Only recently, Avrachenkov *et al.* [11] established that the mean queue lengths of all classes are finite under the usual stability condition, regardless of the higher-order moments of the service requirements. Some of the insensitivity properties of PS only carry over to DPS in various *asymptotic regimes*. We will discuss several such regimes in the following subsections.

3.1 Expected sojourn time for large jobs

In their seminal paper Fayolle, Mitrani & Iasnogorodski [34] showed that the asymptotic (mean) slowdown ratio (the ratio of the conditional mean sojourn time and the service requirement) under DPS is insensitive and independent of the job class:

$$E[R_k(x)] \equiv \frac{E[T_k(x)]}{x} \to \frac{1}{1 - \rho}, \qquad \text{as } x \to \infty. \quad (1)$$

Note that this relation holds with equality for all x in the case of egalitarian PS. The asymptotic insensitivity extends to DPS systems with capacity fluctuating according to a Markovian process, and to other service disciplines such as LAS and SRPT as well [60]. More recently, this property was studied for a broad class of service disciplines by Harchol-Balter *et al.* [39].

Assuming service requirement distributions with *finite variance*, Avrachenkov *et al.* [11] refined (1) for DPS and showed that the expected conditional response time of class k has an asymptote with slope $1/(1 - \rho)$ and a finite bias given by

$$\lim_{x \to \infty} \left(E[T_k(x)] - \frac{x}{1 - \rho} \right) = \frac{\sum_j \lambda_j (1 - \frac{w_k}{w_j}) E[(B_j)^2]}{2(1 - \rho)^2}.$$

It is worth emphasizing that the asymptotic bias with respect to the linear term depends only on the second moments of the service requirement distributions of *other classes* (or, in fact, of classes with different weights). In particular, the second moment of B_k does not impact this result, no matter how large its value. This suggests that the assumption of a finite variance for B_k is likely to be an unnecessary restriction. We will see similar technical restrictions in the following section.

3.2 Heavy-tailed asymptotics

A second asymptotic regime in which DPS exhibits some sort of *insensitivity* has received considerable attention over the last decade. It was shown in [22] that the following asymptotic equivalence — first established for PS by Zwart & Boxma [77] — extends to DPS as well, basically under the condition of regularly varying service requirement distributions with finite variance:

$$\frac{P\{T_k > x\}}{P\{B_k > (1 - \rho)x\}} \to 1, \qquad \text{as } x \to \infty. \quad (2)$$

Again, the "scaling factor" $1 - \rho$ is insensitive and common to all classes. The additional distributional assumption that the variance is finite, is not essential. Indeed, relation (2) was shown to hold for a wider class of service distributions in [21], with an additional condition on the loads. For completeness we list the precise conditions under which the above equivalence has been established. Recall that B_k has the distribution of the service requirements of

class k, B that of the service requirements of an arbitrary customer (i.e., not conditioned on the class) and N is the total number of jobs in the system. Then (2) has been proved under the following conditions:

(i) B_k and B are both regularly varying of index $\nu > 2$ (this ensures a finite variance);

(ii) B_k is regularly varying of index $\nu > 1$ and $P\{N > n\} \leq ab^n$ for all n, with $a < \infty$, $b < 1$;

(iii) B_k is regularly varying of index $\nu > 1$, there exists an $\eta > 0$ so that $E[B^{1+\eta}] < \infty$, and $\max\{w_1, w_2, \ldots, w_K\} \sum_{i=1}^{K} \rho_i / w_i < 1$.

The third set of conditions does not require finite variances of the service requirement distributions, but imposes a seemingly unnecessarily strong condition on the loads. The second condition does not explicitly require finite variances, but to date it is not known whether the queue length decays geometrically, so that it may still implicitly rely on service requirements with finite variance. We are inclined to conjecture that neither a finite variance, nor the restrictive load condition in the third set are necessary for (2) to hold. A further extension of this result to scenarios with a time-varying service capacity can also be found in [21].

For light-tailed service requirements, these insensitivity properties do not hold [21]. For this case, Egorova *et al.* [31] recently studied the logarithmic estimates of the tail of the sojourn time distribution using large-deviations techniques.

3.3 Separation of time scales

Thirdly, the entire joint queue length *distribution* under DPS turns out to be insensitive under a time-scale separation regime. Suppose that the classes are ordered in decreasing order of time scale, i.e., class k operates on a much *faster* time scale than class $k + 1$, for all $k = 1, 2, \ldots, K - 1$. Formally, this can be achieved by replacing the arrival rate of class k by $\lambda_k f_k(r)$ and letting the service requirements of class k be distributed as $B_k / f_k(r)$, where we take the parameter $r \to \infty$ and the functions f_k are such that $f_{k+1}(r)/f_k(r) \to 0$ as $r \to \infty$. Note that in this procedure, the loads ρ_k remain unchanged when r is varied. In the limit, a strict *separation of time scales* occurs, allowing for exact analysis of the queue length distribution. Under the technical assumption of phase-type distributions [43], this yields the following result for $K = 2$. (All limits in this subsection hold as $r \to \infty$, which we omit for compactness of the presentation.)

$$P_r\{N_1 = n_1 | N_2 = n_2\}$$
$$\to \frac{\Gamma(n_1 + \frac{n_2 w_2}{w_1} + 1)}{\Gamma(n_1 + 1)\Gamma(\frac{n_2 w_2}{w_1} + 1)} \rho_1^{n_1} (1 - \rho_1)^{\frac{n_2 w_2}{w_1} + 1}.$$

Here $\Gamma(x) = \int_0^\infty e^{-u} u^{x-1} \mathrm{d}u$ and the subscript r reflects the dependence of the involved random variables on r. In particular,

$$E_r[N_1 | N_2 = n_2]$$

The limiting distribution of the slow class is

$$P_r\{N_2 = n_2\} \to \left(1 - \frac{\rho_2}{1 - \rho_1}\right)\left(\frac{\rho_2}{1 - \rho_1}\right)^{n_2},$$

so that $E_r[N_2] \to \dfrac{\rho_2}{1 - \rho}$, and

$$E_r[N_1] \to \left(\frac{w_2}{w_1}\frac{\rho_2}{1 - \rho} + 1\right)\frac{\rho_1}{1 - \rho_1}.$$

Note that these limiting expressions are entirely insensitive to characteristics of the service requirement distributions other than the

means. Another interesting insight is that the limiting distribution of N_2 (the slow class) is independent of the weights, whereas that of N_1 does depend on the ratio of the weights.

For more than two classes the above scaling leads to an insightful recursion for the limiting queue lengths:

$$E_r[N_i|N_{i+1} = n_{i+1}, \ldots, N_K = n_K]$$
$$\to \left(\frac{\sum_{j=i+1}^{K} n_j w_j}{w_i} + 1\right)\frac{\rho_i}{1 - \sum_{j=1}^{i} \rho_j},$$

and $E_r[N_K] \to \frac{\rho_K}{1 - \rho}$.

3.4 Heavy-traffic regime

The final asymptotic regime that we discuss, giving rise to *"near" insensitivity* is that of a DPS system operating under heavy load conditions, i.e., $\rho = \sum_{k=1}^{K} \rho_k \to 1$. We fix the service requirement distributions and the class probabilities p_k, while increasing the arrival rate, so that $\lambda \uparrow \hat{\lambda} := \left(\sum_k p_k E[B_k]\right)^{-1}$. We denote the limiting loads of all classes by $\hat{\rho}_k = \hat{\lambda} p_k E[B_k]$. As may be expected, in this regime the numbers of jobs in the system of all classes will increase unboundedly. However, when scaled with the *common* factor $1 - \rho$, the joint queue length distribution has a proper limiting distribution.

The DPS queue in heavy traffic was analyzed by Grishechkin [36] assuming finite second moments of the service requirement distributions. Subsequently, assuming exponential service requirement distributions, a direct approach to establish a heavy-traffic limit for the joint queue length distribution was described by Rege & Sengupta [66] and extended to *phase-type* distributions in [44]: When scaling with $1 - \rho$, the queue length vector has a proper distribution as $\rho \to 1$,

$$(1 - \rho)(N_1, N_2, \ldots, N_K) \xrightarrow{d} X \cdot (\frac{\hat{\rho}_1}{w_1}, \frac{\hat{\rho}_2}{w_2}, \ldots, \frac{\hat{\rho}_K}{w_K}),$$

where $\xrightarrow{d}$ denotes convergence in distribution and X is an exponentially distributed random variable with mean

$$E[X] = \frac{\sum_k p_k E[(B_k)^2]}{\sum_k p_k E[(B_k)^2]/w_k},$$

which is equal to 1 when $w_k = 1$ for all k, i.e., in the case of standard PS.

Note that, although the limiting distribution depends on the second moment of the service requirement distributions through $E[X]$, the *impact of the second moment on $E[X]$ is uniformly bounded* and in particular $\min_i\{w_i\} \leq E[X] \leq \max_i\{w_i\}$. This implies that

$$\frac{\min_i\{w_i\}}{w_k} \leq \lim_{\rho \to 1} \frac{(1 - \rho)}{\rho_k} E_\rho[N_k] \leq \frac{\max_i\{w_i\}}{w_k}.$$

Here, the subscript ρ indicates the dependence of the distribution of N_k on ρ.

4. GENERALIZED PROCESSOR SHARING

The GPS discipline operates as illustrated in Figure 2. Traffic belongs to one of K classes, with class k assigned a positive weight $w_k > 0$. When all the classes are backlogged, i.e., have nonempty queues, the total service rate is divided in accordance with the weight factors w_k. When some of the classes are nonbacklogged, the surplus service rate is reallocated among the others in proportion to the weights w_k. It is important to note that, different
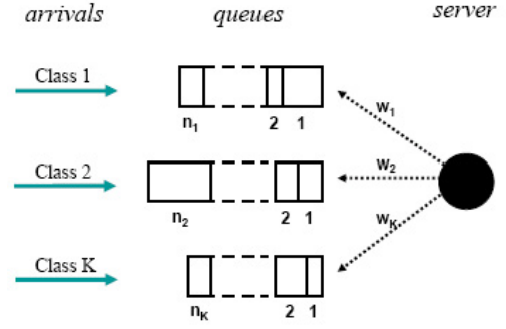

Figure 2: A queue operating under GPS

from DPS, the weight w_k is associated with the *entire queue* belonging to class k and not with each individual job in that queue. As the term Head-of-the-Line (HOL) in one of the original names suggests, it is commonly assumed that only the job or packet at the head of the queue of each class is served. However, this assumption is by no means critical, and the workload behavior is the same for all work-conserving strategies. In fact, the analysis of the GPS discipline extends to situations where the notion of traffic consisting of distinct jobs or packets may not even apply, like in the case of fluid input. In that case, some subtleties may arise since a class may generate traffic at a rate below its service share, and thus may be nonbacklogged and receive service at the same time [30].

The seminal work of Parekh & Gallager [64, 65] on GPS scheduling concerned the derivation of deterministic (worst-case) delay guarantees for leaky-bucket controlled traffic sources. Subsequent papers pursued the evaluation of statistical (average-case) performance measures such as loss probabilities, delay characteristics, and workload distributions. The exact analysis of GPS queues is intractable in all but a few special cases. In fact, in the case of two classes, a GPS system is equivalent to a so-called coupled-processors model, whose solution gives rise to rather formidable boundary value problems [27, 33, 37, 38, 47, 48]. (The GPS discipline also shows resemblance with so-called cycle stealing policies, see for instance [62].) Hence, most of the work has focused on various kinds of bounds and asymptotic approximations. Such asymptotic approximations tend to have a remarkably simple and insightful form, and are particularly valuable because the relevant Quality-of-Service measures typically pertain to the occurrence of extremely rare events with small probabilities.

Yaron & Sidi [73] derived bounds for GPS queues fed by so-called exponentially-bounded burstiness traffic. Bertsimas *et al.* [14], Massoulié [59], and Zhang [75] established large-deviations results for light-tailed traffic sources. Large-buffer asymptotics for heavy-tailed traffic processes were obtained in Borst *et al.* [17], Jelenković & Momčilović [42, 41], Kotopoulos *et al.* [49] and LeLarge [54]. Van Uitert & Borst [72] extended these results to networks of GPS queues. Borst *et al.* [19] analyzed the buffer asymptotics in a two-class GPS system with a mixture of heavy-tailed and light-tailed traffic.

The results in [17, 18, 19] reveal a sharp trichotomy in the qualitative behavior, depending on the traffic intensities and the relative weight values of the various classes. For a wide range of parameters, an individual class with heavy-tailed characteristics is effectively served at a *constant* rate, which is only influenced by the average rates of the other classes. In particular, such a class is essentially immune from the adverse impact of flows with 'heavier'-tailed characteristics. This phenomenon is commonly referred to

as a *reduced-load equivalence*, a term first coined in the context of fluid queues with subexponential activity periods [6], and is also recognized in the sojourn time asymptotics for heavy-tailed DPS queues described in Subsection 3.2. For other parameter regimes however, an individual class (either with heavy-tailed or light-tailed characteristics) may be strongly affected by excessive activity of 'heavier'-tailed flows, and may inherit their traffic characteristics. This phenomenon is sometimes called *induced burstiness*. Finally, for a class with light-tailed characteristics and a sufficiently high weight, the buffer asymptotics involve subtle combination of heavy-tailed and light-tailed large-deviations behavior. This is often referred to as a *reduced-peak equivalence* [20], in analogy with the term reduced-load equivalence. Similar types of qualitatively different regimes in different settings were observed in [23, 24].

While the above results concern so-called *large-buffer asymptotics*, *many-sources asymptotics* may be found in [29, 50, 51]. Mannersalo & Norros [58] developed accurate approximations for queues with Gaussian traffic processes in the many-sources regime. Mandjes & Van Uitert [56] further justified and refined these approximations, and established an interesting connection with tandem queues, see also [57].

All of the above papers focus on evaluating performance metrics for given scheduling weights. The reverse problem of selecting suitable weight factors so as to satisfy given Quality-of-Service requirements and the closely related issue of characterizing the admissible region, have received considerably less attention, see for instance [32, 52, 55, 63, 76].

5. MULTILEVEL PROCESSOR SHARING

Kleinrock [45, Section 4.7] introduced a class of nonanticipating and work-conserving disciplines called Multilevel Processor-Sharing (MLPS) disciplines. An MLPS discipline π is defined by a finite set of level thresholds $a_1 < \cdots < a_N$ defining $N + 1$ levels, $N \geq 0$. A job belongs to level n if its attained service is at least a_{n-1} but less than a_n, where $a_0 = 0$ and $a_{N+1} = \infty$. Between these levels, a strict priority discipline is applied with the lowest level having the highest priority. Within each level n, an internal discipline $D_n \in \{\text{FB}, \text{PS}, \text{FCFS}\}$ is applied.

Below we review recent results related to the mean sojourn time and slowdown ratio comparison with an ordinary PS policy, as well as the expected conditional sojourn time asymptotics, in a stable M/G/1 queue. Recall that $E[T(x)]$ denotes the expected conditional sojourn time of a job with service requirement x. As before, the corresponding slowdown ratio is denoted by $E[R(x)] = E[T(x)]/x$. The mean sojourn time and the mean slowdown ratio are respectively defined by

$$E[T] = \int_0^\infty E[T(x)]\mathrm{d}F(x), \quad E[R] = \int_0^\infty E[R(x)]\mathrm{d}F(x).$$

Finally, let Π denote the family of nonanticipating and work-conserving disciplines.

5.1 Mean sojourn time comparison for IMRL and DMRL service requirements

Define, for all $x \geq 0$,

$$H(x) = \frac{1 - F(x)}{\int_x^\infty (1 - F(y))\, \mathrm{d}y}.$$

This is the inverse of the mean residual lifetime function $1/H(x)$. A service requirement distribution belongs to class IMRL [DMRL] if $1/H(x)$ is increasing [decreasing] for all $x \geq 0$. In this case we use notation $F \in \text{IMRL}[\text{DMRL}]$.

When comparing the mean sojourn time of different disciplines for IMRL or DMRL service requirements in an M/G/1 queue, a key variable is the so-called *level-x (or truncated) workload* [5, 1], which refers to the sum of the remaining service requirements of those jobs in the system whose attained service is less than a given threshold x. By definition, the mean level-x workload $E[V_x^\pi]$ can be expressed as follows for any $\pi \in \Pi$ and $x \geq 0$:

$$E[V_x^\pi] = \int_{0^-}^{x^-} E[B - y \mid B > y]\, \mathrm{d}E[N_y^\pi]. \tag{3}$$

Here $E[N_y^\pi]$ refers to the mean number of those jobs in the system whose attained service requirement is less than y.

If $H(x)$ is monotonic, the mean sojourn times $E[T^\pi]$ and $E[T^{\pi'}]$ in two systems with disciplines $\pi, \pi' \in \Pi$, respectively, may be compared as follows [1, Equation (13)]:

$$E[T^\pi] - E[T^{\pi'}] = -\frac{1}{\lambda} \int_0^\infty (E[V_x^\pi] - E[V_x^{\pi'}]\, \mathrm{d}H(x).$$

This yields the following lemma.

Lemma 1 *Let* $\pi, \pi' \in \Pi$ *such that* $E[V_x^\pi] \leq E[V_x^{\pi'}]$ *for all* $x \geq 0$. *(i) If* $F \in \text{IMRL}$, *then* $E[T^\pi] \leq E[T^{\pi'}]$. *(ii) If* $F \in \text{DMRL}$, *then* $E[T^\pi] \geq E[T^{\pi'}]$.

By this approach, the following result has been obtained, which compares certain MLPS disciplines with the baseline discipline PS.

Theorem 1 *[1, Theorem 1] Let* $\pi \in \text{MLPS}$ *such that the internal disciplines belong to* $\{\text{FB}, \text{PS}\}$. *(i) If* $F \in \text{IMRL}$, *then* $E[T^\pi] \leq E[T^{\text{PS}}]$. *(ii) If* $F \in \text{DMRL}$, *then* $E[T^\pi] \geq E[T^{\text{PS}}]$.

This approach has also been used to demonstrate that FB does *not* minimize the mean sojourn time within the class IMRL [5], contrary to what was believed earlier [68].

5.2 Mean sojourn time comparison for DHR and IHR service requirements

Assume now that the service requirement distribution has density $f(x)$. The hazard rate $h(x)$, $x \geq 0$, is defined by

$$h(x) = \frac{f(x)}{1 - F(x)} = \frac{f(x)}{\int_x^\infty f(y)\, \mathrm{d}y}.$$

A service requirement distribution belongs to class DHR [IHR] if $h(x)$ is decreasing [increasing] for all $x \geq 0$. It is known that DHR $\subseteq$ IMRL and IHR $\subseteq$ DMRL. Thus, we may expect more detailed results for DHR and IHR distributions.

When comparing the mean sojourn times of different disciplines for DHR or IHR service requirements, one should no longer concentrate on the level-x workload but a slightly modified variable called the *unfinished truncated work* [2, 3], which refers to the sum of the remaining truncated service requirements of those jobs in the system whose attained service is less than a given truncation threshold x. By definition, the mean unfinished truncated work $E[U_x^\pi]$ can be expressed as follows for any $\pi \in \Pi$ and $x \geq 0$, cf. (3):

$$E[U_x^\pi] = \int_{0^-}^{x^-} E[\min\{B, x\} - y \mid B > y]\, \mathrm{d}E[N_y^\pi].$$

If $h(x)$ is monotonic, the mean sojourn times $E[T^\pi]$ and $E[T^{\pi'}]$ in two systems with disciplines $\pi, \pi' \in \Pi$, respectively, may be

compared as follows [1, Equation (5)]:

$$E[T^\pi] - E[T^{\pi'}] = -\frac{1}{\lambda} \int_0^\infty (E[U_x^\pi] - E[U_x^{\pi'}]) \, \mathrm{d}h(x). \quad (4)$$

As a consequence, we have the following lemma.

Lemma 2 *Let $\pi, \pi' \in \Pi$ such that $E[U_x^\pi] \leq E[U_x^{\pi'}]$ for all $x \geq 0$. (i) If $F \in$ DHR, then $E[T^\pi] \leq E[T^{\pi'}]$. (ii) If $F \in$ IHR, then $E[T^\pi] \geq E[T^{\pi'}]$.*

By this approach, the following more detailed comparison results have been obtained, giving a partial order with respect to the mean sojourn time among the MLPS disciplines.

Theorem 2 *[4, Theorems 1, 2 and 3] For DHR [IHR] service requirements, the mean sojourn time is decreased [increased] whenever*
(i) an internal discipline is changed from PS to FB, or from FCFS to PS;
(ii) any level with FCFS internal discipline is split into two adjacent FCFS levels; or
(iii) level 1 with PS internal discipline is split into two adjacent PS levels.

Note that splitting a PS level that is not the lowest one is still an open problem. In [4] a conjecture is presented that would be sufficient to prove that for DHR [IHR] service requirements, the mean sojourn time is decreased [increased] if any PS level is split into two adjacent PS levels.

Lemma 2 can also be used to prove the optimality of FB with respect to the mean sojourn time for DHR service requirements [35, 2]. This is due to the fact that FB minimizes the unfinished truncated work process for any x even stochastically [2, Proposition 5]. Earlier proofs [74, 67] have used different approaches.

5.3 Mean slowdown ratio comparison

Define, for all $x \geq 0$,

$$g(x) = \frac{f(x)}{x(1 - F(x))}.$$

Note that for any DHR service requirement distribution, the function $g(x)$ is decreasing. However, for an IHR distribution, this function may be nonmonotonic.

By a simple modification of the argument that leads to Equation (4), we get the following relation. If $g(x)$ is monotonic, the mean slowdown ratios $E[R^\pi]$ and $E[R^{\pi'}]$ in two systems with disciplines $\pi, \pi' \in \Pi$, respectively, may be compared as follows:

$$E[R^\pi] - E[R^{\pi'}] = -\frac{1}{\lambda} \int_0^\infty (E[U_x^\pi] - E[U_x^{\pi'}]) \, \mathrm{d}g(x).$$

This results in the following lemma.

Lemma 3 *Let $\pi, \pi' \in \Pi$ such that $E[U_x^\pi] \leq E[U_x^{\pi'}]$ for all $x \geq 0$. (i) If function $g(x)$ is decreasing for all x, then $E[R^\pi] \leq E[R^{\pi'}]$. (ii) If function $g(x)$ is increasing for all x, then $E[R^\pi] \geq E[R^{\pi'}]$.*

By combining this lemma with the mean unfinished truncated work results found in [4], we get the following new theorem, giving another partial order among the MLPS disciplines, now with respect to the mean slowdown ratio.

Theorem 3 *If function $g(x)$ is decreasing [increasing] for all x, the mean slowdown ratio is decreased [increased] whenever any of the actions (i)–(iii) mentioned in Theorem 2 is performed.*

This approach was originally used in [35] to prove the optimality of FB with respect to the mean slowdown ratio for DHR service requirements.

5.4 Expected conditional sojourn time asymptotics

Let MLPS(PS) denote the set of disciplines that deploy PS as internal discipline in the highest level. In [10] the authors compared the asymptotic expected conditional sojourn time of a discipline $\pi \in$ MLPS(PS) and PS.

Theorem 4 *[10, Proposition 1] Let $\pi \in$ MLPS(PS). Then the response time of the queue has an asymptote with slope $1/(1 - \rho)$ and a positive finite bias K.*

The precise value of K depends on the service requirement distribution and the value of the last threshold a_n (see [10] for a closed-form expression for K). We note that Theorem 4 holds even when the service requirement distribution has an infinite second moment. As a direct consequence of Theorem 4, we conclude that the asymptotic slowdown of any discipline $\pi \in$ MLPS(PS) and PS are exactly the same, that is,

$$\lim_{x \to \infty} \frac{E[T^\pi(x)]}{x} = \lim_{x \to \infty} \frac{E[T^{PS}(x)]}{x} = \frac{1}{1 - \rho}.$$

These results for the asymptotic expected conditional sojourn time and slowdown show that the performance that very large jobs obtain is equivalent under both MLPS(PS) and PS disciplines.

5.5 Optimal choice of the thresholds

From a design point of view, a key issue for the successful implementation of the MLPS discipline lies in the choice of the thresholds. Numerical experiments reported in [4] showed that with DHR service requirements, an MLPS discipline with a very small number of thresholds (typically 2 or 3) but with an appropriate choice of thresholds, can attain a mean sojourn time very close to that achieved by FB. The problem of obtaining analytical expressions for the optimal choice of the thresholds is largely open. As a first step in this direction we refer to [12].

6. CONCLUSION

The Egalitarian Processor-Sharing (PS) discipline arises as a natural paradigm in a variety of practical situations, including resource-sharing in data networks. This area of application has been a driving force for several modifications of the basic PS discipline. We broadly distinguish three reasons for the existence of the modifications discussed in this survey paper: i) describing asymmetric characteristics of existing resource-sharing mechanisms like TCP (embodied in DPS), (ii) modeling deliberate service discrimination among users using WFQ or WRR (both DPS and GPS), and (iii) exploiting variability in the job sizes for performance benefits (MLPS). Given these distinct origins, the various policies have also been analyzed under differing scenarios and using a wide range of analysis tools. We have aimed at bundling the key analytical results obtained for DPS, GPS and MLPS.

A unifying factor amongst these three policies is the need to set appropriate weights or thresholds. However, optimality results to guide these parameter setting have remained rather scarce so far. We believe this issue provides a fertile area for future research.

7. REFERENCES

[1] Aalto, S. (2006). M/G/1/MLPS compared with M/G/1/PS within service time distribution class IMRL. *Math. Meth. Oper. Res.* **64**, 309–325.

[2] Aalto, S. Ayesta, U., Nyberg-Oksanen, E. (2004). Two-level processor-sharing scheduling disciplines: mean delay analysis. In: *Proc. ACM SIGMETRICS / Performance 2004 Conf.*, New York NY, USA, 97–105.

[3] Aalto, S., Ayesta, U., Nyberg-Oksanen, E. (2005). M/G/1/MLPS compared to M/G/1/PS. *Oper. Res. Lett.* **33**, 519–524.

[4] Aalto, S., Ayesta, U. (2006). Mean delay analysis of multi-level processor-sharing disciplines. In: *Proc. IEEE Infocom 2006*, Barcelona, Spain.

[5] Aalto, S., Ayesta, U. (2006). On the nonoptimality of the foreground-background discipline for IMRL service times. *J. Appl. Prob.* **43**, 523–534.

[6] Agrawal, R., Makowski, A.M., Nain, Ph. (1999). On a reduced load equivalence for fluid queues under subexponentiality. *Queueing Systems* **33**, 5–41.

[7] Altman, E., Avrachenkov, K.E., Ayesta, U. (2006). A survey on Discriminatory Processor Sharing. *Queueing Systems* **53**, 53–63. Special Issue on Queueing Models for Fair Resource Sharing.

[8] Altman, E., Jimenez, T., Kofman, D. (2004). DPS queues with stationary ergodic service times and the performance of TCP in overload. In: *Proc. IEEE Infocom 2004*, Hong Kong.

[9] Anantharam, V. (1999). Scheduling strategies and long-range dependence. *Queueing Systems* **33**, 73–89.

[10] Avrachenkov, K.E., Ayesta, U., Brown, P. (2005). Batch arrival processor sharing with application to size-based scheduling. *Queueing Systems* **50**, 459–480.

[11] Avrachenkov, K.E., Ayesta, U., Brown, P., Núñez-Queija, R. (2005). Discriminatory Processor Sharing revisited. In: *Proc. IEEE Infocom 2005*, Miami FL, USA.

[12] Avrachenkov, K., Brown, P., Osipova, N. (2006). Optimal choice of threshold in two-level processor sharing. In: *Euro-NGI Conf. Stochastic Performance Models for Resource Allocation in Communication Systems*, Amsterdam, The Netherlands, 61–64.

[13] Ben Fredj, S., Bonald, T., Proutière, A., Régnié, G., Roberts, J.W. (2001). Statistical bandwidth sharing: a study of congestion at the flow level. In: *Proc. Sigcomm 2001*, 111–122.

[14] Bertsimas, D., Paschalidis, I.Ch., Tsitsiklis, J.N. (1999). Large deviations analysis of the Generalized Processor Sharing policy. *Queueing Systems* **32**, 319–349.

[15] Biersack, E., Schroeder, B. (2007). Scheduling in practice. In this issue.

[16] Bonald, T., Roberts, J.W. (2000). Performance of bandwidth sharing mechanisms for service differentiation in the Internet. In: *Proc. ITC Specialist Seminar on IP Traffic Measurement, Modeling and Management*, Monterey CA, USA, 22.1–22.10.

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

[18] Borst, S.C., Boxma, O.J., Van Uitert, M.J.G. (2003). The asymptotic workload behavior of two coupled queues. *Queueing Systems* **43**, 81–102.

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

[20] Borst, S.C., Zwart, A.P. (2003). A reduced-peak equivalence for queues with a mixture of light-tailed and heavy-tailed input flows. *Adv. Appl. Prob.* **35**, 739–805.

[21] Borst, S.C., Núñez-Queija, R., Zwart A.P. (2006). Sojourn time asymptotics in processor-sharing queues. *Queueing Systems* **53**, 31–51. Special Issue on Queueing Models for Fair Resource Sharing.

[22] Borst, S.C., Van Ooteghem, D.T.M.B., Zwart, A.P. (2005). Tail asymptotics for Discriminatory Processor-Sharing queues with heavy-tailed service requirements. *Perf. Eval.* **61**, 281–298.

[23] Boxma, O.J., Deng, Q., Zwart, A.P. (2002). Waiting-time asymptotics for the M/G/2 queue with heterogeneous servers. *Queueing Systems* **40**, 5–31.

[24] Boxma, O.J., Kurkova, I.A. (2001). The M/G/1 queue with two service speeds. *Adv. Appl. Prob.* **33**, 520–540.

[25] Boxma, O.J., Zwart, A.P. (2007). Tails in scheduling. In this issue.

[26] Cohen, J.W. (1979). The multiple phase service network with Generalized Processor Sharing. *Acta Informatica* **12**, 245–284.

[27] Cohen, J.W., Boxma, O.J. (1983). *Boundary Value Problems in Queueing System Analysis* (North-Holland Publ. Cy., Amsterdam).

[28] Crovella, M., Bestavros, A. (1996). Self-similarity in World Wide Web traffic: evidence and possible causes. In: *Proc. ACM SIGMETRICS '96*, 160–169.

[29] Delas, S., Mazumdar, R.R., Rosenberg, C.P. (2002). Tail asymptotics for HOL priority queues handling a large number of independent stationary sources. *Queueing Systems* **40**, 183–204.

[30] Dupuis, P., Ramanan, K. (1998). A Skorokhod problem formulation and large deviation analysis of a processor sharing model. *Queueing Systems* **28**, 109–124.

[31] Egorova, R., Mandjes, M.R.H., Zwart, A.P. (2006). Sojourn time asymptotics in processor sharing queues with varying service rate. CWI Report PNA-R0611. Submitted for publication.

[32] Elwalid, A.I., Mitra, D. (1999). Design of Generalized Processor Sharing schedulers which statistically multiplex heterogeneous QoS classes. In: *Proc. Infocom '99*, New York NY, USA, 1220–1230.

[33] Fayolle, G., Iasnogorodski, R. (1979). Two coupled processors: the reduction to a Riemann-Hilbert problem. *Z. Wahr. verw. Geb.* **47**, 325–351.

[34] Fayolle, G., Mitrani, I., Iasnogorodski, R. (1980). Sharing a processor among many job classes. *J. ACM* **27**, 519–532.

[35] Feng, H., Misra, V. (2003). Mixed scheduling disciplines for network flows. In: *ACM SIGMETRICS Perf. Eval. Rev.* **31**, 36–39.

[36] Grishechkin, G. (1992). On a relationship between processor-sharing queues and Crump-Mode-Jagers branching processes. *Adv. Appl. Prob.* **24**, 653–698.

[37] Guillemin, F., Mazumdar, R.R., Dupuis, A., Boyer, J. (2003). Analysis of the fluid weighted fair queueing system. *J. Appl. Prob.* **40**, 180–199.

[38] Guillemin, F., Pinchon, D. (2004). Analysis of Generalized Processor Sharing systems with two classes of customers and

exponential services. *J. Appl. Prob.* **41**, 832–858.

[39] Harchol-Balter, M., Sigman, K., Wierman, A. (2002). Asymptotic convergence of scheduling policies with respect to slowdown. *Perf. Eval.* **49** *(Proc. Performance 2002)*, 241–256.

[40] Heyman, D.P., Lakshman, T.V., Neidhardt, A.L. (1997). A new method for analysing feedback-based protocols with applications to engineering Web traffic over the Internet. In: *Proc. ACM SIGMETRICS '97*, Seattle WA, USA, 24–38.

[41] Jelenković, P.R., Momčilović, P. (2002). Finite buffer queue with Generalized Processor Sharing and heavy-tailed input processes. *Comp. Netw.* **40**, 433–443.

[42] Jelenković, P.R., Momčilović, P. (2001). Network multiplexer with Generalized Processor Sharing and heavy-tailed on-off flows. In: *Teletraffic Engineering in the Internet Era, Proc. ITC-17*, 719–730.

[43] Van Kessel, G., Núñez-Queija, R., Borst, S.C. (2005). Differentiated bandwidth sharing with disparate flow sizes. In: *Proc. IEEE Infocom 2005*, Miami FL, USA.

[44] Van Kessel, G., Núñez-Queija, R., Borst, S.C. (2005). Asymptotic regimes and approximations for Discriminatory Processor Sharing. *ACM SIGMETRICS Perf. Eval. Rev.* **32** *(Proc. MAMA 2004 Workshop)*, 44–46.

[45] Kleinrock, L. (1976). *Queueing Systems, Vol. II: Computer Applications*. Wiley, New York.

[46] Kleinrock, L. (1967). Time-shared systems: a theoretical treatment. *J. ACM* **14**, 242–261.

[47] Knessl, C., Morrison, J.A. (2003). Heavy traffic analysis of two coupled processors. *Queueing Systems* **43**, 173–220.

[48] Konheim, A.G., Meilijson, I., Melkman, A. (1981). Processor sharing of two parallel lines. *J. Appl. Prob.* **18**, 952–956.

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

[50] Kotopoulos, C., Mazumdar, R.R. (2002). Buffer occupancy and delay asymptotics in multi-buffered systems with Generalized Processor Sharing handling a large number of independent traffic streams. Preprint, Purdue University.

[51] Kotopoulos, C., Mazumdar, R.R. (2002). Many sources asymptotics for a 2-buffer system with Generalized Processor Sharing. Preprint, Purdue University.

[52] Kumaran, K., Margrave, G.E., Mitra, D., Stanley, K.R. (2000). Novel techniques for the design and control of Generalized Processor Sharing schedulers for multiple QoS classes. In: *Proc. Infocom 2000*, Tel-Aviv, Israel, 932–941.

[53] Leland, W.E., Taqqu, M.S., Willinger, W., Wilson, D.V. (1994). On the self-similar nature of Ethernet traffic (extended version). IEEE/ACM *Trans. Netw.* **2**, 1–15.

[54] Lelarge, M. (2001). Asymptotic behavior of Generalized Processor Sharing queues under subexponential hypothesis. Report RR-4339, INRIA Rocquencourt.

[55] Lieshout, P.M.D., Mandjes, M., Borst, S.C. (2006). GPS scheduling: selection of optimal weights and comparison with strict priorities. In: *Proc. ACM SIGMETRICS / Performance 2006 Conf.*, Saint-Malo, France, 75–86.

[56] Mandjes, M., Van Uitert, M.J.G. (2005). Sample-path large deviations for Generalized Processor Sharing queues with Gaussian inputs. *Perf. Eval.* **61**, 225–256.

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

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

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

[60] Núñez-Queija, R. (2000). *Processor-Sharing Models for Integrated-Services Networks*. PhD Thesis, Eindhoven University of Technology.

[61] O'Donovan, T.M. (1974). Direct solutions of M/G/1 processor-sharing models. *Oper. Res.* **22**, 1232–1235.

[62] Osogami, T., Harchol-Balter, M., Scheller-Wolf, A. (2005). Analysis of cycle stealing with switching times and thresholds. *Perf. Eval.* **61**, 347–369.

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

[64] Parekh, A.K., Gallager, R.G. (1993). A Generalized Processor Sharing approach to flow control in integrated services networks: the single-node case. IEEE/ACM *Trans. Netw.* **1**, 344–357.

[65] Parekh, A.K., Gallager, R.G. (1994). A Generalized Processor Sharing approach to flow control in integrated services networks: the multiple-node case. IEEE/ACM *Trans. Netw.* **2**, 137–150.

[66] Rege, K.M., Sengupta, B. (1996). Queue length distribution for the discriminatory processor-sharing queue. *Oper. Res.* **44**, 653–657.

[67] Righter, R., Shanthikumar, J.G. (1989). Scheduling multiclass single server queueing systems to stochastically maximize the number of successful departures. *Prob. Eng. Inf. Sc.* **3**, 323–333.

[68] Righter, R., Shanthikumar, J.G., Yamazaki, G. (1990). On extremal service disciplines in single-stage queueing system. *J. Appl. Prob.* **27**, 409–416.

[69] Roberts, J.W. (2004). A survey on statistical bandwidth sharing. *Comp. Netw.* **45**, 319–332.

[70] Roberts, J.W., Massoulié, L. (1998). Bandwidth sharing and admission control for elastic traffic. In: *Proc. ITC Specialist Seminar*, Yokohama, Japan.

[71] Van Uitert, M.J.G. (2003). *Generalized Processor Sharing queues*. PhD Thesis, Eindhoven University of Technology.

[72] Van Uitert, M.J.G., Borst, S.C. (2002). A reduced-load equivalence for Generalised Processor Sharing networks with long-tailed input flows. *Queueing Systems* **41**, 123–163.

[73] Yaron, O., Sidi, M. (1994). Generalized Processor Sharing networks with exponentially bounded burstiness arrivals. In: *Proc. IEEE Infocom '94*, Toronto, Canada, 628–634.

[74] Yashkov, S.F. (1987). Processor-sharing queues: Some progress in analysis. *Queueing Systems* **2**, 1–17.

[75] Zhang, Z.-L. (1998). Large deviations and the Generalized Processor Sharing scheduling for a multiple-queue system. *Queueing Systems* **28**, 349–376.

[76] Zhang, Z.-L., Towsley, D., Kurose, J. (1997). Call admission control schemes under Generalized Processor Sharing scheduling. *Telecommun. Syst.* **7**, 125–152.

[77] Zwart, A.P., Boxma, O.J. (2000). Sojourn time asymptotics in the M/G/1 processor sharing queue. *Queueing Systems* **35**, 141–166.