Recent sojourn time results for Multilevel Processor-Sharing scheduling disciplines

Samuli Aalto

TKK Helsinki University of Technology,
PO Box 3000, FI-02015 TKK, Finland,
samuli.aalto@tkk.fi

Urtzi Ayesta

LAAS-CNRS,
7 Avenue du Colonel Roche, 31 077 Toulouse Cedex 4, France,
urtzi@laas.fr

Abstract

Multilevel Processor-Sharing (MLPS) disciplines refer to a family of age-based scheduling disciplines introduced already decades ago. A time-discretized version of an MLPS discipline is applied in the scheduler of the traditional UNIX operating system. In recent years, MLPS disciplines have been used to study the way that packet level scheduling mechanisms impact the performance perceived at the flow level in the Internet. Inspired by this latter application, many new sojourn time results have been discovered for these disciplines in the context of the M/G/1 queue. This paper aims to give a consistent overview of these new results. In addition, it points out some intriguing open problems for further research.

Key Words and Phrases: Queueing theory, Scheduling, Multilevel Processor-Sharing, Sojourn time, M/G/1 queue,

1 Introduction

Scheduling refers to the allocation of available resources among competing demands. First applications of scheduling came from the industrial manufacturing systems. Later on, various computer and communications systems have played a significant role. Scheduling theory concerns various optimal scheduling problems, see e.g. CONWAY et al. (1967); PINEDO (1995). In the beginning, the focus was on the mathematical formulations of static and deterministic scheduling

problems with a given set of n jobs and m machines. Since the 1970's, a school of computer scientists have considered the computational complexity of these problems. Static problems (with the number of jobs and machines fixed) have also been formulated in a stochastic setting with random release and service times. This setting has been attacked by another school of computer scientists that have grasped the worst-case behaviour of various on-line scheduling algorithms via so called competitive analysis. Queueing theorists have concerned the stochastic and dynamic setting, where new customers with random service times enter the system according to some random arrival process, see e.g. KLEINROCK (1976). Their work includes both the performance analysis of various scheduling disciplines and the solution of optimal scheduling problems. Recent years have witnessed a resurgence of this field, see e.g. HARCHOL-BALTER (2007). In particular for modern communications systems, the extremely high variability of the workload underlines the importance of age-based scheduling when optimizing the system performance.

Multilevel Processor-Sharing (MLPS) disciplines refer to a family of age-based scheduling disciplines introduced in KLEINROCK (1976). A time-discretized version of an MLPS discipline, called Multilevel Feedback (MLF) is applied in the scheduler of the traditional UNIX operating system, see e.g. STALLINGS (2005). In recent years, MLPS disciplines have been used to study the way that packet level scheduling mechanisms impact the performance perceived at the flow level in the Internet, see GUO and MATTA (2002); FENG and MISRA (2003); AVRACHENKOV et al. (2004); RAI et al. (2004). Inspired by this latter application, many new sojourn time results have been discovered for these disciplines in the context of the M/G/1 queue, see FENG and MISRA (2003); AVRACHENKOV et al. (2004, 2005, 2007); AALTO et al. (2004, 2005, 2007); AALTO (2006); AALTO and AYESTA (2006a,b, 2007a,b).

This paper aims to give a consistent overview of these new sojourn time results. It consists of the following parts. We start by introducing the MLPS disciplines in Section 2. The classical sojourn time results are presented in Section 3, which is followed by an asymptotic analysis in Section 4. Section 5 consists of optimality results revealing conditions under which some MLPS discipline minimizes the mean sojourn time or the mean slowdown ratio. In Section 6, we compare the performance of different MLPS disciplines. Finally, Section 7 not only summarizes the paper but also picks up some intriguing open problems for further research.

2 Multilevel Processor-Sharing disciplines

Consider a single server queueing system. Scheduling *discipline* (a.k.a. queueing or service discipline) refers to the decision rule that defines how the service capacity is shared among the customers in the system at any time t. Thus, for each customer i, the discipline, denoted by π, specifies the proportion $\sigma_i^\pi(t)$ of the service capacity that the customer is allocated at time t. The *attained service time* (or, briefly, *age*) of customer i is defined as

$$X_i^\pi(t) = \int_0^t \sigma_i^\pi(u)\,\mathrm{d}u,$$

while the *remaining service time* is given by

$$Y_i^\pi(t) = S_i - X_i^\pi(t),$$

where S_i refers to the (total) *service time* of customer i. The *sojourn time* T_i^π of customer i is clearly affected both by its service time S_i and the discipline π in use,

$$T_i^\pi = \inf\{t \geq 0 \mid X_i^\pi(t) = S_i\} - A_i,$$

where A_i refers to the arrival time of customer i. Well-known disciplines are, among others,

- First-Come-First-Served (FCFS), which assigns the whole service capacity to the customer $i \in \mathcal{N}^\pi(t)$ with the smallest arrival time A_i,
- Processor-Sharing (PS), which shares the service capacity evenly among all the customers $i \in \mathcal{N}^\pi(t)$,
- Foreground-Background (FB[1]), which shares the service capacity evenly among those customers $i \in \mathcal{N}^\pi(t)$ with the smallest attained service time $X_i^\pi(t)$, and
- Shortest-Remaining-Processing-Time (SRPT), which assigns the whole service capacity to the customer $i \in \mathcal{N}^\pi(t)$ with the smallest remaining service time $Y_i^\pi(t)$.

Here $\mathcal{N}^\pi(t)$ refers to the set of all customers in the system at time t,

$$\mathcal{N}^\pi(t) = \{i \mid A_i \leq t \text{ and } X_i^\pi(t) < S_i\}.$$

A discipline π is *work-conserving* if it does not idle when there are customers waiting, and *non-anticipating*[2] if the decision rule is independent of the remaining service times. Note that a non-anticipating decision rule may be *age-based* so that the decisions depend on the attained service times. Let Π denote the family of work-conserving and non-anticipating disciplines. FCFS, PS, and FB clearly belong to Π, whereas SRPT does not. In addition, FB is purely age-based.

Multilevel Processor-Sharing (MLPS) disciplines constitute a subfamily of Π, see KLEINROCK (1976). An MLPS discipline π is age-based, and it is characterized by a finite set of level thresholds $a_1 < \cdots < a_N$ that define $N + 1$ different priority levels, $N \geq 0$. A customer belongs to level n if its attained service time 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. Thus, those customers are served first whose attained service time is less than a_1. Within each priority level n, an internal discipline D_n is applied. According to Kleinrock's definition, there are three possible internal disciplines, namely FCFS, PS, and FB. The internal discipline may be different at various levels. Note that an MLPS discipline that applies FB at all levels is, in fact, the same as the ordinary FB discipline.

[1]FB has many aliases like Foreground-Background Processor-Sharing (FBPS), Least-Attained-Service (LAS), Least-Attained-Service-Time (LAST), Shortest Elapsed Time (SET), or Shortest-Elapsed-Processing-Time (SEPT), depending on the context, see NUYENS and WIERMAN (2008).

[2]Non-anticipating disciplines are sometimes called non-anticipative, non-clairvoyant, or blind, again depending on the context.

We refer to an MLPS discipline with level thresholds a_n and internal disciplines D_n by using the notation $D_1 + \ldots + D_{N+1}(a_1, \ldots, a_N)$. For example, PS+PS$(a)$ refers to the two-level discipline with threshold a that applies PS as the internal discipline at both levels.

In this paper we consider the MLPS disciplines in the context of the M/G/1 queue with Poisson arrivals and independent and identically distributed service times with a general distribution. Throughout the paper, λ refers to the arrival rate of new customers, and S stands for a generic service time with a cumulative distribution function denoted by $F(x) = P\{S \leq x\}$, $x \geq 0$. In addition, let $\overline{F}(x) = 1 - F(x)$ denote the corresponding tail distribution function. We assume that the mean service time is finite, $E[S] < \infty$, and the traffic load $\rho = \lambda E[S]$ satisfies the stability condition $\rho < 1$.

3 Classical results for the conditional mean sojourn time

For any discipline $\pi \in \Pi$, let $E[T^\pi|S = x]$ denote the conditional mean sojourn time of a customer with service time x. In addition, let $E[R^\pi|S = x]$ denote the corresponding *slowdown ratio* defined by

$$E[R^\pi|S = x] = \frac{E[T^\pi|S = x]}{x}.$$

Assume now that an MLPS discipline with $N + 1$ levels is applied, and consider a customer whose service time x satisfies $a_{n-1} < x \leq a_n$ for some $n \leq N + 1$. So the customer passes levels $1, \ldots, n - 1$ before leaving the system at level n. Because of strict priority between the levels, the customer enters the final level n at the moment when the attained service time of all the customers in the system is at least a_{n-1}. Note that this time epoch is independent of all the internal disciplines. After entering the final level, the time until the customer leaves the system depends only on the internal discipline D_n of that level. The upper levels have no effect on this departure time due to the strict priority between the levels. On the other hand, if a new customer enters the system, the service of the customers at level n is delayed until the attained service time of all the customers in the system is again at least a_{n-1}, which is independent of all the internal disciplines. Thus, for any MLPS discipline π, the conditional mean sojourn time takes the following form,

$$E[T^\pi|S = x] = t_1(a_{n-1}) + t_2(D_n, a_{n-1}, x, a_n), \quad \text{for all } a_{n-1} < x \leq a_n.$$

The three possible internal disciplines are discussed separately below. Before that we, however, give some elementary results for systems with truncated service times. All the results of this section can be found in KLEINROCK (1976).

Truncated service times Let $x \geq 0$, and replace, for a while, the service times S by their truncated versions $S \wedge x = \min\{S, x\}$. It is easy to see that

$$E[S \wedge x] = \int_0^x \overline{F}(t)\,\mathrm{d}t, \quad E[(S \wedge x)^2] = 2\int_0^x t\,\overline{F}(t)\,\mathrm{d}t.$$

Furthermore, let the truncated load be denoted by

$$\rho_x = \lambda E[(S \wedge x)].$$

Clearly, $\rho_x \leq \rho < 1$ for all x. The mean workload (i.e., unfinished work) for a work-conserving M/G/1 queue with truncated service times is, by the Pollaczek-Khinchin formula,

$$w_x = \frac{\lambda E[(S \wedge x)^2]}{2(1-\rho_x)}.$$

Of course, when $x \to \infty$, we get the ordinary Pollaczek-Khinchin formula,

$$w_\infty = \frac{\lambda E[S^2]}{2(1-\rho)}.$$

Internal discipline FCFS Let us return to the original service times S. First we consider the conditional mean sojourn time for MLPS disciplines within an FCFS level. If there is just one level, then, for all $x \geq 0$,

$$E[T^{\mathrm{FCFS}}|S=x] = w_\infty + x.$$

If π is an MLPS discipline with FCFS applied at level n, then, for all $a_{n-1} < x \leq a_n$,

$$E[T^\pi|S=x] = \frac{w_{a_n} + x}{1-\rho_{a_{n-1}}}.$$

Internal discipline FB Next we consider the conditional mean sojourn time for MLPS disciplines within an FB level. If there is just one level, then, for all $x \geq 0$,

$$E[T^{\mathrm{FB}}|S=x] = \frac{w_x + x}{1-\rho_x}.$$

If π is an MLPS discipline with FB applied at level n, then, for all $a_{n-1} < x \leq a_n$,

$$E[T^\pi|S=x] = E[T^{\mathrm{FB}}|S=x] = \frac{w_x + x}{1-\rho_x}.$$

Internal discipline PS Finally we consider the conditional mean sojourn time for MLPS disciplines within a PS level. If there is just one level, then, for all $x \geq 0$,

$$E[T^{\mathrm{PS}}|S=x] = \frac{x}{1-\rho}.$$

If π is an MLPS discipline with PS applied at level n, then, for all $a_{n-1} < x \leq a_n$,

$$E[T^\pi|S=x] = E[T^{\mathrm{FB}}|S=a_{n-1}] + \frac{\alpha(x-a_{n-1})}{1-\rho_{a_{n-1}}}.$$

Here $\alpha(t)$ refers to the conditional mean sojourn time in a certain M^{X}/G/1 queue with batch arrivals and PS discipline, whose derivative $\alpha'(t)$ satisfies the following integral equation, for all $0 < t < a_n - a_{n-1}$,

$$\begin{aligned}\alpha'(t) &= 1 + 2\lambda E[T^{\mathrm{FB}}|S = a_{n-1}]\overline{F}(a_{n-1}+t) \\ &\quad + \frac{\lambda}{1-\rho_{a_{n-1}}}\int_0^{a_n-a_{n-1}-t} \alpha'(u)\overline{F}(a_{n-1}+t+u)\,\mathrm{d}u \\ &\quad + \frac{\lambda}{1-\rho_{a_{n-1}}}\int_0^{t} \alpha'(u)\overline{F}(a_{n-1}+t-u)\,\mathrm{d}u. \end{aligned} \tag{1}$$

More explicit expressions have been derived only for a slight generalization of the exponential distribution (see KLEINROCK (1976)) and for the hyperexponential distribution (see OSIPOVA (2007)).

4 Asymptotic analysis of the conditional mean sojourn time

In this section we consider the asymptotic analysis of the conditional mean sojourn time $E[T^\pi|S = x]$ and the slowdown ratio $E[R^\pi|S = x]$ as the service time x tends to infinity. Clearly, only the internal discipline D_{N+1} at the highest level matters in this case. Below we concern each discipline separately.

Internal discipline FCFS First we consider the MLPS disciplines π with $D_{N+1} = \mathrm{FCFS}$. If the service time distribution has a finite second moment $E[S^2] < \infty$ (so that $w_\infty < \infty$), the conditional mean sojourn time has clearly an asymptote with slope $1/(1-\rho_{a_N})$ and a constant positive bias, for all $x > a_N$,

$$E[T^\pi|S = x] - \frac{x}{1-\rho_{a_N}} = \frac{w_\infty}{1-\rho_{a_N}}.$$

In addition, the asymptotic slowdown ratio is in this case as follows:

$$\lim_{x\to\infty} E[R^\pi|S = x] = \frac{1}{1-\rho_{a_N}} \le \frac{1}{1-\rho}.$$

On the other hand, $E[T^\pi|S = x] = \infty$ for all the service time distributions with an infinite second moment $E[S^2] = \infty$.

Internal discipline FB Consider now the MLPS disciplines π with $D_{N+1} = \mathrm{FB}$. HARCHOL-BALTER et al. (2002) show that the asymptotic slowdown ratio is as follows,

$$\lim_{x\to\infty} E[R^\pi|S = x] = \frac{1}{1-\rho}.$$

The limit exists even for the service time distributions with an infinite second moment but there is no asymptote for the conditional mean sojourn time $E[T^\pi|S = x]$ as AVRACHENKOV et al. (2004) demonstrates.

Internal discipline PS Finally we consider the MLPS disciplines π with $D_{N+1} = \mathrm{PS}$. An asymptotic analysis of the conditional mean sojourn time is presented in AVRACHENKOV et al. (2005). Below we summarize their results. A natural baseline discipline is now PS for which

$$E[T^{\mathrm{PS}}|S=x] = \frac{x}{1-\rho} \quad \text{and} \quad E[R^{\mathrm{PS}}|S=x] = \frac{1}{1-\rho}.$$

Thus, even for the service time distributions with an infinite second moment, the conditional mean sojourn time of PS has an asymptote with slope $1/(1-\rho)$.

Theorem 4.1 *Let $\pi \in \mathrm{MLPS}$ such that $D_{N+1} = \mathrm{PS}$. The conditional mean sojourn time has an asymptote with slope $1/(1-\rho)$ and a positive (finite) bias,*

$$\lim_{x\to\infty}\left(E[T^{\pi}|S=x] - \frac{x}{1-\rho}\right) = \frac{w_{a_N}(1-\rho_{a_N}) + a_N(\rho-\rho_{a_N})}{(1-\rho)^2}.$$

Corollary 4.2 *Let $\pi \in \mathrm{MLPS}$ such that $D_{N+1} = \mathrm{PS}$. The asymptotic slowdown ratio is the same as for the PS discipline,*

$$\lim_{x\to\infty} E[R^{\pi}|S=x] = \frac{1}{1-\rho}.$$

5 Optimality of MLPS disciplines

As originally shown by SCHRAGE (1968), the mean sojourn time in an M/G/1 queue is minimized by the SRPT discipline. However, the information about the remaining service times is not always available, but one is obliged to only consider the non-anticipating disciplines. Thus, we are looking for an *optimal* discipline $\pi^* \in \Pi$ such that

$$E[T^{\pi^*}] = \min_{\pi\in\Pi} E[T^{\pi}].$$

There are two well-known optimality results among these disciplines. If the service time distribution belongs to the class NBUE, then FCFS is optimal (minimizing the mean sojourn time) as shown by RIGHTER et al. (1990), while FB is optimal for DHR service times, see YASHKOV (1987); RIGHTER and SHANTHIKUMAR (1989). Recently, an additional optimality result was found, giving a condition under which a two-level MLPS discipline FCFS+FB is optimal, see AALTO and AYESTA (2007a,b). In this section, we describe how these results can be derived based on the so-called Gittins index approach developed in GITTINS (1989).

As before, for any $\pi \in \Pi$, let $E[T^{\pi}|S=x]$ denote the conditional mean sojourn time of a customer with service time x, and $E[R^{\pi}|S=x]$ the corresponding slowdown ratio. The mean sojourn time and the mean slowdown ratio have respectively the following expressions:

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

5.1 Service time distribution classes

From here on, we assume that the service time can take any positive value. Thus, $\overline{F}(x) > 0$ for all x. In addition, we assume that the service time distribution is continuous with a density function denoted by $f(x)$. The corresponding hazard (or failure) rate $h(x)$ is as follows:

$$h(x) = \frac{f(x)}{\overline{F}(x)} = \frac{f(x)}{\int_0^\infty f(x+t)\,\mathrm{d}t}.$$

Furthermore, we define, for all x,

$$H(x) = \frac{\int_0^\infty f(x+t)\,\mathrm{d}t}{\int_0^\infty \overline{F}(x+t)\,\mathrm{d}t} = \frac{\overline{F}(x)}{\int_0^\infty \overline{F}(x+t)\,\mathrm{d}t}.$$

Note that $1/H(x)$ is in fact the conditional mean of the remaining service time,

$$E[S - x \mid S > x] = \frac{\int_0^\infty \overline{F}(x+t)\,\mathrm{d}t}{\overline{F}(x)} = \frac{1}{H(x)}.$$

All the following distribution classes (except the last one) are well-known, see e.g. SHAKED and SHANTHIKUMAR (1994). A service time distribution belongs to the class

- *Decreasing Hazard Rate* (DHR) if $h(x)$ is decreasing[3] for all x;
- *Increasing Mean Residual Lifetime* (IMRL) if $1/H(x)$ is increasing for all x;
- *New Worse than Used in Expectation* (NWUE) if $1/H(0) \leq 1/H(x)$ for all x.

Classes *Increasing Hazard Rate* (IHR), *Decreasing Mean Residual Lifetime* (DMRL), and *New Better than Used in Expectation* (NBUE) are defined correspondingly.[4] It is known that

$$\mathrm{DHR} \subset \mathrm{IMRL} \subset \mathrm{NWUE} \quad \text{and} \quad \mathrm{IHR} \subset \mathrm{DMRL} \subset \mathrm{NBUE}.$$

Moreover, the service time distributions in the first three classes have a coefficient of variation greater than 1. Thus, they are more variable than the exponential distribution. Correspondingly, the distributions in the last three classes are less variable than the exponential distribution.

Finally, let $k > 0$. A service time distribution belongs to the class NBUE+DHR(k) if

(i) $1/H(0) \geq 1/H(x)$ for all $x < k$, and

(ii) $h(x)$ is decreasing for all $x > k$.

[3] Throughout the paper we use the terms "decreasing" and "increasing" in their weak form so that the corresponding functions need not be *strictly* monotonic.

[4] Classes DHR and IHR are also called *Decreasing Failure Rate* (DFR) and *Increasing Failure Rate* (IFR), respectively.

This class was introduced in AALTO and AYESTA (2007b). An example is the Pareto distribution with shape parameter $\alpha > 1$ and scale parameter $k > 0$,

$$F(x) = \begin{cases} 0, & 0 \le x < k, \\ 1 - \left(\dfrac{k}{x}\right)^{\alpha}, & x \ge k. \end{cases}$$

Note, however, that another version of the Pareto distribution with shape parameter $\beta > 1$ and scale parameter $b > 0$ that is defined by

$$F(x) = 1 - \left(\frac{1}{1+bx}\right)^{\beta}, \quad x \ge 0,$$

belongs to the class DHR.

5.2 Gittins index

The *Gittins index* of a customer with age x is defined by

$$G(x) = \sup_{\Delta \ge 0} J(x, \Delta),$$

where

$$J(x,\Delta) = \frac{\int_0^{\Delta} f(x+t)\,\mathrm{d}t}{\int_0^{\Delta} \overline{F}(x+t)\,\mathrm{d}t} = \frac{\overline{F}(x) - \overline{F}(x+\Delta)}{\int_0^{\Delta} \overline{F}(x+t)\,\mathrm{d}t}.$$

Note that

$$J(x,0) = \frac{f(x)}{\overline{F}(x)} = h(x) \quad \text{and} \quad J(x,\infty) = \frac{\overline{F}(x)}{\int_0^{\infty} \overline{F}(x+t)\,\mathrm{d}t} = H(x).$$

Furthermore, let

$$\Delta^*(x) = \sup\{\Delta \ge 0 \mid J(x,\Delta) = G(x)\}.$$

By definition,

$$G(x) = J(x, \Delta^*(x)).$$

Note further that $J(x,\Delta)$ is continuous with respect to both arguments. In addition, the one-sided partial derivatives with respect to Δ are defined for any pair (x,Δ),

$$\frac{\partial}{\partial \Delta} J(x,\Delta) = \frac{f(x+\Delta)\int_0^{\Delta} \overline{F}(x+t)\,\mathrm{d}t - \overline{F}(x+\Delta)\int_0^{\Delta} f(x+t)\,\mathrm{d}t}{(\int_0^{\Delta} \overline{F}(x+t)\,\mathrm{d}t)^2}. \tag{2}$$

Lemma 5.1 *If the service time distribution belongs to DHR, then $J(x,\Delta)$ is decreasing with respect to Δ for any fixed x.*

Proof. Let $x \geq 0$. Assume that the service time distribution belongs to DHR. Let $\Delta \geq 0$. Now $h(x+t) \geq h(x+\Delta)$ for all $0 \leq t \leq \Delta$, which is equivalent with

$$\frac{f(x+t)}{f(x+\Delta)} \geq \frac{\overline{F}(x+t)}{\overline{F}(x+\Delta)}. \tag{3}$$

By (2), we have

$$\begin{aligned} \frac{\partial}{\partial \Delta} J(x,\Delta) \leq 0 \quad &\Longleftrightarrow \quad \frac{f(x+\Delta)}{\int_0^\Delta f(x+t)\,\mathrm{d}t} \leq \frac{\overline{F}(x+\Delta)}{\int_0^\Delta \overline{F}(x+t)\,\mathrm{d}t} \\ &\Longleftrightarrow \quad \frac{1}{\int_0^\Delta \frac{f(x+t)}{f(x+\Delta)}\,\mathrm{d}t} \leq \frac{1}{\int_0^\Delta \frac{\overline{F}(x+t)}{\overline{F}(x+\Delta)}\,\mathrm{d}t}. \end{aligned}$$

The claim follows from this by (3). □

Proposition 5.2 *If the service time distribution belongs to DHR, then $G(x)$ is decreasing for all x.*

Proof. Let $x \geq 0$. Assume that the service time distribution belongs to DHR. Then $G(x) = J(x,0) = h(x)$ by Lemma 5.1, and, thus, $G(x)$ is decreasing. □

Similarly, by using the counterpart of Lemma 5.1, we deduce that if the service time distribution belongs to IHR, then $G(x) = J(x,\infty) = H(x)$ and, thus, $G(x)$ is increasing for all x. Below we present a relevant result for the more general class NBUE.

Lemma 5.3 *The service time distribution belongs to NBUE if and only if $J(0,\Delta) \leq J(0,\infty)$ for any Δ.*

Proof. By definition the service time distribution belongs to NBUE if and only if $J(0,\infty) = H(0) \leq H(x) = J(x,\infty)$ for any x. Let $x \geq 0$. Now

$$\begin{aligned} J(0,x) \leq J(0,\infty) \quad &\Longleftrightarrow \quad \frac{1-\overline{F}(x)}{\int_0^x \overline{F}(t)\,\mathrm{d}t} \leq \frac{1}{\int_0^\infty \overline{F}(t)\,\mathrm{d}t} \\ &\Longleftrightarrow \quad \int_x^\infty \overline{F}(t)\,\mathrm{d}t \leq \overline{F}(x) \int_0^\infty \overline{F}(t)\,\mathrm{d}t \\ &\Longleftrightarrow \quad \frac{1}{\int_0^\infty \overline{F}(t)\,\mathrm{d}t} \leq \frac{\overline{F}(x)}{\int_x^\infty \overline{F}(t)\,\mathrm{d}t}. \end{aligned}$$

The claim follows from this, since the last inequality is equivalent with the inequality $J(0,\infty) \leq J(x,\infty)$. □

Proposition 5.4 *If the service time distribution belongs to NBUE, then $G(x) \geq G(0)$ for all x.*

Proof. Let $x \geq 0$. Assume that the service time distribution belongs to NBUE. Then

$$J(x,\infty) = H(x) \geq H(0) = J(0,\infty)$$

by definition, and $G(0) = J(0,\infty)$ by Lemma 5.3. Thus,

$$G(x) \geq J(x,\infty) \geq J(0,\infty) = G(0),$$

which completes the proof. □

The following Proposition is proved in AALTO and AYESTA (2007b).

Proposition 5.5 *If the service time distribution belongs to NBUE+DHR(k), where $k > 0$, then*

(i) $\Delta^*(0) \geq k$,

(ii) $G(x) \geq G(0)$ *for all* $x < \Delta^*(0)$,

(iii) $G(x)$ *is decreasing for all* $x > k$.

If additionally $\Delta^*(0) < \infty$, *then*

(iv) $G(\Delta^*(0)) \leq G(0)$.

5.3 Optimality results

The *Gittins index discipline* π^* assigns the whole service capacity to the customer $i \in \mathcal{N}^{\pi^*}(t)$ with the highest Gittins index,

$$G(X_i^{\pi^*}(t)) = \max_{j \in \mathcal{N}^{\pi^*}(t)} G(X_j^{\pi^*}(t)).$$

Recall that $X_i^{\pi^*}(t)$ denotes the attained service time of customer i. If there are multiple customers with the same highest index, any one of them can be chosen. The Gittins index discipline is clearly work-conserving and non-anticipating, belonging thus to Π.

It is known that the Gittins index discipline π^* is optimal with respect to the mean sojourn time (among the work-conserving and non-anticipating disciplines), see Theorem 3.28 in GITTINS (1989) or Theorem 4.7 in YASHKOV (1992). Together with the three Propositions presented in the previous subsection, this justifies the following optimality results.

Theorem 5.6 *If the service time distribution belongs to*

(i) DHR, then $E[T^{\mathrm{FB}}] = \inf_{\pi \in \Pi} E[T^\pi]$;

(ii) NBUE, then $E[T^{\mathrm{FCFS}}] = \inf_{\pi \in \Pi} E[T^\pi]$;[5]

[5] In fact, not only FCFS but any non-preemptive work-conserving discipline is optimal in this case.

(iii) NBUE+DHR(k), then $E[T^{\mathrm{FCFS+FB}(\Delta^*(0))}] = \inf_{\pi\in\Pi} E[T^\pi]$.

As mentioned above, the first two of these results are already about 20 years old, while the third one was detected recently in AALTO and AYESTA (2007b). A less general version appeared before that in AALTO and AYESTA (2007a).

Instead of the mean sojourn time, FENG and MISRA (2003) consider the minimization of the mean slowdown ratio obtaining the following result. However, the Gittins index cannot be applied for the proof but another approach is needed, see Section 6.3.

Theorem 5.7 *If the service time distribution belongs to DHR, then* $E[R^{\mathrm{FB}}] = \inf_{\pi\in\Pi} E[R^\pi]$.

6 Comparison among MLPS disciplines

In this section we review recent results related to the mean sojourn time and slowdown ratio comparison among the MLPS disciplines. The section is mainly based on the following papers: FENG and MISRA (2003); AALTO et al. (2004); AALTO (2006); AALTO and AYESTA (2006a); AALTO et al. (2007).

6.1 Mean sojourn time comparison for IMRL and DMRL service times

When comparing the mean sojourn time of different disciplines for IMRL or DMRL service times in an M/G/1 queue, a key variable is the so-called *level-x (or truncated) workload* $V_x^\pi(t)$, which refers to the sum of the remaining service times of those customers whose attained service is less than a given threshold x,

$$V_x^\pi(t) = \sum_{i\in\mathcal{N}_x^\pi(t)} (S_i - X_i^\pi(t)),$$

where

$$\mathcal{N}_x^\pi(t) = \{i \in \mathcal{N}^\pi(t) \mid A_i \le t \text{ and } X_i^\pi(t) < x\},$$

see AALTO (2006); AALTO and AYESTA (2006b).

If the conditional mean remaining service time $1/H(x)$ is monotonic (and, thus also $H(x)$), 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:

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

This yields the following lemma.

Lemma 6.1 *Let $\pi, \pi' \in \Pi$ such that $E[V_x^\pi] \le E[V_x^{\pi'}]$ for all $x \ge 0$. If the service time distribution belongs to*

(i) IMRL, then $E[T^\pi] \le E[T^{\pi'}]$;

(ii) DMRL, then $E[T^{\pi}] \geq E[T^{\pi'}]$.

By this approach, the following result has been obtained in AALTO (2006), which compares certain MLPS disciplines with the baseline discipline PS.

Theorem 6.2 *Let* $\pi \in \mathrm{MLPS}$ *such that the internal disciplines belong to* $\{\mathrm{FB}, \mathrm{PS}\}$. *If the service time distribution belongs to*

(i) IMRL, then $E[T^{\pi}] \leq E[T^{\mathrm{PS}}]$*;*

(ii) DMRL, then $E[T^{\pi}] \geq E[T^{\mathrm{PS}}]$.

This approach has also been used in AALTO and AYESTA (2006b) to demonstrate that FB does *not* minimize the mean sojourn time within the class IMRL, contrary to what was earlier believed.

6.2 Mean sojourn time comparison for DHR and IHR service times

When comparing the mean sojourn times of different disciplines for DHR or IHR service times, one should no longer concentrate on the level-x workload but a slightly modified variable called the *unfinished truncated work* $U_x^{\pi}(t)$, which refers to the sum of the remaining truncated service times of those jobs in the system whose attained service is less than a given truncation threshold x,

$$U_x^{\pi}(t) = \sum_{i \in \mathcal{N}_x^{\pi}(t)} ((S_i \wedge x) - X_i^{\pi}(t)),$$

where $S_i \wedge x = \min\{S_i, x\}$, see AALTO et al. (2004, 2005).

If the hazard rate $h(x)$ 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:

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

As a consequence, we have the following lemma.

Lemma 6.3 *Let* $\pi, \pi' \in \Pi$ *such that* $E[U_x^{\pi}] \leq E[U_x^{\pi'}]$ *for all* $x \geq 0$. *If the service time distribution belongs to*

(i) DHR, then $E[T^{\pi}] \leq E[T^{\pi'}]$*;*

(ii) IHR, then $E[T^{\pi}] \geq E[T^{\pi'}]$.

By this approach, the following more detailed comparison results have been obtained in AALTO and AYESTA (2006a), giving a partial order with respect to the mean sojourn time among the MLPS disciplines.

Theorem 6.4 *Let $\pi, \pi' \in \mathrm{MLPS}$ such that π is derived from π' by one of the following operations:*

- *an internal discipline is changed from PS to FB, or from FCFS to PS;*
- *any level with FCFS internal discipline is split into two adjacent FCFS levels; or*
- *level 1 with PS internal discipline is split into two adjacent PS levels.*

Now if the service time distribution belongs to

(i) DHR, then $E[T^{\pi}] \leq E[T^{\pi'}]$;

(ii) IHR, then $E[T^{\pi}] \geq E[T^{\pi'}]$.

Note that splitting a PS level that is not the lowest one is still an open problem. In AALTO and AYESTA (2006a), a conjecture is presented that would be sufficient to prove that for DHR [IHR] service times, the mean sojourn time is decreased [increased] if any PS level is split into two adjacent PS levels.

Lemma 6.3 can also be used to prove the optimality of FB with respect to the mean sojourn time for DHR service times, see FENG and MISRA (2003); AALTO et al. (2004). This is due to the fact that FB minimizes the unfinished truncated work process for any x even stochastically. Earlier proofs used different approaches.

6.3 Mean slowdown ratio comparison

Finally, we compare the MLPS disciplines with respect to the mean slowdown ratio $E[R^{\pi}]$. Define, for all $x \geq 0$,

$$g(x) = \frac{h(x)}{x},$$

where $h(x)$ refers to the hazard rate as before. Note that for any DHR service time distribution, the function $g(x)$ is decreasing. However, for an IHR distribution, this function may be nonmonotonic.

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 6.5 *Let $\pi, \pi' \in \Pi$ such that $E[U_x^{\pi}] \leq E[U_x^{\pi'}]$ for all $x \geq 0$. If $g(x)$ is*

(i) decreasing for all x, then $E[R^{\pi}] \leq E[R^{\pi'}]$;

(ii) increasing for all x, then $E[R^{\pi}] \geq E[R^{\pi'}]$.

By combining this lemma with the results found in AALTO and AYESTA (2006a), the following theorem was discovered in AALTO et al. (2007).

Theorem 6.6 *Let $\pi, \pi' \in \mathrm{MLPS}$ such that π is derived from π' by one of the operations mentioned in Theorem 6.4. If $g(x)$ is*

(i) decreasing for all x, then $E[R^{\pi}] \leq E[R^{\pi'}]$;

(ii) increasing for all x, then $E[R^{\pi}] \geq E[R^{\pi'}]$.

Corollary 6.7 *Let $\pi, \pi' \in \mathrm{MLPS}$ such that π is derived from π' by one of the operations mentioned in Theorem 6.4. If the service time distribution belongs to* DHR*, then $E[R^{\pi}] \leq E[R^{\pi'}]$.*

This approach was originally used in FENG and MISRA (2003) to prove the optimality of FB with respect to the mean slowdown ratio for DHR service times.

6.4 Optimization of the level thresholds

From a designer point of view, a key issue for the successful implementation of an MLPS discipline lies on the choice of the level thresholds. Numerical experiments reported in AALTO and AYESTA (2006a) propose that with DHR service times, an MLPS discipline with only 2 or 3 levels but with an appropriate choice of level thresholds, can provide a mean sojourn time very close to the optimal value achieved by FB.

The problem of obtaining analytical expressions for the optimal choice of the thresholds is largely open. AVRACHENKOV et al. (2007) consider the two-level PS+PS disciplines assuming a hyperexponential distribution,

$$F(x) = 1 - pe^{-\mu_1 x} - (1-p)e^{-\mu_2 x},$$

which belongs to the class DHR. Theorem 6.4 says that, for any a,

$$E[T^{\mathrm{FB}}] \leq E[T^{\mathrm{PS+PS}}(a)] \leq E[T^{\mathrm{PS}}]$$

in this case. AVRACHENKOV et al. (2007) find that, if $\mu_1 \gg \mu_2$, the optimal threshold a^* is well approximated by

$$\tilde{a} = \frac{1}{\mu_1 - \mu_2} \log\left(\frac{\mu_1 - \lambda}{\mu_2(1-\rho)}\right).$$

Numerical results reported in AALTO and AYESTA (2006a) indicate that the above approximation may work well even when μ_1 and μ_2 are of the same order.

7 Summary and open problems

The paper reflects the progress that has been made in recent years in the study of the sojourn time for the MLPS disciplines. This family of disciplines has attracted a lot of attention due to its broad definition, which covers a wide variety of non-anticipating policies, including well known disciplines like PS, FCFS, or FB.

We have seen that under the rather realistic assumption of DHR service time distributions, the mean sojourn time of FB is optimal within the set of non-anticipating policies. Unfortunately, the combination of the difficulties in implementing FB, together with the fear that large jobs might suffer from starvation, has led to a situation where age-based scheduling disciplines are rarely implemented in real systems.

An important conclusion we can draw on the performance of MLPS is that the mean sojourn time of an MLPS discipline with just 2 or 3 priority levels matches very closely the mean sojourn time obtained with FB, while at the same time reducing the degradation received by the largest jobs (see Section 4). Thus, MLPS disciplines might be more suitable in practical implementations than the theoretically optimal policies like FB.

As an important research avenue for the future we believe that in the coming years the performance of scheduling disciplines in networks of queues should deserve more attention. Indeed, the vast majority of researchers consider the simplest case of a single bottleneck scenario, and only partial results exist for the more realistic case, where scheduling disciplines like MLPS might be implemented in various parts of a network.

A challenge also remains for the single bottleneck scenario M/G/1, viz. the completion of Theorem 6.4 to cover the case where a PS level, that is not the lowest one, is split.

References

S. AALTO. M/G/1/MLPS compared with M/G/1/PS within service time distribution class IMRL. *Mathematical Methods of Operations Research*, 64:309–325, 2006.

S. AALTO and U. AYESTA. Mean delay analysis of multi level processor sharing disciplines. In *Proc. of IEEE Infocom 2006*. Barcelona, Spain, 2006a.

S. AALTO and U. AYESTA. On the nonoptimality of the foreground-background discipline for IMRL service times. *Journal of Applied Probability*, 43:523–534, 2006b.

S. AALTO and U. AYESTA. Mean delay optimization for the M/G/1 queue with Pareto type service times. In *Proc. of ACM Sigmetrics 2007*, pages 383–384. San Diego, CA, 2007a.

S. AALTO and U. AYESTA. Optimal scheduling of service requirements with a DHR tail in the M/G/1 queue. Technical Report hal-00166642, HAL, France, 2007b.

S. AALTO, U. AYESTA, S. BORST, V. MISRA, and R. NÚÑEZ-QUEIJA. Beyond processor sharing. *ACM Sigmetrics Performance Evaluation Review*, 34(4):36–43, 2007.

S. AALTO, U. AYESTA, and E. NYBERG-OKSANEN. Two-level processor-sharing scheduling disciplines: mean delay analysis. In *Proc. of ACM Sigmetrics/PERFORMANCE 2004*, pages 97–105. New York, NY, 2004.

S. AALTO, U. AYESTA, and E. NYBERG-OKSANEN. M/G/1/MLPS compared to M/G/1/PS. *Operations Research Letters*, 33:519–524, 2005.

K. AVRACHENKOV, U. AYESTA, and P. BROWN. Batch arrival processor sharing with application to multilevel processor sharing scheduling. *Queueing Systems*, 50:459–480, 2005.

K. AVRACHENKOV, U. AYESTA, P. BROWN, and E. NYBERG. Differentiation between short and long TCP flows: Predictability of the response time. In *Proc. of IEEE INFOCOM 2004*, pages 762–773. Hong Kong, 2004.

K. AVRACHENKOV, P. BROWN, and N. OSIPOVA. Optimal choice of threshold in two level processor sharing. Technical Report RR-6215, INRIA, France, 2007. To appear in: *Annals of Operations Research.*

R.W. CONWAY, W.L. MAXWELL, and L.W. MILLER. *Theory of Scheduling*. Addison-Wesley, Reading, 1967.

H. FENG and V. MISRA. Mixed scheduling disciplines for network flows. *ACM Sigmetrics Performance Evaluation Review*, 31(2):36–39, 2003.

J. GITTINS. *Multi-armed Bandit Allocation Indices*. Wiley, Chichester, 1989.

L. GUO and I. MATTA. Scheduling flows with unknown sizes: approximate analysis. In *Proc. of ACM Sigmetrics 2002*, pages 276–277. Marina del Rey, CA, 2002.

M. HARCHOL-BALTER. Guest Editor's Foreword. *ACM Sigmetrics Performance Evaluation Review*, 34(4):2–3, 2007. Special Issue on New Perspectives in Scheduling.

M. HARCHOL-BALTER, K. SIGMAN, and A. WIERMAN. Asymptotic convergence of scheduling policies with respect to slowdown. *Performance Evaluation*, 49:241–256, 2002.

L. KLEINROCK. *Queueing Systems*, volume II: Computer Applications. Wiley, New York, 1976.

M. NUYENS and A. WIERMAN. The foreground-background queue: a survey. *Performance Evaluation*, 65:286–307, 2008.

N. OSIPOVA. Batch processor sharing with hyper-exponential service time. Technical Report RR-6180, INRIA, France, 2007. To appear in: *Operations Research Letters.*

M. PINEDO. *Scheduling: Theory, Algorithms, and Systems.* Prentice Hall, Englewood Cliffs, 1995.

I.A. RAI, G. URVOY-KELLER, M.K. VERNON, and E.W. BIERSACK. Performance analysis of LAS-based scheduling disciplines in a packet switched network. In *Proc. of ACM Sigmetrics/PERFORMANCE 2004*, pages 106–117. New York, NY, 2004.

R. RIGHTER and J.G. SHANTHIKUMAR. Scheduling multiclass single server queueing systems to stochastically maximize the number of successful departures. *Probability in the Engineering and Informational Sciences*, 3:323–333, 1989.

R. RIGHTER, J.G. SHANTHIKUMAR, and G. YAMAZAKI. On extremal service disciplines in single-stage queueing system. *Journal of Applied Probability*, 27:409–416, 1990.

L.E. SCHRAGE. A proof of the optimality of the shortest remaining processing time discipline. *Operations Research*, 16:687–690, 1968.

M. SHAKED and J.G. SHANTHIKUMAR. *Stochastic Orders and Their Applications.* Academic Press, Boston, 1994.

W. STALLINGS. *Operating Systems: Internals and Design Principles.* Prentice Hall, Upper Saddle River, fifth edition, 2005.

S. YASHKOV. Processor-sharing queues: Some progress in analysis. *Queueing Systems*, 2:1–17, 1987.

S. YASHKOV. Mathematical problems in the theory of shared-processor systems. *Journal of Mathematical Sciences*, 58:101–147, 1992.