Aalto, Samuli; Scully, Ziv

# Minimizing the mean slowdown in the M/G/1 queue

*Published in:*
QUEUEING SYSTEMS

*DOI:*
10.1007/s11134-023-09888-6

Published: 01/08/2023

*Document Version*
Publisher's PDF, also known as Version of record

Queueing Systems
https://doi.org/10.1007/s11134-023-09888-6

# Minimizing the mean slowdown in the M/G/1 queue

**Samuli Aalto**[1] · **Ziv Scully**[2]

Received: 18 February 2023 / Revised: 12 August 2023 / Accepted: 17 August 2023


**Abstract**
We consider the optimal scheduling problem in the M/G/1 queue. While this is a thoroughly studied problem when the target is to minimize the mean delay, there are still open questions related to some other objective functions. In this paper, we focus on minimizing mean slowdown among non-anticipating polices, which may utilize the attained service of the jobs but not their remaining service time when making scheduling decisions. By applying the Gittins index approach, we give necessary and sufficient conditions for the jobs' service time distribution under which the well-known scheduling policies first come first served and foreground background are optimal with respect to the mean slowdown. Furthermore, we characterize the optimal non-anticipating policy in the multi-class case for certain types of service time distributions. In fact, our results cover a more general objective function than just the mean slowdown, since we allow the holding costs for a job to depend on its own service time $S$ via a generic function $c(S)$. When minimizing the mean slowdown, this function reads as $c(x) = 1/x$.

**Keywords** Optimal scheduling · M/G/1 · Slowdown · Gittins index · FCFS · FB

**Mathematics Subject Classification** 60K25 · 90B22 · 90B36 · 68M20

## 1 Introduction

We consider the optimal scheduling problem in the M/G/1 queue. Let $S$ denote the service time of a job. There may be a single class of jobs or multiple classes (even an infinite number of classes). Within each class, the service times are assumed to be independent and identically distributed with a finite mean. Jobs of each class arrive

✉ Samuli Aalto
samuli.aalto@aalto.fi

1 Department of Information and Communications Engineering, Aalto University, Espoo, Finland

2 School of Operations Research and Information Engineering, Cornell University, Ithaca, NY, USA

Published online: 04 September 2023

according to an independent Poisson process. Thus, the aggregate arrival process is a Poisson process, too. The total arrival rate of jobs is denoted by $\lambda$. We assume that the total load $\rho = \lambda E[S] < 1$ in order to have a stable M/G/1 queue. The delay (a.k.a. sojourn time or response time) of a job is denoted by $T^\pi$, where $\pi$ refers to the scheduling policy applied. We assume that the applied scheduling policy allows preemptions.

The optimal scheduling policy depends naturally on the objective function but also on the information available to the scheduler. Policy $\pi$ is said to be *non-anticipating* if it is aware of the arrival time and the attained service of each job in the system, while an *anticipating* scheduler knows even the remaining service times. If the aim is to minimize the mean delay $E[T^\pi]$, then the optimal anticipating scheduling policy is SRPT (Shortest Remaining Processing Time) [18, 23]. In the special case where the service times are deterministic, SRPT coincides with the ordinary FCFS (first come first served) discipline or any other non-preemptive and work-conserving scheduling policy.

The optimal non-anticipating policy minimizing the mean delay, however, depends essentially on the service time distribution. For example, FCFS is optimal when the service time distribution belongs to the family of NBUE (New Better Used in Expectation) distributions. On the other hand, for more variable DHR (Decreasing Hazard Rate) service times the optimal non-anticipating policy is FB[1] (foreground background), which is another well-known non-anticipating scheduling policy [16].

All these results can be justified by utilizing the concept of *Gittins index*. It is known that the optimal non-anticipating policy minimizing the mean delay is the Gittins index policy [2, 3, 8, 19, 21], which always chooses the job with the highest index $G^{\mathrm{del}}(a)$ defined by

$$G^{\mathrm{del}}(a) = \sup_{b>a} \frac{P\{S \le b \mid S > a\}}{E[\min\{S, b\} - a \mid S > a]}, \tag{1}$$

where $S$ denotes the (original) service time and $a$ the (currently) attained service of the job.

While the optimal scheduling problem in the M/G/1 queue is a thoroughly studied problem when the target is to minimize the mean delay, there are still open questions related to other objective functions. In this paper, we focus on minimizing *mean slowdown*[2] $E[T^\pi/S]$, i.e., the expectation of the ratio between the delay and the service time of a job. Among the anticipating policies, the optimal policy is known to be SPTP[3] (Shortest Processing Time Product) [12, 27, 28]. But the optimal non-anticipating scheduler with respect to the mean slowdown has long been an open problem [1, 5, 6, 11]. It is only known that FB is the optimal non-anticipating policy if the service time distribution of all jobs is such the ratio $h(x)/x$ is decreasing [7], where $h(x)$ denotes the hazard rate.

[1] The FB policy chooses always the job with the least attained service. It is also known as LAS (Least Attained Service).

[2] Slowdown is also known as *stretch*.

[3] The SPTP policy chooses always the job with the smallest product of the original and the remaining service times. It is also known as the RS policy [11, 17, 26].

In this paper, we prove that the condition given above is not only sufficient but also necessary for the optimality of FB (see Theorem 2 and Corollary 4 in Sect. 3). As well, we give sufficient and necessary conditions under which FCFS is optimal minimizing the mean slowdown among the non-anticipating policies (see Theorem 1 and Corollary 3 in Sect. 3). Furthermore, we characterize the optimal non-anticipating policy in the multi-class case for certain types of service time distributions (see Theorems 3 and 4 together with Corollaries 5 and 6 in Sect. 4).

Our approach is based on the Gittins index. In fact, we consider even a more general objective function than just the mean slowdown. More precisely said, we assume that the holding costs for a job with service time $S$ accrue at rate $c(S) > 0$. If the aim is to minimize the mean slowdown, then the holding cost rate function $c(x)$ is given by $c(x) = 1/x$. On the other hand, the choice $c(x) = 1$ corresponds to minimizing mean delay. It was recently shown in [19, 21] that the Gittins index approach is applicable even for this kind of a general setting of holding costs: The optimal non-anticipating policy is the index policy that always chooses the job with the highest index $G_c(a)$ defined by[4]

$$G_c(a) = \sup_{b>a} \frac{E[c(S)1_{\{S\le b\}} \mid S > a]}{E[\min\{S, b\} - a \mid S > a]}, \tag{2}$$

where $1_A$ refers to the indicator function of event $A$. Starting from this formula, we are able to prove the results mentioned above.

The rest of the paper is organized as follows. In Sect. 2, we consider a single job and derive certain important properties of the Gittins index function defined in (2). These properties are then utilized in Sects. 3 and 4, where we characterize the optimal scheduling policy with respect to the general holding costs in the single-class and the multi-class cases, respectively. The main results related to minimizing mean slowdown are then illustrated by numerical examples in Sect. 5.

## 2 Properties of the Gittins index

In this section, we consider a single job with service time $S$. The aim is to derive such properties of the Gittins index (2) that enable us (later on in Sect. 3) to characterize for which type of service time distributions FCFS or FB are optimal with respect to the general holding costs. Thus, we assume here that the holding costs for the job accrue at rate $c(S)$, which depends on its own service time $S$. We assume that the cost rate function $c(x)$ is right-continuous with left limits.

The service time $S$ is assumed to have a general distribution with the cumulative distribution function denoted by

$$F(x) = P\{S \le x\}, \quad x \ge 0.$$

[4] A historical note: (2) was actually first discovered in 1972 by von Olivier [25]. However, [25] proves the optimality of the resulting index policy only relative to other index policies. The question of whether non-index policies could do better was open until [21]. See [21, Section II-B] for further discussion.

Let $\bar{F}(x)$ denote the corresponding tail distribution function,

$$\bar{F}(x) = 1 - F(x), \quad x \geq 0,$$

and assume that $\bar{F}(x) > 0$ for all $x \geq 0$.[5] In addition, we assume that the service time distribution has a right-continuous density function $f(x)$ with left limits.[6] Let $h(x)$ denote the corresponding *hazard rate* function,

$$h(x) = \frac{f(x)}{\bar{F}(x)}, \quad x \geq 0. \tag{3}$$

In addition, we introduce the following auxiliary functions that depend both on the service time distribution and the cost rate function $c(x)$:

$$\begin{aligned} h_c(x) &= c(x)h(x), \quad x \geq 0, \\ H_c(x) &= \frac{E[c(S) \mid S > x]}{E[S - x \mid S > x]} = \frac{\int_x^\infty h_c(y)\bar{F}(y)\,\mathrm{d}y}{\int_x^\infty \bar{F}(y)\,\mathrm{d}y}, \quad x \geq 0. \end{aligned} \tag{4}$$

Since $c(x)$ and $f(x)$ are right-continuous with left limits, $h(x)$ and $h_c(x)$ are also such functions. In addition, $H_c(x)$ is, by construction, a continuous function. Note also that if the aim is to minimize mean delay, which corresponds to $c(x) = 1$ for all $x$, then $h_c(x)$ equals the hazard rate function, $h_c(x) = h(x)$, and $H_c(x)$ equals the inverse of the so-called *mean residual lifetime* function, $H_c(x) = 1/E[S - x \mid S > x]$.

Let us now rewrite the *Gittins index* (2) for this job as follows:

$$G_c(a) = \sup_{b>a} J_c(a, b), \quad a \geq 0, \tag{5}$$

where $a$ denotes the attained service of the job and $J_c(a, b)$ refers to the following *efficiency function*:

$$J_c(a, b) = \frac{E[c(S)1_{\{S \leq b\}} \mid S > a]}{E[\min\{S, b\} - a \mid S > a]} = \frac{\int_a^b h_c(y)\bar{F}(y)\,\mathrm{d}y}{\int_a^b \bar{F}(y)\,\mathrm{d}y}, \quad a < b. \tag{6}$$

Note that

$$\lim_{b \to a^+} J_c(a, b) = h_c(a), \quad \lim_{b \to \infty} J_c(a, b) = H_c(a), \tag{7}$$

where $h_c(a)$ and $H_c(a)$ are defined in (4). Note also that, with a fixed $a$, the function $J_c(a, b)$ is continuous with respect to $b$ for any $b > a$. By (7), it can also be

[5] This assumption is made for ease of exposition. Similar results are achievable for distributions with a finite upper bound $t_F = \sup\{x \geq 0 : \bar{F}(x) > 0\}$ for the service times (such as the uniform distribution). The statements and proofs are essentially the same, except the domain $[0, \infty)$ is replaced by $[0, t_F)$.

[6] One of our main results, namely Theorem 1, can also be shown for distributions without a density, such as discrete distributions. But our other results require a density function, so for ease of exposition, we focus on the case where a density exists.

continuously continued by defining

$$J_c(a,a) = h_c(a), \quad J_c(a,\infty) = H_c(a). \tag{8}$$

Define finally

$$b_c^*(a) = \max\{b \in [a,\infty] : J_c(a,b) = G_c(a)\}, \quad a \geq 0, \tag{9}$$

to be the (largest) maximizer $b$ of $J_c(a,b)$.

### 2.1 Properties related to the $H_c$-function

In this section, we study connections between the function $H_c(x)$ defined in (4) and the corresponding Gittins index $G_c(x)$. In particular, we derive sufficient and necessary conditions under which $G_c(x) = H_c(x)$ for some $x$.

**Lemma 1** *Let $b > a$. Now $H_c(b) \geq H_c(a)$ if and only if $J_c(a,\infty) \geq J_c(a,b)$.*

***Proof*** By (4), we have

$$\begin{aligned}
&H_c(b) \geq H_c(a)\\
&\iff \frac{\int_b^\infty h_c(y)\bar{F}(y)\,\mathrm{d}y}{\int_b^\infty \bar{F}(y)\,\mathrm{d}y} \geq \frac{\int_a^\infty h_c(y)\bar{F}(y)\,\mathrm{d}y}{\int_a^\infty \bar{F}(y)\,\mathrm{d}y}\\
&\iff \left(\int_b^\infty h_c(y)\bar{F}(y)\,\mathrm{d}y\right)\left(\int_a^\infty \bar{F}(y)\,\mathrm{d}y\right) \geq\\
&\qquad \left(\int_a^\infty h_c(y)\bar{F}(y)\,\mathrm{d}y\right)\left(\int_b^\infty \bar{F}(y)\,\mathrm{d}y\right)\\
&\iff \left(\int_b^\infty h_c(y)\bar{F}(y)\,\mathrm{d}y\right)\left(\int_a^\infty \bar{F}(y)\,\mathrm{d}y\right) \geq\\
&\qquad \left(\int_a^\infty h_c(y)\bar{F}(y)\,\mathrm{d}y\right)\left(\int_a^\infty \bar{F}(y)\,\mathrm{d}y - \int_a^b \bar{F}(y)\,\mathrm{d}y\right).
\end{aligned}$$

By rearranging the terms in last inequality, we get

$$\begin{aligned}
&H_c(b) \geq H_c(a)\\
&\iff \left(\int_a^\infty h_c(y)\bar{F}(y)\,\mathrm{d}y\right)\left(\int_a^b \bar{F}(y)\,\mathrm{d}y\right) \geq\\
&\qquad \left(\int_a^\infty h_c(y)\bar{F}(y)\,\mathrm{d}y - \int_b^\infty h_c(y)\bar{F}(y)\,\mathrm{d}y\right)\left(\int_a^\infty \bar{F}(y)\,\mathrm{d}y\right)\\
&\iff \frac{\int_a^\infty h_c(y)\bar{F}(y)\,\mathrm{d}y}{\int_a^\infty \bar{F}(y)\,\mathrm{d}y} \geq \frac{\int_a^\infty h_c(y)\bar{F}(y)\,\mathrm{d}y - \int_b^\infty h_c(y)\bar{F}(y)\,\mathrm{d}y}{\int_a^b \bar{F}(y)\,\mathrm{d}y},
\end{aligned}$$

from which we get, by (6),

$$\begin{aligned} & H_c(b) \geq H_c(a) \\ \iff & \frac{\int_a^\infty h_c(y)\bar{F}(y)\,\mathrm{d}y}{\int_a^\infty \bar{F}(y)\,\mathrm{d}y} \geq \frac{\int_a^b h_c(y)\bar{F}(y)\,\mathrm{d}y}{\int_a^b \bar{F}(y)\,\mathrm{d}y} \\ \iff & J_c(a,\infty) \geq J_c(a,b). \end{aligned}$$

This completes the proof. □

**Proposition 1** *Let $a \geq 0$. The following three statements are equivalent:*

(i) $H_c(x) \geq H_c(a)$ *for all* $x > a$;
(ii) $G_c(x) \geq G_c(a)$ *for all* $x > a$;
(iii) $G_c(a) = H_c(a)$.

***Proof*** Note first that, by (5) and (7), the equivalence between (i) and (iii) follows immediately from Lemma 1. Below we prove the equivalence between (ii) and (iii) in two parts.

1° Assume first that $G_c(a) = H_c(a)$ so that also claim (i) true. Let $x > a$. It follows from (i) that $H_c(x) \geq H_c(a)$. On the other hand, $G_c(x) \geq H_c(x)$ by (5) and (7). Thus,

$$G_c(x) \geq H_c(x) \geq H_c(a) = G_c(a).$$

2° Assume now that $G_c(x) \geq G_c(a)$ for all $x > a$. By (9), the claim that $G_c(a) = H_c(a)$ is equivalent to claim $b^*(a) = \infty$. We will prove this latter claim by contradiction.

Consider what happens if $b^*(a) < \infty$. Then $J_c(a,x) < J_c(a,b^*(a))$ for all $x \in (b^*(a),\infty]$. Thus, there are $d > 0$ and $M \in (b^*(a),\infty)$ such that $J_c(a,\infty) < d < J_c(a,b^*(a))$ and $J_c(a,x) \leq d$ for all $x \in [M,\infty]$. On the other hand, since $J_c(a,x)$ is continuous (with respect to $x$) and $J_c(a,M) \leq d < J_c(a,b^*(a))$, there is $m \in (b^*(a),M]$ such that $J_c(a,m) = d$. There is also $m^* \in [m,M]$ such that

$$J_c(a,m^*) = \sup_{x\in[m,M]} J_c(a,x).$$

Clearly, for all $x \in [m^*,\infty]$,

$$J_c(a,x) \leq J_c(a,m^*). \tag{10}$$

Let then $x \in (m^*,\infty]$. Now, by defining $p \in (0,1)$ so that

$$p = \frac{\int_a^{m^*} \bar{F}(y)\,\mathrm{d}y}{\int_a^x \bar{F}(y)\,\mathrm{d}y},$$

it follows from (10) that

$$J_c(a,x) = pJ_c(a,m^*) + (1-p)J_c(m^*,x) \geq pJ_c(a,x) + (1-p)J_c(m^*,x),$$

which implies that, for any $x \in (m^*, \infty]$,

$$J_c(m^*, x) \leq J_c(a, x) \leq J_c(a, m^*). \tag{11}$$

By continuity, we also have

$$J_c(m^*, m^*) \leq J_c(a, m^*). \tag{12}$$

From (11) and (12) it follows that

$$G_c(m^*) = \sup_{x \geq m^*} J_c(m^*, x) \leq J_c(a, m^*) < J_c(a, b^*(a)) = G_c(a),$$

which contradicts our assumption that $G_c(x) \geq G_c(a)$ for all $x > a$. Thus, necessarily $b^*(a) = \infty$. □

By Proposition 1, we get the following immediate consequence describing further the connections between the service time distribution and the corresponding Gittins index.

**Corollary 1** *Let $a \geq 0$. The following three statements are equivalent:*

(i) *$H_c(x)$ is increasing for all $x > a$;*
(ii) *$G_c(x)$ is increasing for all $x > a$;*
(iii) *$G_c(x) = H_c(x)$ for all $x \geq a$.*

Note that, in this paper, we use the terms 'increasing' and 'decreasing' in their weak forms. So, the functions $H_c(x)$ and $G_c(x)$ in the claims above are not required to be strictly increasing.

### 2.2 Properties related to the $h_c$-function

In this section, we study connections between the function $h_c(x)$ defined in (4) and the corresponding Gittins index $G_c(x)$. In particular, we derive sufficient and necessary conditions under which $G_c(x) = h_c(x)$ for some $x$.

**Lemma 2** *Let $b > a$. If $h_c(x)$ is decreasing for all $x \in [a, b)$, then $J_c(a, x)$ is decreasing (with respect to $x$) for all $x \in [a, b)$.*

***Proof*** Let $x \in (a, b)$. Now $h_c(y) \geq h_c(x)$ for all $y \in [a, x)$. Thus,

$$J_c(a, x) = \frac{\int_a^x h_c(y)\bar{F}(y)\,\mathrm{d}y}{\int_a^x \bar{F}(y)\,\mathrm{d}y} \geq h_c(x).$$

Since, by (6),

$$\frac{\partial}{\partial x} J_c(a, x) = (h_c(x) - J_c(a, x)) \frac{\bar{F}(x)}{\int_a^x \bar{F}(y)\,\mathrm{d}y},$$

the claim follows clearly from the previous inequality. □

**Proposition 2** *Let $a \geq 0$. The following two statements are equivalent:*

(i) *$h_c(x)$ is decreasing for all $x \geq a$.*
(ii) *$G_c(x) = h_c(x)$ for all $x \geq a$.*

***Proof*** 1° Assume first that $h_c(x)$ is decreasing for all $x \geq a$. Let $x \geq a$. It follows from Lemma 2, when letting $b \to \infty$, that $J_c(x, y)$ is decreasing with respect to $y$ for any $y \geq x$ so that

$$J_c(x, x) = \sup_{y>x} J_c(x, y).$$

However, this is equivalent to the claim that

$$G_c(x) = h_c(x).$$

2° Assume now that $G_c(x) = h_c(x)$ for all $x \geq a$. Consider what happens if $h_c(x)$ is not decreasing for all $x \geq a$. Here we need to study two separate cases (2.1° and 2.2° below).

2.1° Assume first that $h_c(x)$ is not decreasing for all $x \geq a$ since there is an interval $[m, M)$ where $h_c(x)$ is strictly increasing. Thus, $h_c(x) > h_c(m)$ for all $x \in (m, M)$. But now

$$h_c(m) = \frac{\int_m^M h_c(m)\bar{F}(y)\,\mathrm{d}y}{\int_m^M \bar{F}(y)\,\mathrm{d}y} < \frac{\int_m^M h_c(y)\bar{F}(y)\,\mathrm{d}y}{\int_m^M \bar{F}(y)\,\mathrm{d}y} = J_c(m, M),$$

which implies that

$$G_c(m) \geq J_c(m, M) > h_c(m).$$

However, this contradicts our assumption that $G_c(x) = h_c(x)$ for all $x \geq a$.

2.2° Assume now that $h_c(x)$ is not decreasing for all $x \geq a$ since there is a jump up at some point $x_0 > a$. Thus, $h_c(x_0^-) < h_c(x_0^+)$. Let $d_1$ and $d_2$ be such that $h_c(x_0^-) < d_1 < d_2 < h_c(x_0^+)$. Since $h_c(x)$ is right-continuous with left limits, there are $\delta_1$ and $\delta_2$ such that $h_c(x) < d_1$ for all $x \in (x_0 - \delta_1, x_0)$ and $h_c(x) > d_2$ for all $x \in (x_0, x_0 + \delta_2)$. Let then $x \in (x_0 - \delta_1, x_0)$, and define

$$p(x) = \frac{\int_x^{x_0} \bar{F}(y)\,\mathrm{d}y}{\int_x^{x_0+\delta_2} \bar{F}(y)\,\mathrm{d}y}.$$

Clearly, $p(x) \in (0, 1)$, and we have

$$\begin{aligned} J_c(x, x_0 + \delta_2) &= p(x)J_c(x, x_0) + (1 - p(x))J_c(x_0, x_0 + \delta_2) \\ &= p(x)J_c(x, x_0) + (1 - p(x))\frac{\int_{x_0}^{x_0+\delta_2} h_c(y)\bar{F}(y)\,\mathrm{d}y}{\int_{x_0}^{x_0+\delta_2} \bar{F}(y)\,\mathrm{d}y} \\ &> p(x)J_c(x, x_0) + (1 - p(x))d_2. \end{aligned}$$

Since $p(x) \to 0$ and $J_c(x, x_0) \to h_c(x_0^-)$ as $x \to x_0$ and $d_2 > d_1$, it follows from the previous inequality that there is $\delta_1' \in (\delta_1, 0)$ such that

$$J_c(x_0 - \delta_1', x_0 + \delta_2) > d_1 > h_c(x_0 - \delta_1').$$

However, this contradicts our assumption that $G_c(x) = h_c(x)$ for all $x \geq a$. □

**Lemma 3** *Let $a \geq 0$. If $G_c(x) \leq G_c(a)$ for all $x \geq a$, then*

$$G_c(a) = h_c(a).$$

***Proof*** Assume that $G_c(x) \leq G_c(a)$ for all $x \geq a$.

1° If $b^*(a) = a$, then, by (9) and (8), we have

$$G_c(a) = J_c(a, b^*) = J_c(a, a) = h_c(a).$$

2° Now assume that $b^*(a) > a$. In addition, let $x \in (a, b^*(a))$, and define

$$p(x) = \frac{\int_a^x \bar{F}(y)\,\mathrm{d}y}{\int_a^{b^*(a)} \bar{F}(y)\,\mathrm{d}y}.$$

Clearly, $p(x) \in (0, 1)$. Note also that we may write

$$G_c(a) = J_c(a, b^*(a)) = p(x)J_c(a, x) + (1 - p(x))J_c(x, b^*(a)).$$

Since $J_c(a, x) \leq G_c(a)$ by definition, we have

$$\begin{aligned} G_c(a) &= p(x)J_c(a, x) + (1 - p(x))J_c(x, b^*(a)) \\ &\leq\ p(x)G_c(a) + (1 - p(x))J_c(x, b^*(a)), \end{aligned}$$

which implies that

$$G_c(a) \leq J_c(x, b^*(a)).$$

On the other hand, due to our assumption, we have

$$J_c(x, b^*(a)) \leq G_c(x) \leq G_c(a).$$

Thus, necessarily

$$G_c(x) = G_c(a).$$

In addition, since $J_c(x, b^*(a)) \leq G_c(x)$, we have

$$\begin{aligned} G_c(a) &= p(x)J_c(a, x) + (1 - p(x))J_c(x, b^*(a)) \\ &\leq p(x)J_c(a, x) + (1 - p(x))G_c(x) \\ &= p(x)J_c(a, x) + (1 - p(x))G_c(a), \end{aligned}$$

which implies that

$$G_c(a) \leq J_c(a, x).$$

However, this is possible only if

$$J_c(a, x) = G_c(a).$$

Since this is true for any $x \in (a, b^*(a))$, it implies that

$$h_c(a) = J_c(a, a) = \lim_{x \to a} J_c(a, x) = G_c(a),$$

which completes the proof. □

**Proposition 3** *Let $a \geq 0$. The following two statements are equivalent:*

(i) *$G_c(x)$ is decreasing for all $x \geq a$.*
(ii) *$G_c(x) = h_c(x)$ for all $x \geq a$.*

***Proof*** 1° Assume first that $G_c(x)$ is decreasing for all $x \geq a$. Let $x \geq a$. Since $G_c(y)$ is decreasing for all $y \geq x$, it follows immediately from Lemma 3 that $G_c(x) = h_c(x)$.

2° Assume now that $G_c(x) = h_c(x)$ for all $x \geq a$. It follows immediately from Proposition 2 that $h_c(x)$, and thus also $G_c(x)$, is decreasing for all $x \geq a$. □

By Propositions 2 and 3, we get the following immediate consequence further describing the connections between the Gittins index and the service time distribution.

**Corollary 2** *Let $a \geq 0$. The following three statements are equivalent:*

(i) *$h_c(x)$ is decreasing for all $x \geq a$.*
(ii) *$G_c(x)$ is decreasing for all $x \geq a$.*
(iii) *$G_c(x) = h_c(x)$ for all $x \geq a$.*

## 3 Characterization of the optimal policy in the single-class case

In this section, we consider the single-class case, i.e., all jobs have the same service time distribution function $F(x)$. We reveal the properties the service time distribution should have in order for FCFS or FB to be the optimal non-anticipating policy with respect to the generalized holding costs introduced in Sect. 1. As before, let $c(x)$ denote the corresponding holding cost rate function, which is now common to all the jobs. Recall also our assumptions made in Sect. 2 that the density function $f(x)$ of the service time distribution and the cost rate function $c(x)$ are right-continuous with left limits.

**Definition 1** Let $\mathcal{I}$ denote the set of current jobs in the system. A scheduling policy belongs to the *MAS* class if it chooses the job with the *most attained service*,

$$i^* = \arg\max_{i \in \mathcal{I}} a_i. \tag{13}$$

Note that MAS is a whole family of scheduling policies consisting of all non-anticipating policies that are work-conserving and non-preemptive. In particular, FCFS is such a policy. Other well-known examples are non-preemptive LCFS (Last Come First Served) and SIRO (Service In Random Order).

**Definition 2** Let $\mathcal{I}$ denote the set of current jobs in the system. The *FB* policy chooses the job with the *least attained service*,

$$i^* = \arg\min_{i \in \mathcal{I}} a_i. \tag{14}$$

**Theorem 1** *Assume the single-class case. Any scheduling policy belonging to the MAS class (including FCFS) minimizes the generalized holding costs among the non-anticipating policies if and only if*

$$H_c(a) \geq H_c(0) \quad \textit{for all } a \geq 0, \tag{15}$$

*where $H_c(a)$ is defined in (4).*

***Proof*** The result follows immediately from Proposition 1 with choice $a = 0$ and the optimality of the Gittins index policy [19, 21]. □

**Theorem 2** *Assume the single-class case. The FB policy minimizes the generalized holding costs among the non-anticipating policies if and only if $h_c(a)$ is decreasing for all $a \geq 0$, where $h_c(a)$ is defined in (4).*

***Proof*** The result follows immediately from Corollary 2 with choice $a = 0$ and the optimality of the Gittins index policy [19, 21]. □

### 3.1 Minimizing mean slowdown in the single-class case

We now spell out the implications of Theorems 1 and 2 for the specific case of minimizing mean slowdown $E[T^\pi/S]$, which corresponds to holding cost rate function $c(x) = 1/x$. Let us denote the corresponding $h_c$-function by

$$h^{\mathrm{sld}}(x) = \frac{h(x)}{x}, \quad x \geq 0, \tag{16}$$

which is called the *scaled hazard rate* in the sequel. In addition, we denote the corresponding $H_c$-function by

$$H^{\mathrm{sld}}(x) = \frac{E[1/S \mid S > x]}{E[S - x \mid S > x]} = \frac{\int_x^\infty \frac{h(y)}{y} \bar{F}(y)\,\mathrm{d}y}{\int_x^\infty \bar{F}(y)\,\mathrm{d}y}, \quad x \geq 0, \tag{17}$$

and the corresponding Gittins index $G_c(x)$ by $G^{\mathrm{sld}}(x)$. By Theorems 1 and 2, we have the following characterizations for the optimality of the MAS and FB policies.

**Corollary 3** *Assume the single-class case. Any scheduling policy belonging to the MAS class (including FCFS) minimizes the mean slowdown among the non-anticipating policies,*

$$E[T^{\mathrm{MAS}}/S] = \min_{\pi} E[T^{\pi}/S],$$

*if and only if*

$$H^{\mathrm{sld}}(a) \geq H^{\mathrm{sld}}(0) \quad \textit{for all } a \geq 0. \tag{18}$$

This result was already anticipated in [1], where it is presented as a conjecture without any proof. It is easy to show that the family of service time distributions satisfying (18) is a (proper) subset of the NBUE distributions, which were mentioned in Sect. 1. As an example of a distribution that belongs to NBUE but does not satisfy condition (18) serves any Weibull distribution with shape parameter $k \in [1, 2)$.

**Corollary 4** *Assume the single-class case. The FB policy minimizes the mean slowdown among the non-anticipating policies,*

$$E[T^{\mathrm{FB}}/S] = \min_{\pi} E[T^{\pi}/S],$$

*if and only if the scaled hazard rate $h(a)/a$ is decreasing for all $a \geq 0$.*

Feng and Misra [7] already proved that FB minimizes the mean slowdown among the non-anticipating policies if the scaled hazard rate $h(x)/x$ of the service time distribution is a decreasing function of $x$. Such distributions include clearly all DHR distributions, for which $h(x)$ is required to be decreasing. Here we complete the result by proving that this condition is not only sufficient but also necessary for the optimality of FB.

## 4 Characterization of the optimal policy in the multi-class case

In this section, we assume that there are multiple job classes and the scheduler is aware of the class of each job. Let $F_j(x)$ denote the service time distribution function and $c_j(x)$ the holding cost rate function of class $j$. As before, we assume that the density function $f_j(x)$ of the service time distribution and the cost rate function $c_j(x)$ are right-continuous with left limits for all classes $j$. Our aim is to characterize, for certain types of service time distributions, the optimal non-anticipating policy with respect to the generalized holding costs introduced in Sect. 1.

Fix index $i$ for a while and consider job $i$. Let $j$ refer to its class. In line with (4), we define job $i$'s $h_c$-function $h_{c,i}(x)$ and $H_c$-function $H_{c,i}(x)$ as follows:

$$\begin{aligned} h_{c,i}(x) &= c_j(x)h_j(x), \quad x \geq 0, \\ H_{c,i}(x) &= \frac{E[c_j(S_j) \mid S_j > x]}{E[S_j - x \mid S_j > x]} = \frac{\int_x^{\infty} h_{c,i}(y)\bar{F}_j(y)\,\mathrm{d}y}{\int_x^{\infty} \bar{F}_j(y)\,\mathrm{d}y}, \quad x \geq 0, \end{aligned} \tag{19}$$

where $h_j(x)$ refers to the hazard rate function of service time distribution function $F_j(x)$.

**Definition 3** Let $\mathcal{I}$ denote the set of current jobs in the system. The *MAX-H* policy chooses the job that maximizes the current value of the $H_c$-function,

$$i^* = \arg\max_{i \in \mathcal{I}} H_{c,i}(a_i), \tag{20}$$

where $a_i$ is the attained service of job $i$ and $H_{c,i}(\cdot)$ refers to its $H_c$-function as defined in (19).

**Definition 4** Let $\mathcal{I}$ denote the set of current jobs in the system. The *MAX-h* policy chooses the job that maximizes the current value of the $h_c$-function,

$$i^* = \arg\max_{i \in \mathcal{I}} h_{c,i}(a_i), \tag{21}$$

where $a_i$ is the attained service of job $i$ and $h_{c,i}(\cdot)$ refers to its $h_c$-function as defined in (19).

Note that, if the aim is to minimize mean delay, which corresponds to $c_j(x) = 1$ for all job classes $j$, then MAX-H is the same as the *SERPT* (Shortest Expected Remaining Processing Time) policy, since function $H_{c,i}(x)$ equals, in this case, the inverse of the mean residual lifetime function, $H_{c,i}(x) = 1/E[S_j - x \mid S_j > x]$. Correspondingly, MAX-h is the same as the *HHR* (Highest Hazard Rate) policy, since $h_{c,i}(x)$ is the hazard rate function of the job $i$'s service time distribution in this special case.

Note also that functions $H_{c,i}(x)$ and $h_{c,i}(x)$ are common to all jobs $i$ belonging to the same class, say $j$. Therefore, we may, as well, refer to them by $H_{c,j}(x)$ and $h_{c,j}(x)$, respectively, without any danger for confusion.

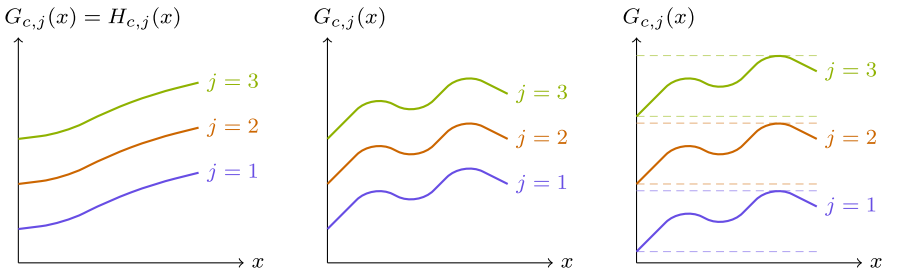
**Theorem 3** *Assume the multi-class case. The MAX-H policy minimizes the generalized holding costs among the non-anticipating policies if class-wise functions $H_{c,j}(x)$ are increasing for all classes $j$.*

***Proof*** The result follows immediately from Corollary 1 with choice $a = 0$ and the optimality of the Gittins index policy [19, 21]. □

**Theorem 4** *Assume the multi-class case. The MAX-h policy minimizes the generalized holding costs among the non-anticipating policies if and only if class-wise functions $h_{c,j}(x)$ are decreasing for all classes $j$.*

***Proof*** The result follows immediately from Corollary 2 with choice $a = 0$ and the optimality of the Gittins index policy [19, 21]. □

The precondition of Theorem 4 is essentially the multi-class analogue of the precondition of Theorem 2. However, the precondition of Theorem 3 is *stricter* than the multi-class analogue of the precondition of Theorem 1. Below, we discuss why this difference occurs. See Fig. 1 for an accompanying illustration.




(a) $H_{c,j}(x)$ increasing (b) $H_{c,j}(x)$ minimized at $x = 0$ (c) $H_{c,j}(x)$ minimized at $x = 0$ and non-overlapping

**Fig. 1** **a** If $H_{c,j}(x)$ is increasing for all classes $j$, then MAX-H is optimal. **b** But if we weaken this condition to $H_{c,j}(x)$ being minimized at $x = 0$ for all classes $j$, then MAX-H may be suboptimal. **c** However, if additionally the $H_{c,j}(x)$ values of different classes $j$ occupy non-overlapping intervals, then MAX-H is optimal after all

For MAX-H to be optimal, in the single-class case, we require only that $H_c(x)$ is minimized at $x = 0$, whereas in the multi-class case, we require $H_{c,j}(x)$ to be increasing at all $x \geq 0$. The issue with MAX-H when we only have $H_{c,j}(x)$ minimized at $x = 0$ is that while MAX-H correctly prioritizes jobs that have not begun service (because $H_{c,j}(0) = G_{c,j}(0)$ by Proposition 1), it may incorrectly prioritize jobs that have begun service. This is not a problem in the single-class case, as all that matters is that jobs that have begun service have priority over jobs that have not yet begun service. But in the multi-class case, we may need to compare a class $j$ job in service to a class $j'$ job that has not yet begun service.

However, there are cases where MAX-H is optimal even when the precondition of Theorem 3 is not satisfied. It turns out that if the class-wise Gittins index functions $G_{c,j}(x)$ are minimized at $x = 0$ but have non-overlapping values, i.e.,

$$G_{c,j}(x) \geq G_{c,j'}(y) \quad \text{for all } x, y \geq 0 \text{ and all } j < j',$$

then the class-wise $H_{c,j}(x)$ functions are also minimized at $x = 0$ and have non-overlapping values, thanks to Proposition 1 and the fact that $H_{c,j}(x) \leq G_{c,j}(x)$. In this non-overlapping case, illustrated in Fig. 1c, the Gittins index policy reduces to preemptive class-based priority: class $j$ has priority over class $j'$ for $j < j'$, with jobs served in FCFS order[7] within each class. But MAX-H reduces to the same policy, so it is also optimal.

We have shown that the precondition of Theorem 3 is sufficient but not necessary for MAX-H to be optimal. Exactly stating the necessary condition, or even just a more lenient sufficient condition, seems to be difficult. For example, one might consider the condition that the $H_{c,j}(x)$ functions are minimized at $x = 0$ and have non-overlapping values. But this is neither necessary, because we might have $G_{c,j}(x) = H_{c,j}(x)$ whenever there is overlap; nor sufficient, because while non-overlapping Gittins index functions $G_{c,j}(x)$ suffice, this is not implied by non-overlapping $H_{c,j}(x)$ functions.

[7] Or, more generally, using any MAS policy (Definition 1) within each class.

### 4.1 Minimizing mean slowdown in the multi-class case

We now spell out the implications of Theorems 3 and 4 to minimizing the mean slowdown, which corresponds to holding cost rate functions $c_j(x) = 1/x$ for all job classes $j$. We have the following characterizations for the optimality of the MAX-H and MAX-h policies as direct consequences of Theorems 3 and 4, respectively.

**Corollary 5** *Assume the multi-class case. The MAX-H policy minimizes the mean slowdown among the non-anticipating policies,*

$$E[T^{\mathrm{MAX-H}}/S] = \min_{\pi} E[T^{\pi}/S],$$

*if class-wise functions $H_j^{\mathrm{sld}}(x)$ are increasing for all classes $j$, where $H_j^{\mathrm{sld}}(x)$ is defined (in line with (17)) as follows*

$$H_j^{\mathrm{sld}}(x) = \frac{E[1/S_j \mid S_j > x]}{E[S_j - x \mid S_j > x]} = \frac{\int_x^{\infty} \frac{h_j(y)}{y} \bar{F}_j(y)\,\mathrm{d}y}{\int_x^{\infty} \bar{F}_j(y)\,\mathrm{d}y}, \quad x \geq 0.$$

**Corollary 6** *Assume the multi-class case. The MAX-h policy minimizes the mean slowdown among the non-anticipating policies,*

$$E[T^{\mathrm{MAX-h}}/S] = \min_{\pi} E[T^{\pi}/S],$$

*if and only if class-wise scaled hazard rate functions $h_j(x)/x$ are decreasing for all classes $j$.*

## 5 Numerical examples

In this section, we illustrate numerically the main results related to the minimization of the mean slowdown. Thus, we assume that the holding cost rate is given by $c(x) = 1/x$ for all jobs. In addition, we give examples of the corresponding Gittins index $G^{\mathrm{sld}}(a)$ as a function of attained service $a$.

For the illustration, we use the *Weibull service time distribution*, for which $E[S] = \frac{1}{\mu}\Gamma(1 + \frac{1}{k})$ and the tail distribution function is given by

$$\bar{F}(x) = e^{-(\mu x)^k},$$

where $k > 0$ is the shape parameter and $\mu > 0$ the scale parameter. With $k = 1$, we have the exponential distribution as a special case.

The scaled hazard rate for a Weibull$(k, \mu)$ distribution reads as

$$\frac{h(x)}{x} = k\mu^2(\mu x)^{k-2}$$

and the corresponding $H_c$-function as

$$H^{\mathrm{sld}}(x) = \frac{\int_x^\infty k\mu^2(\mu y)^{k-2}e^{-(\mu y)^k}\,\mathrm{d}y}{\int_x^\infty e^{-(\mu y)^k}\,\mathrm{d}y}.$$

Note that, for $k = 2$, both the scaled hazard rate $h(x)/x$ and the $H_c$-function $H^{\mathrm{sld}}(x)$ reduce to the same constant value for all $x > 0$:

$$\frac{h(x)}{x} = H^{\mathrm{sld}}(x) = 2\mu^2.$$

In addition, we note that the scaled hazard rate $h(x)/x$ is decreasing (satisfying the condition of Corollary 4) when $k \leq 2$, and the $H_c$-function $H^{\mathrm{sld}}(x)$ is increasing (satisfying the condition of Corollary 3) when $k \geq 2$.

The behavior of the Weibull distribution with different shape parameter values $k$ is illustrated in Fig. 2. The shape parameter takes values $k \in \{1, 2, 3, 4\}$, and the scale parameter is chosen to be $\mu = \Gamma(1+\frac{1}{k})$ so that the mean service time remains constant $E[S] = 1$. In the top panel, we have drawn the scaled hazard rate $h(a)/a$ as a function of attained service $a$. In the middle and bottom panels, there are corresponding curves for the $H_c$-function $H^{\mathrm{sld}}(a)$ and the Gittins index $G^{\mathrm{sld}}(a)$, respectively. Note that, in line with Corollaries 1 and 2, the Gittins index $G^{\mathrm{sld}}(a)$ is equal to the $H_c$-function $H^{\mathrm{sld}}(a)$ for all $a \geq 0$ when $k \in \{2, 3, 4\}$ and equal to the scaled hazard rate $h(a)/a$ for all $a \geq 0$ when $k \in \{1, 2\}$.

***Example 1*** Consider first the single-class case where all jobs have the same Weibull service time distribution with shape parameter $k$ and scale parameter $\mu = \Gamma(1 + \frac{1}{k})$. In Fig. 3, we have drawn the mean slowdown with loads $\rho = 0.5$ (upper panel) and $\rho = 0.8$ (lower panel) as a function of inverse shape parameter $1/k$ for the scheduling policies FCFS and FB based on the following known formulas:

$$E[T^{\mathrm{FCFS}}/S] = 1 + \frac{\lambda E[S^2]E[1/S]}{2(1-\rho)},$$
$$E[T^{\mathrm{FB}}/S] = \int_0^\infty \left(\frac{1}{1-\rho(x)} + \frac{\lambda m_2(x)}{2x(1-\rho(x))^2}\right) f(x)\,\mathrm{d}x,$$

where $f(x)$ refers to the density function,

$$f(x) = k\mu^2(\mu x)^{k-1}e^{-(\mu x)^k},$$

$\rho(x) = \lambda E[\min\{S, x\}]$, and $m_2(x) = E[\min\{S, x\}^2]$. For comparison, we have also drawn the mean slowdown for the PS (Processor Sharing) policy based on the following known formula:

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

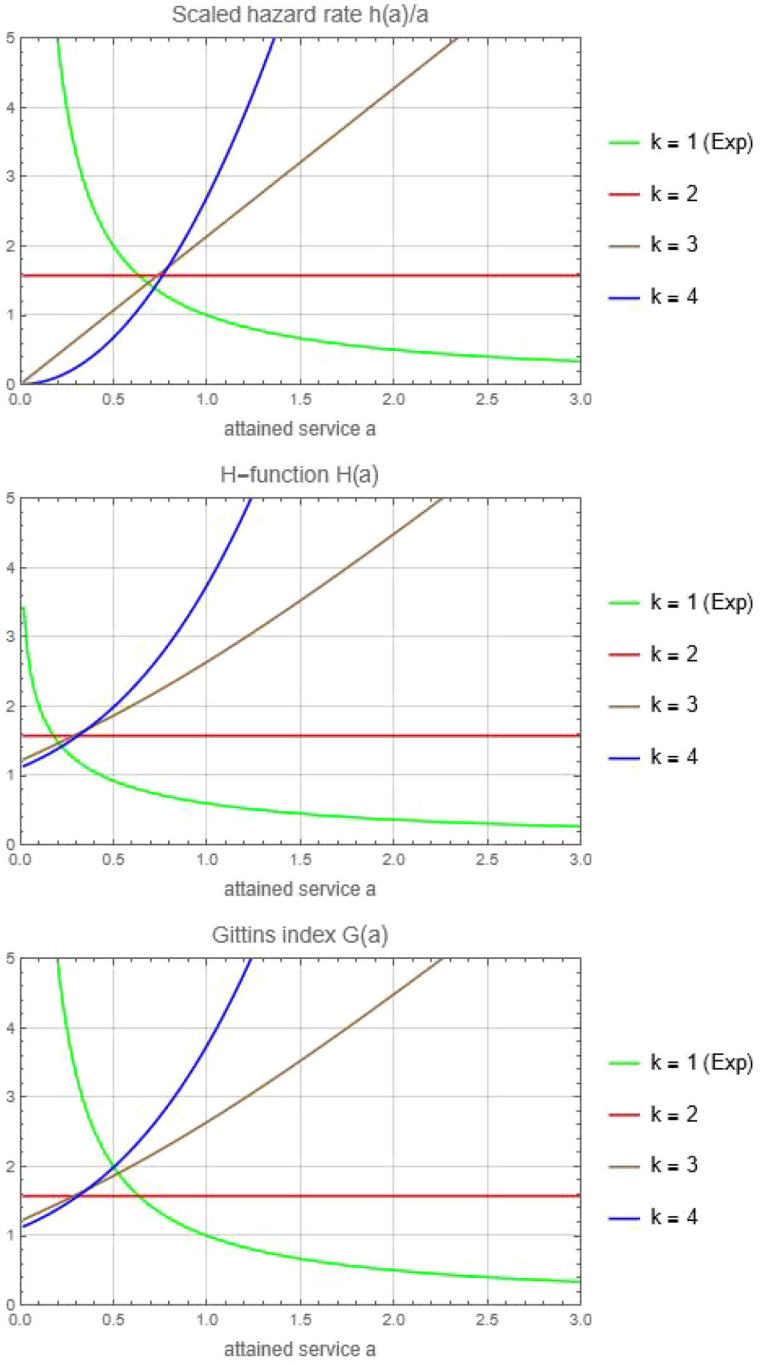



**Fig. 2** Weibull service time distribution: Scaled hazard rate $h(a)/a$ (top), $H_c$-function $H^{\text{sld}}(a)$ (middle), and Gittins index $G^{\text{sld}}(a)$ (bottom) as a function of attained service $a$ for shape parameter values $k \in \{1, 2, 3, 4\}$

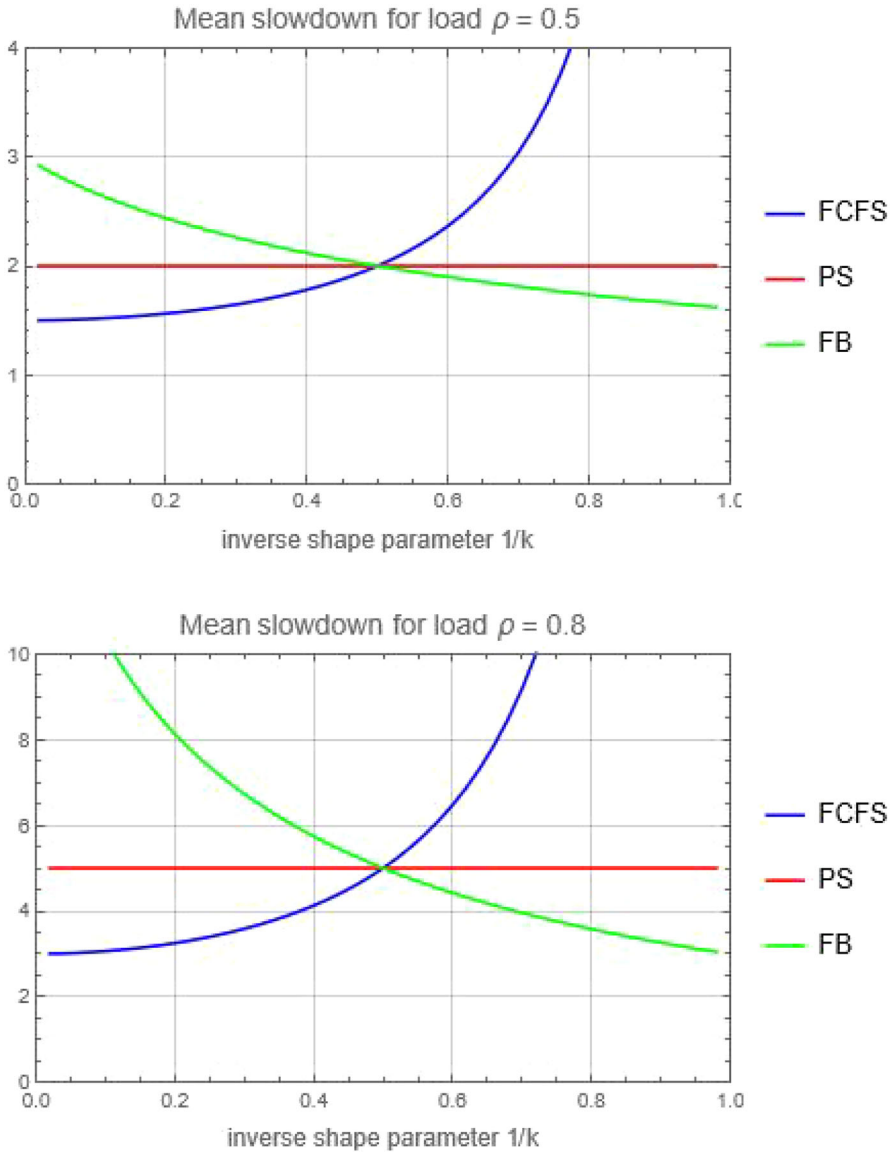



**Fig. 3** Single-class case (Example 1): Mean slowdown with loads $\rho = 0.5$ (upper) and $\rho = 0.8$ (lower) as a function of inverse shape parameter $1/k$ for scheduling policies FCFS, PS, and FB

The shape parameter takes now continuously values $k \in (1, \infty)$ so that $1/k \in (0, 1)$. In the upper panel, the load takes value $\rho = 0.5$, and in the lower one, we have $\rho = 0.8$. Note that, in line with Corollary 3, FCFS is optimal when $1/k \leq 1/2$. However, when $1/k \geq 1/2$, the performance of FCFS becomes soon very bad as $1/k$ increases. In fact, the mean slowdown of the FCFS scheduling policy approaches $\infty$ as $1/k \to 1$. On the other hand, while FB is optimal when $1/k \geq 1/2$ (in line with Corollary 4),

its performance decreases remarkably when $1/k \leq 1/2$, the degradation being even worse with a higher load.

***Example 2*** Next we consider the multi-class case with two job classes. The service times for each class $j \in \{1, 2\}$ follow the Weibull distribution with shape parameter $k_j$ and scale parameter $\mu_j = \Gamma(1 + \frac{1}{k_j})$. Thus, $E[S] = E[S_1] = E[S_2] = 1$. We choose $k_1 = 2$ and $k_2 = 4$ so that the condition of Corollary 5 is satisfied, meaning the Gittins index policy reduces to MAX-H. In addition, we use the same arrival rate for both classes: $\lambda_1 = \lambda_2 = \lambda/2$, where $\lambda$ denotes the total job arrival rate. Based on simulations, we have estimated the mean slowdown for MAX-H, FCFS, PS, and FB with various load levels $\rho = \lambda E[S]$. In each simulation run with a fixed scheduling policy and load, we have gathered the system statistics until there are $10^6$ job arrivals. The results are presented in Fig. 4. In the top panel, we have the estimated total mean slowdown for all jobs as a function of load $\rho$. In the middle and bottom panels, there are corresponding curves for the jobs of classes 1 and 2, respectively. Note that, in line with Corollary 5, MAX-H is better than the other scheduling policies for the total mean slowdown. In addition, FCFS is consistently second best, and FB performs the worst in this example, as expected.

Interestingly, the classwise results for the two best scheduling policies, MAX-H and FCFS, are very different: MAX-H "prefers" class 1 to class 2, whereas FCFS behaves just the opposite. The intuition for this is as follows. As shown in Fig. 2, under MAX-H, the initial index of class 1 ($k_1 = 2$) jobs is greater than that of class 2 jobs ($k_2 = 4$), thus explaining MAX-H's preference for class 1 jobs. This makes sense: class 1 jobs are more likely to be very small than class 2 jobs. Treating all jobs the same, as in FCFS, thus leads to worse mean slowdown for class 1 jobs.

The question remains: given the very different classwise preferences, how come MAX-H and FCFS have such similar overall mean slowdown? We believe this is due to the fact that while a class 2 ($k_2 = 4$) job may, when it first arrives, have worse Gittins index than a class 1 ($k_1 = 2$) job, after just a small amount of service, a class 2 job reaches an index higher than that of class 1 jobs (see Fig. 2). This means that the average fraction of time that FCFS is serving a job other than the one of maximal Gittins index is relatively small, consisting entirely of the short first segments of class 2 jobs.

***Example 3*** In the last example, we again consider the multi-class case with two job classes where the service times for both classes follow the Weibull distribution with unit mean, $E[S] = E[S_1] = E[S_2] = 1$. Now we choose the shape parameters as follows: $k_1 = 1$ and $k_2 = 2$. Thus, the condition of Corollary 6 is satisfied, meaning the Gittins index policy reduces to MAX-h in this case. As in the previous example, we use the same arrival rate for both classes: $\lambda_1 = \lambda_2 = \lambda/2$, where $\lambda$ denotes the total job arrival rate. Based on simulations, we have estimated the mean slowdown for scheduling policies MAX-h, PS, and FB with various load levels $\rho = \lambda E[S]$. FCFS is left out since its performance is much worse than the other three policies in this case. In each simulation run with a fixed scheduling policy and load, we have gathered the system statistics until there are $10^6$ job arrivals. The results are presented in Fig. 5. In the top panel, we have the estimated total mean slowdown for all jobs

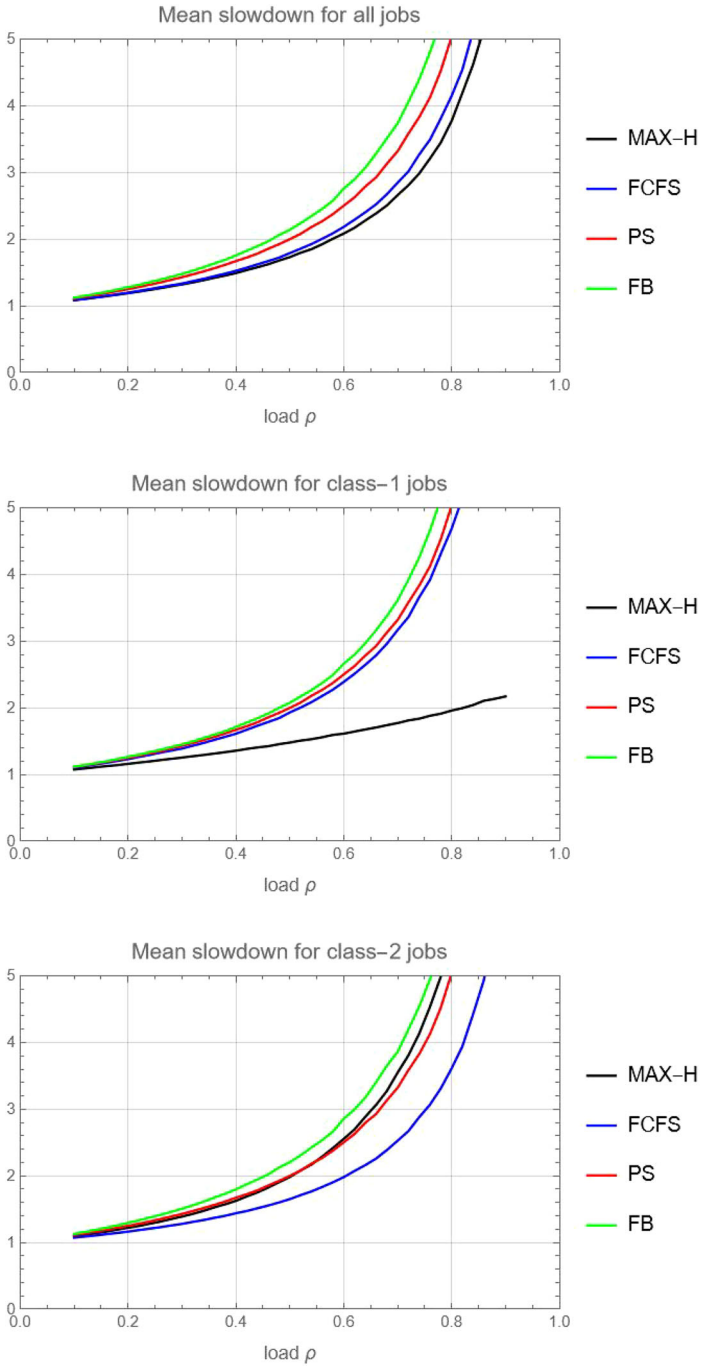



**Fig. 4** Multi-class case (Example 2): Simulated mean slowdown of all jobs (top), class-1 jobs (middle), and class-2 jobs (bottom) as a function of load $\rho$ for scheduling policies MAX-H, FCFS, PS, and FB

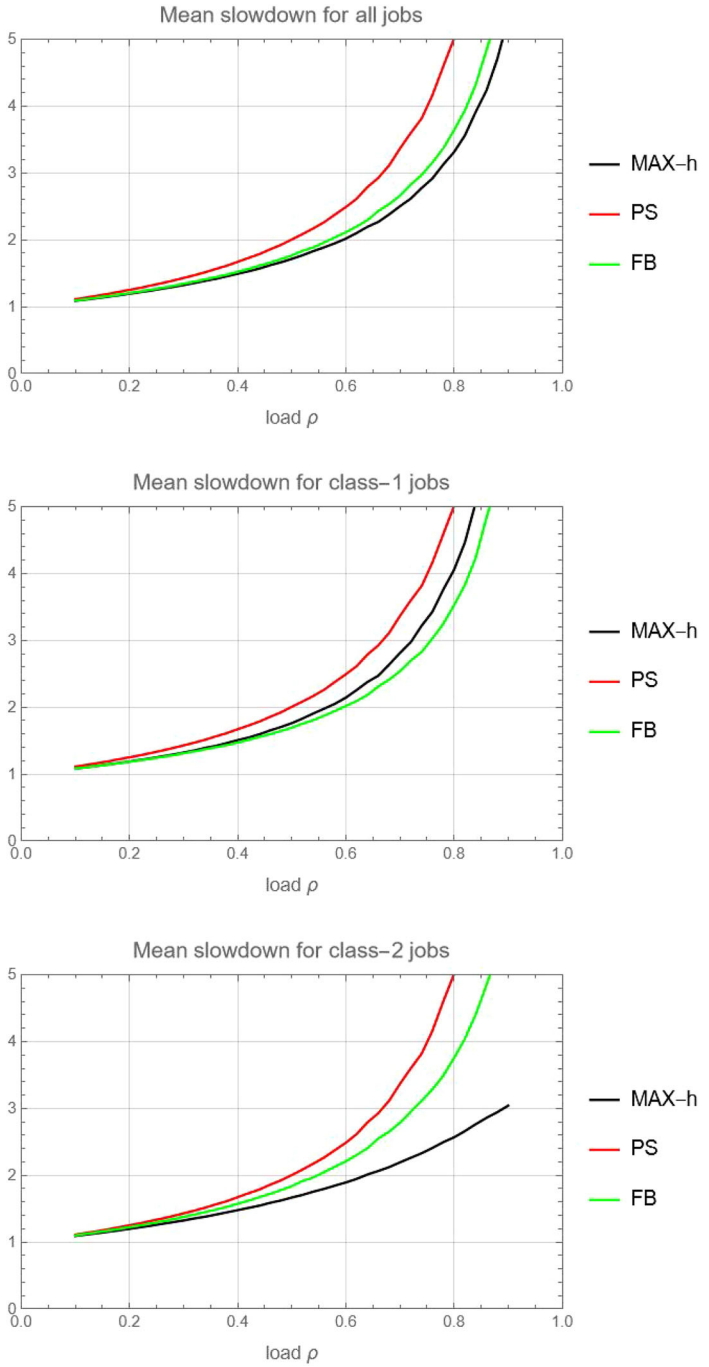



**Fig. 5** Multi-class case (Example 3): Simulated mean slowdown of all jobs (top), class-1 jobs (middle), and class-2 jobs (bottom) as a function of load $\rho$ for scheduling policies MAX-h, PS, and FB

as a function of load $\rho$. In the middle and bottom panels, there are corresponding curves for the jobs of classes 1 and 2, respectively. Note that, in line with Corollary 6, MAX-h is better than the other scheduling policies for the total mean slowdown. In addition, FB is consistently better than PS. Note also that the classwise results for the two best policies, MAX-h and FB, are very different: MAX-h "prefers" clearly class 2 to class 1, whereas FB gives roughly similar performance to both classes.

## 6 Conclusion and discussion

We considered the optimal scheduling problem in the M/G/1 queue with rather general holding costs, which cover, for example, the minimization of the mean slowdown. To determine the optimal scheduling rule among the non-anticipating policies, which are aware of the attained services of the jobs but not on their remaining service times, we applied the Gittins index approach. In the single-class case, we found the necessary and sufficient conditions under which the FCFS rule (or any other work-conserving and non-preemptive scheduling policy) is optimal (Theorem 1). In addition, we found the necessary and sufficient conditions under which the FB rule is optimal (Theorem 2). In the-multi class case, where the scheduler can identify the class of each job, we derived the necessary and sufficient conditions under which the MAX-H and MAX-h rules (Definitions 3 and 4, respectively) are optimal (Theorems 3 and 4, respectively). To prove these optimality results, we needed the following technical assumptions: the service time distributions have a right-continuous density function with left limits and the holding cost rates are right-continuous functions of the service time with left limits.

There are a number of directions that could be fruitful to explore in future work. Recently, several "near optimality" results for the Gittins index or related policies have been shown for the constant holding cost setting ($c(x) = 1$), so a natural question is whether such results hold with general holding cost functions. One example is multiserver systems: Scully et al. [20] prove mean delay bounds for the constant-holding-cost Gittins index in the M/G/$k$, which Grosof et al. [10] extend to the case where some jobs occupy multiple servers at once during service. Can we prove analogous multiserver performance bounds for the Gittins index with general holing costs? Coming back to the single-server setting, another question is whether we need the full power of the Gittins index to have good performance. Scully et al. [22] show that a variant of SERPT is a constant-factor approximation for mean delay in the M/G/1. Can we show that (a variant of) MAX-H, which is the natural generalization of SERPT to general holding costs, is a constant-factor approximation for mean holding cost?

Another potential future direction has to do with extending the ideas behind the Gittins index to even more general objective functions. For example, there are many objective functions which demand a *time-varying* holding cost, such as metrics related to deadlines. Yu et al. [29] show how such problems can be viewed through the lens of restless bandits and thus approached using the Whittle index, a generalization of the Gittins index. The Whittle index has been used for other queue scheduling problems [4, 9, 15], including recent work on the age-of-information metric [13, 14, 24]. Unlike the Gittins index, we should not expect the Whittle index to yield optimal policies in

general, but it often yields policies that are in some sense asymptotically optimal. For the Gittins index, we now have a general theory of its optimality in the M/G/1. Can we develop a similarly general theory of the Whittle index's asymptotic optimality?

**Acknowledgements** This work was done in part while Ziv Scully was visiting the Simons Institute for the Theory of Computing; and in part while he was a postdoc at Harvard and MIT, where he was supported by NSF Grant Nos. DMS-2023528 and DMS-2022448. We also thank the Associate Editor and the reviewers of the paper for their useful comments, which helped us improve the paper.

**Funding** Open Access funding provided by Aalto University.



## References

1. Aalto, S.: Minimizing the mean slowdown in a single-server queue. Queueing Syst. **100**, 373–375 (2022)
2. Aalto, S., Ayesta, U., Righter, R.: On the Gittins index in the M/G/1 queue. Queueing Syst. **63**, 437–458 (2009)
3. Aalto, S., Ayesta, U., Righter, R.: Properties of the Gittins index with application to optimal scheduling. Probab. Eng. Inf. Sci. **25**, 269–288 (2011)
4. Ansell, P.S., Glazebrook, K.D., Niño-Mora, J., O'Keeffe, M.: Whittle's index policy for a multi-class queueing system with convex holding costs. Math. Methods Oper. Res. **57**, 21–39 (2003)
5. Bansal, N., Dhamdhere, K., Könemann, J., Sinha, A.: Non-clairvoyant scheduling for minimizing mean slowdown. Algorithmica **40**, 305–318 (2004)
6. Becchetti, L., Leonardi, S.: Non-clairvoyant scheduling to minimize the average flow time on single and parallel machines. In: Proceedings of ACM STOC, pp. 94–103 (2001)
7. Feng, H., Misra, V.: Mixed scheduling disciplines for network flows. ACM Sigmetrics Perform. Evaluat. Rev. **31**(2), 36–39 (2003)
8. Gittins, J.: Multi-armed Bandit Allocation Indices. Wiley, New York (1989)
9. Glazebrook, K.D., Lumley, R.R., Ansell, P.S.: Index heuristics for multiclass M/G/1 systems with nonpreemptive service and convex holding costs. Queueing Syst. **45**, 81–111 (2003)
10. Grosof, I., Scully, Z., Harchol-Balter, M., Scheller-Wolf, A.: Optimal scheduling in the multiserver-job model under heavy traffic. In: Proceedings of the ACM on Measurement and Analysis of Computing Systems **6**, article 51 (2022)
11. Harchol-Balter, M.: Open problems in queueing theory inspired by datacenter computing. Queueing Syst. **97**, 3–37 (2021)
12. Hyytiä, E., Aalto, S., Penttinen, A.: Minimizing slowdown in heterogeneous size-aware dispatching systems. In: Proceedings of ACM Sigmetrics/Performance, pp. 29–40 (2012)
13. Kadota, I., Sinha, A., Uysal-Biyikoglu, E., Singh, R., Modiano, E.: Scheduling policies for minimizing age of information in broadcast wireless networks. IEEE/ACM Trans. Netw. **26**(6), 2637–2650 (2018)
14. Maatouk, A., Kriouile, S., Assad, M., Ephremides, A.: On the optimality of the Whittle's index policy for minimizing the age of information. IEEE Trans. Wirel. Commun. **20**, 1263–1277 (2021)
15. Niño-Mora, J.: Dynamic priority allocation via restless bandit marginal productivity indices. TOP **15**, 161–198 (2007)
16. Nuyens, M., Wierman, A.: The foreground-background queue: A survey. Perform. Eval. **65**, 286–307 (2004)

17. Nuyens, M., Wierman, A., Zwart, B.: Preventing large sojourn times using SMART scheduling. Oper. Res. **56**, 88–101 (2008)
18. Schrage, L.: A proof of the optimality of the shortest remaining processing time discipline. Oper. Res. **16**, 687–690 (1968)
19. Scully, Z.: A new toolbox for scheduling theory. Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA (2022)
20. Scully, Z., Grosof, I., Harchol-Balter, M.: The Gittins policy is nearly optimal in the M/G/k under extremely general conditions. In: Proceedings of the ACM on Measurement and Analysis of Computing Systems **4**, article 43 (2020)
21. Scully, Z., Harchol-Balter, M.: The Gittins policy in the M/G/1 queue. In: Proceedings of WiOpt (2021)
22. Scully, Z., Harchol-Balter, M., Scheller-Wolf, A.: Simple near-optimal scheduling for the M/G/1. In: Proceedings of the ACM on Measurement and Analysis of Computing Systems **4**, article 11 (2020)
23. Smith, D.: A new proof of the optimality of the shortest remaining processing time discipline. Oper. Res. **26**, 197–199 (1978)
24. Tripathi, V., Modiano, E.: A Whittle index approach to minimizing functions of age of information. In: Proceedings of Allerton Conference on Communication, Control, and Computing, pp. 1160–1167 (2019)
25. von Olivier, G.: Kostenminimale Prioritäten in Wartesystemen vom Typ M/G/1 [Cost-minimum priorities in queueing systems of type M/G/1]. Elektronische Rechenanlagen **14**, 262–271 (1972)
26. Wierman, A., Harchol-Balter, M., Osogami, T.: Nearly insensitive bounds for SMART scheduling. In: Proceedings of ACM Sigmetrics, pp. 205–216 (2005)
27. Yang, S., de Veciana, G.: Size-based adaptive bandwidth allocation: optimizing the average QoS for elastic flows. In: Proceedings of IEEE Infocom, pp. 657–666 (2002)
28. Yang, S.J., de Veciana, G.: Enhancing both network and user performance for networks supporting Best Effort traffic. IEEE/ACM Trans. Netw. **12**, 349–360 (2004)
29. Yu, Z., Xu, Y., Tong, L.: Deadline scheduling as restless bandits. IEEE Trans. Autom. Control **63**, 2343–2358 (2018)