Aalto, Samuli

# Minimizing the mean slowdown in a single-server queue

*Published in:*
QUEUEING SYSTEMS

*DOI:*
10.1007/s11134-022-09777-4

Published: 01/04/2022

*Document Version*
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

*Please cite the original version:*
Aalto, S. (2022). Minimizing the mean slowdown in a single-server queue. *QUEUEING SYSTEMS*, *100*(3-4), 373-375. https://doi.org/10.1007/s11134-022-09777-4

**Queueing Systems Vol. 100**
Special Issue: 100 views on queues

# Minimizing the mean slowdown in a single-server queue

**Samuli Aalto**

February 15, 2022

## 1 Introduction

We consider the optimal scheduling problem in a single server queue. Let $S_i$ and $T_i$ denote the service time and the delay (a.k.a. sojourn time or response time) of job $i$. Service times are assumed to be independent and identically distributed with a finite mean $E[S]$. In the open version of the problem, jobs arrive according to a Poisson process at rate $\lambda < 1/E[S]$ so that we have a stable M/G/1 queue. On the other hand, in the closed version, there is a finite number of jobs at time 0 and no further arrivals. The optimal scheduling discipline depends naturally on the objective function but also on the information available to the scheduler. The scheduler is said to be non-anticipating if it knows the arrival times and the attained service of each job in the system, while an anticipating scheduler knows even the remaining service times of the jobs in the system.

If the aim is to minimize the mean delay $E[T]$, the optimal anticipating scheduling policy (for both versions of the problem) is SRPT (Shortest Remaining Processing Time) [10]. 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 than Used in Expectation) distributions, but the worst possible work-conserving scheduling policy if the service times are DHR (Decreasing Hazard Rate). On the other hand, FB (Foreground Background), which is another well-known scheduling policy [9], behaves just opposite: It is optimal when the service time distribution belongs to DHR, but the worst possible among work-conserving policies if service times are NBUE.

Samuli Aalto
Department of Communications and Networking, Aalto University, Finland
E-mail: samuli.aalto@aalto.fi

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 (for both versions of the problem) is the Gittins index policy [7,6,1,2,11], which always chooses the job $i$ with the highest index $G_i(a_i)$ defined by

$$G_i(a_i) = \sup_{b>a_i} \frac{P\{S_i \le b \mid S_i > a_i\}}{E[\min\{S_i,b\} - a_i \mid S_i > a_i]}, \tag{1}$$

where $a_i$ denotes the (known) current attained service of job $i$. Let us use abbreviation GIMD for this policy. As said above, GIMD coincides with FCFS (or any other non-preemptive and work-conserving scheduling policy) if the service times are NBUE, and it coincides with FB when the service times are DHR [1,2].

## 2 Problem statement

Instead of mean delay, we focus in this letter on the optimal scheduling problem where the objective function is the *mean slowdown* $E[\frac{T}{S}]$, i.e., the expectation of the ratio between the delay and the service time of a job. The optimal anticipating scheduling policy minimizing the mean slowdown (for both versions of the problem) is known to be SPTP (Shortest Processing Time Product) [13,8], which is called RS in [12]. But the optimal non-anticipating scheduler with respect to the mean slowdown has long been an open problem [4,3].

The current state of the art is as follows: Let $h(x)$ denote the hazard rate of the service time distribution. Feng and Misra [5] proved that FB is the optimal non-anticipating policy when the modified hazard rate $\tilde{h}(x) = \frac{h(x)}{x}$ of the service time distribution is decreasing. Such distributions include clearly all DHR distributions, for which $h(x)$ is required to be decreasing. Recently, Scully and Harchol-Balter [11] showed that the Gittins index approach is applicable even when minimizing the mean slowdown: The optimal non-anticipating policy is the index policy that always chooses the job $i$ with the highest index $\tilde{G}_i(a_i)$ defined by

$$\tilde{G}_i(a_i) = \sup_{b>a_i} \frac{E[\frac{1_{\{S_i \le b\}}}{S_i} \mid S_i > a_i]}{E[\min\{S_i,b\} - a_i \mid S_i > a_i]}, \tag{2}$$

where $1_A$ refers to the indicator function of event $A$.

Now the challenge is to characterize the optimal non-anticipating policy more explicitly. For example, which properties the service time distribution should have in order that the ordinary FCFS policy is optimal minimizing the mean slowdown among the non-anticipating scheduling policies? This family of distributions certainly includes the deterministic distributions since, in this case, FCFS coincides with SPTP. But is it a strict subset of NBUE distributions as could be expected? A natural starting point for answering these questions is given by Equation (2).

## 3 Discussion

When minimizing the mean delay among the non-anticipating policies, the exponential service time distribution, for which the hazard rate $h(x)$ is constant, is a kind of watershed: It belongs both to NBUE and DHR so that FCFS and FB are equally good in this case. In fact, all work-conserving policies which are non-anticipating are equally good and optimal for the exponential distribution.

Interestingly, the Weibull service time distribution with shape parameter $\alpha = 2$, for which the modified hazard rate $\tilde{h}(x)$ is constant, seems to behave similarly when the target is to minimize the mean slowdown among the non-anticipating policies: It is numerically easy to demonstrate that FCFS and FB are equally good in this case. Moreover, when $\alpha \leq 2$, the Weibull distribution has the required property that implies the optimality of FB among the non-anticipating policies, i.e., $\tilde{h}(x)$ is decreasing. So, our guess is that FCFS is the optimal non-anticipating policy when $\alpha \geq 2$. This is in line with the following claim, which is a kind of counterpart of Theorem 4 in [1].

*Conjecture 1* FCFS minimizes the mean slowdown among the non-anticipating policies if and only if the service time distribution has the following property:

$$\tilde{H}(x) \geq \tilde{H}(0), \quad \text{for all } x \geq 0, \tag{3}$$

where function $\tilde{H}(x)$ is defined by

$$\tilde{H}(x) = \frac{\int_x^\infty \tilde{h}(y) P\{S > y\}\,\mathrm{d}y}{\int_x^\infty P\{S > y\}\,\mathrm{d}y}. \tag{4}$$

Proving this claim is an open problem that we would like to solve.

## References


1. S. Aalto, U. Ayesta, and R. Righter. On the Gittins index in the M/G/1 queue. *Queueing Systems*, 63:437–458, 2009.
2. S. Aalto, U. Ayesta, and R. Righter. Properties of the Gittins index with application to optimal scheduling. *Probability in the Engineering and Informational Sciences*, 25:269–288, 2011.
3. N. Bansal, K. Dhamdhere, J. Könemann, and A. Sinha. Non-clairvoyant scheduling for minimizing mean slowdown. *Algorithmica*, 40:305–318, 2004.
4. L. Becchetti and S. Leonardi. Non-clairvoyant scheduling to minimize the average flow time on single and parallel machines. In *Proc. of ACM STOC*, pages 94–103, 2001.
5. H. Feng and V. Misra. Mixed scheduling disciplines for network flows. *ACM Sigmetrics Performance Evaluation Review*, 31(2):36–39, 2003.
6. J. Gittins, K. Glazebrook, and R. Weber. *Multi-armed Bandit Allocation Indices*. Wiley, second edition, 2011.
7. J. C. Gittins. *Multi-armed Bandit Allocation Indices*. Wiley, 1989.
8. E. Hyytiä, S. Aalto, and A. Penttinen. Minimizing slowdown in heterogeneous size-aware dispatching systems. In *Proc. of ACM Sigmetrics/Performance*, pages 29–40, 2012.
9. M. Nuyens and A. Wierman. The foreground–background queue: A survey. *Performance Evaluation*, 65:286–307, 2004.
10. L. E. Schrage. A proof of the optimality of the shortest remaining processing time discipline. *Operations Research*, 16:687–690, 1968.
11. Z. Scully and M. Harchol-Balter. The Gittins policy in the M/G/1 queue. In *Proc. of WiOpt*, 2021.
12. A. Wierman, M. Harchol-Balter, and T. Osogami. Nearly insensitive bounds for SMART scheduling. In *Proc. of ACM Sigmetrics*, pages 205–216, 2005.
13. S. J. Yang and G. de Veciana. Enhancing both network and user performance for networks supporting Best Effort traffic. *IEEE/ACM Transactions on Networking*, 12:349–360, 2004.