Mean Delay Optimization for the M/G/1 Queue with Pareto Type Service Times

Samuli Aalto
TKK Helsinki University of Technology, Finland
samuli.aalto@tkk.fi

Urtzi Ayesta
LAAS-CNRS, France
urtzi@laas.fr

Categories and Subject Descriptors

C.4 [**Performance of Systems**]: Performance Attributes; F.2.2 [**Nonnumerical Algorithms and Problems**]: Sequencing and Scheduling; G.3 [**Probability and Statistics**]: Queueing Theory

General Terms

Performance

Keywords

Mean delay, M/G/1, scheduling, Pareto distribution, Gittins index

1. INTRODUCTION

We consider the optimal scheduling problem in the M/G/1 queue with the objective to minimize the mean delay (*i.e.*, sojourn time). We assume that jobs are served according to a work conserving and nonanticipating (scheduling) discipline π. A discipline is work conserving if it does not idle when there are jobs waiting, and nonanticipating if the remaining service times of jobs are unknown for the scheduler. Let Π denote the family of such disciplines. In particular we note that the Shortest-Remaining-Processing-Time (SRPT) discipline does not belong to Π.

It is known that for Decreasing Hazard Rate (DHR) service times, the Foreground-Background (FB) discipline is optimal in Π, whereas the First-Come-First-Served (FCFS) discipline minimizes the mean delay for the service time distributions that belong to the New Better than Used in Expectation (NBUE) class [6, 4, 5, 1]. FB is an age-based discipline giving full priority to the job with least amount of attained service, see [3, Section 4.6].

Our goal is to shed light on the impact of designing an optimal scheduling discipline when the hazard rate of the service time distribution is nonmonotone. In particular, we concentrate on distributions for which the hazard rate is constant (zero or positive) for small values while decreasing for larger values. An example is the Pareto distribution,

$$P\{S > x\} = \left(\frac{k}{x}\right)^{\alpha}, \quad x \geq k > 0, \tag{1}$$

which has been used to model, for example, flow sizes in the Internet.

SIGMETRICS'07, June 12–16, 2007, San Diego, California, USA.
ACM 978-1-59593-639-4/07/0006.

In this extended abstract we give the following result for these service time distributions. (Due to the strict page limit, all the proofs are left for a later complete paper.) The optimal discipline is a combination of FCFS and FB disciplines. More precisely, it is an age-based discipline which gives full priority to the jobs with attained service less than some fixed threshold θ^*. These priority jobs are served in the FCFS manner. If there are no jobs with attained service less than θ^*, the job with least amount of attained service will be served. In addition, we show that the optimal threshold θ^* depends on the service time distribution, but not on the arrival rate. Thus, for a given service time distribution, the optimal discipline remains the same independent of the load of the system. We use notation $\mathrm{FCFS} + \mathrm{FB}(\theta^*)$ for this discipline. It belongs to the class of Multilevel Processor-Sharing (MLPS) disciplines, see [3, Section 4.7]. The problem is solved by applying the so called Gittins index.

2. SERVICE TIME DISTRIBUTIONS

Consider an M/G/1 queue with arrival rate λ, mean service time $E[S]$, and load $\rho = \lambda E[S] < 1$. Let $F(x) = P\{S \leq x\}$, $x \geq 0$, denote the cumulative distribution of the service time of any job. Define $\overline{F}(x) = 1 - F(x)$, and assume that $\overline{F}(x) > 0$ for all x. In addition, we only consider the distributions with a density function $f(x)$ that is right-continuous with left-limits. The hazard rate $h(x)$ is defined by

$$h(x) = \frac{f(x)}{\overline{F}(x)} = \frac{f(x)}{\int_0^\infty f(x+y)\,dy}.$$

A service time distribution belongs to the class DHR if $h(x)$ is decreasing for all x, *i.e.*, $h(x) \geq h(y)$ whenever $x \leq y$.

In this paper we introduce a new class of service times, called CDHR(k) (first Constant and then Decreasing Hazard Rate with threshold k). Let $k > 0$. A service time distribution belongs to the class CDHR(k) if

A1: $h(x)$ is constant for all $x < k$,

A2: $h(x)$ is decreasing for all $x \geq k$.

In fact, we focus on those CDHR(k) service times for which the following additional assumption holds:

A3: $h(0) < h(k)$.

Under this additional assumption, the service time distribution does not belong to the class DHR.

3. OPTIMAL DISCIPLINE

As mentioned, we utilize the approach developed by Gittins [2]. In our M/G/1 context, the *Gittins index* $G(a)$ is defined for any job as a function of its attained service a by

$$G(a) = \sup_{\Delta \geq 0} J(a, \Delta),$$

where

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

In addition, for any $a \geq 0$, let

$$\Delta^*(a) = \inf\{\Delta \geq 0 \mid J(a,\Delta) = G(a)\}.$$

By definition, $G(a) = J(a, \Delta^*(a))$ for all a. Note further that $J(a,0) = h(a)$.

Let $\pi^* \in \Pi$ be such that service is always given to the job with the highest Gittins index. We call this discipline the *Gittins discipline*. It is known that the Gittins discipline π^* is optimal with respect to the mean delay for an M/G/1 queue, see [2, Theorem 3.28], [7, Theorem 4.7].

In our main result given below, by studying properties of the Gittins index, we characterize the optimal scheduling discipline for the service time distributions of type CDHR(k).

Theorem 1 *Assume that the service time distribution belongs to* CDHR(k).

(i) If A3 is not satisfied, then FB *is optimal.*

(ii) If A3 is satisfied, then FCFS + FB(θ^*) *is optimal, where* $\theta^* = \Delta^*(0) > k$.

Note that the optimal threshold $\theta^* = \Delta^*(0)$ depends only on the service time distribution.

4. NUMERICAL EXAMPLE

As an illustration of Theorem 1(ii), we consider the Pareto distribution (1) with parameters $k = 1$ and $\alpha = 2$.

The function $G(a)$ is plotted in the upper panel of Figure 1. In this case $\Delta^*(0) = 2 + \sqrt{3} = 3.732$, and

$$G(0) = G(\Delta^*(0)) = h(\Delta^*(0)) = \frac{2}{2+\sqrt{3}} = 0.536.$$

Note further that $G(a)$ is decreasing for all $a \geq 1$.

Let us now determine the structure of the Gittins discipline based on $G(a)$. Assume that the queue is empty and that a new job arrives. Obviously this job will start being served immediately. From Figure 1, we see that

$$G(a) > G(0), \quad \text{for all } 0 < a < \Delta^*(0).$$

Thus, independently of the arrival process, the service of this job will not be interrupted until it gets $\Delta^*(0)$ units of service. Once this happens, if no new job has previously arrived, the original job will continue being served. But if there is a new job waiting, it will replace the original one in service. Finally, consider the situation that all jobs in the system have obtained more units of service than $\Delta^*(0)$. From Figure 1 we observe that the Gittins index decreases as the attained service increases. Thus, among these jobs, the full priority is given to the job with least amount of service. In other words, these jobs will be served according to the FB discipline. Thus, we conclude that, in this example, the Gittins discipline is FCFS + FB($\Delta^*(0)$), which is in line with Theorem 1(ii).

In the lower panel of Figure 1 we have plotted the mean delay ratio $\overline{T}^{\mathrm{FCFS+FB}(\theta^*)}/\overline{T}^{\mathrm{FB}}$ for different loads ρ. This illustrates the performance gain when the FB discipline is replaced with the optimal one. Note that, even though the optimal threshold θ^* does not depend on the load, the gain obtained by the optimal discipline does depend. The maximum gain of 18% is achieved with load $\rho = 0.8$. It is also interesting to know that, with this Pareto service time distribution, $\overline{T}^{\mathrm{FCFS}} = \infty$ for any load $\rho > 0$.

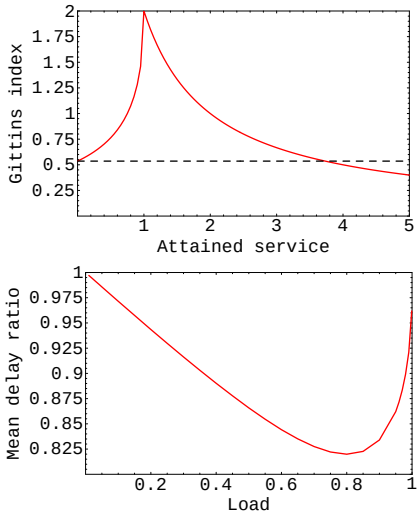

Figure 1: Pareto distribution with parameters $k = 1$, $\alpha = 2$. *Upper panel*: Gittins index $G(a)$ as a function of attained service a with the horizontal line equal to $G(0)$. *Lower panel*: Mean delay ratio $\overline{T}^{\mathrm{FCFS+FB}(\theta^*)}/\overline{T}^{\mathrm{FB}}$ as a function of load ρ.

5. REFERENCES

[1] S. Aalto and U. Ayesta. On the nonoptimality of the foreground-background discipline for IMRL service times. *Journal of Applied Probability*, 43:523–534, 2006.

[2] J. Gittins. *Multi-armed Bandit Allocation Indices.* Wiley, Chichester, 1989.

[3] L. Kleinrock. *Queueing Systems, Volume II: Computer Applications.* Wiley, New York, 1976.

[4] R. Righter and J. 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.

[5] R. Righter, J. Shanthikumar, and G. Yamazaki. On extremal service disciplines in single-stage queueing system. *Journal of Applied Probability*, 27:409–416, 1990.

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

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