Fluid Limits for Shortest Job First with Aging

Yonatan Shadmi∗
shdami@campus.technion.ac.il

Abstract

We investigate fluid scaling of single server queueing systems under the shortest job first with aging (SJFA) scheduling policy. We use the measure-valued Skorokhod map to characterize the fluid limit for SJFA queues with a general aging rule and establish convergence results to the fluid limit. We treat in detail examples of linear and exponential aging.

AMS subject classifications: 60K25, 60G57, 68M20

Keywords: measure-valued Skorokhod map, measure-valued processes, fluid limits, shortest job first, aging

1 Introduction

It is well known that prioritizing jobs by their size is optimal in the sense of minimizing the number of jobs in system at any point in time ([16], [19]). This could be important in systems with limited buffer size. As the average number of jobs in the queue and the average waiting time are related by Little’s law, minimizing the number of jobs in the buffer is equivalent to minimizing the average waiting time ([17], [12]). However, these kind of policies may cause starvation and deny resource access from large sized jobs. To avoid this undesired phenomenon, more dynamic approaches have been suggested, that take into consideration the sojourn time as well as the job size. In particular, the priority of a job is dictated initially by its size, and is updated as time elapses. This is done in practice ([2], [8]) and is common in the literature of computer science ([18], [4], [6]), and is known as aging. In this case, even large sized jobs will eventually have higher priority than smaller jobs with later arrival time. Scaling limits of such policies have not been treated before, even though both fluid and diffusion limits of SJF were extensively studied. This paper analyzes and characterizes fluid limits of SJFA for the first time; the main tool is the measure valued Skorokhod map from [1].

Prominent examples of aging rules are linear ([18]), exponential ([2], [8]), highest response ratio next (HRRN) ([4]) and the fair sojourn protocol (FSP) ([6]). Among the many aging rules that our analysis covers, the linear is the simplest one and most of the examples in this paper consider this rule. Though somewhat more cumbersome, the exponential rule is used in practice in UNIX ([2], [8]), which motivates us to study it in details as well. HRRN and FSP are essentially different and are explained later. The exponential aging rule in UNIX is implemented as follows (as described in [8], where other details and examples can be found). One employs the rule

\[ \text{pri} = \text{cpu\_use} + \text{base} + \text{nice} \]

∗Viterbi Faculty of Electrical Engineering, Technion – Israel Institute of Technology, Haifa 32000, Israel
where: \textit{pri} stands for the priority value of a job at the current time, and \textbf{lower priority value} corresponds to \textbf{higher priority}; \textit{cpu\_use} stands for recent CPU usage and is the term responsible of the aging. This value increases linearly in time only when the process is running (discretely on every clock interrupt), and is divided by two every second; \textit{base} is the base priority for user processes; \textit{nice} is an optional additional term, if a user is willing to reduce its own priority and be nice to others. In our analysis, we ignore the terms \textit{base} and \textit{nice}, as well as the linear increase of \textit{cpu\_use} which is irrelevant in non-preemptive policies because it happens only while the process is running and not while it waits for service. Note that the updating rule of \textit{cpu\_use} is a discrete time version of exponential aging due to the division by two every second of the priority value.

The linear and exponential aging rules are the simplest among the above examples, as they belong to a class of deterministic aging rules with non-intersecting aging trajectories. Here, by aging trajectories we mean trajectories on the time-priority plane, describing how a job priority is updated with time. This means that jobs do not exchange places in the priority ordering as time progresses. A generalization of these two rules is the class of aging rules that can be described through an ordinary differential equation. In this paper, we consider this class of aging rules.

Job size based priority systems is an active research topic in the operations research literature since the work of Schrage and Miller [17], [16] (see also [19] for a simpler proof of optimaility). In particular, some research effort has been devoted to scaling limits of such policies. Fluid limits of the preemptive version, the shortest remaining processing time policy, were shown to exist, be unique, and were characterized in [7]. To describe the evolution of the system, a measure valued process was used, one that counts the number of jobs in the queue with job size in some set.

Diffusion scaling limits for \textit{shortest remaining processing time} (SRPT), the preemptive version of the SJF, appeared shortly after in [9] under heavy traffic assumption, again using measure valued processes as state descriptors. The measure counting the number of jobs in the queue with size in some set is shown to converge to a Dirac measure, concentrated on the rightmost point of the support of the job size distribution. The magnitude of the limiting Dirac delta is a reflected Brownian motion divided by the largest possible job size (zero if the job size is unbounded). A followup of this work is [15], where non-standard diffusion limit was taken to show a generalized state space collapse - the queue length and the workload converge to the same process.

In [1], Atar et al. prove convergence to the fluid limit of several priority queues, including SRPT and SJF. They introduce a measure valued Skorokhod map, and use it to map the arrival process and the server effort process to a measure valued state descriptor. This technique allows considering also time-varying arrival and service distributions.

Most of the work on such policies studies SRPT, however, Gromoll and Keutel argue in [10] that, at least on fluid scale, the performance of SRPT and SJF is the same (they converge to the same fluid limit). See also [14], [11] and [3] for more comparisons between SJF and SRPT.

Though the literature on SJF and SRPT and their asymptotic behaviour is rich, there is no mentioning of aging in this context. With aging, the system state changes not only when jobs arrive and depart, but also vary with time, as the priority values undergo aging. These dynamics prevent a direct use of the measure valued Skorokhod map as in [1].

Our objective here is to expand the existing asymptotic analysis to include SJFA with deterministic aging and non-overlapping trajectories. The idea is to first identify a coordinate transformation that transforms the system into an equivalent system, and then use the measure valued Skorokhod map with measures defined on the new coordinate system. As an easy example, for linear aging, let \( S_i \) and \( \tau_i \) be the \( i \)-th job’s size and arrival time, resp. Whereas SJFA prioritizes according to \( S_i - (t - \tau_i) \), this is equivalent to prioritizing according to \( S_i + \tau_i \), which is independent of time. For a general aging rule, this transformation requires further effort. With these means we are able to prove convergence to the fluid limit and characterize the limit.
Indeed, linear and exponential aging rules belong to the class of aging rules that are described by an ordinary differential equation, as mentioned above. However, we have also mentioned the HRRN and the FSP, which are important policies that avoid starvation, but do not fall in the category we cover in this work. In the HRRN policy, all jobs are assigned on arrival the same priority value, but this value increases with a rate which is the inverse of the job size. Thus, the aging trajectories do intersect, making it impossible to use our method. In the FSP, there is a virtual processor sharing server and jobs enter service in the order of their virtual departure process. In this setting, the aging trajectories are stochastic; they depend on the system state. This is a much more interesting and complicated situation which, we assume, requires much more sophisticated tools to analyze. This motivates us to seek further tools to cover other settings.

The paper is organized as follows. In §2 we introduce the queueing model and state the main theorem. The proof is given in §3. §4 provides four examples with different arrival distributions and aging rules.

Notation

For any Polish space $S$, let $\mathbb{D}_S$ denote the space of càdlàg functions from $\mathbb{R}_+$ to $S$, equipped with the Skorokhod J1 metric. Let $\mathbb{C}$ denote its subspace of continuous functions. If $S = \mathbb{R}_+$ we simply write $\mathbb{D}$ and $\mathbb{C}$. Let $\mathbb{D}^\uparrow$ denote the subspace of non-decreasing elements of $\mathbb{D}$. Let $\mathcal{M}$ denote the space of finite Borel measures on $\mathbb{R}$ equipped with the Lévy metric, which induces the topology of weak convergence, and let $\mathcal{M}_0$ denote its subspace of atomless Borel measures on $\mathbb{R}$. It is well known that $\mathcal{M}$ is a Polish space, and so is $\mathbb{D}^\uparrow \mathcal{M}$. Let $\mathbb{D}^\uparrow \mathcal{M}$ denote the subspace of non-decreasing elements of $\mathbb{D} \mathcal{M}$, where an element $\nu$ of $\mathbb{D} \mathcal{M}$ is said to be non-decreasing if $t \mapsto \int f d\nu_t$ is non-decreasing for every non-negative continuous and bounded function $f$.

2 SJFA

2.1 Queueing model

Consider a sequence of single server queues, indexed by $N \in \mathbb{N}$, operating under the SJFA with the same aging policy. Whenever the server is available, it admits the highest priority job into service, where the priority of each job is as follows. The system keeps track of each job’s priority value, where lower value means higher priority. Each job is assigned an initial value on arrival which is the job size. The value, $g$, varies smoothly with time according to an aging rule. To formalize this, we order the jobs according to their arrival, and denote by $\tau^N_i$ and $S^N_i$ the arrival time and job size, resp., of the $i$-th job to enter the $N$-th system. The fact that the initial priority value is the job size is expressed by

$$g(\tau^N_i) = S^N_i. \tag{1}$$

The aging rule is modeled by the differential equation

$$\frac{dg}{ds} = f(g, s), \quad s \geq \tau^N_i. \tag{2}$$

In the context of aging, $f$ is assumed to be negative, as the priority of a job increases as times progresses (i.e. $g$ is decreasing). The linear and exponential aging rules correspond to the choices $f(g, s) = -c$ and $f(g, s) = -\lambda g$, resp., for constants $c, \lambda > 0$.

In fact, under certain conditions, Equation (2) can be solved backwards as well to obtain a solution for all $s$ in $\mathbb{R}_+$. This fact is used in this work and, considering Equations (1) and (2) jointly, the aging...
Figure 1: Illustration of aging trajectories in the cases of linear aging (left hand side) and exponential aging (right hand side). The arrows illustrate the flow of a priority value on the time-priority plane. The lines describe the trajectories that the priority values follow.

trajectory of the i-th job is described by

\[ \begin{align*}
\frac{d}{ds} g &= f(g, s), \quad s \in \mathbb{R}_+ , \\
g(\tau_i^N) &= S_i^N .
\end{align*} \tag{3} \]

We require that \( f \) is Lipschitz in the first argument. Then this problem admits a unique solution on \( \mathbb{R}_+ \) (\cite[Theorem 6]{Theorem 6}), that is, a unique aging trajectory \( g: \mathbb{R}_+ \to \mathbb{R} \) satisfying \( g(\tau_i^N) = S_i^N \). This way, any point \((x, t) \in \mathbb{R}_+^2\) has exactly one trajectory passing through it, denoted \( g(x, t) \). In particular, \( g_i^N := g(S_i^N, \tau_i^N) \) is the unique function \( h: \mathbb{R}_+ \to \mathbb{R} \) such that \( \frac{d}{ds} h = f(h(s), s), s \in \mathbb{R}_+ , \) and \( h(\tau_i^N) = S_i^N \). The trajectories of the linear and exponential aging rules are, resp., \( g(x, t)(s) = x - c(s - t) \) and \( g(x, t)(s) = x \exp(-\lambda (s - t)) \) (see Figure 1). We denote by \( \theta_i^N \) the departure time of the \( i \)-th job, but stress that in our model the priority value continues to change according to the aging rule even after the job has departed from the system.

We introduce the following processes, underlying the queueing model in this section.

\( \alpha^N : \quad \alpha_i^N (-\infty, x] \) is the work to have arrived at the queue by time \( t \) with priority value in \( (-\infty, x] \) at time \( t \),

\( \mu^N : \quad \mu_i^N (t) \) is the total work the server can complete by time \( t \), if it is never idle,

\( \xi^N : \quad \xi_i^N (-\infty, x] \) is the work in the queue at time \( t \) with priority value in \( (-\infty, x] \) at time \( t \),

\( \beta^N : \quad \beta_i^N (-\infty, x] \) is the work completed by the server by time \( t \), with priority value in \( (-\infty, x] \) at time \( t \).

The sample paths of \( \alpha^N, \xi^N \) and \( \beta^N \) are in \( D_M \), and \( \mu^N \) has sample paths in \( D^\uparrow \). Mathematically, \( \alpha^N, \beta^N \) and \( \xi^N \) are given by

\[ \begin{align*}
\alpha_i^N (-\infty, x] &= \sum_{i=1}^{\infty} S_i^N \mathbb{1}_{\{\tau_i^N \leq t\}} \mathbb{1}_{\{g_i^N(t) \leq x\}} \\
\beta_i^N (-\infty, x] &= \sum_{i=1}^{\infty} S_i^N \mathbb{1}_{\{\theta_i^N \leq t\}} \mathbb{1}_{\{g_i^N(t) \leq x\}} ,
\end{align*} \tag{5} \]

\[ \begin{align*}
\xi_i^N (-\infty, x] &= \alpha_i^N (-\infty, x] - \beta_i^N (-\infty, x] = \sum_{i=1}^{\infty} S_i^N \mathbb{1}_{\{\tau_i^N \leq \theta_i^N \}} \mathbb{1}_{\{g_i^N(t) \leq x\}}.
\end{align*} \tag{6} \]
The initial state of the buffer is already contained in \( \alpha_N^N \). We assume that \( \mu \) is given by \( \mu(t) = \int_0^t m^N(s) \, ds \), where \( m^N(s) \) is the service rate at time \( s \) satisfying \( m^N(s) \geq m_0 > 0 \) for all \( s \). Define the busyness process \( B^N(t) = \mathbb{1}\{\text{The server is busy at time } t\} \); then the work done by the server by time \( t \) is \( T^N(t) = \int_0^t m^N(s) \, B^N(s) \, ds \), and \( v^N = \mu^N - T^N \) is the lost work due to idleness. Let the process \( J(t) \) denote the residual work of the job in service at time \( t \). It then holds that
\[
\beta^N_t [0, \infty) = T^N(t) + J^N(t) - J^N(0).
\]

At any time when the server is ready to serve a new job, the scheduler admits to service the job with the highest priority, i.e. with the smallest priority value. If the queue is empty and the server is idle, the next job to arrive will be immediately admitted.

### 2.2 Main result

We provide sufficient conditions for \( (\bar{\xi}^N, \bar{\beta}^N, \bar{\iota}^N) \) to weakly converge as a result, and characterize this weak limit by a formula.

**Assumption 1.** The function \( f \) in (3) is continuous as a function of two variables, and is globally Lipschitz in the first argument on \( \mathbb{R}^2_+ \), i.e. satisfies
\[
|f(x_1, t) - f(x_2, t)| \leq L |x_1 - x_2|
\]
for some \( L > 0 \) for all \( (x_1, t) \) and \( (x_2, t) \) in \( \mathbb{R} \times \mathbb{R}_+ \).

Assumption 1 guaranties the existence of a unique path, \( g(s) \), that solves the ODE (5, Theorem 6)]
\[
\begin{align*}
\frac{d}{ds} g &= f(g, s), \quad s \in \mathbb{R}_+, \\
g(t) &= x,
\end{align*}
\]
for any given \( x \) and \( t \). The uniqueness of the path is on \( \mathbb{R} \times \mathbb{R}_+ \) in the sense that there is only one solution passing through any point \((x, t) \in \mathbb{R} \times \mathbb{R}_+ \). As a consequence, two solution paths are either the same path, or do not intersect.

**Theorem 1.** Let the aging rule be as in (3). Suppose that Assumption 1 holds and that \( (\bar{\alpha}^N, \bar{\mu}^N) \Rightarrow (\alpha, \mu) \in C^1_{M_0} \times C^\uparrow \). Denote \( \Xi(t, x) := \alpha_t \left( -\infty, g_{(x,0)}(t) \right) \). Then \( (\bar{\xi}^N, \bar{\beta}^N, \bar{\iota}^N) \Rightarrow (\xi, \beta, \iota) \), where for all \( x \) and \( t \) > 0
\[
\beta_t(x, \infty) + \iota(t) = -\inf_{s \in [0,t]} \{ (\Xi(s, g_{(x,0)}(0)) - \mu(s)) \wedge 0 \}, \tag{8}
\]
\[
\xi_t(-\infty, x) = \Xi(t, g_{(x,0)}(0)) - \mu(t) - \inf_{s \in [0,t]} \{ (\Xi(s, g_{(x,0)}(0)) - \mu(s)) \wedge 0 \}. \tag{9}
\]

Above, (9) characterizes the measure valued process \( \xi \), and (8) characterizes both \( \iota \) and \( \beta \) by taking the limit \( x \to \infty \) and then subtracting \( \iota(t) \).

In the proof given in the next section, the characterization of the limit is expressed by other means. The somewhat complicated notation above is used in order to avoid introducing a transformation of coordinates.

### 3 Proof

The fluid model is the unique triplet \( (\xi', \beta', \iota) \in D_M \times D_M^\uparrow \times D^\uparrow \) satisfying for given data \( (\alpha', \mu) \in D_M^\uparrow \times D^\uparrow \) the following relations.
1. \( \xi' [0, x] = \alpha' [0, x] - \mu + \beta' (x, \infty) + \iota, \)
2. \( \xi' [0, x] = 0 \, d\beta' (x, \infty) - a.s., \)
3. \( \xi' [0, x] = 0 \, d\iota - a.s., \)
4. \( \beta' [0, \infty) + \iota = \mu. \)

This was introduced as the measure-valued Skorokhod problem (MVSP) in [1], followed by an existence and uniqueness theorem ([1 Proposition 2.8]).

Recall that the sample paths of \( \alpha^N \) and \( \beta^N \) belong to \( D_M \) but not to \( D_M^\uparrow \), so it is not yet clear how the MVSP is relevant to our model.

The MVSP defines a map \((\xi', \beta', \iota) = \Theta (\alpha', \mu), \) referred to as the measure-valued Skorokhod map (MVSM). Namely, for each \( \psi \in \mathbb{R} \times \mathbb{R}^+ \), the MVSP is defined by

\[
\psi \mapsto \bar{\mu}^N_{\psi(\alpha'), \psi(\beta')} \in D_M^\uparrow, \text{ where } \bar{\mu}^N_{\psi(\alpha'), \psi(\beta')} = (\Gamma_1(\psi), \Gamma_2(\psi)),
\]

is the Skorokhod map

\[
\Gamma_1(\psi)(t) = \psi(t) - \inf_{s \in [0,t]} (\psi(s) \land 0), \quad \Gamma_2(\psi)(t) = - \inf_{s \in [0,t]} (\psi(s) \land 0).
\]

Note that (8) and (9) have this form exactly with \( \psi = \Xi (\cdot, g(x,t) (0)) - \mu. \)

**Proof of Theorem 2.** The proof opens with the definition of a coordinate transformation, \( \varphi : \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R} \times \mathbb{R}^+ \), that maps a point \( (x, t) \in \mathbb{R} \times \mathbb{R}^+ \) to a point \( (x', t') = \varphi (x, t) \in \mathbb{R} \times \mathbb{R}^+ \), such that \( \alpha' \) and \( \beta' \) do belong to \( D_M^\uparrow \), where \( \alpha' \) and \( \beta' \) are the weak limits of the measures defined by

\[
\alpha'_{\psi(x,t)} [0, x'] (x, t) = \alpha^N_{\psi(x,t)} (-\infty, x), \quad \beta'_{\psi(x,t)} [0, x'] (x, t) = \beta^N_{\psi(x,t)} (-\infty, x).
\]

The coordinate transformation \( (x', t') = \varphi (x, t) \) is based on the uniqueness of the path \( g(x,t) \) due to Assumption 1 for any point \( (x, t) \), we assign the point \( (x', t') \), given by

\[
\begin{cases}
  t' = t, \\
  x' = g(x,t) (0).
\end{cases}
\]

The purpose of this transformation is to transform the aging trajectories in such a way that the priorities do not vary with time in the new system. The inverse transformation is given by

\[
\begin{cases}
  t = t', \\
  x = g(x',0) (t').
\end{cases}
\]

Note that \( x' \geq 0 \) because \( g(t) \) is assumed to be decreasing.

It will be convenient to define a function \( F : D_M \to D_M \) such that \( \nu' = F (\nu) \) is defined by

\[
\nu'_{\psi(x,t)} [0, x'] = \nu_{\psi(\alpha', \beta')} (-\infty, x (x', t')).
\]

By the inverse transformation, the relation \( \nu = F^{-1} (\nu') \) corresponds to the relation

\[
\nu_{\psi(-\infty, x)} (x, t) = \nu_{\psi(x', t')} [0, x'] (x, t).
\]

Note that \( x' \mapsto \nu'_{\psi(x,t)} [0, x'] \) is right-continuous for all \( t > 0 \) because

\[
|\nu'_{\psi(x,t)} [0, x' + \varepsilon] - \nu'_{\psi(x,t)} [0, x']| = \nu_{\psi(x', t)} (x (x', t), x (x' + \varepsilon, t)).
\]
and, by Theorem 6 in [5],

\[ |x(x' + \epsilon, t) - x(x', t)| = |g(x_{x' + \epsilon, t}) (t) - g(x'_{x', t}) (t)| \leq \epsilon e^{|t|} \to 0. \]

In addition, it is monotone as \( x' \to x \to \nu_t (-\infty, x) \) is. We will show later that \( \alpha' \) has indeed sample paths in \( \mathbb{D}^\uparrow_M \), and thus the fluid model can be defined through \( (\xi', \beta', t) = \Theta(\alpha', \mu) \), and \( \xi = F^{-1} (\xi') \), \( \beta = F^{-1} (\beta') \).

Recall the model description from \([2.1]\). Note that under Assumption \([\mathcal{A}]\) the following conditions are equivalent:

\[ g_i^N (t) \leq x = g(x,t) (t) \iff S_i^N := g_i^N (0) \leq g(x,t) (0) = x'. \]

With this, by Equations \([5]-[7]\), \( \alpha^N, \beta^N \) and \( \xi^N \) take the form

\[ \alpha_t^N (-\infty, x) = \sum_{i=1}^\infty S_i^N \mathbb{1}_{\{t_i^N \leq t\}} \mathbb{1}_{\{s_i^N \leq x\}} = \alpha'_{t'} [0, x'], \]

\[ \beta_t^N (-\infty, x) = \sum_{i=1}^\infty S_i^N \mathbb{1}_{\{t_i^N \leq t\}} \mathbb{1}_{\{s_i^N \leq x\}} = \beta'_{t'} [0, x'], \]

\[ \xi_t^N (-\infty, x) = \sum_{i=1}^\infty S_i^N \mathbb{1}_{\{t_i^N \leq t < \theta_i^N \}} \mathbb{1}_{\{s_i^N \leq x\}} = \xi'_{t'} [0, x']. \]

Now, \( t' \to \beta'_{t'} (x', \infty) \) is monotone and \( \alpha'_{t'} \in \mathbb{D}^\uparrow_M \).

Under Assumption \([\mathcal{A}]\) the priority condition is equivalent to the condition

\[ \begin{cases} 
\theta_i^N \geq \tau_j^N \\
\theta_j^N \geq \tau_j^N \\
g_i^N (0) < g_j^N (0) 
\end{cases} \implies \theta_i^N < \theta_j^N. \]

This implies

\[ \int \xi_i^N [0, x'] d\beta_t^N (x', \infty) = \sum_{j=1}^\infty \xi_i^N_{\theta_{j}^N} [0, x'] \left( \beta_{\theta_{j}^N} (x', \infty) - \beta_{\theta_j^N} (x', \infty) \right) \]

\[ = \sum_{j=1}^\infty \sum_{i=1}^\infty S_i^N \mathbb{1}_{\{t_i^N \leq \theta_j^N < \theta_i^N \}} \mathbb{1}_{\{s_i^N \leq x\}} S_j^N \mathbb{1}_{\{g_j^N (0) > x\}} \]

\[ = 0. \]

Moreover, the non-idling property implies

\[ \xi_t^N [0, x'] = 0 \quad dt^N \quad a.s. \]

In addition

\[ \xi_t^N [0, x'] = \alpha_t^N [0, x'] - \beta_t^N [0, x'] = \alpha_t^N [0, x'] - \mu^N (t) + \nu^N (t) - J^N (t) + J^N (0) + \beta_t^N (x', \infty), \]

and note that \( \alpha^N \in \mathbb{D}^\uparrow_M \) and \( \mu^N + J^N - J (0) \in \mathbb{D}^\uparrow \). The later is justified by writing

\[ \mu^N + J^N - J (0) = (\mu^N - T^N) + (T^N + J^N - J (0)) = t^N + \beta^N [0, \infty). \]
The above holds for the scaled processes as well, thus \( \left( \xi_t^N, \beta_t^N, \bar{\gamma}_t^N \right) = \Theta' \left( \alpha_t^N, \bar{\mu}^N + \bar{J}_t^N - J_t^N \right) \).

By Lemma 1 and the continuous mapping theorem, it follows that \( \bar{\alpha}^N \Rightarrow \alpha \) if and only if \( \bar{\alpha}_t^N \Rightarrow \alpha' = F(\alpha) \in \mathbb{C}_M \). In addition,

\[
\alpha_t' \left( \{x'\} \right) = \alpha_t' [0, x'] - \alpha_t' [0, x'] = \alpha_t (-\infty, x] - \alpha_t (-\infty, x) = \alpha_t (\{x\}) = 0,
\]

so \( \alpha_t' \) is also atomless. Finally, \( \mathbb{D}_M^\uparrow \) is a closed subset of \( \mathbb{D}_M^\downarrow \) (Lemma 2.4), and each \( \alpha_t^N \) is in \( \mathbb{D}_M^\downarrow \), so we conclude that \( \alpha' \) is in \( \mathbb{C}_M^\uparrow \).

We now have the same setting as in Theorem 5.13 from \( \mathbb{H} \), by which \( \left( \xi_t^N, \beta_t^N, \bar{\gamma}_t^N \right) \Rightarrow \Theta' (\alpha', \mu) \).

Finally, by Lemma 1, we conclude \( \xi^N \Rightarrow \xi = F^{-1} (\xi') \), \( \beta^N \Rightarrow \beta = F^{-1} (\beta') \).

\[\square\]

**Lemma 1.**

1. The functions \( F \) and \( F^{-1} \), restricted to \( \mathbb{D}_M ([0, T]), T > 0 \), are continuous.

2. The function \( F \), restricted to \( \mathbb{D}_M ([0, T]), T > 0 \), sends elements of \( \mathbb{C}_M \) to \( \mathbb{C}_M \).

**Proof.**

1. Denote by \( B (\nu, \varepsilon) \) the open ball in \( \mathbb{D}_M ([0, T]) \) with radius \( \varepsilon \), centered around \( \nu \). We show that \( F (B (\nu, \varepsilon)) \subset B (\nu', e^{LT} \varepsilon), \nu' = F (\nu) \). Take any \( \lambda \in B (\nu, \varepsilon) \) and denote \( \lambda' = F (\lambda) \), then there exists a strictly increasing mapping \( \gamma : [0, T] \to [0, T] \), with \( \gamma (0) = 0, \gamma (T) = T \) and \( \sup_{t \in [0, T]} |\gamma (t) - t| < \varepsilon \), such that for any \( t \in [0, T] \) and any \( x \)

\[
\lambda_t (-\infty, x - \varepsilon] - \varepsilon \leq \gamma_{\nu (t)} (-\infty, x] \leq \lambda_t (-\infty, x + \varepsilon] + \varepsilon,
\]

which is just \( \sup_{t \in [0, T]} d_L \left( \nu_{\gamma (t)}, \lambda_t \right) < \varepsilon \).

Making yet another use of Assumption \( \mathbb{H} \), Theorem 6 in \( \mathbb{H} \) yields

\[
\begin{align*}
x'(x, t) - x' (x - \varepsilon, t) &= g(x, t) (0) - g(x, t - \varepsilon, t) (0) \leq \varepsilon e^{LT} \leq \varepsilon e^{LT}, \\
x'(x + \varepsilon, t) - x'(x, t) &= g(x + \varepsilon, t) (0) - g(x, t) (0) \leq \varepsilon e^{LT} \leq \varepsilon e^{LT}.
\end{align*}
\]

Using the above inequalities, for any \( t' \in [0, T] \) and \( x \),

\[
\begin{align*}
\nu_{\gamma (t')} [0, x] &= \nu_{\gamma (t)} (-\infty, x] \leq \lambda_t (-\infty, x + \varepsilon] + \varepsilon \leq \lambda_{\nu (t')} [0, x' + \varepsilon e^{LT}] + \varepsilon e^{LT}, \\
\nu_{\gamma (t')} [0, x'] &= \nu_{\gamma (t)} (-\infty, x] \geq \lambda_t (-\infty, x - \varepsilon] - \varepsilon \geq \lambda_{\nu (t')} [0, x' - \varepsilon e^{LT}] - \varepsilon e^{LT}.
\end{align*}
\]

Hence, \( \sup_{t' \in [0, T]} d_L \left( \nu_{\gamma (t')}, \lambda_t \right) < \varepsilon e^{LT} \).

Continuity for \( F^{-1} \) follows by similar arguments.

2. Just as in the first part of the Lemma, if for all \( t > 0 \) and \( x \)

\[
\nu_{t + h} [0, x - \varepsilon] - \varepsilon \leq \nu_t [0, x] \leq \nu_{t + h} [0, x' + \varepsilon] + \varepsilon,
\]

then for all \( t > 0 \) and \( x' > 0 \)

\[
\nu_{t + h} [0, x' - \varepsilon e^{LT}] - \varepsilon e^{LT} \leq \nu_t' [0, x'] \leq \nu_{t + h}' [0, x' + \varepsilon e^{LT}] + \varepsilon e^{LT}.
\]

This means that for any \( \varepsilon > 0 \), choosing \( h \) small enough such that \( d_L (\nu_{t + h}, \nu_t) < \delta, \delta = \varepsilon e^{LT} \) (such \( h \) exists as we assume \( \nu \in \mathbb{C}_M \)), implies \( d_L \left( \nu_t', \nu_t' \right) < \varepsilon \).
4 Examples

In this section we provide insight through simplifications and examples. The examples assume some workload arrival distribution of a fluid system, and provide explicit solutions. Recall that for a pair of data \((\alpha, \mu) \in \mathbb{C}_{M_0} \times \mathbb{C}_t\), the fluid solution is obtained by first applying the appropriate change of coordinates \(x' = g(x,t)(0)\) and then finding the measure valued process in \(\mathbb{C}_{M_0}\), \(\alpha' = F(\alpha)\) (defined by \(\alpha'_t[0,x'] = \alpha'_t(-\infty,x)\)). Then one applies the MVSM \((\xi', \beta', \iota) = \Theta(\alpha', \mu)\) and finally \(\xi = F^{-1}(\xi')\), \(\beta = F^{-1}(\beta')\). The entire procedure is summarized in the formulation of Theorem 1 through Equations (8) and (9),

\[
\begin{align*}
\beta_t(x, \infty) + \iota(t) &= - \inf_{s \in [0,t]} \{ (\Xi(s, g(x,t)(0)) - \mu(s)) \wedge 0 \}, \\
\xi_t(-\infty, x) &= \Xi(t, g(x,t)(0)) - \mu(t) - \inf_{s \in [0,t]} \{ (\Xi(s, g(x,t)(0)) - \mu(s)) \wedge 0 \},
\end{align*}
\]

with \(\Xi(t, x) = \alpha_t(-\infty, g(x,0)(t))\). In many cases, it is more natural to describe the arrival process through the instantaneous arrival distribution rather than through the cumulative arrival process \(\alpha\). The instantaneous arrival distribution describes the distribution of either the job sizes or the workload (each can be derived from the other) of the arriving work at each instant. Therefore, it is useful to have a relation connecting \(\alpha\) with this distribution. Let \(\pi_t\) be the instantaneous workload arrival distribution at time \(t\), then \(\alpha_t(-\infty, x) = \int_0^t \pi_s \left[0, g(x,t)(s)\right] ds\). All the following examples are of this form, where various instantaneous workload arrival distributions are considered. First we present a simplification of the formula for \(\xi\) and \(\beta\) under certain conditions. There appear four examples afterwards. The first three examples are of linear aging and the last is of exponential aging. In the first example, the instantaneous workload arrival distribution is uniform over the interval \([0,1]\), in the second example it is uniform over a time-varying interval, and in the two last examples it is Pareto distributed.

A simplification

Assume for simplicity that the arrival rate is larger than the service rate; in this case, \(\iota = 0\). We then may guess the following solution on the prime plane:

\[
\begin{align*}
\beta'_t[0,x] &= \alpha'_t[0,x] \wedge \mu(t), \\
\xi'_t[0,x] &= \alpha'_t[0,x] - \beta'_t[0,x] = \left[\alpha'_t[0,x] - \mu(t)\right]^+.
\end{align*}
\] (13)

Note that this is not always a valid guess, as it is not clear that

\[
\beta'_t(x', \infty) = \mu(t) - \beta'_t[0,x'] = \left[\mu(t) - \alpha'_t[0,x']\right]^+
\]

is non-decreasing in \(t\). However, it is true in many cases, e.g. when the job size distribution, and the arrival and service rates are time invariant. To verify that this guess is the true solution whenever \(\beta'_t(x', \infty)\) is non-decreasing, observe that either \(\xi'_t[0,x']\) or \(\beta'_t(x', \infty)\) is zero, Hence, \(\int_0^t \xi'_t[0,x'] d\beta'_t(x', \infty) = 0\), and by definition \(\beta'_t[0,\infty] = \mu(t)\) and \(\xi'_t[0,x'] = \alpha'_t[0,x'] - \beta'_t[0,x']\). Thus, this guess is the unique solution to the MVSP with data \((\alpha', \mu)\).

The interpretation of the guess (13) is that there is a process \(x^*(t) = \inf\{ x : \alpha'_t[0,x] \geq \mu(t) \} \) such that the entire mass below \(x^*(t)\) was served by time \(t\) and the entire mass above it is in the queue at time \(t\). The guess (13) is equal to the fluid solution if \(t \mapsto x^*(t)\) is non-decreasing. Note that the aging "pushes" \(x^*(t)\) to the right, thus does not invalidates the guess.

This simplification is used in the first example in this section to illustrate its usefulness. We do not use this simplification in the other examples.
Linear aging with uniformly distributed workload

We begin by analyzing the case of uniformly distributed workload, specifically on $[0, x]$, with constant arrival rate. This corresponds for $\pi_s[0, x] = 1 \land (x \lor 0)$. The aging rule in this example is chosen to be the linear aging rule. This is a rather simple example where closed form solutions can be found. For linear aging, the cumulative workload arrival process is derived from $\pi$ by $\alpha_t (-\infty, x] = \int_0^t \pi_s [0, x + t - s] \, ds$. We can compute

$$\alpha_t (-\infty, x] = \begin{cases} 
    t & \text{if } x > 1, \\
    t + x - \frac{x^2}{2} - \frac{1}{2} & \text{if } 0 < x < 1 \text{ and } x > 1 - t, \\
    xt + \frac{t^2}{2} & \text{if } 0 < x < 1 \text{ and } x < 1 - t, \\
    x + t - \frac{1}{2} & \text{if } x < 0 \text{ and } x > 1 - t, \\
    0 & \text{if } x < 0 \text{ and } x < -t, \\
    \frac{(x+t)^2}{2} & \text{if } x < 0 \text{ and } -t < x < 1 - t,
\end{cases}$$

and, consequently, compute

$$\alpha'_t [0, x'] = \begin{cases} 
    t & \text{if } x' > 1 + t, \\
    x' - \frac{(x'-t)^2}{2} - \frac{1}{2} & \text{if } t < x' < 1 + t \text{ and } x' > 1, \\
    x't - \frac{t^2}{2} & \text{if } t < x' < t + 1 \text{ and } x' < 1, \\
    x' - \frac{1}{2} & \text{if } x' < t \text{ and } x' > 1, \\
    0 & \text{if } x' < t \text{ and } x' < 0, \\
    \frac{x'^2}{2} & \text{if } x' < t \text{ and } 0 < x' < 1.
\end{cases}$$

If we choose for example $\mu(t) = t/2$, we can obtain $x^*(t) = (t + 1)/2$, which is increasing with $t$. Hence, the guess (13) is valid. To simplify the expressions of $\xi'_t$ and $\beta'_t$, we consider times $t > 1$. Then the solution is given by

$$\beta'_t [0, x'] = \begin{cases} 
    \frac{t}{2} & \text{if } x' > \frac{1+t}{2}, \\
    x' - \frac{1}{2} & \text{if } 1 < x' < \frac{t+1}{2}, \\
    0 & \text{if } x' < 0, \\
    \frac{x'^2}{2} & \text{if } 0 < x' < 1,
\end{cases}$$

$$\xi'_t [0, x'] = \begin{cases} 
    \frac{t}{2} & \text{if } x' > 1 + t, \\
    x' - \frac{(x'-t)^2}{2} - \frac{1}{2} - \frac{t}{2} & \text{if } t < x' < 1 + t \text{ and } x' > 1, \\
    x' - \frac{1+t}{2} & \text{if } t > x' > \frac{t+1}{2}, \\
    0 & \text{if } x' < \frac{t+1}{2}.
\end{cases}$$

For the sake of completeness, we give the explicit expression for $\xi_t (-\infty, x]$,

$$\xi_t [0, x] = \begin{cases} 
    \frac{t}{2} & \text{if } x > 1, \\
    x + t - \frac{x^2}{2} - \frac{1}{2} - \frac{t}{2} & \text{if } 0 < x < 1 \text{ and } x > 1 - t, \\
    x + \frac{1-t}{2} & \text{if } 0 > x > \frac{1-t}{2}, \\
    0 & \text{if } x < \frac{1-t}{2}.
\end{cases}$$

Figure 2(a) shows the evolution of $\xi_t (-\infty, x]$ with time. In this figure, and in all the other as well, the system is assumed to be empty at time $t = 0$. 
(a) $\xi_t (-\infty, x]$ when the aging is linear and the workload arrival distribution is uniform over the interval $[0, 1]$.

(b) $\xi_t (-\infty, x]$ when the aging is linear and the workload arrival distribution is uniform with time-varying support.

(c) $\xi_t (-\infty, x]$ when the aging is linear and the workload arrival distribution is Pareto with parameter $\eta = 1.2$.

(d) $\xi_t (-\infty, x]$ when the aging is exponential and the workload arrival distribution is Pareto with parameters $\eta = 1.2$ and $\lambda = 0.1$.

Figure 2: Examples of $\xi_t (-\infty, x]$ for different aging rules and different workload arrival distribution.
To ease the notation, the dependency of \( \int_0^a \) specific choice of interval \([0, t]\) an instantaneous workload distribution is again uniform, but its support is now a periodically-time-varying arrival distribution is periodic. This is exactly the setting in the following example. In this example, the instantaneous workload distribution is again uniform, but its support is now a periodically-time-varying interval \([0, a(t)]\). This corresponds to \( \pi_t [0, x] = a(s) \land (x \lor 0) \). Our aging rule is still linear. We solve for a specific choice of \( a(t) \), a triangular wave- a piecewise linear periodic function as illustrated in Figure 3.

To describe the solution, we define the following.

\[
\begin{align*}
N(x, t) &= 2 \left\lfloor \frac{x + t}{2} \right\rfloor, \\
A_1(x, t) &= 1 + \left\lfloor \frac{x + t}{2} \right\rfloor, \\
s^*_1(x, t) &= 2(x + t - A_1(x, t)), \\
A_2(x, t) &= \frac{1}{2} - \left\lfloor \frac{x + t}{2} \right\rfloor, \\
s^*_2(x, t) &= 2(x + t - A_2(x, t)).
\end{align*}
\]

To ease the notation, the dependency of \( x \) and \( t \) is omitted, and we simply write \( N, A_i, s^*_i, i = 1, 2 \). Let \( D_1 = \{(x, t) : N < x + t < N + 1/2\} \) and \( D_2 = \{(x, t) : N + 1/2 < x + t < N + 2\} \). Solving the integral \( \int_0^t \pi_s [0, x + t - s] ds \) gives

\[
\alpha_t(-\infty, x) =
\begin{cases}
\frac{3}{4} [t] + t - [t] - \frac{(t-|t|)^2}{4} & \text{if } t \leq s^* \text{ and } |t| \text{ is odd} \\
\frac{3}{4} [t] + \frac{t-|t|}{2} + \frac{(t-|t|)^2}{4} & \text{if } t \leq s^* \text{ and } |t| \text{ is even} \\
\frac{3}{4} \left[s^*_1\right] + s^*_1 - \left[s^*_1\right] - \frac{\left(s^*_1-\left[s^*_1\right]\right)^2}{4} + (x + t)(t - s^*_1) - \frac{t^2 - s^*_1^2}{2} & \text{if } 0 \leq s^*_1 < t \leq x + t \text{ and } D_1 \\
\frac{3}{4} \left[s^*_2\right] + \frac{s^*_2 - \left[s^*_2\right]}{2} + \frac{(s^*_2 - \left[s^*_2\right])^2}{4} + (x + t)(t - s^*_2) - \frac{t^2 - s^*_2^2}{2} & \text{if } 0 \leq s^*_2 < t \leq x + t \text{ and } D_2 \\
\frac{xt + \frac{t^2}{2}}{(x+t)^2} & \text{if } s^* < 0 < t < x + t \\
\frac{3}{4} \left[s^*_1\right] + s^*_1 - \left[s^*_1\right] - \frac{\left(s^*_1-\left[s^*_1\right]\right)^2}{4} + (x + t)(t + x - s^*_1) - \frac{(t+x)^2 - s^*_1^2}{2} & \text{if } 0 < s^*_1 < x + t < t \text{ and } D_1 \\
\frac{3}{4} \left[s^*_2\right] + \frac{s^*_2 - \left[s^*_2\right]}{2} + \frac{(s^*_2 - \left[s^*_2\right])^2}{4} + (x + t)(t + x - s^*_2) - \frac{(t+x)^2 - s^*_2^2}{2} & \text{if } 0 < s^*_2 < x + t < t \text{ and } D_2 \\
0 & \text{if } x + t < 0.
\end{cases}
\]

![Figure 3: Time-varying upper boundary of the workload distribution support.](image-url)
In the above, the condition $s^* > t$ should be interpreted as either $s_1^* > t$ and $D_1$, or $s_2^* > t$ and $D_2$; and likewise for $s^* < 0$.

It is now possible to find $\alpha'$ and $x^*(t)$, though obtaining $\xi'$, $\beta'$ and $\iota'$ is rather tedious even for such $\mu(t)$ for which the guess (13) is valid, and we do not continue with that line. The general solution, as mentioned at the beginning of the section, takes the form

$$\xi_t(-\infty, x) = \Xi(t, x + t) - \mu(t) - \inf_{s \in [0, t]} \{ (\Xi(s, x + t) - \mu(s)) \land 0 \} ,$$

with $\Xi(t, x) = s_t(-\infty, x - t)$. Figure 2(b) shows the evolution of $\xi_t(-\infty, x]$ with time. Note the difference between 2(a) and 2(b). The fact that $\pi_s$ is periodic is reflected in the wiggliness of the graph.

**Pareto distributed workload**

The Pareto distribution is a heavy tailed distribution, used often in queueing theory to model internet packet inter-arrival times ([13]). This distribution is of a special interest in this section because we can find explicit expressions for $\alpha$ when the workload is Pareto distributed, both for linear and for exponential aging. We begin with the linear aging setting, and compare the results later with the exponential aging setting. For the Pareto distribution of the arriving workload we choose the parameters 1 and $\eta$, which correspond to $\pi_s(0, x] = (1 - x^{-\eta}) 1_{(x \geq 1)}$. In this case we can calculate $\alpha$ and obtain

$$\alpha_t[0, x] = \int_0^t \pi_s[0, x + t - s] ds = \int_0^t (1 - (x + t - s)^{-\eta}) 1_{(x + t - s \geq 1)} ds =
\begin{cases}
0 & x < 1 - t,
\frac{x + t - 1 + \frac{(x+t)^{1-\eta} - 1}{\eta}}{1 - x \leq 1,}
\frac{x + t + \frac{(x+t)^{1-\eta} - x^{1-\eta}}{\eta}}{1 < x.}
\end{cases}$$

Now one can then find

$$\alpha'_t[0, x'] = \begin{cases}
0 & x' < 1,
\frac{x' - 1 + \frac{x'^{1-\eta} - 1}{\eta}}{1 \leq x' \leq 1 + t,}
\frac{t + \frac{x'^{1-\eta} - (x' - t)^{1-\eta}}{\eta}}{1 + t < x'.}
\end{cases}$$

The solution takes again the form

$$\xi_t(-\infty, x] = \Xi(t, x + t) - \mu(t) - \inf_{s \in [0, t]} \{ (\Xi(s, x + t) - \mu(s)) \land 0 \} ,$$

with $\Xi(t, x) = s_t(-\infty, x - t)$. Figure 2(c) shows the evolution of $\xi_t(-\infty, x]$ with time, when we used $\eta = 1.2$. Figure 4 shows the influence of the Pareto parameter $\eta$ on the shape of the graphs.

**Exponential aging**

We continue with an example in a setting with exponential aging. For this example we choose again the Pareto distribution with parameters 1 and $\eta$ as the instantaneous workload arrival distribution correspond-
Figure 4: $\xi_t (-\infty, x]$ when the aging is linear and the workload arrival distribution is Pareto, after 5 time units.

\[ \alpha_t [0, x] = \int_0^t \pi_s [0, x e^{-\lambda (s-t)}] ds = \int_0^{t+\lambda^{-1} \log(x \wedge 1)} \left( 1 - x^{-\eta} e^{\lambda \eta (s-t)} \right) ds \]

\[ = \begin{cases} 
0 & x \leq e^{-\lambda t} \\
t + \frac{1}{\lambda} \log(x) - \frac{x^{-\eta}}{\lambda \eta} (x^\eta - e^{-\lambda \eta t}) & e^{-\lambda t} < x < 1, \\
t - \frac{x^{-\eta}}{\lambda \eta} (1 - e^{-\lambda \eta t}) & x \geq 1, 
\end{cases} \]

$\alpha'_t [0, x'] = \alpha_t [0, x' e^{-\lambda t}] = \begin{cases} 
0 & x' \leq 1, \\
\frac{1}{\lambda} \log(x') - \frac{1-x'^{-\eta}}{\lambda \eta} & 1 < x' < e^\lambda, \\
t - \frac{x'^{-\eta}}{\lambda \eta} (e^{\lambda \eta} - 1) & x' \geq e^\lambda.
\end{cases}$

The solution takes the form

\[ \xi_t (-\infty, x] = \Xi \left( t, xe^{\lambda t} \right) - \mu \left( t \right) - \inf_{s \in [0,t]} \left\{ \left( \Xi \left( s, xe^{\lambda s} \right) - \mu \left( s \right) \right) \wedge 0 \right\}, \]

with $\Xi \left( t, x \right) = \alpha_t (-\infty, xe^{-\lambda t}]$.

Figure 2(d) shows the evolution of $\xi_t (-\infty, x]$ with time for this example with the choice $\mu(t) = t/2$. It is interesting to compare 2(c) with 2(d), both sharing the same workload arrival distribution while differing in their aging rules.

References

[1] R. Atar, A. Biswas, H. Kaspi, and K. Ramanan. A Skorokhod map on measure-valued paths with applications to priority queues. *Ann. Appl. Probab.*, 28(1):418–481, 02 2018.

[2] M. Bach, A. . T, AT, and T. I. S. (Firm). *The Design of the UNIX Operating System*. Prentice-Hall international editions. Prentice-Hall, 1986.

[3] N. Bansal and M. Harchol-Balter. Analysis of SRPT scheduling: investigating unfairness. In *Proceedings of the Joint International Conference on Measurements and Modeling of Computer Systems (SIGMETRICS 2001, Cambridge, MA, USA, June 16-20, 2001)*, pages 279–290, United States, 2001. Association for Computing Machinery, Inc.
[4] H. S. Behera, B. K. Swain, A. K. Parida, and G. Sahu. A New Proposed Round Robin with Highest Response Ratio Next (rrhrrn) Scheduling Algorithm for Soft Real Time Systems. 2012.

[5] G. Birkhoff and G. Rota. Ordinary Differential Equations. Introductions to higher mathematics. Wiley, 1989.

[6] M. Dell’Amico, D. Carra, M. Pastorelli, and P. Michiardi. Revisiting Size-Based Scheduling with Estimated Job Sizes. In 2014 IEEE 22nd International Symposium on Modelling, Analysis Simulation of Computer and Telecommunication Systems, pages 411–420, 2014.

[7] D. G. Down, H. C. Gromoll, and A. L. Puha. Fluid Limits for Shortest Remaining Processing Time Queues. Mathematics of Operations Research, 34(4):880–911, 2009.

[8] D. Feitelson. Notes on Operating Systems. The Hebrew University of Jerusalem, 2011.

[9] H. C. Gromoll, u. Kruk, and A. L. Puha. Diffusion Limits for Shortest Remaining Processing Time Queues. Stochastic Systems, 1(1):1–16, 2011.

[10] M. Gromoll H.C., Keutel. Invariance of fluid limits for the shortest remaining processing time and shortest job first policies. Queueing Syst, 70:145–164, 02 2012.

[11] M. Harchol-Balter, K. Sigman, and A. Wierman. Asymptotic Convergence of Scheduling Policies with Respect to Slowdown. Performance Evaluation, 49, 07 2003.

[12] J. D. C. Little. A Proof for the Queuing Formula: L = λw. Operations Research, 9(3):383–387, 1961.

[13] S. Mirtchev and R. Goleva. Evaluation of Pareto/D/1/K queue by simulation. 01 2008.

[14] M. Nuyens and B. Zwart. A large-deviations analysis of the G/G/1 SRPT queue. Queueing Systems: Theory and Applications, 54(2):85–97, 2006.

[15] A. L. Puha. Diffusion limits for shortest remaining processing time queues under nonstandard spatial scaling. Ann. Appl. Probab., 25(6):3381–3404, 12 2015.

[16] L. Schrage. Letter to the Editor—A Proof of the Optimality of the Shortest Remaining Processing Time Discipline. Operations Research, 16(3):687–690, 1968.

[17] L. E. Schrage and L. W. Miller. The Queue M / G /1 with the Shortest Remaining Processing Time Discipline. Operations Research, 14(4):670–684, 1966.

[18] A. Silberschatz, P. Galvin, and G. Gagne. Operating System Concepts. Windows XP update. Wiley, 2005.

[19] D. R. Smith. Technical Note—A New Proof of the Optimality of the Shortest Remaining Processing Time Discipline. Operations Research, 26(1):197–199, 1978.