The equivalence between processor sharing and service in random order

S.C. Borst[*,**,†], O.J. Boxma[*,**], J.A. Morrison[†], R. Núñez Queija[*,**]

[*]CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

[**]Department of Mathematics & Computer Science
Eindhoven University of Technology
P.O. Box 513, 5600 MB Eindhoven, The Netherlands

[†]Bell Laboratories, Lucent Technologies
P.O. Box 636, Murray Hill, NJ 07974, USA

Abstract

In this note we explore a useful equivalence relation for the delay distribution in the G/M/1 queue under two different service disciplines: (i) PS (Processor Sharing); and (ii) ROS (Random Order of Service). We provide a direct probabilistic argument to show that the sojourn time under PS is equal (in distribution) to the waiting time under ROS of a customer arriving to a non-empty system. We thus conclude that the sojourn time distribution for PS is related to the waiting-time distribution for ROS through a simple multiplicative factor, which corresponds to the probability of a non-empty system at an arrival instant. We verify that previously derived expressions for the sojourn time distribution in the M/M/1 PS queue and the waiting-time distribution in the M/M/1 ROS queue are indeed identical, up to a multiplicative constant. The probabilistic nature of the argument enables us to extend the equivalence result to more general models, such as the M/M/1/K queue and $\cdot$/M/1 nodes in product-form networks.

1 Introduction

In this note we study a useful equivalence result for the delay distribution in the G/M/1 queue, which is one of the most celebrated models in queueing theory. The G/M/1 model may be described as follows. Customers arrive to a single server according to some renewal process of rate λ, and require independent exponentially distributed amounts of service with parameter μ. For stability, the offered load should not exceed the service capacity, which may be expressed as $\rho < 1$, with $\rho := \lambda/\mu$ denoting the traffic intensity.
It is well-known that the distribution of the queue length $\mathbf{N}$ at arrival epochs (i.e. the total number of customers present, either waiting or being served) has a simple geometric form [6, 11],

$$\mathbb{P}\{\mathbf{N} = n\} = (1-\sigma)\sigma^n, \qquad n = 0, 1, 2, \ldots. \tag{1}$$

Here $z = \sigma$ is the unique, real zero inside the unit circle of the function $z - \alpha(\mu(1-z))$, with $\alpha(\cdot)$ representing the Laplace-Stieltjes Transform (LST) of the inter-arrival time distribution. Note that σ is the probability that the system is non-empty at an arrival instant. In case of a Poisson arrival process, we have $\alpha(z) = \lambda/(\lambda + z)$, so that $\sigma = \rho$. The geometric queue length distribution in (1) holds irrespective of the service discipline, as long as it is restricted to operate obliviously of the actual service times, see for instance [21]. In contrast, the distribution of the customer *delay* (waiting time or sojourn time) crucially depends on the service discipline that is used. For the FCFS (First-Come First-Served) discipline, it is well-known [6] that the sojourn time has an exponential distribution,

$$\mathbb{P}\{\mathbf{V} > t\} = \mathrm{e}^{-\mu(1-\sigma)t}, \qquad t \geq 0.$$

If the arrival process is Poisson so that $\sigma = \rho$, then the above formula takes the form $\mathbb{P}\{\mathbf{V} > t\} = \mathrm{e}^{-(\mu-\lambda)t}$.

Besides FCFS, a second major service mechanism is the Processor Sharing (PS) discipline. In a PS queue, the service rate is equally shared among all customers present. Thus, when there are $n \geq 1$ customers present, each customer receives service at rate $1/n$. Originally, the PS paradigm emerged as an idealization of round-robin scheduling mechanisms in time-shared computer systems [13, 14]. In recent years, the PS discipline has attracted renewed interest as a convenient modeling abstraction for bandwidth sharing protocols in high-speed networks [15, 20]. The performance of such computer-communication systems as perceived by the users is largely determined by the response time of tasks, or the transfer time of documents. Therefore, the sojourn time distribution in PS queues has been extensively investigated.
Initiated by Kleinrock's analysis of the M/M/1 PS queue [13, 14], many studies in the literature have focused on the analysis of the sojourn time *conditioned* on the service requirement. Extensions of such an analysis to generally distributed service requirements, multiple servers, and more general sharing disciplines were pursued in [5, 24, 25]. However, determining the sojourn time *distribution* in PS queues turned out to be a rather challenging problem.
For the M/M/1 PS queue, Coffman *et al.* [4] first derived a closed-form expression for the LST of the sojourn time distribution conditioned on the service requirement and the number of customers seen upon arrival. Sengupta and Jagerman [26] found an alternative expression for the LST of the distribution of the sojourn time conditioned only on the number of customers seen upon arrival. Building on [4], Morrison [18] established an expression for the *distribution* function of the sojourn time. For results on the sojourn time distribution in M/G/1 PS queues, we refer to the survey papers [28, 29].
The sojourn time distribution in G/M/1 PS queues has received less attention in the literature. Ramaswami [23] characterized the LST of the sojourn time distribution by a differential equation and determined the first two moments of the distribution. Jagerman and Sengupta [9] gave explicit expressions for the LST, and derived a heavy-traffic limit distribution under proper scaling, showing that, in the limit, the sojourn time is distributed as the product of two independent exponentially distributed random variables. The sojourn time in the 'repair' node (with PS discipline) of the machine-repairman model was examined by Mitra [16]. Extensions to multiple customer classes, both in the moderate and in the heavy-traffic regime, were considered by Mitra and Morrison [17, 19].

Comparing the results of Ramaswami [23] with those of Cohen [6, p. 444], Cohen [7] observed a simple relationship in the G/M/1 queue between the *sojourn* time distribution for PS and the *waiting*-time distribution for *Random Order of Service* (ROS). As the name suggests, in a ROS queue, customers are served in random order: whenever a service is completed, the next customer to be served is selected uniformly at random from the customers present, if any. An arriving customer which finds the server idle, is taken into service immediately. The ROS discipline provides a reasonable modeling assumption for, e.g., contention phenomena in distributed multi-access systems, see [3], where the results of the present paper are used.
Specifically, Cohen [7] proved that the sojourn time distribution in a G/M/1 queue for PS is related to the waiting-time distribution for ROS through a simple multiplicative factor:

$$\mathbb{P}\{\mathbf{V}_{ps} > t\} = \frac{1}{\sigma}\mathbb{P}\{\mathbf{W}_{ros} > t\}, \qquad t \geq 0, \tag{2}$$

with, as before, $\sigma := \mathbb{P}\{\mathbf{N} > 0\} = \mathbb{P}\{\mathbf{W}_{ros} > 0\}$ representing the probability that the system is non-empty at an arrival epoch. Note that the proportionality relation (2) may equivalently be expressed as follows:

$$\mathbb{P}\{\mathbf{V}_{ps} > t\} = \mathbb{P}\{\mathbf{W}_{ros} > t \,|\, \mathbf{W}_{ros} > 0\}, \qquad t \geq 0.$$

Cohen's proof of (2) relies on the fact that the transform of the delay distribution satisfies in both cases the same differential equation, which possesses a unique solution. In the present paper, we give a direct probabilistic proof of relation (2). The probabilistic nature of our proof allows us to extend the relationship to more general settings, including the above-mentioned machine-repairman model.
The remainder of the paper is organized as follows. In Section 2 we give a probabilistic proof of the equivalence relation (2). For the case of a Poisson arrival process, we verify in Section 3 that, indeed, the expressions derived by Flatto [8] for the waiting time distribution under ROS and by Morrison [18] for the sojourn time distribution under PS are identical, up to a multiplicative factor ρ. In Section 4 we discuss some straightforward extensions of the equivalence result, paving the way for the generalization to product-form networks in Section 5.

2 Probabilistic coupling

In contrast with the approach of Cohen [7], our proof of (2) relies on a direct probabilistic coupling argument. The key insight is that whenever a service completion occurs in the PS system, each of the customers present is equally likely to be the one that departs due to the memorylessness property of the exponential distribution. In that respect, the pool of customers competing for service under PS behaves exactly as the pool of customers waiting for service under ROS.

In order to formalize the above insight, let us focus on two tagged customers, X and Y. Customer X arrives in a PS system to find n customers present, and joins the pool of customers being served, which then consists of $n + 1$ customers. Customer Y arrives in a ROS system to find $n + 1$ customers present, one of which is being served, and thus joins the pool of *waiting* customers, which then contains $n + 1$ customers.

We now construct sample paths for the subsequent evolution of the two systems on a joint probability space such that: (i) customer X leaves the PS system at the same time as customer Y leaves the waiting pool to be taken into service in the ROS system; (ii) viewed in isolation, the evolution of each of the two systems follows the correct probabilistic laws. This is accomplished by coupling each customer in the PS system with exactly one customer in the waiting pool of the ROS system ensuring that, in particular, customers X and Y are coupled. Additionally, we couple the sequences of arrivals and service completions in the two systems as follows. Let A_1, A_2, ... be a sequence of i.i.d. random variables drawn from the distribution of the inter-arrival time. At the time instants determined by the sequence A_k, a customer arrives to each of the two systems. This pair of customers are immediately coupled. Let D_1, D_2, ... be a sequence of i.i.d. random variables drawn from the exponential distribution with parameter μ. At inter-departure times governed by the sequence D_k, a pair of customers are selected uniformly at random from all pairs present (i.e., one customer from the service pool in the PS system and the customer coupled to it in the waiting pool of the ROS system are selected). In the PS system, the customer that belongs to the selected pair departs from the system, whereas in the ROS system, the selected customer replaces the customer in service, which leaves the system. The above process is continued until the pair of customers X and Y are selected.

Thus, by construction, customer X departs from the PS system at the same time as customer Y starts service in the ROS system. In addition, it is easily verified that the evolution follows the correct probabilistic laws of each of the two systems in isolation. In fact, the above sample-path construction is a direct translation of the operational rules of the ROS system. For the PS system, the remaining service requirements of all customers are independent and exponentially distributed due to the memoryless property of the exponential distribution. As a result, given a service completion, all of the customers present are equally likely to be the one that leaves the system.

In conclusion, if we denote the sojourn time of a customer that arrives to a PS queue with n other customers competing for service by $\mathbf{V}_{ps}(n)$, and the waiting time of a customer that arrives to a ROS queue with n other customers waiting for service and one *additional customer* in service by $\mathbf{W}_{ros}(n)$, then the above sample-path construction proves that $\mathbf{V}_{ps}(n)$ and $\mathbf{W}_{ros}(n)$ are equal in distribution:

$$\mathbf{V}_{ps}(n) \stackrel{\mathrm{d}}{=} \mathbf{W}_{ros}(n), \qquad n = 0, 1, 2, \ldots. \tag{3}$$

Let us now turn to $\mathbf{V}_{ps}$, the (unconditional) sojourn time in the PS queue, and $\mathbf{W}_{ros}$, the (unconditional) waiting time in the ROS queue. For the PS queue we have, conditioning on the number of customers present at an arrival epoch,

$$\mathbb{P}\{\mathbf{V}_{ps} > t\} = \sum_{n=0}^{\infty} \mathbb{P}\{\mathbf{N} = n\}\mathbb{P}\{\mathbf{V}_{ps} > t \,|\, \mathbf{N} = n\} = \sum_{n=0}^{\infty} (1-\sigma)\sigma^n \mathbb{P}\{\mathbf{V}_{ps}(n) > t\}. \tag{4}$$

For the ROS queue, observing that the waiting time equals 0 exactly when there are no customers present at an arrival instant,

$$\begin{aligned} \mathbb{P}\{\mathbf{W}_{ros} > t\} &= \sum_{n=0}^{\infty} \mathbb{P}\{\mathbf{N} = n+1\}\mathbb{P}\{\mathbf{W}_{ros} > t \,|\, \mathbf{N} = n+1\} \\ &= \sigma \sum_{n=0}^{\infty} (1-\sigma)\sigma^n \mathbb{P}\{\mathbf{W}_{ros}(n) > t\}. \end{aligned} \tag{5}$$

The proportionality result (2) then follows from (3)–(5).

Remark 2.1 *In the PS literature, many studies have focused on the sojourn time conditioned on the service requirement of the customer. In fact, one of the attractive features of the PS discipline is that customers with smaller service requirements tend to experience smaller delays (the sojourn time is in fact stochastically increasing in the service requirement of a customer). It is worth emphasizing therefore that the equivalence result does not extend to the delays when conditioned on the service requirement. Observe that for the ROS discipline, the waiting time is independent of the service requirement.*

3 Special case: the M/M/1 queue

For the case of a Poisson arrival process, expressions for the distributions of $\mathbf{W}_{ros}$ and $\mathbf{V}_{ps}$ have been derived by Flatto [8] and Morrison [18], respectively.
In [8], time is normalized such that arrivals occur at unit rate, i.e., $\lambda = 1$. Thus, Formula (1.1) of [8] gives $\mathbb{P}\{\mathbf{W}_{ros} > t/\lambda\}$, or, equivalently:

$$\mathbb{P}\{\mathbf{W}_{ros} > t\} = \frac{2(1-\rho)}{\rho} \int\limits_{\phi=0}^{\pi} \frac{\mathrm{e}^{(2\xi(\phi)-\phi)\cot\phi}}{\mathrm{e}^{\pi\cot\phi}+1} \frac{\mathrm{e}^{-[1-(2/\sqrt{\rho})\cos\phi+1/\rho]\lambda t}}{[1-(2/\sqrt{\rho})\cos\phi+1/\rho]^2} \sin\phi \mathrm{d}\phi, \tag{6}$$

where

$$\xi(\phi) = \arctan(\frac{\sin\phi}{\cos\phi - \sqrt{\rho}}), \qquad 0 \leq \xi(\phi) \leq \pi.$$

On the other hand, in [18], time is normalized such that the mean service requirement equals unity, i.e., $\mu = 1$. For the generic case, Formula (2.20) of [18] takes the form:

$$\mathbb{P}\{\mathbf{V}_{ps} > t\} = \frac{2}{1-\rho} \int_{\theta=0}^{\pi} \frac{\mathrm{e}^{-\theta\left[2\sqrt{\rho}-(1+\rho)\cos\theta\right]/\left[(1-\rho)\sin\theta\right]-(1-\rho)^2\mu t/(1+\rho-2\sqrt{\rho}\cos\theta)}}{1+\mathrm{e}^{-\pi\left[2\sqrt{\rho}-(1+\rho)\cos\theta\right]/\left[(1-\rho)\sin\theta\right]}} \sin\theta \mathrm{d}\theta. \tag{7}$$

In the Appendix, we use the above integral expressions to verify the equivalence relation

$$\mathbb{P}\{\mathbf{W}_{ros} > t\} = \rho\, \mathbb{P}\{\mathbf{V}_{ps} > t\}, \qquad t \geq 0. \tag{8}$$

Via the proportionality relation (8), all results for the M/M/1 PS queue in [18] carry over to the M/M/1 ROS queue, and vice versa. In particular, Flatto [8] analyzes the asymptotic tail behavior of the waiting-time distribution in the M/M/1 ROS queue,

$$\mathbb{P}\{\mathbf{W}_{ros} > t\} \sim \kappa(\lambda t)^{-5/6} \exp(-\eta\lambda t - \gamma(\lambda t)^{1/3}), \tag{9}$$

where

$$\begin{aligned}
\kappa &= 2^{2/3}3^{-1/2}\pi^{5/6}\rho^{17/12}\frac{1+\sqrt{\rho}}{(1-\sqrt{\rho})^3}\exp(\frac{1+\sqrt{\rho}}{1-\sqrt{\rho}}), \\
\eta &= (1/\sqrt{\rho}-1)^2, \\
\gamma &= 3\left(\frac{\pi}{2}\right)^{2/3}\rho^{-1/6}.
\end{aligned}$$

Through relation (8), the above results directly yield the asymptotic tail behavior of the sojourn time distribution in the M/M/1 PS queue.

4 Model extensions

The probabilistic nature of the proof in Section 2 allows for various model extensions. In this section we describe how the equivalence result may be extended to two somewhat related models: the M/M/1/K queue and the machine-repairman model.

M/M/1/K queue

Consider an M/M/1/K queue with arrival rate λ and service rate μ. Arriving customers that find the maximum number of K customers present, are blocked and lost. It is well-known that the queue length $\mathbf{N}^K$ at arrival epochs then retains a (truncated) geometric form,

$$\mathbb{P}\{\mathbf{N}^K = n\} = \frac{(1-\rho)\rho^n}{1-\rho^{K+1}}, \qquad n = 0, 1, \ldots, K,$$

with $\rho := \lambda/\mu$ denoting the traffic intensity as before. Obviously, $\mathbf{N}^K$ has the same distribution for both service disciplines PS and ROS. For the queue length as observed by arrivals that are not blocked (denoted by $\widehat{\mathbf{N}}^K$) we have, for $n = 0, 1, \ldots, K-1$,

$$\mathbb{P}\{\widehat{\mathbf{N}}^K = n\} = \mathbb{P}\{\mathbf{N}^K = n \,|\, \mathbf{N}^K \leq K-1\} = \mathbb{P}\{\mathbf{N}^{K-1} = n\}. \tag{10}$$

As we will show, the proportionality relation between the sojourn time distribution for PS and the waiting-time distribution for ROS is preserved in the presence of a finite buffer. Now, however, the sojourn time in a PS system of capacity K is related to the waiting time in a ROS system of capacity $K+1$. Both the sojourn time in the PS system and the waiting time in the ROS system only concern customers that are not blocked upon arrival. Specifically, let $\mathbf{V}_{ps}^K(n)$ be the sojourn time of a customer which arrives to find $n \leq K-1$ customers present in a PS system of capacity K. Let $\mathbf{W}_{ros}^{K+1}(n)$ be the waiting time of a customer in a ROS system of capacity $K+1$ which arrives to find $n \leq K-1$ customers waiting plus one additional customer in service. Using similar probabilistic coupling techniques as in Section 2, it may be concluded that

$$\mathbf{V}_{ps}^K(n) \stackrel{d}{=} \mathbf{W}_{ros}^{K+1}(n), \qquad n = 0, 1, 2, \ldots, K-1. \tag{11}$$

We have, for the PS queue,

$$\mathbb{P}\{\mathbf{V}_{ps}^K > t\} = \sum_{n=0}^{K-1} \mathbb{P}\{\widehat{\mathbf{N}}^K = n\}\mathbb{P}\{\mathbf{V}_{ps}^K(n) > t\}, \tag{12}$$

and, for the ROS queue,

$$\mathbb{P}\{\mathbf{W}_{ros}^{K+1} > t\} = \sum_{n=0}^{K-1} \mathbb{P}\{\widehat{\mathbf{N}}^{K+1} = n+1\}\mathbb{P}\{\mathbf{W}_{ros}^{K+1}(n) > t\}. \tag{13}$$

Combining (10)–(13), we obtain

$$\mathbb{P}\{\mathbf{V}_{ps}^K > t\} = \frac{1}{\sigma(K)}\mathbb{P}\{\mathbf{W}_{ros}^{K+1} > t\}, \qquad t \geq 0,$$

where

$$\sigma(K) := \mathbb{P}\{\widehat{\mathbf{N}}^{K+1} > 0\} = \rho\frac{1-\rho^K}{1-\rho^{K+1}}$$

is the probability that a non-blocked customer in the M/M/1/$K+1$ ROS queue must wait before entering service.

Machine-repairman model

The machine-repairman model consists of a single repairman responsible for maintaining a pool of K machines which experience random failures according to identical but independent renewal processes of rate λ. Repair times are exponentially distributed with parameter μ. For notational convenience, define $\nu := \mu/\lambda$. It is well-known that the queue length at the repair node *at arrival epochs* has the form

$$\mathbb{P}\{\widehat{\mathbf{N}}^K = n\} = \frac{\dfrac{\nu^{K-n-1}}{(K-n-1)!}}{\sum\limits_{k=0}^{K-1} \dfrac{\nu^k}{k!}}, \qquad n = 0, 1, \ldots, K-1.$$

The proportionality relation between the sojourn time distribution for PS and the waiting-time distribution for ROS extends to the repair node in the above model as well. Again, however, the sojourn time in a PS system with population size K is related to the waiting time in a ROS system with population size $K+1$. Specifically, let $\mathbf{V}_{ps}^K(n)$ be the sojourn time of a customer which arrives to find $n \leq K-1$ customers present at the repair node in a PS system with population size K. Let $\mathbf{W}_{ros}^{K+1}(n)$ be the waiting time of a customer in a ROS system with population size $K+1$ which arrives to find $n \leq K-1$ customers waiting at the repair node plus one additional customer in service. Mimicking the probabilistic coupling arguments of Section 2, it may be deduced that

$$\mathbf{V}_{ps}^K(n) \stackrel{d}{=} \mathbf{W}_{ros}^{K+1}(n), \qquad n = 0, 1, 2, \ldots, K-1,$$

which after similar manipulations as before yields

$$\mathbb{P}\{\mathbf{V}_{ps}^K > t\} = \frac{1}{\sigma(K)} \mathbb{P}\{\mathbf{W}_{ros}^{K+1} > t\}, \qquad t \geq 0, \tag{14}$$

with

$$\sigma(K) := \mathbb{P}\{\widehat{\mathbf{N}}^{K+1} > 0\} = \frac{\sum\limits_{k=0}^{K-1} \dfrac{\nu^k}{k!}}{\sum\limits_{k=0}^{K} \dfrac{\nu^k}{k!}}.$$

If the renewal processes governing the failures are non-Poisson, then the above proof does not directly apply. However, the model may still be viewed as a closed queueing network consisting of two queues: a single-server queue representing the repairman, and an infinite-server queue modeling the operational machines. In the next section, we extend the proportionality result to a class of product-form networks which includes the above model as a special case, thus generalizing (14) to non-Poisson renewal processes.

5 Product-form networks

In this section we indicate how the equivalence result may be extended to $\cdot$/M/1 nodes in product-form networks. We adopt the setting described by Baskett *et al.* [2], which allows for nodes with FCFS, PS, ample service, and LCFS (Last-Come First-Served). For FCFS nodes, the results are restricted to a single class of customers at that node and exponentially distributed service requirements (but service rates may be state-dependent). All three other service disciplines allow for service requirements with phase-type distributions (in fact, general distributions [1, 10]). This only highlights the most important elements; readers are referred to [2] for more details.
Note that this setting can be extended, allowing the FCFS nodes to be replaced with any non-preemptive service discipline operating obliviously of the service requirements, for instance ROS. The queue length process obeys the same probabilistic law for any such discipline. The only difference is that customers might overtake (in various ways) within the node, but, all of them being of the same type, this does not alter the evolution of the entire network in terms of the number of customers of each type at the various nodes.

Consider a network as in [2] with one of the FCFS nodes with exponential services, let's say node 0, replaced by a ROS node. We call this the "ROS network". The service rate at node 0 is denoted by μ. (In [2] the service rate at FCFS nodes may depend on the number of customers at that node, but this is not incorporated in the main result. Although some forms of state-dependent service rates could be included in the analysis, we shall not do so.) We *impose* that all customers that can visit the ROS node are of the same class, i.e., they all follow the same route through the network and share the same service distributions at all nodes. If there are external arrivals to this route, we assume that they occur according to a Poisson process, independent of the state of the network (this can be generalized too). The rate at which external arrivals occur at other routes may however depend on the number of customers traveling through the particular route. If the route through node 0 is closed (i.e., customers on this route never leave the system), we denote the number of customers on this route by $K+1$. For notational convenience we shall write $K=\infty$ if the route is open. Note however that, in the latter case, the number of customers actually present on the route may vary between zero and infinity, whereas if $K<\infty$, this number always equals $K+1$. We now construct a "PS network" from the ROS network, changing the service discipline at node 0 from ROS to PS and, if $K<\infty$, reducing the number of customers on this route in the PS network by one.
Focus on a particular configuration of customers of all types in the PS network (i.e., focus on a particular *state* of the PS network) with $n \geq 0$ customers at node 0 and call this configuration $\underline{z}$. Associate with it the configuration $\underline{z}+\underline{e}_0$ in the corresponding ROS network obtained by adding one customer to node 0. Let $\mathbf{Z}^K$ and $\mathbf{Z}^{K+1}$ be distributed as the configurations (in equilibrium) of customers in the PS and ROS networks, respectively. When the route through node 0 is open, the two random variables have the same distribution and, using our convention that $K=\infty$ in this case, we denote both of them by $\mathbf{Z}^\infty$. By [2, Thm. p. 253] we have,

$$\frac{\mathbb{P}\{\mathbf{Z}^{K+1}=\underline{z}+\underline{e}_0\}}{\mathbb{P}\{\mathbf{Z}^K=\underline{z}\}}=\frac{C_{K+1}\,d(\underline{z}+\underline{e}_0)\,(1/\mu)^{n+1}}{C_K\,d(\underline{z})\,(1/\mu)^n}=:\sigma(K+1). \tag{15}$$

In both networks we have chosen to normalize time such that the rate of external arrivals at node 0 equals 1. Then, the function $d(\cdot)$ depends only on the numbers of customers on the routes that do not traverse node 0 and, therefore, $d(\underline{z}) = d(\underline{z} + \underline{e}_0)$. Since C_K and C_{K+1} are normalization constants, we have that $\sigma(K+1) = C_{K+1}/(\mu C_K)$ is a constant, independent of the configuration $\underline{z}$.

Let us observe both networks at moments when a customer makes a transition into node 0. Let $\widehat{\mathbf{Z}}^K$ be distributed as the configuration of *other customers* in the PS network at such transition moments, i.e., not counting the customer making the transition. Similarly, $\widehat{\mathbf{Z}}^{K+1}$ is distributed as the configuration in the ROS network at moments of transitions into node 0. Now recall the so-called arrival-departure property: when a customer arrives at a certain queue, it sees the network in equilibrium if the customer is traveling on an open route; and the customer sees the network as if it is in equilibrium with one less customer on the route if the route is closed, see Sevcik and Mitrani [27]. By this property we have that $\mathbb{P}\{\widehat{\mathbf{Z}}^K = \underline{z}\} = \mathbb{P}\{\mathbf{Z}^{K-1} = \underline{z}\}$. Hence, using (15),

$$\mathbb{P}\{\widehat{\mathbf{Z}}^{K+1} = \underline{z} + \underline{e}_0\} = \sigma(K)\,\mathbb{P}\{\widehat{\mathbf{Z}}^K = \underline{z}\}. \tag{16}$$

Adding over all possible configurations $\underline{z}$ in the PS network, we conclude (as before) that $\sigma(K)$ equals the probability that a customer arriving at the ROS node finds the server busy. We can interpret (16) as follows: with probability $1 - \sigma(K)$, a customer arriving at the ROS node is immediately taken into service; otherwise, with probability $\sigma(K)$, it sees each possible configuration (not counting the customer in service at the ROS node) with the same probability as a customer arriving at node 0 in the PS network.

Let us now focus on two customers, customer X arrives at node 0 in the PS network and sees configuration $\underline{z}$, and customer Y arrives at node 0 in the ROS network and sees the 'corresponding' configuration $\underline{z} + \underline{e}_0$. As before, we couple each customer in the ROS network to exactly one customer in the PS network, except for the additional customer in service at node 0 of the ROS network. Customers X and Y are coupled to each other. Every two coupled customers are located at the corresponding nodes in the two networks, they belong to the same customer class, are at the same stage of service and will follow the same route through the network, simultaneously undergoing the same service phases at each of the subsequently visited nodes. Let $\mathbf{V}_{ps}(\underline{z})$ be the sojourn time of X at node 0 in the PS network and let $\mathbf{W}_{ros}(\underline{z} + \underline{e}_0)$ be the waiting time of Y at node 0 in the ROS network. Since the two networks evolve according to the same probabilistic law as long as the ROS node is not empty, we can (again) conclude that the sojourn times of X and Y in their respective pools are equal:

$$\mathbf{V}_{ps}(\underline{z}) \stackrel{\mathrm{d}}{=} \mathbf{W}_{ros}(\underline{z} + \underline{e}_0).$$

As before, let $\mathbf{V}_{ps}$ be the sojourn time of an arbitrary customer in the PS queue and let $\mathbf{W}_{ros}$ be the waiting time of an arbitrary customer in the ROS queue. Now, using (16) and the interpretation given below that equation, we find, by de-conditioning over all possible states $\underline{z}$ seen upon arrival by customer X and the corresponding states $\underline{z} + \underline{e}_0$ seen by Y (noting that a customer that enters an empty ROS node does not have to wait):

$$\begin{aligned}
\mathbb{P}\{\mathbf{W}_{ros} > t\} &= 0 + \sum_{\underline{z}} \mathbb{P}\{\widehat{\mathbf{Z}}^{K+1} = \underline{z} + \underline{e}_0\}\,\mathbb{P}\{\mathbf{W}_{ros}(\underline{z} + \underline{e}_0) > t\} \\
&= \sum_{\underline{z}} \sigma(K)\mathbb{P}\{\widehat{\mathbf{Z}}^K = \underline{z}\}\,\mathbb{P}\{\mathbf{V}_{ps}(\underline{z}) > t\} \\
&= \sigma(K)\mathbb{P}\{\mathbf{V}_{ps} > t\},
\end{aligned}$$

which proves the proportionality result for product-form networks.

Appendix

In this appendix, we verify the equivalence relation (8) using the integral expressions for the distributions of the waiting time in the M/M/1 ROS queue and the sojourn time in the M/M/1 PS queue as given by Formulas (6) and (7), respectively. In order to rewrite Formula (7) into the form of Formula (6), we first derive some useful identity relations. For given $\theta \in [0, \pi]$, let $\phi \in [0, \pi]$ be such that

$$1 - \sqrt{\rho}\,\mathrm{e}^{-i\theta} = \frac{1-\rho}{1-\sqrt{\rho}\,\mathrm{e}^{i\phi}}. \tag{17}$$

Then the following identity relations hold

$$\begin{aligned}
1+\rho-2\sqrt{\rho}\cos\theta &= \frac{(1-\rho)^2}{1+\rho-2\sqrt{\rho}\cos\phi}, &(18)\\
\frac{\sin\theta}{1-\rho}\mathrm{d}\theta &= -\frac{(1-\rho)\sin\phi}{(1+\rho-2\sqrt{\rho}\cos\phi)^2}\mathrm{d}\phi, &(19)\\
\frac{2\sqrt{\rho}-(1+\rho)\cos\theta}{(1-\rho)\sin\theta} &= \cot\phi. &(20)
\end{aligned}$$

Relation (19) follows from (18) by differentiation. To arrive at Relation (20), we use the additional relations

$$\sin\theta = \frac{(1-\rho)\sin\phi}{1+\rho-2\sqrt{\rho}\cos\phi}, \qquad \cos\theta = \frac{2\sqrt{\rho}-(1+\rho)\cos\phi}{1+\rho-2\sqrt{\rho}\cos\phi}.$$

The latter relations are obtained by equating real parts, as well as imaginary parts, in (17). Substituting (18)-(20) into (7) and multiplying the numerator and denominator of the integrand by $\exp(\pi\cot\phi)$, we obtain

$$\mathbb{P}\{\mathbf{V}_{ps} > t\} = 2(1-\rho)\int_{\phi=0}^{\pi} \frac{\exp\left\{(\pi-\theta)\cot\phi - (1+\rho-2\sqrt{\rho}\cos\phi)\mu t\right\}}{\Big(1+\exp\{\pi\cot\phi\}\Big)\Big(1+\rho-2\sqrt{\rho}\cos\phi\Big)^2}\sin\phi\,\mathrm{d}\phi. \tag{21}$$

Now, using the fact that

$$\tan(\phi-\theta) = \frac{\tan\phi-\tan\theta}{1+\tan\phi\tan\theta} = \frac{2\sin\phi(\cos\phi-\sqrt{\rho})}{(\cos\phi-\sqrt{\rho})^2-\sin^2\phi} = \frac{2\frac{\sin\phi}{\cos\phi-\sqrt{\rho}}}{1-\left(\frac{\sin\phi}{\cos\phi-\sqrt{\rho}}\right)^2},$$

and

$$\tan(\phi-\theta) = \tan(\pi+\phi-\theta) = \frac{2\tan\left(\frac{\pi+\phi-\theta}{2}\right)}{1-\tan^2\left(\frac{\pi+\phi-\theta}{2}\right)},$$

we have

$$\pi-\theta = 2\arctan\left(\frac{\sin\phi}{\cos\phi-\sqrt{\rho}}\right) - \phi. \tag{22}$$

Substituting (22) into (21), we find that $\mathbb{P}\{\mathbf{W}_{ros} > t\} = \rho\,\mathbb{P}\{\mathbf{V}_{ps} > t\}$, where $\mathbb{P}\{\mathbf{W}_{ros} > t\}$ is given by (6), which completes the proof.

References

[1] Barbour, A.D. (1976). Networks of queues and the method of stages. *Adv. Appl. Prob.* **8**, 584–591.

[2] Baskett, F., Chandy, K.M., Muntz, R.R., Palacios, F.G. (1975). Open, closed, and mixed networks of queues with different classes of customers. *J. ACM* **22**, 248–260.

[3] Boxma, O.J., Denteneer, D., Resing, J.A.C. (2001). Some models for contention resolution in cable networks. EURANDOM Report 2001-037.

[4] Coffman, E.G., Jr., Muntz, R.R., Trotter, H. (1970). Waiting time distribution for processor-sharing systems. *J. ACM* **17**, 123–130.

[5] Cohen, J.W. (1979). The multiple phase service network with generalised processor sharing. *Acta Informatica* **12**, 245–284.

[6] Cohen, J.W. (1982). *The Single Server Queue* (2nd edition). North-Holland, Amsterdam.

[7] Cohen, J.W. (1984). On processor sharing and random service (Letter to the editor). *J. Appl. Prob.* **21**, 937–937.

[8] Flatto, L. (1997). The waiting time distribution for the random order service M/M/1 queue. *Ann. Appl. Prob.* **7**, 382–409.

[9] Jagerman, D.L., Sengupta, B. (1991). The GI/M/1 processor-sharing queue and its heavy traffic analysis. *Comm. Statist. – Stoch. Models* **7**, 379–395.

[10] Kelly, F.P. (1979). *Reversibility and Stochastic Networks.* Wiley, Chichester.

[11] Kendall, D.G. (1953). Stochastic processes occurring in the theory of queues and their analysis by the method of the embedded Markov chain. *Ann. Math. Stat.* **24**, 338–354.

[12] Kingman, J.F.C. (1962). On queues in which customers are served in random order. *Proc. Camb. Phil. Soc.* **58**, 79–91.

[13] Kleinrock, L. (1964). Analysis of a time-shared processor. *Nav. Res. Log. Quarterly* **11**, 59–73.

[14] Kleinrock, L. (1967). Time-shared systems: a theoretical treatment. *J. ACM* **14**, 242–261.

[15] Massoulié, L., Roberts, J.W. (1999). Bandwidth sharing: Objectives and algorithms. In: *Proc. IEEE INFOCOM '99*, 1395–1403.

[16] Mitra, D. (1981). Waiting time distributions from closed queueing network models of shared-processor systems. In: *Proc. Performance '81*, ed. F.J. Kylstra, North-Holland, Amsterdam, 113–131.

[17] Mitra, D., Morrison, J.A. (1983). Asymptotic expansions of moments of the waiting time in closed and open processor-sharing systems with multiple job classes. *Adv. Appl. Prob.* **15**, 813–839.

[18] Morrison, J.A. (1985). Response-time distribution for a processor-sharing system. *SIAM J. Appl. Math.* **45**, 152–167.

[19] Morrison, J.A., Mitra, D. (1985). Heavy-usage asymptotic expansions for the waiting time in closed processor-sharing systems with multiple classes. *Adv. Appl. Prob.* **17**, 163–185.

[20] Núñez Queija, R. (2000). *Processor-Sharing Models for Integrated-Services Networks.* Ph.D. thesis, Eindhoven University of Technology, ISBN 90-646-4667-8 (also available from the author upon request).

[21] Núñez Queija, R. (2001). Note on the GI/GI/1 queue with LCFS-PR observed at arbitrary times. *Prob. Eng. Inf. Sc.* **15**, 179–187.

[22] Ott, T.J. (1984). The sojourn time distributions in the M/G/1 queue with processor sharing. *J. Appl. Prob.* **21**, 360–378.

[23] Ramaswami, V. (1984). The sojourn time in the GI/M/1 queue with processor sharing. *J. Appl. Prob.* **21**, 437–442.

[24] Sakata, M., Noguchi, S., Oizumi, J. (1969). Analysis of a processor-shared queueing model for time-sharing systems. *In: Proc. 2nd Hawaii Int. Conf. System Sciences*, 625–628.

[25] Sakata, M., Noguchi, S., Oizumi, J. (1971). An analysis of the M/G/1 queue under round-robin scheduling. *Oper. Res.* **19**, 371–385.

[26] Sengupta, B., Jagerman, D.L. (1985). A conditional response time of the M/M/1 processor-sharing queue. *AT&T Techn. J.* **64**, 409–421.

[27] Sevcik, K.C., Mitrani, I. (1981). The distribution of queuing network states at input and output instants. *J. ACM* **28**, 358–371.

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

[29] Yashkov, S.F. (1992). Mathematical problems in the theory of processor-sharing queueing systems. *J. Soviet Math.* **58**, 101–147.