# Priorities at the end of service

Onno Boxma * Yoav Kerner †

**Abstract**

Consider an $M/G/1$ queue with $N$ customer priority classes, all with the same service requirement distribution. The special feature of the model is, that the assignment of a job is done *at the end* of the service period – it is assigned to a customer of the highest priority class then present. Two variations of this scheme are studied: (i) the stoppable server case, in which the server only works when there are customers, and (ii) the unstoppable case, in which the server always works but scraps a job when at its completion there is no customer to receive it. For both variants we determine the probability generating function of the steady-state joint queue length distribution, as well as the Laplace-Stieltjes transform of the sojourn time distribution of each class. This is done in detail for $N = 2$ customer classes, and more globally for general $N$.

## 1 Introduction

We consider an $M/G/1$ queue with multiple customer classes, with a complete priority ordering across the classes. Unlike many service systems, the service itself is not customer dependent, but is viewed as the production of a standard unit. Our key assumption is that the assignment of a job is done *at the end* of the service period; the job is then assigned to the oldest customer of the highest priority class then present. Notice that this receiving customer may not yet have been present at the beginning of the service.

*Department of Mathematics and Computer Science, Eindhoven University iof Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands; o.j.boxma@tue.nl

†Department of Industrial Engineering and Management, Ben Gurion University of the Negev, Beer-Sheva, 84105 Israel; kerneryo@bgu.ac.il

Although assignment of a job at the end of a service has hardly received any attention in the queueing literature, it is quite natural in several settings. One may think of a hamburger restaurant, and also of various production systems where items are produced, partly on demand but also to add to the stock (Make-to-Order versus Make-to-Stock). In such settings one can distinguish two different server behaviors. Variant 1 is the stoppable server, where the server idles when there are no customers waiting. Variant 2 is the unstoppable server, where the server always works – even when there are no customers waiting. In the latter case, it is possible that even when the job is finished, there still is no customer. We then assume that this job is scrapped (alternatively, such a job might be added to the stock and used later when a new request arrives; see [1, 2] for studies of such a queueing/inventory model, but without priorities). Scrapping instead of adding to the inventory may be natural in the case of food orders, or when maintaining an inventory is costly.

**Related literature.** There is a huge literature on priority queues; see, e.g., the books of Harchol-Balter [5], Haviv [6], Jaiswal [8], Kleinrock [9] and Takagi [10]. In queueing models with priorities, one can distinguish many important categories. We mention queues with changing priorities (e.g., polling models); queues in which the actual, or remaining, service times play a role in the priority scheme (like shortest-remaining-job-first); and queues with several customer classes with fixed priorities among the classes. Our paper falls in the latter category. In this category one usually distinguishes between preemptive and non-preemptive priority. In our model there is no preemption, but one might argue that our priority scheme lies between preemptive and non-preemptive priority. When there are high-priority customers present at the end of a service, the system acts like a standard non-preemptive priority queue; but when there are no high-priority customers present at the end of a service and such a customer arrives during the next service, then in a sense it preempts/takes over the service of a customer of lower priority.

To the best of our knowledge, the priority models of the present paper, in which a job is assigned to a customer at the end of a service, were only studied by Haviv and Kerner [7]. They obtained the mean sojourn time of each customer class, for both the stoppable and unstoppable server model. They also compare the mean waiting time of customers of class $i$ in the stoppable case with that of the standard Head-of-the-line (HOL) preemptive priority model: $W_i - W_i^{HOL}$ is shown to be monotonically decreasing in $i$, i.e., class 1 benefits the most when HOL is replaced by the stoppable server policy, and

class $N$ loses the most.

**Main contribution.** The main contribution of the present paper is to demonstrate how the joint queue length distribution right after service completion epochs can be determined, and how this can be used to obtain the sojourn time LST of each customer class. We do this in detail for $N = 2$, for both the stoppable and unstoppable case, and we outline the procedure in some detail for $N = 3$ and then more globally for general $N$.

**Organization of the paper.** We give a detailed model description in Section 2. Section 3 is devoted to the stoppable case; in successive subsections, we consider $N = 2$, $N = 3$ and general $N$. Section 4 treats the unstoppable case. Section 5 briefly discusses the model where the server, at a service completion that leaves the system empty, stops with probability $q$ and (unstoppably) produces a job with probability $1 - q$.

# 2 Model description

We consider a multi-class $M/G/1$ priority queue. There are $N$ classes of customers. The arrival process of class $i$ is Poisson with rate $\lambda_i$, $i = 1, ..., N$, and independent of anything else. The total arrival rate is $\lambda := \sum_{i=1}^{N} \lambda_i$, and $\phi_i := \lambda_i/\lambda$ denotes the fraction of customers that is of class $i$, $i = 1, \ldots, N$. All service times, for all classes, are independent and identically distributed. A generic service time is denoted by $B$, and its Laplace-Stieltjes transform (LST) by $b(\cdot)$.

Priority is given to classes with lower index; in particular, class-1 customers have the highest priority. Unlike other priority regimes, in our model the priority is applied at the *end* of service. That is, whenever a service is completed, the customer to which it is assigned is the oldest one from the class with highest priority then present (notice that assignment to the oldest one is relevant when it comes to our study of sojourn times; for our queue length study, it is irrelevant). We consider two variants of this service policy. In the first one, when the system is empty after a service completion, the server stops serving until the next arrival; we refer to this as the *stoppable* server model. In the second one, when the system is empty after a service completion, the server continues performing a service. If no arrival occurs during that service, the service is wasted. Otherwise, it is assigned to the class with highest priority among the arrivals. We refer to this variant as the *unstoppable* server model. Finally, when we study sojourn times, we assume

that within each class, service is First-Come-First-Served.

# 3 The stoppable server model

We start with some notation. Let $X_i^{(n)}$ be the number of class-$i$ customers that are in the system right after the $n$-th service completion, and let the state be $\underline{X}^{(n)} = \left(X_i^{(n)}\right)_{i=1}^N$. Also, let $A_i^{(n)}$ be the number of class-$i$ arrivals during the $n$-th service, and, as above, let $\underline{A}^{(n)} = \left(A_i^{(n)}\right)_{i=1}^N$.
The transition from the $n$-th service completion to the next service completion *depends on the state right before the $(n+1)$-st service completion*. This state itself is a function of the state at the $n$-th service completion and the arrivals during the service period. Thus, we define $i_{min}(n) = \arg\min\{i|X_i^{(n)} + A_i^{(n+1)} \geq 1\}$. Note that after the $n$-th service completion, the next customer to receive service is from class $i_{min}(n)$. The transitions of the process are as follows. If a positive $i_{min}(n)$ exists then

$$\underline{X}^{(n+1)} = \underline{X}^{(n)} + \underline{A}^{(n+1)} - e_{i_{min}(n)}, \tag{1}$$

with $e_j$ being the $j$-th unit vector. Otherwise, $\underline{X}^{(n+1)} = (0, 0, ...0) =: \underline{0}$.

## 3.1 Steady-state distribution

In this subsection we investigate the joint steady-state distribution of the numbers of customers in the various classes. Comparison with an ordinary $M/G/1$ queue immediately shows that this steady-state distribution exists iff $\rho := \lambda \mathrm{E}B < 1$. We write $\underline{z} := (z_1, ..., z_N)$ and we denote the vector with the first $i$ components being 0, and the rest $z_{i+1}, ...z_N$, by $z[i]$. Let

$$F(\underline{z}) = \lim_{n\to\infty} \mathrm{E}\left(\prod_{i=1}^N z_i^{X_i^{(n)}}\right). \tag{2}$$

Also, denote the $N$-dimensional probability generating function (PGF) of the number of arriving customers from the various classes during a single service period by

$$\beta(\underline{z}) := \mathrm{E}\left(\prod_{i=1}^N z_i^{A_i^{(n)}}\right) = \int_0^\infty e^{-t\sum_{i=1}^N \lambda_i(1-z_i)} \mathrm{d}P(B < t) = b(\sum_{i=1}^N \lambda_i(1-z_i)). \tag{3}$$

Taking the multi-dimensional generating function of both sides in (1), and then taking the limit $n \to \infty$, yields

$$F(\underline{z}) = \sum_{i=1}^{N} \frac{(F(z[i-1]) - F(z[i]))\,\beta(z[i-1]) + (F(z[i]) - F(\underline{0}))\,(\beta(z[i-1]) - \beta(z[i]))}{z_i}$$
$$+F(\underline{0}) \left( \phi_1 \beta(\underline{z}) + \sum_{i=2}^{N} \phi_i z_i \left( \sum_{j=1}^{i-1} \frac{\beta(z[j-1]) - \beta(z[j])}{z_j} + \frac{\beta(z[i-1])}{z_i} \right) \right). \tag{4}$$

The $i$-th element of the first row in (4) refers to a departure of a class-$i$ customer. Within it, the first term represents the event that the highest prioritized customers left behind at the last departure are from class $i$, and there were no higher priority arrivals during the service. The second term represents the event that the system is not empty at the last departure, those present are from classes less prioritized than $i$, and the most prioritized customers who arrived during the service are from class $i$. The second row in (4) represents the event that the last departure left the system empty. We there distinguish between the first arrival being of class 1 or of one of the classes $2, \ldots, N$. In Subsection 3.2 we present the solution for $F(\underline{z})$ for the special case $N = 2$ in much detail; Subsection 3.3 more globally considers $N = 3$, after which we are ready to present the structure for the general case, in Subsection 3.4.

## 3.2 The case $N = 2$

For the case of $N = 2$ classes, (4) simplifies to

$$\begin{aligned} F(z_1, z_2) &= \frac{1}{z_1} \left[ F(z_1, z_2) - F(0, z_2) \right] \beta(z_1, z_2) \\ &+ \left[ F(0, z_2) - F(0, 0) \right] \frac{\beta(z_1, z_2) - \beta(0, z_2)}{z_1} \\ &+ \frac{1}{z_2} \left[ F(0, z_2) - F(0, 0) \right] \beta(0, z_2) + F(0, 0) \phi_1 \beta(z_1, z_2) \\ &+ z_2 F(0, 0) \phi_2 \frac{\beta(z_1, z_2) - \beta(0, z_2)}{z_1} + F(0, 0) \phi_2 \beta(0, z_2). \end{aligned} \tag{5}$$

Multiplying both sides of (5) by $z_1 z_2$ and moving the $F(z_1, z_2)$ term in the righthand side to the lefthand side yields

$$F(z_1, z_2)(z_1 - \beta(z_1, z_2))z_2 = F(0, z_2)(z_1 - z_2)\beta(0, z_2) + F(0, 0)L(z_1, z_2), \tag{6}$$

where

$$L(z_1, z_2) := z_2\beta(z_1, z_2)\left\{\phi_1 z_1 + \phi_2 z_2 - 1\right\} + (z_2 - z_1)\beta(0, z_2)\left(1 - \phi_2 z_2\right). \quad (7)$$

At this stage we remark that $F(z, z)$, the PGF of the steady-state total number of customers $X_1+X_2$ just after a service completion, should equal the PGF of the steady-state number of customers, just after a service completion, in the $M/G/1$ queue with arrival rate $\lambda_1 + \lambda_2$ and generic service time $B$. Taking $z_1 = z_2 = z$ in (6) gives

$$F(z, z) = F(0, 0)\frac{(1 - z)\beta(z, z)}{\beta(z, z) - z}, \quad (8)$$

where substitution of $z = 1$ readily yields $F(0, 0) = 1-\rho$; all this is indeed in agreement with the classic result for the queue length in $M/G/1$, cf. Equation (II.4.17) of [4].

Now let us determine $F(0, z_2)$ from (6) (realizing that just plugging in $z_1 = 0$ in (6) gives a useless identity). For this, consider the factor $z_1 - \beta(z_1, z_2)$ in the lefthand side of that equation. Observe that, for any $z_2$ such that $|z_2| \leq 1$, $z_1 - \beta(z_1, z_2)$ has a unique zero $\mu(z_2)$ inside the unit circle; and that zero is given by

$$\mu(z_2) := \mathrm{E}\left(e^{-\lambda_2(1-z_2)P_1}\right), \quad (9)$$

where $P_1$ denotes the busy period in an $M/G/1$ queue with arrival rate $\lambda_1$ and generic service time $B$ (i.e., the system we study but without class-2 customers); cf. [4], p. 250. Inserting $z_1 = \mu(z_2)$ in (6) yields 0 in the lefthand side (as $F(z_1, z_2)$ is analytic for $|z_1| \leq 1, |z_2| \leq 1$), and hence the righthand side should equal 0 as well. Thus,

$$\frac{F(0, z_2)}{F(0, 0)} = \frac{\mu(z_2)z_2\left\{\phi_1\mu(z_2) + \phi_2 z_2 - 1\right\}}{[z_2 - \mu(z_2)]\beta(0, z_2)} + 1 - \phi_2 z_2. \quad (10)$$

Having $F(0, 0) = 1 - \rho$ provides $F(0, z_2)$. Going back to (6), we obtain $F(z_1, z_2)$; after some rewriting and simplification,

$$F(z_1, z_2) = \frac{1 - \rho}{z_1 - \beta(z_1, z_2)} \times \quad (11)$$
$$\left[\frac{(z_1 - z_2)\mu(z_2)}{z_2 - \mu(z_2)}\left(\phi_1\mu(z_2) + \phi_2 z_2 - 1\right) + \beta(z_1, z_2)\left(\phi_1 z_1 + \phi_2 z_2 - 1\right)\right].$$

Taking $z_2 = 1$ in (11) yields the marginal PGF of number of class-1 customers just after a departure. Introducing $\rho_i := \lambda_i \mathrm{E}B$ for $i = 1, 2$, we obtain

after some calculations (where we use that $\mu(1) = 1$ and $\mu'(1) = \lambda_2 \mathrm{E}P_1 = \lambda_2 \mathrm{E}B/(1-\rho_1) = \frac{\rho_2}{1-\rho_1}$):

$$F(z_1, 1) = \frac{(1-\rho)(1-z_1)}{\beta(z_1,1) - z_1}[\frac{\rho_1}{\rho}\beta(z_1,1) + \frac{\rho_2}{\rho}\frac{1}{1-\rho}], \tag{12}$$

or, after some rearranging,

$$F(z_1, 1) = \frac{(1-\rho_1)(1-z_1)}{\beta(z_1,1) - z_1}\beta(z_1,1)[\frac{\rho_1}{\rho}\frac{1-\rho}{1-\rho_1} + \frac{\rho_2}{\rho}\frac{1}{\beta(z_1,1)(1-\rho_1)}]. \tag{13}$$

As a consequence of this rearranging, the term outside the square brackets represents the PGF of the number of customers just after a service completion when $\lambda_2 = 0$, i.e., in an $M/G/1$ queue with arrival rate $\lambda_1$ and generic service time $B$, again cf. (II.4.17) of [4]. This decomposition is helpful in obtaining moments, because all queue length moments for that $M/G/1$ queue are known. Differentiation of $F(z_1, 1)$ gives, in particular,

$$\mathrm{E}X_1 = \frac{\lambda_1^2 \mathrm{E}B^2}{2(1-\rho_1)} + \phi_1 \frac{1-\rho}{1-\rho_1}\rho_1. \tag{14}$$

Observe that, compared to the standard $M/G/1$ queue with only class 1 (which would have the same first term but $\rho_1$ as second term) the mean is smaller. The explanation is that the presence of class-2 customer here is relevant for class-1 customers (unlike in an ordinary preemptive priority queue), and actually can help: the server may start serving when only class-2 customers are present, but the completed service is assigned to a class-1 customer if at least one of those has arrived in the meantime.
Also observe that the probability of having zero class-1 customers right after a departure equals

$$F(0, 1) = (1-\rho)[\frac{\rho_1}{\rho} + \frac{\rho_2}{\rho}\frac{1}{\beta(0,1)(1-\rho)}]. \tag{15}$$

We now turn to the determination of the sojourn time LST of class-1 customers. For this, introduce $Y_1$, the steady-state number of class-1 customers present immediately after the departure of a class-1 customer.The probability that a departure is of class 1 equals $\phi_1$, so that

$$F(z_1, 1) = \mathrm{E}z^{X_1} = \phi_1 \mathrm{E}z^{Y_1} + \phi_2. \tag{16}$$

Hence, after some rewriting (in order to get the PGF of the number of customers in the $M/G/1$ queue with only class-1 customers as the factor before the brackets):

$$\mathrm{E}z^{Y_1} = \frac{(1-\rho_1)(1-z)}{\beta(z,1)-z}\beta(z,1)\left[\frac{1-\rho}{1-\rho_1} + \frac{\rho_2}{1-\rho_1}\frac{\frac{1-\beta(z,1)}{(1-z)\rho_1}}{\beta(z,1)}\right]. \tag{17}$$

**Remark** The latter shows that $Y_1$ can can be composed as a sum of two random variables. The first is the number of customers in the $M/G/1$ queue with only class-1 customers, and the second is a mixture between a mass point at 0 and another random variable, with weights $\frac{1-\rho}{1-\rho_1}$ and $\frac{\rho_2}{1-\rho_1} = 1 - \frac{1-\rho}{1-\rho_1}$.

The distributional form of Little's law can be applied here to obtain the LST of the steady-state sojourn time of a class-1 customer: the class-1 customers present right after the departure of a class-1 customer are exactly those class-1 customers who have arrived during the sojourn time of that customer (because class-1 customers are served in order of their arrival). Hence we have $\mathrm{E}Y_1 = \lambda_1 \mathrm{E}S_1$ and $\mathrm{E}z^{Y_1} = \mathrm{E}\mathrm{e}^{-\lambda_1(1-z_1)S_1}$. The LST of the sojourn time $S_1$ thus follows from (17):

$$\mathrm{E}\mathrm{e}^{-\omega S_1} = \frac{(1-\rho_1)\omega \mathrm{E}\mathrm{e}^{-\omega B}}{\omega - \lambda_1 + \lambda_1 \mathrm{E}\mathrm{e}^{-\omega B}}\left[\frac{1-\rho}{1-\rho_1} + \frac{\rho_2}{1-\rho_1}\frac{\mathrm{E}\mathrm{e}^{-\omega B^{res}}}{\mathrm{E}\mathrm{e}^{-\omega B}}\right], \tag{18}$$

where $B^{res}$ denotes a residual service time, having LST $\mathrm{E}\mathrm{e}^{-\omega B^{res}} = \frac{1-\mathrm{E}\mathrm{e}^{-\omega B}}{\omega \mathrm{E}B}$; and

$$\mathrm{E}S_1 = \frac{(\lambda_1+\lambda_2)\mathrm{E}B^2}{2(1-\rho_1)} + \frac{1-\rho}{1-\rho_1}\mathrm{E}B. \tag{19}$$

The latter result is in agreement with Theorem 3.1 of [7].

We have written $\mathrm{E}\mathrm{e}^{-\omega S_1}$ such that, in the term before the square brackets, one may recognize (cf. p. 256 of [4]) the LST of the sojourn time $S_1(M/G/1)$ in an $M/G/1$ queue with only class-1 customers (i.e., $\lambda_2 = 0$):

$$\mathrm{E}\mathrm{e}^{-\omega S_1} = \mathrm{E}\mathrm{e}^{-\omega S_1(M/G/1)}\left[\frac{1-\rho}{1-\rho_1} + \frac{\rho_2}{1-\rho_1}\frac{\mathrm{E}\mathrm{e}^{-\omega B^{res}}}{\mathrm{E}\mathrm{e}^{-\omega B}}\right]. \tag{20}$$

The term within brackets is the LST of a mixture between 0 and a RV that can be positive as well as negative – and zero when $B$ is exponentially distributed, in which case $S_1$ is distributed as $S_1(M/M/1)$.
Rewriting (19), we see that

$$\mathrm{E}S_1 = \mathrm{E}S_1(M/G/1) + \frac{\rho_2}{1-\rho_1}[\mathrm{E}B^{res} - \mathrm{E}B]. \tag{21}$$

**Remark** We notice the seemingly paradoxical result that the presence of class-2 customers actually *shortens* the mean sojourn time of class-1 customers if $\mathrm{E}B^{res} < \mathrm{E}B$. The explanation is that sometimes no class-1 customer is present at the start of a service (where at least one class-2 customer is present), but still that job is claimed by a class-1 customer who arrives before the service is completed. Of course, the condition $\mathrm{E}B^{res} < \mathrm{E}B$ is equivalent to the coefficient of variation of $B$ being smaller than 1. Thus, we can say that the existence of class-2 customers is beneficial for class-1 customers if the service time has a smaller coefficient of variation than one, and harms them otherwise.

Let us now turn to the distribution of $X_2$ and the sojourn time of class-2 customers. Taking $z_1 = 1$ in (11) yields the marginal PGF of the number of class-2 customers just after a departure (notice that the first factor is the PGF of a nonnegative random variable):

$$F(1,z_2) = \frac{\rho_2(1-z_2)}{\beta(1,z_2)-1} \times \frac{1-\rho}{\rho_2}\left[\frac{\mu(z_2)}{z_2-\mu(z_2)}\left[\phi_1(1-\mu(z_2))+\phi_2(1-z_2)\right]+\phi_2\beta(1,z_2)\right]. \tag{22}$$

Differentiation gives us the mean number of class-2 customers just after a departure:

$$\mathrm{E}X_2 = \frac{\mathrm{E}B^2}{2}\left(\frac{\lambda^2}{1-\rho}-\frac{\lambda_1^2}{1-\rho_1}\right)+\phi_2\frac{(\rho_1+\rho(1-\rho_1))}{1-\rho_1}. \tag{23}$$

One could also have obtained this result by observing that $EX_1+EX_2$ equals $\frac{\lambda^2 EB^2}{2(1-\rho)}+\rho$, the known mean number of customers in an $M/G/1$ queue with arrival rate $\lambda$.

We next determine the LST of the sojourn time $S_2$ of class-2 customers. For that purpose, consider $Y_2$, the number of class-2 customers immediately after a class-2 departure. The only two terms in the righthand side of (5) that correspond to a class-2 departure are the third and the last term. Hence

$$\mathrm{E}z^{Y_2} = \frac{\frac{1}{z}\left[F(0,z)-F(0,0)\right]\beta(0,z)+F(0,0)\phi_2\beta(0,z)}{[F(0,1)-F(0,0)]\beta(0,1)+F(0,0)\phi_2\beta(0,1)}. \tag{24}$$

It can readily be verified, using (15), that the denominator of its righthand side equals $\phi_2$, as it should be; this is the probability that an arbitrary

departure is a class-2 departure. Substituting (10) into (24), we get

$$\mathrm{E}z^{Y_2} = \frac{(1-\rho)\mu(z)}{\mu(z)-z}[\frac{\lambda_1}{\lambda_2}(1-\mu(z))+1-z]. \tag{25}$$

We can also rewrite (25) as

$$\mathrm{E}z^{Y_2} = \frac{(1-\rho)\mu(z)(1-z)}{(1-\rho_1)(\mu(z)-z)}\left[\rho_1\frac{1-\mu(z)}{(1-z)\lambda_2\mathrm{E}P_1}+1-\rho_1\right] \tag{26}$$

The latter implies that $Y_2$ can be composed as a sum of two random variables. The first is the number of customers in the $M/G/1$ queue with arrival rate $\lambda_2$ and generic service time $P_1$ (so a service time which itself is a busy period of an $M/G/1$ queue with only class-1 customers) and the second is a mixture of the number of class-2 arrivals during the residual part (the overshoot) $P_1^{res}$ of a busy period $P_1$ (w.p. $\rho_1$) and 0 (w.p. $1-\rho_1$).
Having thus obtained an expression for $\mathrm{E}z^{Y_2}$, the distributional form of Little's law gives the LST of the sojourn time $S_2$ of class-2 customers:

$$\mathrm{E}\mathrm{e}^{-\omega S_2} = \mathrm{E}(1-\frac{\omega}{\lambda_2})^{Y_2} = \frac{(1-\rho)\omega\mathrm{E}[\mathrm{e}^{-\omega P_1}]}{\omega-\lambda_2+\lambda_2\mathrm{E}[\mathrm{e}^{-\omega P_1}]}\left[\rho_1\mathrm{E}\mathrm{e}^{-\omega P_1^{res}}+1-\rho_1\right], \tag{27}$$

where $\mathrm{E}\mathrm{e}^{-\omega P_1^{res}} = \frac{1-\mathrm{E}\mathrm{e}^{-\omega P_1}}{\omega\mathrm{E}P_1}$. Differentiation gives

$$\mathrm{E}S_2 = \frac{(\lambda_1+\lambda_2)\mathrm{E}B^2}{2(1-\rho_1)(1-\rho)}+\frac{\mathrm{E}B}{1-\rho_1}, \tag{28}$$

in agreement with Theorem 3.1 of [7].

**Remark** As already observed by Haviv and Kerner [7], the *conservation of work* principle holds: the mean amount of work in the stoppable $N$-class system equals the mean amount of work in the corresponding $M/G/1$ queue with arrival rate $\lambda$. This implies that a conservation law holds: Equating the mean amounts of work in our system and in the $M/G/1$ queue yields

$$\mathrm{E}B\sum_{i=1}^{N}\mathrm{E}X_i^{wait}+\rho\frac{\mathrm{E}B^2}{2\mathrm{E}B} = \frac{\lambda\mathrm{E}B^2}{2(1-\rho)}, \tag{29}$$

where $X_i^{wait}$ denotes the number of waiting class-$i$ customers. Using Little's law, one can replace $\mathrm{E}X_i^{wait}$ by $\lambda_i(\mathrm{E}S_i-\mathrm{E}B)$, and (29) then results in the

conservation law

$$\begin{aligned}\sum_{i=1}^{N} \rho_i \mathrm{E}S_i &= \frac{\lambda \mathrm{E}B^2}{2(1-\rho)} - \rho \frac{\mathrm{E}B^2}{2\mathrm{E}B} + \rho \mathrm{E}B \\ &= \rho[\frac{\lambda \mathrm{E}B^2}{2(1-\rho)} + \mathrm{E}B], \end{aligned} \tag{30}$$

which is indeed satisfied by (19) and (28). Observe that the term inside the brackets in (30) equals the mean sojourn time E$S$ in the $M/G/1$ queue with arrival rate $\lambda$ and generic service time $B$.

**Remark** We are also interested in the steady-state queue length PGF of class-$i$ customers at an *arbitrary* epoch, $i = 1, 2$. Just like in the ordinary $M/G/1$ queue, we have the following fundamental equalities in steady state (for $i = 1, 2$): The distribution of the number of class-$i$ customers just after a class-$i$ departure has the same distribution as the number of class-$i$ customers just before a class-$i$ arrival; and that also equals the distribution of the number of class-$i$ customers at an arbitrary epoch. The former equality follows by a step argument (just as many one-step increments as one-step decrements, in the long run, of class-$i$ customers); the latter equality follows by PASTA. Hence the marginal steady-state queue length PGF's are given by (17) and (25).

However, such a step argument does not hold in two dimensions, and the steady-state joint PGF of $Y_1$ and $Y_2$ does not easily follow from $F(z_1, z_2)$. We refer to [3] for a multi-class step argument in a particular multi-class $M/G/1$ queues *without* priorities.

### 3.3 The case $N = 3$

For the case of $N = 3$ customer classes, (4) simplifies to

$$\begin{aligned}
F(z_1,z_2,z_3) &= \frac{1}{z_1}\left[F(z_1,z_2,z_3)-F(0,z_2,z_3)\right]\beta(z_1,z_2,z_3)\\
&+ \left[F(0,z_2,z_3)-F(0,0,0)\right]\frac{\beta(z_1,z_2,z_3)-\beta(0,z_2,z_3)}{z_1}\\
&+ \frac{1}{z_2}\left[F(0,z_2,z_3)-F(0,0,z_3)\right]\beta(0,z_2,z_3)\\
&+ \left[F(0,0,z_3)-F(0,0,0)\right]\frac{\beta(0,z_2,z_3)-\beta(0,0,z_3)}{z_2}\\
&+ \frac{1}{z_3}\left[F(0,0,z_3)-F(0,0,0)\right]\beta(0,0,z_3)\\
&+ F(0,0,0)\phi_1\beta(z_1,z_2,z_3)\\
&+ \phi_2 z_2 F(0,0,0)\frac{\beta(z_1,z_2,z_3)-\beta(0,z_2,z_3)}{z_1}\\
&+ \phi_2 F(0,0,0)\beta(0,z_2,z_3)\\
&+ \phi_3 z_3 F(0,0,0)\frac{\beta(z_1,z_2,z_3)-\beta(0,z_2,z_3)}{z_1}\\
&+ \phi_3 z_3 F(0,0,0)\frac{\beta(0,z_2,z_3)-\beta(0,0,z_3)}{z_2}\\
&+ \phi_3 F(0,0,0)\beta(0,0,z_3). \qquad (31)
\end{aligned}$$

After rearranging and cancelling terms, (31) becomes

$$\begin{aligned}
&F(z_1,z_2,z_3)z_2z_3(z_1-\beta(z_1,z_2,z_3)) = F(0,z_2,z_3)\beta(0,z_2,z_3)z_3(z_1-z_2)\\
+\;&F(0,0,z_3)\beta(0,0,z_3)z_1(z_2-z_3)\\
+\;&F(0,0,0)\left\{\beta(z_1,z_2,z_3)z_2z_3\left(\phi_1z_1+\phi_2z_2+\phi_3z_3-1\right)\right.\\
+\;&\beta(0,z_2,z_3)z_3\left(z_2-z_1\right)\left(1-\phi_2z_2-\phi_3z_3\right)\\
+\;&\left.\beta(0,0,z_3)z_2z_3\left(-1+\phi_3(z_1-z_3)\right)\right\}. \qquad (32)
\end{aligned}$$

Now, as in the case where $N = 2$, we observe that for any $z_2, z_3$ such that $|z_2|, |z_3| \leq 1$, the function $z_1 - \beta(z_1, z_2, z_3)$ has a unique zero inside the unit disc $|z_1| \leq 1$. Furthermore, this unique zero equals $\mu_1(z_2, z_3)$, with

$$\mu_1(z_2,z_3) = \mathrm{E}\left(e^{-(\lambda_2(1-z_2)+\lambda_3(1-z_3))P_1}\right), \qquad (33)$$

with $P_1$ again denoting a busy period of the $M/G/1$ queue with arrival rate $\lambda_1$ and generic service time $B$. Now, for $z_1 = \mu_1(z_2, z_3)$, the lefthand side of (32) equals 0, and hence the righthand side as well. Thus, for $z_1 = \mu_1(z_2, z_3)$, we have

$$\begin{aligned} & F(0, z_2, z_3)\beta(0, z_2, z_3)z_3(z_2 - \mu_1(z_2, z_3)) \qquad (34) \\ = \;& F(0, 0, z_3)\beta(0, 0, z_3)\mu_1(z_2, z_3)(z_2 - z_3) \\ + \;& F(0, 0, 0)\left\{\beta(\mu_1(z_2, z_3), z_2, z_3)z_2z_3(\phi_1\mu_1(z_2, z_3) + \phi_2 z_2 + \phi_3 z_3 - 1)\right. \\ + \;& \beta(0, z_2, z_3)z_3(z_2 - \mu_1(z_2, z_3))(1 - \phi_2 z_2 - \phi_3 z_3) \\ + \;& \left.\beta(0, 0, z_3)z_2z_3(-1 + \phi_3(\mu_1(z_2, z_3) - z_3))\right\}. \end{aligned}$$

Define $\mu_2(z_3) := \mathrm{E}\left(e^{-\lambda_3(1-z_3)P_2}\right)$, with $P_2$ a busy period of the $M/G/1$ queue with arrival rate $\lambda_2$ and generic service time $P_1$ (i.e., itself a busy period of an $M/G/1$ queue). Now, observe that for any $z_3$ such that $|z_3| \leq 1$, the function $z_2 - \mu_1(z_2, z_3)$ has a unique root $z_2$ within the unit disc $|z_2| \leq 1$. Furthermore, this root equals $\mu_2(z_3)$. So, for $z_2 = \mu_2(z_3)$, the lefthand side of (34) equals 0, and hence the righthand side equals 0 as well. This determines $F(0, 0, z_3)$, as we already know that $F(0, 0, 0) = 1 - \rho$. Inserting the thus obtained expression for $F(0, 0, z_3)$ into (34) provides $F(0, z_2, z_3)$. Inserting the values of $F(0, z_2, z_3)$ and $F(0, 0, z_3)$ into (32) provides $F(z_1, z_2, z_3)$.

## 3.4 The case of general $N$

Having described in some detail how $F(z_1, z_2, z_3)$ can be determined for $N = 3$, we are now ready to outline the procedure for determining $F(z_1, z_2, \ldots, z_N)$ for general $N$. We remind the reader that $z[i] = (0, \ldots, 0, z_{i+1}, \ldots, z_N)$ with 0 as its first $i$ elements, so that $F(z[i]) = F(0, \ldots, 0, z_{i+1}, \ldots, z_N)$.
*Step 0:* Move the term $F(z[0]) = F(\underline{z})$ in (4) from the righthand side to the lefthand side, and multiply by $\prod_{i=1}^{N} z_i$. Simplify the righthand side by observing that terms $F(z[i])\beta(z[i-1])$ cancel in the numerator. We thus arrive at

$$\begin{aligned} & F(\underline{z})(z_1 - \beta(\underline{z}))\prod_{j=2}^{N} z_j = \sum_{i=1}^{N-1} F(z[i])\beta(z[i])(z_i - z_{i+1}) \prod_{j\neq i,i+1}^{N} z_j \\ - \;& F(\underline{0})\sum_{i=1}^{N} \frac{\beta(z[i-1]) - \beta(z[i])}{z_i}\prod_{j=1}^{N} z_j + F(\underline{0})L(\underline{z})\prod_{j=1}^{N} z_j, \qquad (35) \end{aligned}$$

where $F(\underline{0})L(\underline{z})$ is given by the last line of (4).
*Step 1:* Observe that in the lefthand side the term $z_1 - \beta(z_1, \ldots, z_N)$ has for all $|z_i| \leq 1$, $i = 2, \ldots, N$, a unique zero inside the unit disc $|z_1| \leq 1$. This unique zero equals $\mu_1(z_2, \ldots, z_N)$, with

$$\mu_1(z_2, \ldots, z_N) = \mathrm{E}(\mathrm{e}^{-\sum_{i=2}^{N} \lambda_i (1-z_i) P_1}); \tag{36}$$

$P_1$ again denotes a busy period of the $M/G/1$ queue with arrival rate $\lambda_1$ and generic service time $B$. For $z_1 = \mu_1(z_2, \ldots, z_N)$, the righthand side should also be zero. That results in

$$\begin{aligned}
& F(z[1])\beta(z[1])(z_2 - \mu_1(z_2, \ldots, z_N)) \prod_{i=3}^{N} z_i \qquad (37)\\
= & \sum_{i=2}^{N-1} F(z[i])\beta(z[i])(z_i - z_{i+1})\mu_1(z_2, \ldots, z_N) \prod_{j \neq 1,i,i+1}^{N} z_j \\
- & F(\underline{0}) \left[ \frac{\beta(\mu_1(z_2, \ldots, z_N), z_2, \ldots, z_N) - \beta(z[1])}{\mu_1(z_2, \ldots, z_N)} + \sum_{i=2}^{N} \frac{\beta(z[i-1]) - \beta(z[i])}{z_i} \right] \\
\times & \mu_1(z_2, \ldots, z_N) \prod_{j=2}^{N} z_j \\
+ & F(\underline{0}) L(\underline{z}^{(1)}) \mu_1(z_2, \ldots, z_N) \prod_{j=2}^{N} z_j,
\end{aligned}$$

where $\underline{z}^{(1)} = (\mu_1(z_2, \ldots, z_N), z_2, \ldots, z_N)$.
*Step 2:* Observe that the term $z_2 - \mu_1(z_2, \ldots, z_N)$ in this lefthand side has for all $|z_i| \leq 1$, $i = 3, \ldots, N$, a unique zero inside the unit disc $|z_2| \leq 1$. This unique zero equals $\mu_2(z_3, \ldots, z_N)$, with

$$\mu_2(z_3, \ldots, z_N) = \mathrm{E}(\mathrm{e}^{-\sum_{i=3}^{N} \lambda_i (1-z_i) P_2}); \tag{38}$$

$P_2$ denotes a busy period of the $M/G/1$ queue with arrival rate $\lambda_2$ and generic service time $P_1$.
For $z_2 = \mu_2(z_3, \ldots, z_N)$, the righthand side should also be zero. That results in an equation with terms involving $F(z[i])$, $i = 3, \ldots, N$, in the righthand side, and as lefthand side
$F(z[2])\beta(z[2])(z_3 - \mu_2(z_3, \ldots, z_N))\mu_1(\mu_2(z_3, \ldots, z_N), z_3, \ldots, z_N) \prod_{i=4}^{N} z_i$.

*Step j:* Continuing like this, in step $j$, $j = 3, \ldots, N-1$, the term $z_j - \mu_{j-1}(z_j, \ldots, z_N)$ has a unique zero $\mu_j(z_{j+1}, \ldots, z_N)$ with

$$\mu_j(z_{j+1}, \ldots, z_N) = \mathrm{E}(\mathrm{e}^{-\sum_{i=j+1}^{N} \lambda_i (1-z_i) P_j}); \tag{39}$$

$P_j$ denotes a busy period of an $M/G/1$ queue with arrival rate $\lambda_j$ and generic service time $P_{j-1}$, the busy period in an $M/G/1$ queue with arrival rate $\lambda_{j-1}$ and generic service time $P_{j-2}$.

This continues until, in step $N-1$, we thus express $F(z[N]) = F(0, \ldots, 0, z_N)$ in $F(0, \ldots, 0) = 1 - \rho$. Thus $F(z[N])$ is known; and we subsequently work our way backwards, expressing $F(z[N-1])$ into $F(z[N])$, then $F(z[N-2])$ into $F(z[N-1])$ and $F(z[N])$, etc. Finally, $F(z_1, \ldots, z_N)$ is obtained. From there we can determine performance measures like $\mathrm{E}(z^{Y_i})$ and, by the distributional form of Little's law, the sojourn time LST of class-$i$ customers.

# 4 The unstoppable server model

In this section we again consider the single-server model of $N$ customer classes with independent Poisson($\lambda_i$) arrival processes, $i = 1, \ldots, N$, and generic service time $B$ for all classes. However, there is one difference with the stoppable case: when the system becomes empty, the (unstoppable) server does not idle but immediately begins a new service, with generic service time $B$. At service completion there are $N+1$ options. Option (i): At least one class-1 customer is present; the just completed job is for the longest waiting class-1 customer, who immediately leaves. Option (ii): No class-1 customer is present, but at least one class-2 customer; the just completed job is for the longest waiting class-2 customer, who immediately leaves; etc. for options (iii)-($N$). Option ($N+1$): There are no customers present. Then the just completed job is scrapped. The stability condition for this system obviously again is $\rho = \sum_{i=1}^{N} \lambda_i \mathrm{E}B < 1$, which we assume to hold. We again write $\underline{z} := (z_1, ..., z_N)$ and we denote the vector with the first $i$ components being 0, and the rest $z_{i+1}, ...z_N$, by $z[i]$.

## 4.1 Steady-state distribution

Let $U(\underline{z}) = U(z_1, \ldots, z_N)$ denote the PGF of the steady-state vector of numbers of customers of classes $1, \ldots, N$ immediately after a service completion,

in the unstoppable case. In the same way as (4) was derived, we obtain for $U(\underline{z})$:

$$U(\underline{z}) = \sum_{i=1}^{N} \frac{(U(z[i-1]) - U(z[i]))\,\beta(z[i-1]) + (U(z[i]) - U(\underline{0}))\,(\beta(z[i-1]) - \beta(z[i]))}{z_i}$$

$$+U(\underline{0}) \left[ \sum_{i=1}^{N} \frac{\beta(z[i-1]) - \beta(z[i])}{z_i} + \beta(\underline{0}) \right]. \tag{40}$$

The first line of (40) is exactly the same as the first line of (4), and that makes sense: the stoppable and unstoppable policy only differ when a service completion leaves the system empty, and that case is covered by the second line of (4) (for stoppable) and the seond line of (40) (for unstoppable). Just like in (4), the $i$-th element of the first line in (40) refers to a departure of a class-$i$ customer. Within it, the first term represents the event that the highest prioritized customers left behind at the last departure are from class $i$, and there were no higher priority arrivals during the service. The second term represents the event that the system is not empty at the last departure, those present are from classes less prioritized than $i$, and the most prioritized customers who arrived during the service are from class $i$. In the second line of (40), the $i$-th term represents the event that a class-$i$ customer is the highest-priority customer who has arrived during a service, $i = 1, \ldots, N$ (which actually is a service that started in an empty system); and the very last term corresponds to no arrival during that service (and hence the job is scrapped). It is readily seen that some terms cancel against each other. However, the present set-up allows us a systematic collection of all terms. In Subsection 4.2 we present the solution for $U(\underline{z})$ for the special case $N = 2$ in much detail, and then we shall simplify (40); Subsection 4.3 more globally considers $N = 3$, after which we are ready to present the structure for the general case, in Subsection 4.4.

## 4.2 The case of $N = 2$

For the case of $N = 2$ classes, (40) simplifies to

$$\begin{aligned} U(z_1,z_2) &= \frac{1}{z_1}\left[U(z_1,z_2)-U(0,z_2)\right]\beta(z_1,z_2) \\ &+ \left[U(0,z_2)-U(0,0)\right]\frac{\beta(z_1,z_2)-\beta(0,z_2)}{z_1} \\ &+ \frac{1}{z_2}\left[U(0,z_2)-U(0,0)\right]\beta(0,z_2) \\ &+ U(0,0)\left[\frac{\beta(z_1,z_2)-\beta(0,z_2)}{z_1}+\frac{\beta(0,z_2)-\beta(0,0)}{z_2}+\beta(0,0)\right]. \end{aligned} \tag{41}$$

In comparison with the six terms of the righthand side of Equation (5), the first three ones are the same, but the three terms corresponding to an empty system after a service completion are different. Multiplication by $z_1z_2$ and regrouping gives us

$$U(z_1,z_2)[z_1-\beta(z_1,z_2)]z_2 = U(0,z_2)(z_1-z_2)\beta(0,z_2)+U(0,0)z_1(z_2-1)\beta(0,0). \tag{42}$$

$U(z,z)$, the PGF of the steady-state total number of customers $X_1+X_2$ just after a service completion, is given by

$$U(z,z) = U(0,0)\beta(0,0)\frac{(1-z)}{\beta(z,z)-z}, \tag{43}$$

where substitution of $z = 1$ yields that $U(0,0) = (1-\rho)/\beta(0,0)$. Notice that the PGF of the number of customers in an ordinary $M/G/1$ queue with arrival rate $\lambda$ and generic service time $B$, in which the server *only works when there is work*, is given by $U(z,z)\beta(z,z)$.

We next determine $U(0,z)$ from (42). Just like in the stoppable case, we exploit the fact that the factor $z_1-\beta(z_1,z_2)$ in the lefthand side of (42) has a unique zero $z_1 = \mu(z_2) = \mathrm{E}e^{-\lambda_2(1-z_2)P_1}$ for every $z_2$ with $|z_2| \le 1$. The righthand side of (42) should also be zero for such $z_1 = \mu(z_2)$, and hence

$$U(0,z_2) = (1-\rho)\frac{\mu(z_2)(1-z_2)}{(\mu(z_2)-z_2)\beta(0,z_2)}; \tag{44}$$

furthermore, the probability of having no class-1 customers at the end of a service is $U(0,1) = (1-\rho_1)/\beta(0,1)$. Substituting (44) into (42) we obtain,

after some rewriting and simplifying:

$$U(z_1, z_2) = \frac{(1-\rho)(1-z_2)}{z_1 - \beta(z_1, z_2)} \frac{z_1 - \mu(z_2)}{\mu(z_2) - z_2}. \tag{45}$$

The marginal PGF of number of class-1 customers just after a service completion follows by taking $z_2 = 1$ in (45) and using that $\mu(1) = 1$ and $\mu'(1) = \rho_2/(1-\rho_1)$:

$$U(z_1, 1) = \frac{(1-\rho_1)(1-z_1)}{\beta(z_1, 1) - z_1}. \tag{46}$$

It should be observed that $\lambda_2$ does not appear in this expression; $\mathrm{E}z^{X_1}$ is not at all affected by the presence of type-2 customers (as long as $\rho_2 < 1 - \rho_1$). Differentiation yields

$$\mathrm{E}X_1 = \frac{\lambda_1^2 \mathrm{E}B^2}{2(1-\rho_1)}. \tag{47}$$

Comparison with (14) for the stoppable case shows that the mean number of class-1 customers just after a service completion is smaller in the unstoppable case.

Again like in the stoppable case, we want to relate $X_1$ to $Y_1$, the steady-state number of class-1 customers present immediately after the departure of a class-1 customer (cf. (16)). This time, the probability that a departure is of class-1 does not equal $\phi_1$ but $\rho_1$ (because a job is scrapped if the system is empty at the end of a service completion)

$$U(z_1, 1) = \mathrm{E}z^{X_1} = \rho_1 \mathrm{E}z^{Y_1} + 1 - \rho_1. \tag{48}$$

Hence

$$\mathrm{E}z^{Y_1} = \frac{1-\rho_1}{\rho_1} \frac{1 - \beta(z, 1)}{\beta(z, 1) - z}. \tag{49}$$

The distributional form of Little's law subsequently gives the LST of the sojourn time $S_1$:

$$\mathrm{E}\mathrm{e}^{-\omega S_1} = \frac{(1-\rho_1)\omega}{\omega - \lambda_1 + \lambda_1 \mathrm{E}\mathrm{e}^{-\omega B}} \frac{1 - \mathrm{E}\mathrm{e}^{-\omega B}}{\omega \mathrm{E}B}. \tag{50}$$

Recognizing the expression for the LST's of the waiting time $W_1(M/G/1)$ and sojourn time $S_1(M/G/1)$ in an ordinary $M/G/1$ queue with only class-1

customers (i.e., $\lambda_2 = 0$ and the server only works when there is work), we can write

$$\mathrm{E}e^{-\omega S_1} = \mathrm{E}e^{-\omega W_1(M/G/1)}\mathrm{E}e^{-\omega B^{res}} = \mathrm{E}e^{-\omega S_1(M/G/1)}\frac{\mathrm{E}e^{-\omega B^{res}}}{\mathrm{E}e^{-\omega \mathrm{B}}}. \tag{51}$$

Observe that

$$\mathrm{E}S_1 = \mathrm{E}S_1(M/G/1) + \mathrm{E}B^{res} - \mathrm{E}B. \tag{52}$$

Unlike the stoppable case, the presence of class-2 customers does not affect the mean sojourn time of class-1 customers; but the fact that the server always works does affect this mean (in a positive way if the coefficient of variation of $B$ is smaller than one). We can rewrite (52) into

$$\mathrm{E}S_1 = \frac{1}{1-\rho_1}\frac{\mathrm{E}B^2}{2\mathrm{E}B}, \tag{53}$$

which is in agreement with Theorem 4.3 of [7].

We now turn to the distribution of $X_2$ and the sojourn time $S_2$ of class-2 customers. Taking $z_1 = 1$ in (45) yields the marginal PGF of the number of class-2 customers just after a service completion:

$$U(1, z_2) = \frac{(1-\rho)(1-z_2)}{1-\beta(1,z_2)}\frac{1-\mu(z_2)}{\mu(z_2)-z_2}. \tag{54}$$

In particular, $U(1,0) = \frac{1-\rho}{1-\beta(1,0)}\frac{1-\mu(0)}{\mu(0)}$, and

$$\mathrm{E}X_2 = \frac{\lambda_2 \mathrm{E}B^2}{2\mathrm{E}B}[\frac{1}{(1-\rho)(1-\rho_1)} - 1]. \tag{55}$$

We next consider $Y_2$, the number of class-2 customers immediately after a class-2 departure. The only two terms in the righthand side of (41) that correspond to a class-2 departure are the third and the fifth term. Hence

$$\begin{aligned} \mathrm{E}z^{Y_2} &= \frac{\frac{U(0,z)-U(0,0)}{z}\beta(0,z) + U(0,0)\frac{\beta(0,z)-\beta(0,0)}{z}}{[U(0,1)-U(0,0)]\beta(0,1) + U(0,0)[\beta(0,1)-\beta(0,0)]} \\ &= \frac{U(0,z_2)\beta(0,z_2)/z_2 - U(0,0)\beta(0,0)/z_2}{U(0,1)\beta(0,1) - U(0,0)\beta(0,0)}. \end{aligned} \tag{56}$$

Simplifying the denominator to $\rho_2$ (which indeed is the probability that a service completion corresponds to a class-2 departure) and substituting (44), we obtain

$$\mathrm{E}z^{Y_2} = \frac{1-\rho}{\rho_2}\frac{1-\mu(z)}{\mu(z)-z}. \tag{57}$$

To this equation we can apply the distributional form of Little's law, which here amounts to $\mathrm{E}z^{Y_2} = \mathrm{E}\mathrm{e}^{-\lambda_2(1-z_2)S_2}$. Hence, realizing that $\mu(1-\omega/\lambda_2) = \mathrm{E}\mathrm{e}^{-\omega P_1}$ with $P_1$ the busy period in an $M/G/1$ queue with arrival rate $\lambda_1$ and generic service time $B$:

$$\begin{aligned}\mathrm{E}\mathrm{e}^{-\omega S_2} &= \frac{1-\rho}{\rho_2}\frac{1-\mathrm{E}\mathrm{e}^{-\omega P_1}}{\mathrm{E}\mathrm{e}^{-\omega P_1}-(1-\frac{\omega}{\lambda_2})} \\ &= \frac{1-\rho}{1-\rho_1}\frac{\omega}{\omega-\lambda_2+\lambda_2\mathrm{E}\mathrm{e}^{-\omega P_1}}\frac{1-\mathrm{E}\mathrm{e}^{-\omega P_1}}{\omega \mathrm{E}P_1}. \end{aligned} \tag{58}$$

The latter expression shows that $S_2$ can be written as the sum of two independent random variables. The first one is the waiting time in an $M/G/1$ queue with arrival rate $\lambda_2$ and generic service time a busy period $P_1$ of another $M/G/1$ queue, with arrival rate $\lambda_1$ and generic service time $B$; and the second one is the residual of $P_1$. Either from this observation, or by differentiating (58), we obtain the following expression for the mean sojourn time of class-2 customers:

$$\mathrm{E}S_2 = \frac{\mathrm{E}B^2}{2\mathrm{E}B(1-\rho)(1-\rho_1)}, \tag{59}$$

in agreement with Theorem 4.3 of [7].

**Remark** Just as in the stoppable case, $Y_i$ is also distributed as the steady-state number of class-$i$ customers just before a class-$i$ arrival, and as the steady-state number of class-$i$ customers at an arbitrary epoch ($i = 1, 2$).

**Remark** In Equation (9) of [7] it is shown, for general $N$, that one has the following conservation law for a weighted sum of the mean sojourn times:

$$\sum_{i=1}^{N} \rho_i \mathrm{E}S_i = \rho[\frac{\lambda \mathrm{E}B^2}{2(1-\rho)} + \mathrm{E}B^{res}]. \tag{60}$$

This should be compared with (30) for the stoppable case. Our expressions in (53) and (59) (for $\mathrm{E}S_1$ and $\mathrm{E}S_2$ in the $N = 2$ case) indeed satisfy the conservation law (60).

## 4.3 The case $N = 3$

For the case of $N = 3$ customer classes, (40) simplifies to

$$\begin{aligned} U(z_1,z_2,z_3) &= \frac{1}{z_1}\left[U(z_1,z_2,z_3)-U(0,z_2,z_3)\right]\beta(z_1,z_2,z_3) \\ &+ \left[U(0,z_2,z_3)-U(0,0,0)\right]\frac{\beta(z_1,z_2,z_3)-\beta(0,z_2,z_3)}{z_1} \\ &+ \frac{1}{z_2}\left[U(0,z_2,z_3)-U(0,0,z_3)\right]\beta(0,z_2,z_3) \\ &+ \left[U(0,0,z_3)-U(0,0,0)\right]\frac{\beta(0,z_2,z_3)-\beta(0,0,z_3)}{z_2} \\ &+ \frac{1}{z_3}\left[U(0,0,z_3)-U(0,0,0)\right]\beta(0,0,z_3) \\ &+ U(0,0,0)[\frac{\beta(z_1,z_2,z_3)-\beta(0,z_2,z_3)}{z_1}+\frac{\beta(0,z_2,z_3)-\beta(0,0,z_3)}{z_2} \\ &+ \frac{\beta(0,0,z_3)-\beta(0,0,0)}{z_3}+\beta(0,0,0)]. \end{aligned} \tag{61}$$

After rearranging and cancelling terms, (61) becomes

$$\begin{aligned} & U(z_1,z_2,z_3)z_2z_3(z_1-\beta(z_1,z_2,z_3)) = U(0,z_2,z_3)\beta(0,z_2,z_3)z_3(z_1-z_2) \\ + \;& U(0,0,z_3)\beta(0,0,z_3)z_1(z_2-z_3) \\ + \;& U(0,0,0)\beta(0,0,0)z_1z_2(z_3-1). \end{aligned} \tag{62}$$

The solution procedure is exactly the same as in Section 3.3 for the stoppable case with $N = 3$. Consider the unique zero $z_1 = \mu_1(z_2, z_3)$ of $z_1-\beta(z_1, z_2, z_3)$. Observe that the righthand side of (62) is zero for such $z_1$, yielding a relation in which $U(0, z_2, z_3)$ is expressed in $U(0, 0, z_3)$ and $U(0, 0, 0)$. Next consider the unique zero $z_2 = \mu_2(z_3)$ of $z_2-\mu_1(z_2, z_3)$. Observe that the righthand side of the above-mentioned relation is zero for such $z_2$, yielding a relation in which $U(0, 0, z_3)$ is expressed in $U(0, 0, 0)$, which itself equals $(1-\rho)/\beta(0, 0, 0)$. The latter follows either from the normalization condition $U(1, 1, 1) = 1$ or by the probabilistic argument that a fraction $1 - \rho$ of the jobs is wasted, while the waste probability also equals $U(0, 0, 0)\beta(0, 0, 0)$.

## 4.4 The case of general $N$

We remind the reader that the only difference between the righthand sides of (4) and (40) is in their last line, featuring $F(\underline{0})$ versus $U(\underline{0})$. That implies that

the $N-1$ steps, as outlined in Section 3.4, can be copied until we finally arrive at a formula that expresses $U(z[N]) = U(0,\ldots,0,z_N)$ into $U(0,\ldots,0) = (1-\rho)/\beta(0,\ldots,0)$. Thus $U(z[N])$ is known, and we subsequently work our way backwards, expressing $Uz[N-1])$ into $U(z[N])$, then $U(z[N-2])$ into $U(z[N-1])$ and $U(z[N])$, etc., until finally $U(z_1,\ldots,z_N)$ is obtained. From there one can again determine various performance measures like the sojourn time LST's.

# 5 Closing remarks

A straightforward generalization of the stoppable and unstoppable model is the following. Consider a policy such that whenever a service is completed and the system is empty, the server stops with probability $q$ and does not stop with probability $1-q$. It is readily seen that the PGF $H(\underline{z})$ of the joint steady-state distribution of the numbers of customers in the various classes satisfies an equation that is basically the same as Equations (4) and (40): $H(\underline{z})$ is expressed as the sum of three terms, in which the first term is the same as the first term in the righthand side of both (4) and (40) (of course with $F(\cdot)$ respectively $U(\cdot)$ replaced by $H(\cdot)$), while the second (resp. third) term is $qH(\underline{0})$ (resp. $(1-q)H(\underline{0})$) times the second term, in large brackets, in the righthand side of (4) (resp. (40)). For its analysis, follow the same sequential process as described in Subsection 3.4 (and briefly in Subsection 4.4). Because the first term in the righthand side is the same as in (4) and (40), one thus arrives, in exactly the same way as for the stoppable and unstoppable variants, at a relation between $H(0,\ldots,0,z_N)$ and $H(0,\ldots,0)$; and the mixture of elements related to an empty system at service completion replaces the associated elements in (4) (and in (40)) also in the solution. Also, the fraction of service completions of class $i$ of course remains $\rho_i$, $i = 1, ..., N$. Thus, the mean sojourn time for each class in this mixed model is a weighted average of the means in the stoppable and unstoppable model, with weights $q$ and $1-q$, respectively. The optimization problem of choosing the value of $q \in [0,1]$ that minimizes the mean sojourn time of an arbitrary customer thus also becomes trivial: simply choose $q = 1$ when $\frac{\mathrm{E}B^2}{2\mathrm{E}B} - \mathrm{E}B > 0$, and otherwise choose $q = 0$. This follows from the two conservation laws (30) and (60), which in combination with the above imply that the mean sojourn

time of an arbitrary customer in the mixed case is given by

$$\sum_{i=1}^{N} \frac{\lambda_i}{\lambda} \mathrm{E}S_i = \sum_{i=1}^{N} \frac{\rho_i}{\rho} \mathrm{E}S_i = \frac{\lambda \mathrm{E}B^2}{2(1-\rho)} + q\mathrm{E}B + (1-q)\mathrm{E}B^{res}. \qquad (63)$$

**Acknowledgment**. The research of Onno Boxma was supported by the NWO Gravitation project NETWORKS, grant number 024.002.003.

# References

[1] H. Albrecher, O. Boxma, R. Essifi and R. Kuijstermans (2017). A queueing model with randomized depletion of inventory. *Probability in the Engineering and Informational Sciences* **31**, 43-59.

[2] O.J. Boxma, R. Essifi and A.J.E.M. Janssen (2016). A queueing/inventory and an insurance risk model. *Advances in Applied Probability* **48**, 1139-1160.

[3] O.J. Boxma and T. Takine (2003). The $M/G/1$ FIFO queue with several customer classes. *Queueing Systems* **45**, 185-189.

[4] J.W. Cohen (1982). *The Single Server Queue*. North-Holland Publ. Cy., Amsterdam; 2nd ed.

[5] M. Harchol-Balter (2013). *Performance Modeling and Design of Computer Systems – Queueing Theory in Action.* Cambridge University Press, New York.

[6] M. Haviv (2013). *Queues – A Course in Queueing Theory*. Springer, New York.

[7] M. Haviv and Y. Kerner (2022). Queueing with priorities and standard service: Stoppable and unstoppable servers. *Stochastic Models* **38**, 503-514.

[8] N.K. Jaiswal (1968). *Priority Queues*. Academic Press, New York.

[9] L. Kleinrock (1976). Queueing Systems, Vol. 2: Computer Applications.

[10] H. Takagi (1991). *Queueing Analysis: A Foundation of Performance Evaluation. Volume 1: Vacation and Priority Systems.* North-Holland Publ. Cy., Amsterdam.