

Strategies for a Centralized Single Product Multi-Class $M/G/1$ Make-to-Stock Queue

Hossein Abouee-Mehrizi[†], Barış Balcıoğlu[§], Opher Baron[†]

[†] University of Toronto,
Joseph L. Rotman School of Management
105 St. George Street, Toronto, M5S 3E6, CANADA,
H.Abouee07@Rotman.Utoronto.Ca, Opher.Baron@Rotman.Utoronto.Ca

[§] University of Toronto,
Department of Mechanical and Industrial Engineering
5 King's College Rd., Toronto, ON M5S 3G8, CANADA,
baris@mie.utoronto.ca

Abstract

Make-to-stock queues are typically investigated in the $M/M/1$ settings. For centralized single-item systems, the Inventory Rationing (IR) policy is established as optimal and the Strict Priority (SP) policy is shown to be a practical compromise, balancing cost and ease of implementation. However, when service time is general, i.e., for $M/G/1$ queues, the optimal policy is still unknown. In these settings, we derive the optimal cost and control for the IR and SP policies. Dynamic programming, the common tool used in investigating the IR policy in make-to-stock queues, is less practical when service time is general. We, in contrast, focus on the *customer composition*: the proportion of customers of each class to the total number of customers in the queue. This focus is beneficial because the number of customers in $M/G/1$ queues is invariant for any non-idling and non-anticipating policy. To characterize customer composition we consider a series of two-priority $M/G/1$ queues where their first service time in each busy period is different from standard service times, i.e., this first service time is exceptional. We characterize the required exceptional first service times and the exact solution of such queues.

Keywords and Phrases: Make-to-Stock, $M/G/1$ queue, priority classes, customer composition, inventory rationing, strict priority policy

1 Introduction

Market segmentation and customer differentiation are widely accepted as ways to increase profitability. A common approach in differentiating among customers is to provide different service levels for different customer classes. For example, in a make-to-stock system, service level is often measured by product availability on the shelf. In this case, the service level is directly influenced by allocation policies and inventory levels. Important research and managerial questions are whether customer classes requesting the same product should be prioritized and if so how to prioritize them? We analyze inventory control strategies for a supplier that keeps a centralized inventory to serve a single product to different classes of customers. We model the underlying production system as an $M/G/1$ queue.

Production and inventory control of such systems can be managed using many policies. Broadly, inventory control policies can be characterized by whether different customer types are prioritized, and whether allocation decisions are made when production starts or are postponed until production is completed. In this paper, we focus on postponing the allocation decision since it reduces system costs compared to not postponing.

Accordingly, and similar to earlier literature, we consider the following three centralized inventory control policies that use a base-stock level control for their production decision:

FCFS Under a *First Come First Served* policy, completed items are given to the customer whose order has waited the longest time in the system. The only decision is with respect to the production control, i.e., choosing the base-stock level that minimizes the operation cost of the system. In the $M/G/1$ settings, this policy and its optimal base-stock level can be determined following Sanajian and Balcioglu (2009).

IR Policy Under an *Inventory Rationing* policy, there are non-decreasing threshold inventory levels R_r , $r = 1, \dots, n + 1$ with $R_1 = 0$ and $R_{n+1} = S$. If the inventory level, I , is between R_{r+1} and $R_r + 1$ i.e., $R_r < I \leq R_{r+1}$, only demand requests of classes 1 to r are satisfied on an FCFS basis. When there is no positive-stock, a finished product is

allocated to the highest-priority customer backlogged (in an FCFS fashion within this class). If the inventory level is between R_{r+1} and $R_r + 1$, even if there are pending orders from classes $r + 1$ to n , the completed product is placed in inventory. When the inventory reaches R_{n+1} , the base-stock level, production stops. de Véricourt et al. (2002) provide the optimal rationing levels and the system cost for this policy for the $M/M/1$ queueing systems, and Gayon et al. (2009) propose a heuristic to approximate these levels for systems with Erlangian service times.

SP Policy The *Strict Priority* policy is a special case of the IR policy when $R_1 = R_2 = \dots = R_n = 0$. That is, under this policy, as long as there is stock in the centralized inventory, demand requests are satisfied on an FCFS basis. When there are backlogs, a finished product is allocated to the highest-priority customer among those with pending orders in the system. For the $M/M/1$ queue, de Véricourt et al. (2001) provide the optimal base-stock level and optimal cost for this policy.

Remark 1 *The only information required to evaluate the system's control and cost under any of these three policies is the number of customers of different types in the system. These policies ignore additional information such as the length of time since the production of the current item began. While for the $M/M/1$ settings, such additional information is useless (due to the memoryless property), it might be both available and valuable in $M/G/1$ settings. For example, both Ha (2000) and Gayon et al. (2009) consider Erlangian service times and use information on production status. Indeed, not using additional information might increase the costs of these policies relative to the optimal control policy; however, this keeps their implementation simple and increases their practicality.*

We focus on centralized systems because inventory pooling and its potential benefits have been established in the literature. The first paper to investigate inventory pooling is Eppen (1979). Recently, Benjaafar et al. (2005) compare the decentralized with centralized FCFS systems. For discussion of the advantages of pooling when supply is ample i.e., unlimited, see Gerchack and He (2003) and references therein.

The first papers to discuss an inventory rationing problem in a centralized make-to-stock system with different classes of customers are by Ha (1997a and 1997b). For exponentially distributed production times, Poisson arrivals and lost sales, Ha (1997a) shows that the inventory rationing (IR) policy is optimal. Ha (1997b) extends this work to the backlog case with two classes of customers and shows that a stationary critical-level policy is optimal. de Véricourt et al. (2002) show that the IR policy is the optimal policy for the $M/M/1$ make-to-stock queues. Ha (2000) considers an $M/E_k/1$ make-to-stock system with lost sales, where E_k denotes k -stage Erlang service time, and characterizes the optimal stock allocation policy. Applications of rationing levels have been also investigated when supply is ample; see Arslan et al. (2007) and references therein. Benjaafar et al. (2007) investigate a system with two classes of customers with both backlog and lost sales. They show that the optimal policy can be described by three state-dependent thresholds: a production base-stock level and two order-admission levels, one for each class. de Véricourt et al. (2001) analyze a centralized make-to-stock system with a single item serving several customer classes. They introduce the strict priority (SP) policy and then compare the FCFS, SP, and IR policies for an $M/M/1$ queueing system, demonstrating that inventory rationing policy outperforms the other two.

Table 1: Strategies considered in the literature on centralized inventory and allocation policies for single product multi-class queues with backlog

	IR	FCFS	SP
Ha (1997b)	$M/M/1$		
de Véricourt et al. (2001)	$M/M/1$	$M/M/1$	$M/M/1$
de Véricourt et al. (2002)	$M/M/1$		
Benjaafar et al. (2007)	$M/M/1$		
Gayon et al. (2009)	$M/E_k/1$	$M/E_k/1$	$M/E_k/1$
This Paper	$M/G/1$	$M/G/1$	$M/G/1$

The different inventory and allocation policies for centralized make-to-stock queues, along with some of the papers discussing them, are summarized in Table 1 where we also present

the strategies analyzed in this paper. As seen in Table 1, we consider three strategies for a centralized single product multi-class $M/G/1$ system. We focus on cases where the product allocation is postponed to the end of production and then allocated to one of the customers, possibly according to the customer priority. Note that this allocation does not change the total inventory level, but may reduce costs. This is similar to the idea behind the last-come-first-served (LCFS) backlog allocation analyzed by Baron (2008).

1.1 Summary and Outline

This paper considers the three most common control policies for a centralized single product multi-class $M/G/1$ queue, namely FCFS, IR, and SP. While the characterization of the optimal FCFS policy in this setting is known, we are the first to characterize these solutions for the IR and SP policies.

Observe that in the IR system, the rate of change of the inventory level varies dynamically due to the rationing levels. We establish that the original system can be thought of as a multi-class $M/G/1$ queue with state dependent arrival rates. We therefore develop the exact closed form of the relevant Laplace Transforms (LT) for such a queue. These LT's provide an exact analysis of the multi-class $M/G/1$ system operating under the IR and SP policies (where the product allocation is postponed to the end of production).

To solve the multi-class $M/G/1$ queue with state dependent arrival rates, we focus on *customer composition*, i.e., the proportion of each customer class in the total number of customers in this queue. Note that because the latter is invariant for every non-idling and non-anticipating policy, the different controls only change this customer composition. (Recall, a *non-anticipating* policy uses only information on past scheduling and allocation decisions; with a *non-idling* policy, if there are customers in the system, the server is busy. For more rigorous definitions, see e.g., Bertsimas, 2007. In our context, this invariance means that if the inventory is below the base-stock level, the server is busy, and the server idles only when the inventory is at or above the base-stock level.)

To express the customer composition under the IR and SP policies, we consider a series of queues. In these queues, the first service time in each busy period is different from following service times. As in Takagi (1991, page 32), we call such queues, queues with an exceptional first service time in a busy period. We then show that a careful choice of the exceptional first service times equates their customer composition to that of the original $M/G/1$ system.

Using the customer composition, we derive the optimal rationing and base-stock levels for the IR and SP policies for the original $M/G/1$ system, i.e., with **general** service times. This is a substantial extension of previous work that expressed these optimal controls when service times are exponential (or Erlang). Earlier literature characterized the optimal policies in Markovian settings using dynamic programming. However, when the service time is not exponential, this approach is less practical. For example, Gayon et al. (2009) highlighted the difficulty in finding the optimal controls when the number of customer types is large in $M/E_k/1$ settings. In contrast, the customer composition methodology employs a series of queues and therefore, as we demonstrate numerically, allows the solution of systems with a large number of customer types.

The main contributions of this paper are in both the queueing and operations areas: From a queueing perspective, we demonstrate the concept of customer composition and its tractability by analyzing a series of multi-priority $M/G/1$ queues with state dependent arrival rates. This analysis requires the exact solution of a multi-priority class $M/G/1$ queue with postponement and an exceptional first service time in a busy period. In addition to deriving this solution, we show that the ratio of expected number of customers of any class to the expected total number of customers in such queues is independent of the exceptional first service time and depends only on the first moment of the service and arrival processes. We also develop the LT of the residual service time seen by a high-priority arrival to a queue with a specific number of high-priority customers. From an operations perspective, focusing on customer composition allows us to accurately characterize the optimal rationing and base-stock levels for the SP and IR system in an $M/G/1$ queueing system (as well as the implied optimal costs of these policies). Moreover, we show that the cost of the SP system is

equivalent to the cost of an FCFS system with an appropriately defined backlog cost. This equivalence results in a very simple algorithm that expresses the optimal base-stock level and its implied cost for the SP system. Our theoretical and numerical results support the applicability of both the SP and IR policies for single product multi-class $M/G/1$ systems.

This paper extends the literature on make-to-stock queues that focus on $M/M/1$ queues by allowing for the much more realistic case of general service times. Furthermore, we believe that our approach of focusing on customer composition and expressing it using a series of queues with exceptional first service times can be used to solve additional questions of interest such as dynamic scheduling of queues. We leave this subject as well as additional investigation of make-to-stock $M/G/1$ queues (e.g., with lost sales) to further research.

As discussed above, our solution for the SP and IR policies relies on (i) the exact analysis of a multi-priority $M/G/1$ queue with postponement and an exceptional first service time in a busy period, and (ii) characterizing the relevant exceptional first service times. Because the derivation of both is intricate, we defer it to the end of the paper, presenting it in Section 4. In Section 2, we present the multi-class $M/G/1$ system considered and introduce the terminology used in the paper. In Section 3, we derive the optimal rationing levels, base-stock levels, and cost of the SP and IR policies. This section also includes numerical results comparing the FCFS, SP and IR policies for two, five, and ten customer classes. (A more detailed numerical analysis focusing on examples with two customer classes can be found in Abouee-Mehrizi et al., 2010.) All proofs appear in Appendix A.

2 Modeling a Single Product Multi-Class $M/G/1$

The single product multi-class $M/G/1$ system which we consider has a supplier that produces a single product and caters to demand arising from n distinguishable classes. We assume that demand of each class r (type r demand) follows a Poisson process with rate λ_r , $r = 1, 2, \dots, n$. We use the terms type r and class r interchangeably. We model the production stage as a centralized single server queue with general i.i.d production times with a LT denoted by $\tilde{b}(\cdot)$

with a mean $1/\mu$ and a second moment m_2 .

We assume that unsatisfied demand is backlogged. Thus, for stability, we assume that $\rho := \lambda/\mu < 1$, where $\lambda = \sum_{r=1}^n \lambda_r$. The backlog cost of class r is b_r per unit backlogged per unit time. Without loss of generality, we assume that $b_1 > b_2 > \dots > b_n$ (if two distinct classes have the same backlog cost, we aggregate them to a single class). Customers are prioritized according to their backlog costs, i.e., classes 1 to n from highest to lowest. The system incurs a holding cost of h per unit per unit time.

This model gives rise to a multi-class system where the server can work on one production order at a time. For this problem, we consider a centralized continuously-reviewed inventory system. We use a production control according to a base-stock level, S . Thus, production stops, and the server becomes idle when the inventory level reaches S . We consider three different systems, corresponding to three different production control policies: the FCFS, SP, and IR systems. (From now on, we use these short terminologies, e.g., “SP system” rather than “multi-class single-item $M/G/1$ make-to-stock system with postponement of the product allocation to the end of production under an SP control policy.”)

Let $I(t)$ denote the inventory level at time t in a system (the exact system will be clear from the context), and note that we allow $I(t) < 0$ that implies a backlog in the system. Let $B_r(t)$ be the number of type r backlogs in the system. In the FCFS and SP systems, if any class is backlogged at time t , we have $I(t) < 0$; then $I(t) < 0$ implies a backlog of size $|I(t)|$. However, in the IR system, one can have positive inventory on hand while some customer classes are backlogged; thus, $I(t) > 0$ and $\sum_{r=2}^n B_r(t) > 0$ is possible.

A standard method to express $I(t)$ in a single class production system with base-stock level control, when only $I(t) < 0$ implies a backlog, is to consider the *shortfall process* $N(t) := S - I(t)$, e.g., Baron (2008) and references therein. Then, $N(t)$ is identical to the number of orders in an $M/G/1$ queue facing demand, production, and allocation processes that are identical to the ones faced by the original $M/G/1$ system. When using this shortfall queue formulation, each demand request in the system is mapped to an order arrival in the queue. And each product allocation, when production is completed, is mapped to an order

departure in the shortfall queue. That is, the sample paths of $I(t)$ and $N(t)$ change at the same times following the same event realizations. For example, the idle period of the shortfall queue is terminated when the inventory level in the $M/G/1$ system drops below its base-stock level. A shortfall $N(t) \leq S$ implies that the inventory in the system has $S - N(t)$ units; a shortfall $N(t) > S$ implies a backlog of $|S - N(t)| = N(t) - S$ units.

We use a reasoning similar to the one that guides the use of the shortfall queue when analyzing the three systems mentioned above. That is, we derive the cost of each such system by analyzing a multi-class $M/G/1$ queue with the same priorities and allocation rules as that of the original system.

An important observation with respect to the shortfall process, $N(t)$, is that it is invariant under all non-idling and non-anticipating control policies. Because $N(t)$ is invariant, for any system we have:

$$N(t) = (S - I(t)^+) + \sum_{r=1}^n B_r(t), \quad (1)$$

where $(x)^+ := \max(0, x)$.

Earlier we defined *customer composition* as the proportion of each customer class out of the total number of customers in a queue. Given Eq. (1), knowing the customer composition resulting from specific priorities and allocation rules in this queue is sufficient to represent the cost of this control for the relevant system. To express the relevant customer compositions in the SP and IR systems when they have a backlog, we construct multi-class single-item $M/G/1$ queues with postponement of allocation where their first service time in every busy period is exceptional. We name these queues “backlog queues” for short. We will elaborate upon the ideas of customer composition and backlog queues in the next section.

3 The Costs and Optimization of the Three Policies

We provide the exact cost of the SP and IR systems in Sections 3.2 and 3.3, respectively. For the sake of completeness, in Section 3.1 we give the optimal control and corresponding

cost for the FCFS system. We theoretically and numerically compare the performances of these three systems in Section 3.4.

3.1 The FCFS Policy

Recall that $N(t)$ denotes the number of orders in the shortfall queue at time t . Let $p(i) := P(N = i)$ be the steady-state probability of having i orders in the shortfall queue.

Because all customers are treated the same, the average backlog cost per customer is $b^F := \sum_{r=1}^n \lambda_r b_r / \lambda$. Therefore, for a given base-stock level S , the average cost for the FCFS policy is

$$C_F(S) := h \sum_{i=0}^S (S - i)p(i) + b^F \sum_{i=S+1}^{\infty} (i - S)p(i), \quad (2)$$

and letting $F(i) := \sum_{j=0}^i p(j)$, the optimal base-stock level, S^{F^*} , that minimizes this cost is, see e.g., Veatch and Wein (1996),

$$S^{F^*} = \min\{i : F(i) > b^F / (h + b^F)\}. \quad (3)$$

We find $C_F(S^{F^*})$ as in Sanajian and Balcioğlu (2009). Specifically, we numerically invert the probability generating function of the distribution of the total number of customers in an FCFS single class $M/G/1$ queue. This function can be found, e.g., in Buzacott and Shanthikumar (1993, p. 61), by substituting Eq. (3.73) in Eq. (3.72). We set the backlog cost for this single class queue as the average backlog cost b^F . We then find the optimal base-stock level for the FCFS policy using Eq. (3) and its implied cost using Eq. (2).

3.2 The SP Policy

In this section, we express the cost of the SP system with a base-stock level S . When the number of orders in its shortfall queue exceeds S , i.e., $N(t) > S$, the SP system has backlogs. Let $p_r(i)$ denote the steady-state probability of having i backlogs from class r . The average

cost for the SP system is

$$C_{SP}(S) := h \sum_{i=0}^S (S-i)p(i) + \sum_{r=1}^n b_r \sum_{i=0}^{\infty} ip_r(i) = h \sum_{i=0}^S (S-i)p(i) + \sum_{r=1}^n b_r E[B_r], \quad (4)$$

where $E[B_r]$ is the expected number of backlogs of type r .

Observe that because the holding cost is independent of the classes, the shortfall queue is sufficient to express the holding cost in this system. When $N(t) > S$, the inventory in the system has $N(t) - S$ backlogs. But because the backlog costs differ among classes, the shortfall queue is not sufficient to express these costs. Tracking $E[B_r]$, the expected backlog of class r in the SP system, requires tracking the number of customers of this class in a corresponding multi-class $M/G/1$ queue given that the total number of customers in this queue is S or higher. We next obtain $E[B_r]$ by constructing the SP backlog (SPB) queue. We then use $E[B_r]$ to characterize the optimal SP control policy and its corresponding cost.

3.2.1 The SP Backlog Queue

To obtain $E[B_r]$, we focus on the shortfall queue and construct the SPB queue. While, as explained in Section 2, every event in the SP system is mapped to an event in the SP shortfall queue (an $M/G/1$ queue with the same allocation rules), we do not know how to directly derive their probabilistic description during periods with no inventory. Therefore, we use the SPB queue. To differentiate between queues, we use the terms job in the SPB queue, customers in the SP system, and orders in the shortfall queue. We denote the duration of time when there is no inventory in the system by D . (A rigorous definition of D will be soon provided.) We construct the SPB queue such that its job composition would be identical in distribution to the composition of orders in the shortfall queue for $t \in D$. A major step in this construction is to establish that this equality in distribution would hold if the first service time in a busy period of the SPB queue were distributed as the residual service time seen by an arrival that finds S orders in the shortfall queue.

We assume w.l.o.g that $N(0) = 0$, let $\tau_0^U := 0$, and define for $m = 1, 2, \dots$ the stopping

times

$$\begin{aligned}\tau_m^D &:= \inf \{t : N(t) = S | t > \tau_{m-1}^U\}, \\ \tau_m^U &:= \inf \{t : N(t) = S - 1 | t > \tau_m^D\}.\end{aligned}\tag{5}$$

The events $\{N(t) = S\}$ and $\{N(t) = S - 1\}$ in the shortfall queue are identical to the events $\{I(t) = 0\}$ and $\{I(t) = 1\}$ in the SP system, respectively. Thus, τ_m^D is the m^{th} customer arrival instant when the inventory level goes **Down** from 1 to 0, and τ_m^U is the m^{th} production completion instant when the inventory level goes **Up** from 0 to 1 (both since time 0). We let $D := \lim_{M \rightarrow \infty} \cup_{m=1}^M [\tau_m^D, \tau_m^U)$. Observe that D denotes the durations where the shortfall queue has S or more orders in it and that because $\rho < 1$

$$\begin{aligned}\lim_{M \rightarrow \infty} \cup_{m=1}^M [\tau_m^D, \tau_m^U) / \tau_M^D &= \lim_{M \rightarrow \infty} \cup_{m=1}^M [\tau_m^D, \tau_m^U) / \tau_M^U \\ &= \sum_{i=S}^{\infty} p(i) = 1 - F(S - 1).\end{aligned}\tag{6}$$

We construct the SPB queue such that its job composition would match the customer composition in the system when there is no inventory, i.e., for $t \in D$. By Eq. (6), this duration takes $(1 - F(S - 1))$ percent of the time in steady-state. We require for example that the probability that the server in the SPB queue is idle normalized by $(1 - F(S - 1))$ equals the probability that there are S orders in the shortfall queue, $P(N(t) = S)$.

We define $\tau_{m,0}^I := \tau_m^D$ and for each $m = 1, 2, \dots$ and $i = 1, 2, \dots$ we let

$$\begin{aligned}\tau_{m,i}^B &:= \min \{ \inf \{t : N(t) = S + 1 | t \geq \tau_{m,i-1}^I\}, \tau_m^U \}, \\ \tau_{m,i}^I &:= \min \{ \inf \{t : N(t) = S | t \geq \tau_{m,i}^B\}, \tau_m^U \}.\end{aligned}\tag{7}$$

When these stopping times differ from τ_m^U , they divide the m^{th} interval of D into durations starting at $\tau_{m,i}^I$ when there is nonnegative **Inventory**, i.e., $N(t) = S$, and into durations starting with $\tau_{m,i}^B$ when there is **Backlog**, i.e., $N(t) > S$. We will soon see that for $i \geq 1$, $\tau_{m,i}^B$ and $\tau_{m,i}^I$ correspond also to the beginning of **Busy** and **Idle** periods, respectively, in the SPB queue (when $\tau_{m,i}^B$ and $\tau_{m,i}^I$ differ from τ_m^U). Note that once $\tau_{m,i}^B = \tau_m^U$ for some i , the m^{th} interval of D ends, and we have $\tau_{m,j}^B = \tau_{m,j}^I = \tau_m^U$ for each $j \geq i$. That is,

$\tau_{m,i}^B = \tau_m^U$ happens when the first event following $\tau_{m,i-1}^I$ is a production completion. In this case, $N(\tau_{m,i}^B) = N(\tau_m^U) = S - 1$. (See definition 2 in Baron, 2008, for a similar infinite series of stopping times that are all identical after some m .) For example, it might be that $N(t) = S$ during all of the m^{th} interval of D . In such cases $\tau_{m,1}^B = \tau_m^U$, which implies that there were no backlogs within this interval.

It is important to note that **every** order arrival that sees S orders in the shortfall queue causes backlog and appears at some $\tau_{m,i}^B < \tau_m^U$. Similarly, **every** production completion that decreases the number of orders in the shortfall queue to S and ends a period with positive backlog happens at some $\tau_{m,i}^I < \tau_m^U$. So that for some m

$$\begin{aligned} N(t) = S &\Leftrightarrow t \in [\tau_{m,i}^I, \tau_{m,i+1}^B) \quad i \geq 0 \\ N(t) > S &\Leftrightarrow t \in [\tau_{m,i}^B, \tau_{m,i}^I) \quad i \geq 1. \end{aligned} \tag{8}$$

We next construct the SPB queue by considering the shortfall queue for $t \in D$ and build equivalences between both queues' (a) allocation, (b) arrival, and (c) service processes.

Step (a): at the end of each service completion, the SPB queue will remove the oldest job with the smallest r index. Thus, the SPB queue is a priority queue.

Step (b): we let the arrival process of jobs of type r to the SPB queue follow a Poisson process with rate λ_r , $r = 1, 2, \dots, n$. This construction, together with the memoryless property ensures that the arrival processes of both the shortfall and the SPB queues are identical (in distribution).

Note that both steps (a) and (b) are simple and could, in principle, be done while mapping each event in the system to an event in the SPB queue (and in the shortfall queue) for each $t \in D$. However, completing step (c) is more involved, and from this point on, we were unable to maintain such mapping.

Step (c): we recall Eq. (8) and note that $\tau_{m,i}^B$ denotes the time of the next event following $t = \tau_{m,i-1}^I$ for $i \geq 1$, i.e., the time of the next event whenever $N(t) = S$. This event can be either a new order arrival or a production completion. We construct the service processes in the SPB queue while considering if this event is an arrival or a completion. We use the

term “corresponding” to describe parallelism between actions (i.e., arrivals or completions) in the SPB queue and in the shortfall queue; however, we emphasize that these actions do not happen at the same time. We start at a time that corresponds to $\tau_{m,i}^I$ with an idle server in the SPB queue.

The event at $\tau_{m,i}^B$ is an order arrival: Using Eq.s (5) and (7) we find $\tau_{m,i}^B < \tau_m^U$ and a corresponding job arrival would initiate the busy period in the SPB queue. By Eq. (8) this order sees S orders in front of it. At this order arrival, there is already an active production in the shortfall queue. When this production is completed, it is the first production completion in the interval $[\tau_{m,i}^B, \tau_{m,i}^I) \in D$ (possibly at $\tau_{m,i}^I$), and just before the completion, there would be $N > S$ orders in the shortfall queue. That is, the first production time in any interval $[\tau_{m,i}^B, \tau_{m,i}^I)$ is from the first arrival during this interval, at $\tau_{m,i}^B$, until the end of a production time that started before $\tau_{m,i}^B$. The time to complete this first service time is obviously different in distribution from a regular production time (unless production times are exponential).

Recalling the discussion leading to Eq. (8), we make an important observation: **every** arrival that sees S orders in the shortfall queue appears at some $\tau_{m,i}^B < \tau_m^U$, and should correspond to a first arrival in a busy period in the SPB queue. The time until the first order completion after $\tau_{m,i}^B < \tau_m^U$ is the residual service time seen by these arrivals, i.e., the residual service time seen by arrivals that see S orders in the shortfall queue. Because the service times of such arrivals are different from regular service times, all arrivals that are the first jobs in a busy period of the SPB queue face an exceptional service time. We set the lengths of these exceptional service times (in the SPB queue) as the equilibrium residual service times observed by order arrivals that see S orders (in the shortfall queue). We let $\tilde{b}_S^{SP}(\cdot)$ denote the equilibrium LT of this residual service time. This construction ensures that the lengths of service times faced by first jobs in busy periods of the SPB queue are identical in distribution to the lengths of the residual production times of the first orders during $[\tau_{m,i}^B, \tau_{m,i}^I) \in D$ in the shortfall queue.

When an exceptional first service time is over, if there are other jobs in the SPB queue,

all service times until the SPB queue clears all its jobs (corresponding to $N(\tau_{m,i}^I) = S$), follow a regular service distribution, with a LT $\tilde{b}(\cdot)$. After the SPB clears all jobs and becomes idle, the next event can be either a production completion or a new job arrival. In the case of a new job arrival (corresponding to an arrival that sees S orders in the shortfall queue), another busy period starts in the SPB queue with an exceptional service time with LT $\tilde{b}_S^{SP}(\cdot)$.

The event at $\tau_{m,i}^B$ is a production completion: Then, $\tau_{m,i}^B = \tau_m^U$, i.e., $\tau_{m,i}^B \notin D$, $N(\tau_{m,i}^B) = S - 1$ and the interval corresponding to $[\tau_{m,i-1}^I, \tau_m^U) = [\tau_{m,i-1}^I, \tau_{m,i}^B)$ would be a segment when the SPB queue is idle. The shortfall queue can have a sequence of, say k , $[\tau_{m,0}^I, \tau_{m,1}^B) = [\tau_m^D, \tau_m^U)$ idle segments until the event that follows $\tau_{m+k}^D = \tau_{m+k,0}^I$ is an order arrival and $\tau_{m+k,1}^B < \tau_{m+k}^U$. We observe that, because of the memoryless property, the total length of such a combination of several segments is still *i.i.d.*, and it follows an exponential distribution with parameter λ . Moreover, the sum of all these segments corresponds to an idle period in the SPB queue. As follows from step (b), this idle period is exponentially distributed with parameter λ as well.

The discussion of step (c) above highlights the correspondence between order arrivals that see S orders in the shortfall queue at times $\tau_{m,i}^B$ and job arrivals that start the busy period in the SPB queue, and between order completions that end a period with backlog at times $\tau_{m,i}^I$ and the end of the busy period in the SPB queue. With this correspondence in mind, the superscripts B and I in $\tau_{m,i}^B < \tau_m^U$ and $\tau_{m,i}^I < \tau_m^U$ relate these stopping times to the beginning of **Busy** and **Idle** periods in the SPB queue, respectively.

To summarize: our construction in steps (a-c) indicates that the SPB queue is an $M/G/1$ priority-queue with postponement and an exceptional first service time in a busy period. This exceptional service time has a LT $\tilde{b}_S^{SP}(\cdot)$ identical to the LT of the equilibrium residual service times observed by an arrival to the shortfall queue that sees S orders in front of it. Most importantly, by this construction, the job composition in the SPB queue is identical in distribution to the order composition of the shortfall process during $t \in D$.

3.2.2 Deriving the Optimal SP Policy

Because solving the multi-priority $M/G/1$ queue with postponement and an exceptional first service time in a busy period is intricate, we defer this analysis to Section 4. However, characterizing the customer composition in such queues is an essential building block in our analysis of the SP and IR systems. Thus, this customer composition is given in the next theorem (the proof of which requires the derivations shown in Section 4.)

Theorem 1 *Customer composition:* *The ratio of expected number of type r customers $E[N_r]$ to the expected number of total customers $E[N]$ in an $M/G/1$ queue with n priority classes, postponement of product allocation, exceptional first service times in a busy period with a LT $\tilde{b}_1(\cdot)$, and other service times with a LT $\tilde{b}(\cdot)$ is*

$$\frac{E[N_r]}{E[N]} := \frac{\lambda_r}{\lambda} \frac{1 - \rho}{(1 - \rho_r^+)(1 - \rho_{r-1}^+)}, \quad (9)$$

where $\lambda_r^+ = \sum_{i=1}^r \lambda_i$ and $\rho_r^+ = \lambda_r^+/\mu$ for $r = 1, \dots, n$.

Observe that, surprisingly, the ratio in Eq. (9) is independent of $b_1(\cdot)$, and this ratio only depends on the first moments of the queue's arrival and service processes.

Let μ_1 and $m_2^{\frac{1}{2}}$ denote the rate and second moment of $\tilde{b}_S^{SP}(\cdot)$ (both μ_1 and $m_2^{\frac{1}{2}}$ can be found using Theorem 12). If ρ_b denotes the server's utilization in the SPB queue, then:

Observation 1 *The expected number of customers in an $M/G/1$ queue with an exceptional first service time in a busy period is*

$$E[N] = (1 - \rho_b)\lambda \frac{\lambda^2 m_2^{\frac{1}{2}}/\mu_1 + (1 - \rho)(\lambda m_2^{\frac{1}{2}} + 2/\mu_1)}{2(1 - \rho)^2}. \quad (10)$$

Returning to the SP system, we have:

Theorem 2 *Given base-stock level S , the long-run average cost under the SP policy is*

$$C_{SP}(S) = h \sum_{i=0}^{S-1} (S - i)p(i) + (1 - F(S - 1))E[N^1] \sum_{r=1}^n b_r \frac{\lambda_r}{\lambda} \frac{1 - \rho}{(1 - \rho_r^+)(1 - \rho_{r-1}^+)}, \quad (11)$$

where $E[N^1]$ is given by the right hand side of Eq. (10).

de Véricourt et al. (2001) show that the optimal cost of the SP system in the $M/M/1$ settings can be obtained by considering an FCFS single class $M/M/1$ queue with a specific backlog cost. In Theorem 3, we extend this result to the $M/G/1$ system and show that the specific backlog cost only depends on the first moment of the (regular) service time.

Theorem 3 *Optimal SP policy:* *The cost of the SP policy with base-stock level S is the same as that of an FCFS single class $M/G/1$ queue with weighted backlog cost:*

$$b^{SP} = \sum_{r=1}^n \frac{\lambda_r(1-\rho)b_r}{\lambda(1-\rho_r^+)(1-\rho_{r-1}^+)}. \quad (12)$$

Thus, the cost of the SP policy can be written as

$$C_{SP}(S) = h \sum_{i=0}^{S-1} (S-i)p(i) + b^{SP} \sum_{i=S}^{\infty} (i-S)p(i), \quad (13)$$

and the optimal base-stock level S^{SP*} that minimizes Eq. (13) is

$$S^{SP*} = \min\{i : F(i) > b^{SP}/(h + b^{SP})\}. \quad (14)$$

Observe that according to Theorem 3, finding the optimal base-stock level and cost of the SP system requires only the solution of a standard single class FCFS $M/G/1$ queue. Specifically, we do not need to solve the SPB queue or characterize its exceptional first service time. Therefore, we find $C_{SP}(S^{SP*})$ as we found $C_F(S^{F*})$, by setting the backlog cost to b^{SP} , as given in Eq. (12), and finding S^{SP*} and its corresponding cost using Eqs (14) and (13), respectively.

3.3 The Inventory Rationing Policy

Let $C_{IR} := C(R_1 = 0, R_2, \dots, R_{n+1} = S)$ be the long-run average cost of the IR system given rationing levels R_1, R_2, \dots, R_{n+1} . In this section, we present an exact tractable expression for this cost. The idea in developing this expression is similar to the one used for analyzing the

SP policy. Specifically, we derive the customer composition within each relevant inventory range, $I \in (R_i, R_{i+1}]$ for $i = 0, \dots, n$ and $I \leq 0$, by considering a properly defined backlog queue.

To simplify the exposition, we present a less technical treatment of this policy than in the case of the SP policy. For example, we do not define the stopping times that represent when the customer composition in the system changes (and the system analysis shifts from one backlog queue to another).

We first establish:

Corollary 1 *We can assume without loss of generality that $R_r > R_{r-1}$ for $r = 2, \dots, n + 1$.*

3.3.1 The IR Backlog Queues

Here we construct a series of backlog queues for each class $r = 1, \dots, n + 1$ (the class $n + 1$ backlog queue would be the standard FCFS queue). The job composition in each class r backlog queue would equal in distribution to the relevant customer composition in the IR system when $I \in (R_{r-1}, R_r]$. For the SP policy, once the customer composition has changed, we construct the SPB queue to find the required job composition by carefully constructing its (a) allocation, (b) arrival, and (c) service processes. The resulting SPB queue is a priority $M/G/1$ queue with postponement and an appropriately defined exceptional first service in a busy period. We follow steps (a)-(c) below, formulating the class r backlog queue as a two-priority $M/G/1$ queue with postponement and an exceptional first service time in a busy period.

In step (a) for the SPB queue, the allocation of finished products is made in accordance with a strict priority. In the IR system, in contrast, once the customer composition changes, only the priority of a single class of customers changes and all other classes are treated as before. For example, once the inventory falls below R_{n-1} , class 1, ..., class $n - 2$ remains high-priority, receiving items from inventory upon arrivals, class n remains low-priority, such that its backlog increases upon arrivals, and only class $n - 1$ priority changes from high to

low. Thus, we set the r^{th} backlog queue as a two-priority queue.

In step (b) for the SPB queue, we set the arrival processes of all customer classes as Poisson with rate λ_r . For the IR case, we set the arrival processes as Poisson as well. To set the arrival rates of the low and high-priority queues we use the following observation.

Observation 2 *For any class $r = 2, \dots, n$, once the inventory level in the original system decreases to R_r , type r customers become low-priority until the inventory climbs to $R_r + 1$ again. During these periods the inventory level might downcross R_j for other classes $j < r$, making them low-priority customers and backlogging their demand. It may even be that all stock is depleted and all demand is backlogged. However, before the inventory climbs to level R_r , the system first clears the backlogs of classes $j < r$. In other words, from the point of view of class r , classes 1 to $r - 1$ remain a single class of high-priority customers as long as $I(t) \leq R_r$. Similarly, as long as $I(t) \leq R_r$, classes $j > r$ are low-priority and, therefore, their arrivals do not affect the system times experienced by classes $j \leq r$.*

Using Observation 2, all the required information on the backlog of class r customers can be obtained from the class r backlog queue. This implies that we do not need to include the arrivals from this class as part of the low-priority arrivals to the j backlog queues for $j < r$. We thus let the high- and low-priority customers arrival rates at the r^{th} backlog queue be $\lambda_{r-1}^+ := \sum_{i=1}^{r-1} \lambda_i$ and λ_r , respectively.

In step (c), we set the service process of this SPB queue to have an exceptional first service time in a busy period and regular service times otherwise. We use a similar logic for the backlog queues used in the analysis of the IR system. Orders that arrive when $I = R_r$, i.e., first orders after the customer composition changes, observe a different service time. Thus, the backlog queues will have an exceptional first service time in a busy period and regular service times otherwise.

For the SPB queue, we set the distribution of the exceptional first service time as the residual service time observed by an arrival that sees S orders in the shortfall queue; thus, the exceptional first service time depends on all orders in the system when the busy period

starts. In the SP system, all arrivals reduce the inventory towards 0, i.e., the level where the customer composition changes.

However, for the class r backlog queue, only high-priority jobs in the class $r + 1$ backlog queue correspond to customers that may reduce the inventory in the system to R_r ; thus, the exceptional first service time depends only on the number of high-priority jobs observed by a high-priority arrival to the class $r + 1$ backlog queue. Consider high-priority job arrivals in the class $r + 1$ backlog queue that see $R_{r+1} - R_r$ high-priority jobs. **Each** such arrival corresponds to a customer that decreases the inventory in the system to $R_r - 1$ or increases the backlog of the class r customers. Moreover, **every** such high-priority arrival also corresponds to jobs that start the busy period in the class r backlog queue. Therefore, the first service time in a busy period in the class r backlog queue is the residual service time observed by a high-priority arrival that sees $R_{r+1} - R_r$ high-priority jobs in the $r + 1$ backlog queue.

With this choice of the exceptional first service time in a busy period, the class r backlog queue is a two priority $M/G/1$ queue with high- and low-priority customer arrival rates λ_{r-1}^+ and λ_r , respectively. We define $\Delta_r := R_{r+1} - R_r$ for $r = 1, \dots, n$ and let $\tilde{b}_{\Delta_r}^r(s)$ denote the LT of the residual service time observed by a high-priority arrival to the $r + 1$ backlog queue that sees Δ_r jobs in this queue. (With this notation, $\tilde{b}_S^n(\cdot)$ is identical to $\tilde{b}_S^{SP}(\cdot)$, the LT of the exceptional first service time in a busy period for the SPB queue.) We set the LT of the exceptional first service time in the r backlog queue to $\tilde{b}_{\Delta_r}^r(s)$ and the LT of a regular service time to $\tilde{b}(s)$.

For completeness, we think of the shortfall queue of the IR system as the $n + 1$ backlog queue. We let $\lambda_{n+1} := 0$ and set the class $(n + 1)$ backlog queue to be the shortfall queue, i.e., we set its first exceptional service time as a regular service time (with a LT $\tilde{b}_0^{n+1}(s) = \tilde{b}(s)$). This implies that all orders of type 1 to n in the shortfall queue form a single high-priority class of jobs in the class $(n + 1)$ backlog queue.

Note that because the customer composition does not change as the inventory level decreases below $R_1 = 0$, we can calculate the backlog of class 1 customers from the class 2 backlog queue. However, as shown in Theorem 1, finding the expected number of customers

Table 2: Relations between backlog queues and the IR system

r^{th} backlog queue	$(r + 1)^{th}$ backlog queue	The IR system
Queue is relevant:	Once the total number of high-priority jobs in the $(r + 1)^{th}$ backlog queue increases to Δ_r .	when $I(t) \leq R_r$
The first service time in a busy period corresponds to:	The residual service time of a high-priority job that sees Δ_r high-priority jobs in this queue upon arrival.	The residual service time of a customer arrival of classes $1 \dots r$ that finds $I(t) = R_r$ and $B_r(t) = 0$.
The busy period starts (and the idle period ends) with a job arrival that corresponds to:	A high-priority job arrival to this queue that sees Δ_r high-priority jobs upon arrival.	A customer arrival that decreases $I(t)$ to $R_r - 1$ when $B_r(t) = 0$ or increases $B_r(t)$ to 1 when $I(t) = R_r$.
The busy period ends (and the idle period starts), which corresponds to:	A service completion that reduces the total number of high-priority jobs in this queue to Δ_r .	When the inventory increases to R_r while $B_r(t) = 0$ or when the class r backlog decreases to 0 (this can only happen while $I(t) = R_r$).
Low-priority customers:	The lowest high-priority jobs in this queue.	Customers of class r
High-priority customers:	All but the lowest high-priority jobs in this queue.	Customers of classes 1 to $r - 1$.

in a backlog queue can be done in closed form without deriving these probabilities. To reduce the computational burden, then, we consider the class 1 backlog queue. The exceptional first service time for this queue is the residual service time seen by a high-priority arrival to the class 2 backlog queue that sees Δ_1 high-priority customer in service, i.e., its LT is $\tilde{b}_{\Delta_1}^1(s)$.

To relate the different backlog queues to one another and to the system, we summarize the main relations in Table 2.

We further note that as in the SPB queue, the idle periods of the class $r = 2, \dots, n + 1$

backlog queues may correspond to a collection of several idle segments. Again, due to the memoryless property, the total length of these idle periods is exponential.

To summarize, the constructions presented in this section ensure that the job composition in these backlog queues is the same as the customer composition in the IR system (on the relevant ranges). Thus, properly weighting the probabilities in these backlog queues gives the cost of the IR system. We next derive this cost explicitly.

3.3.2 The Cost of the IR Policy

Here we express C_{IR} using the backlog queues defined above. Let $P_r^h(i)$ denote the steady-state probability of having i high-priority customers in the class r backlog queue. Let $P_r^l(i)$ denote the steady-state probability of having i low-priority orders in this queue. Furthermore, let $F_r^h(i) := \sum_{j=0}^i P_r^h(j)$ denote the cumulative distribution function (CDF) for the high-priority class and $\bar{F}_r^h(i) = 1 - F_r^h(i)$.

When there are $i = 0, \dots, \Delta_n - 1$ orders in the shortfall queue (equivalently i high-priority jobs in the class $(n+1)$ backlog queue), the IR system has $R_{n+1} - i$ units in inventory. Then, $\bar{F}_{n+1}^h(\Delta_n - 1)$ percent of the time the IR system has a $I(t) \leq R_n$.

The class n backlog queue models the IR system between the instants when the inventory level drops to R_n and climbs to $R_n + 1$, i.e., $\bar{F}_{n+1}^h(\Delta_n - 1)$ percent of the time, and is analogous to the SPB queue in Section 3.2. This queue treats type 1 to $n - 1$ jobs as a single high-priority class, type n jobs as the low-priority class, and has $\tilde{b}_{\Delta_n}^n(s)$ as the LT of the first service time in its busy periods. With probability $\bar{P}_n^h(i)$ there are $i = 0, \dots, \Delta_n - 1$ high-priority jobs in this queue corresponding to $R_n - i$ units of inventory in the IR system. For $i = 0, \dots, \Delta_{n-1} - 1$, with probability $P_n^h(i) \bar{F}_{n+1}^h(\Delta_n - 1)$ there are $R_{n+1} - \Delta_n - i$ units of inventory in the system. Similarly, during $\bar{F}_{n+1}^h(\Delta_n - 1) \bar{F}_n^h(\Delta_{n-1} - 1)$ percent of the time, the IR system has an inventory level less than (or equal to) R_{n-1} .

Next, we generalize this approach. Consider the class r backlog queue for $r = 1, \dots, n$, which models the system between the instants when the inventory level drops to R_r and

climbs to $R_r + 1$, i.e., with probability

$$\lim_{t \rightarrow \infty} Pr\{I(t) \leq R_r\} = \prod_{i=r+1}^{n+1} \bar{F}_i^h(\Delta_{i-1} - 1). \quad (15)$$

So during $\prod_{i=r+1}^{n+1} \bar{F}_i^h(\Delta_{i-1} - 1)$ percent of the time, the existence of j class r jobs in the class r backlog queue corresponds to j type r backlogs in the IR system, and the expected backlog cost for class r customers is

$$\prod_{i=r+1}^{n+1} \bar{F}_i^h(\Delta_{i-1} - 1) [b_r \sum_{j=0}^{\infty} j P_r^l(j)], \quad (16)$$

where $b_r \sum_{j=0}^{\infty} j P_r^l(j)$ is the backlog cost of class r given that $I(t) \leq R_r$. Theorem 4 below gives the closed form expression for the expected number of backlogs for each class r .

To calculate the holding cost, we recall Eq. (15). For $r = 2, \dots, n+1$, the holding cost over $R_{r-1} < I(t) \leq R_r$ is

$$\prod_{i=r+1}^{n+1} \bar{F}_i^h(\Delta_{i-1} - 1) [h \sum_{j=0}^{\Delta_{r-1}-1} (R_r - j) P_r^h(j)],$$

where $h \sum_{j=0}^{\Delta_{r-1}-1} (R_r - j) P_r^h(j)$ is the expected holding cost given that $R_{r-1} < I(t) \leq R_r$.

Changing the index from r to $r-1$ gives for any $r = 1, \dots, n$,

$$\prod_{i=r+2}^{n+1} \bar{F}_i^h(\Delta_{i-1} - 1) [h \sum_{j=0}^{\Delta_r-1} (R_{r+1} - j) P_{r+1}^h(j)]. \quad (17)$$

Let $\rho_b := \lambda_r^+ \mu / (\mu_1 \mu - \lambda_r^+ (\mu_1 - \mu))$ be the utilization in the r^{th} backlog queue where the rate μ_1 (as well as the second moment, m_2^1) of the exceptional first service time in the class r backlog queue can be found using its LT, $\tilde{b}_{\Delta_r}^r(s)$. The latter can be derived based on Algorithm 1. (For notational convenience and because the connection is clear, we omit the superscript r from ρ_b , μ_1 , and m_2^1 in the class r backlog queue.) Let $E[N^r]$ denote the expectation of the total number of customers in the class r backlog queue. Then,

Theorem 4 *The expected number of backlogs for class r is*

$$E[B_r] = \sum_{j=0}^{\infty} j P_r^l(j) = E[N^r] \frac{\lambda_r}{\lambda_r^+} \frac{1 - \rho_r^+}{(1 - \rho_r^+)(1 - \rho_{r-1}^+)}, \quad (18)$$

where $\rho_r^+ = \lambda_r^+/\mu$ for $r = 1, \dots, n$, $\rho_0^+ = 0$ and

$$E[N^r] = (1 - \rho_b)\lambda_r^+ \frac{(\lambda_r^+)^2 m_2/\mu_1 + (1 - \rho)(\lambda_r^+ m_2^1 + 2/\mu_1)}{2(1 - \rho)^2}. \quad (19)$$

To summarize, with $\prod_{i=n+2}^{n+1} \bar{F}_i^h(\Delta_{i-1} - 1) := 1$ and using Eq.s (16), (17) and Theorem 4:

Theorem 5 *The long-run average cost of the IR policy is*

$$\begin{aligned} C_{IR} = & h \sum_{r=1}^n [\{ \prod_{i=r+2}^{n+1} \bar{F}_i^h(\Delta_{i-1} - 1) \} \sum_{j=0}^{\Delta_r-1} (R_{r+1} - j) P_{r+1}^h(j)] \\ & + \sum_{r=1}^n b_r [\{ \prod_{i=r+1}^{n+1} \bar{F}_i^h(\Delta_{i-1} - 1) \} E[B_r]], \end{aligned} \quad (20)$$

where $E[B_r]$ is given in Eq. (18).

3.3.3 Searching for the Optimal IR Policy

To calculate the cost in Theorem 5, note that the class $(n + 1)$ backlog queue is an FCFS $M/G/1$ queue with an arrival rate $\lambda = \sum_{i=1}^n \lambda_i$. The probabilities $P_{n+1}^h(i)$ can be obtained as before, similar to Sanajian and Balcioglu (2009). The other probabilities, $P_r^h(i)$, require solving the relevant backlog queues. The probability generating function of these probabilities is $\Pi_h(z)$, which can be numerically inverted to obtain them. The LT of the exceptional first service time in a busy period can be calculated using Algorithm 1, and Theorem 9 that uses this LT, gives $\Pi_h(z)$. Both the algorithm and $\Pi_h(z)$ are presented in Section 4. Finally as demonstrated in Theorem 4, $E[B_r]$ can be obtained without computing $P_r^l(i)$. With the exact cost C_{IR} calculated using this procedure for given rationing levels, we can search over different vectors of $(R_1, R_2, \dots, R_{n+1})$ to find the optimal rationing levels and the corresponding cost.

To find C_{IR}^* , the optimal cost for the IR system, we search over $R_{n+1} = S$ by varying $R_{n+1} = 1, \dots, M$. For each R_{n+1} we look for the optimal rationing levels $0 = R_1 \leq \dots \leq R_n \leq R_{n+1}$. We let $M = \min\{i > S^{SP*} + 2 : C_{IR}(i) > C_{IR}(i-1) > C_{IR}(i-2)\}$, where S^{SP*} denote the optimal base-stock levels under the SP policy. We note that, in all our numerical results, we found that the optimal $R_{n+1} \leq S^{SP*}$.

3.4 Comparison of the Three Policies

We next compare the IR, SP, and FCFS $M/G/1$ systems. As before, we let $C_F(S^{F*})$, $C_{SP}(S^{SP*})$ and C_{IR}^* denote the optimal cost of the the FCFS, SP and IR systems, respectively.

3.4.1 Theoretical Comparison

Note that the SP control is a special case of the IR control and that the customer composition in the SP system leads to lower backlog costs than in the FCFS system while maintaining the same holding cost. Observation 3 below summarizes this and provides theoretical support for the use of the IR and SP policies rather than the FCFS policy in $M/G/1$ make-to-stock queues. The observation is given without a more detailed proof.

Observation 3 *In $M/G/1$ systems with centralized inventory and postponement of the allocation decisions to the end of the service completions, we have*

$$C_{IR}^* \leq C_{SP}(S^{SP*}) \leq C_{CF}(S^{F*}).$$

3.4.2 Numerical Comparison

We showed in Observation 3 that the IR policy minimizes the cost among the FCFS, SP, and IR policies. Here, we demonstrate that our methodology can be used to find the optimal control and cost for 2, 5 and 10 customer classes. A more detailed numerical analysis of the performance of the FCFS, SP, and IR policies can be found in Abouee-Mehrzi et al. (2010). Since $M/M/1$ make-to-stock systems have been investigated before (de Véricourt et al. 2001), we consider two service times with a squared-coefficient of variation (variance to squared mean ratio) $cv^2 \neq 1$: (i) deterministic, with a mean of 1 and $cv^2 = 0$, and (ii) the 2-stage Mixed Generalized Erlang (MGE2) distribution with $cv^2 = 2$, $MGE2(\mu_1 = 1.05523, \mu_2 = 0.09477, a_1 = 0.99504)$ (Altiok, 1997, p. 42–43), a mean of 1 and density

$$f(y) = \frac{(1 - a_1)\mu_1 - \mu_2}{\mu_1 - \mu_2} \mu_1 e^{-\mu_1 y} + \frac{a_1 \mu_1}{\mu_1 - \mu_2} \mu_2 e^{-\mu_2 y}.$$

We vary the parameters $\rho = 0.8, 0.9$ while maintaining the arrival rates equal $\lambda_r = \rho/n$, letting $b_r = n - r + 1$, $r = 1, \dots, n$ (i.e., $b_n = 1$) and $h = 0.1$. This gives a total of 24 cases. For each test we calculate the ratios

$$\Delta SP := \frac{C_{SP}(S^{SP*}) - C_{IR}^*}{C_{IR}^*} \times 100, \quad \Delta F := \frac{C_{CF}(S^{F*}) - C_{IR}^*}{C_{IR}^*} \times 100. \quad (21)$$

Table 3 presents the results of these numerical experiments and demonstrates that the optimal centralized IR and SP policies are tractable using our methodology even when the number of classes is 10. Furthermore, using these policies can significantly reduce costs, compared to the optimal FCFS policy.

Table 3: ΔSP and ΔF for multiple classes of customers

cv^2	ρ	$n = 2$		$n = 5$		$n = 10$	
		ΔSP	ΔF	ΔSP	ΔF	ΔSP	ΔF
0	0.8	0.00	9.73	1.80	22.45	4.11	18.38
	0.9	0.00	12.00	2.36	26.48	5.62	30.14
2	0.8	0.00	11.93	1.72	21.95	3.11	20.39
	0.9	0.00	13.00	1.98	27.85	4.02	29.87

4 Required Queueing Analysis

In this section, we derive the required analytical results to express the costs for the IR and SP policies. Specifically, in Section 4.1 we derive, $\tilde{w}_r(s)$, the LT of the system time of type r customers in an n class multi-priority $M/G/1$ queue with an exceptional first service time in a busy period when product allocation is postponed to the end of production. Then, in Section 4.2 Algorithm 1 outputs the LT of the exceptional first service time in the busy periods for the backlog queues defined in Section 3.

4.1 A Multi-Priority $M/G/1$ Queue with Exceptional First Service Times in a Busy Period

In this section, we consider a multi-priority $M/G/1$ queue with an exceptional first service time in a busy period when product allocation is postponed to the end of production. Following Chapter 3 of Takagi (1991) and Chapter 8 of Conway et al. (1967) wherever possible, we obtain $\tilde{w}_r(s)$, the LT of the density function of the system time of class r customers. (The models in Takagi, 1991 or Conway et al., 1967 consider systems without postponement; thus, their results cannot be used directly to study the IR and SP policies.) To obtain $\tilde{w}_r(s)$, we consider a system with two-priority classes in Section 4.1.1. In Section 4.1.2, we obtain $\Pi_h(z)$, the probability generating function of the number of high-priority customers left in the two-priority class system by a departing high-priority customer. We then relate $\Pi_h(z)$ to $\tilde{w}_r(s)$.

Two concepts we use are the busy period and the delayed busy period in $M/G/1$ queues. Consider an $M/G/1$ queue with an arrival rate λ and service time with LT given by $\tilde{b}(s)$, respectively. The LT of the length of the busy period in this queue, $\theta(s)$, is given by (see e.g., Eq. (7) in Conway et al., 1967, p. 150)

$$\theta(s) = \tilde{b}(s + \lambda(1 - \theta(s))). \quad (22)$$

Often the relevant busy period starts after a delay. As in Conway et al. (1967), we use delay cycle to represent the sum of the delay and the delay busy period, i.e., the busy period following this delay. We denote the LT's of the delay, the delay busy period, and the delay cycle by $\tilde{w}(s)$, $\theta_b(s)$, and $\theta_c(s)$, respectively. With the same service and arrival rate as in the $M/G/1$ queue we have, see e.g., Eqs. (8) and (9) in Conway et al. (1967, p. 151):

$$\begin{aligned} \theta_b(s) &= \tilde{w}(\lambda(1 - \theta(s))), \\ \theta_c(s) &= \tilde{w}(s + \lambda(1 - \theta(s))). \end{aligned} \quad (23)$$

4.1.1 A Markov-Chain Representation for the Two-Priority Class System

In this section, we consider a two-priority $M/G/1$ queue with an exceptional first service time where high- and low-priority customer arrival rates are λ_h and λ_l , respectively, such that $\lambda = \lambda_h + \lambda_l$. We solve this queue following Takagi (1991). We focus on the discrete stochastic process \mathbf{M}^h where $\{M_n^h, n = 1, 2, \dots\}$ is the number of high-priority customers left behind by the n^{th} departing customer (either high- or low-priority) in the two-priority class system. Let π_k be the steady-state probability that an arbitrary departure leaves k high-priority customers behind.

When v_k and w_k denote the probabilities of having k high-priority arrivals during a service time with LT's $\tilde{b}(s)$ and $\tilde{b}_1(s)$, respectively, we have

$$W(z) = \sum_{k=0}^{\infty} w_k z^k = \tilde{b}_1(\lambda_h(1-z)), \quad (24)$$

$$V(z) = \sum_{k=0}^{\infty} v_k z^k = \tilde{b}(\lambda_h(1-z)). \quad (25)$$

Like the analysis of the Markov chain embedded at departures for the $M/G/1$ queue (Gross and Harris, 1998, p. 214), p_{jk} , the transition probabilities of \mathbf{M}^h for $k \geq j-1$, $j \geq 1$ are

$$p_{jk} = P\{M_{n+1}^h = k | M_n^h = j\} = v_{k-j+1}, \quad k \geq j-1, j \geq 1. \quad (26)$$

However, when $j = 0$ there are no high-priority customers in the system at the last departure instant, and, \mathbf{M}^h is no longer Markovian. We therefore consider a different stochastic process $\widetilde{\mathbf{M}}^h$ that is Markovian and tractable. We construct the transition probabilities of $\widetilde{\mathbf{M}}^h$ such that its steady-state probabilities $\tilde{\pi}_k$'s are identical to π_k 's. The proof of the theorem below uses $1 - \rho_b$ to denote the probability that the server is idle. Then, $\pi_0 - (1 - \rho_b)$ is the probability that there are only low-priority customers in the system. Due to PASTA, the mean of service time is $1/\mu$ with probability ρ_b , and $1/\mu_1$ with probability $1 - \rho_b$, thus

$$\rho_b = \frac{\lambda\mu}{\mu_1\mu + \lambda(\mu - \mu_1)}. \quad (27)$$

Theorem 6 *The steady-state probabilities of $\widetilde{\mathbf{M}}^h$ and \mathbf{M}^h are identical:*

$$\tilde{\pi}_k = \pi_k, \text{ for } k = 0, 1, \dots$$

4.1.2 Deriving the Generating Functions

To derive the generating functions, as in Chapter 3 of Takagi (1991), we require the expected length of time that the server works with the aim of satisfying low-priority customer demand. This is the sum of service times that start to satisfy low-priority customers but are taken over by high-priority customers and the final service time during which no high-priority customers arrive. Conway et al. (1967, p. 169) call this the *gross processing time* and define it as “the total amount of time that a job actually spends on the machine.” Let A be the r.v. corresponding to the gross processing time.

Theorem 7 *The expected gross processing time in the two-priority $M/G/1$ queue with an exceptional first service time in a busy period and allocation postponement is*

$$E[A] = \rho_b E[A_1] + (1 - \rho_b)(\tilde{b}_1(\lambda_h)E[A_2] + (1 - \tilde{b}_1(\lambda_h))(E[A_3] + E[A_1])), \quad (28)$$

where with $\tilde{b}_1'(s) := d\tilde{b}_1(s)/ds$

$$E[A_1] = \frac{1 - \tilde{b}(\lambda_h)}{\lambda_h \tilde{b}(\lambda_h)}, \quad E[A_2] = -\frac{\tilde{b}_1'(\lambda_h)}{\tilde{b}_1(\lambda_h)}, \quad E[A_3] = \frac{\lambda_h \tilde{b}_1'(\lambda_h) + (1 - \tilde{b}_1(\lambda_h))}{\lambda_h (1 - \tilde{b}_1(\lambda_h))}.$$

To derive the probability generating functions, we need to first express π_0 , which is more involved than in Takagi (1991). Considering only the high-priority departures, let κ_0 denote the steady-state probability that a departing high-priority customer leaves no high-priority customers behind if we consider only the high-priority departures.

Theorem 8 *The fraction of departures leaving no high-priority customers behind is*

$$\pi_0 = 1 - \frac{\lambda_h}{\lambda}(1 - \kappa_0) = 1 - \frac{\lambda_h(\rho_b - E[A])}{\lambda}. \quad (29)$$

Now, using π_0 and following Takagi (1991) we show

Lemma 1 *The probability generating function of the number of high-priority customers left in the two-priority class system by an arbitrary departure is*

$$\begin{aligned} \Pi(z) = & \frac{(1 - \rho_b)V(z)}{V(z) - z} + \frac{(\lambda_h z + \lambda_l)(1 - \rho_b)W(z)}{\lambda(z - V(z))} + \frac{(1 - \rho_b)\lambda_l(w_0(z - 1))}{\lambda(z - V(z))} \\ & + \frac{(\pi_0 - (1 - \rho_b))v_0(z - 1)}{z - V(z)}. \end{aligned} \quad (30)$$

Using Lemma 1, we have

Theorem 9 *In the two-priority class system, the probability generating function of the number of high-priority customers left behind after the departure of a high-priority customer is*

$$\begin{aligned} \Pi_h(z) = & \frac{\lambda(1 - \rho_b)V(z)}{\lambda_h z(V(z) - z)} \left[z - \frac{(\lambda_h z + \lambda_l)W(z) + \lambda_l w_0(z - 1)}{\lambda} - \frac{(\pi_0 - (1 - \rho_b))v_0(z - 1)}{1 - \rho_b} \right] \\ & + \frac{\lambda(1 - \rho_b)}{\lambda_h z} \left[\frac{(\lambda_h z + \lambda_l)W(z)}{\lambda} - \frac{w_0 \lambda_l}{\lambda} - \frac{(\pi_0 - (1 - \rho_b))v_0}{(1 - \rho_b)} \right]. \end{aligned} \quad (31)$$

Letting $\tilde{w}_h(s)$ denote the Laplace transform of the system time density function of the high-priority customers, we have

Theorem 10 *The LT of the system time density function of the high-priority customers in the two-priority class system is*

$$\begin{aligned} \tilde{w}_h(s) = & \frac{\tilde{b}(s)(1 - \rho_b)(\lambda_l w_0 - \lambda) + (\pi_0 - (1 - \rho_b))v_0 \lambda(\tilde{b}(s) - 1)}{\lambda_h(1 - \tilde{b}(s)) - s} \\ & + \frac{(1 - \rho_b)(\tilde{b}_1(s)(\lambda - s) - \lambda_l w_0)}{\lambda_h(1 - \tilde{b}(s)) - s}. \end{aligned} \quad (32)$$

Corollary 2 below follows from Theorem 10.

Corollary 2 *The LT of the system time density function in a single class FCFS M/G/1 queue with an exceptional first service time in a busy period is*

$$\tilde{w}(s) = \frac{(1 - \rho_b)(\lambda(\tilde{b}(s) - \tilde{b}_1(s)) + s\tilde{b}_1(s))}{s - \lambda(1 - \tilde{b}(s))}. \quad (33)$$

Corollary 2 is, of course, identical to Eq. (2.39) in Takagi (1991, p. 129). We now return to the original system involving n priority classes.

Theorem 11 *The LT of the system time density function of the type r customers in the $M/G/1$ with an exceptional first service time in a busy period is*

$$\tilde{w}_r(s) = \tilde{w}_h(s + \lambda_{r-1}^+(1 - \theta_{r-1}^+(s))), \quad (34)$$

where

$$\theta_{r-1}^+(s) = \tilde{b}(s + \lambda_{r-1}^+(1 - \theta_{r-1}^+(s))). \quad (35)$$

Corollary 3 *The LT of the density function of the system time of class r customers in a multi-priority $M/G/1$ queue with postponement but without an exceptional service time is given by Eq. (34) with*

$$\tilde{w}_h(s) = \frac{(\tilde{b}(s) - 1)v_0[\pi_0\lambda - (1 - \rho)\lambda_r^+] - (1 - \rho)s\tilde{b}(s)}{\lambda_r^+(1 - \tilde{b}(s)) - s}, \quad (36)$$

where $\lambda_r^+ = \sum_{i=1}^r \lambda_i$, $\lambda_r^- = \sum_{i=r+1}^n \lambda_i$, $v_0 = \tilde{b}(\lambda_r^+)$ and π_0 can be found from Eq. (29).

Observation 4 *Observe that for class $r = n$, we have $\tilde{w}_n(s) = \tilde{w}_h(s + \lambda_{n-1}^+(1 - \theta_{n-1}^+(s)))$ and note that in this case, $\lambda_l = 0$ and $\lambda_h = \lambda$. We have a single class and as in Corollary 2, we get $\tilde{w}_h(s) = \tilde{w}(s)$ as given in Eq. (33).*

4.2 Exceptional First Service Time in a Two-Priority $M/G/1$ Queue

In this section, we derive the LT of the residual service time seen by a high-priority arrival in a two-priority $M/G/1$ queue with an exceptional first service time in a busy period that finds j high-priority customers in the system, $\tilde{b}_j^r(s)$. This LT is later employed in Algorithm 1 to obtain the required exceptional first service time LT for the next backlog queues as discussed under the IR policy in Section 3.3.

The derivation of $\tilde{b}_j^r(s)$ in Theorem 12 combines our approach focusing on high-priority arrivals to appropriately backlog queues, with the approach of Kerner (2008) used to derive the LT of the residual service time in an FCFS $M/G/1$ queue.

Theorem 12 For a two-priority class $M/G/1$ queue with exceptional first service time in a busy period, the LT of the residual service time at an arrival of a high-priority customer seeing j high-priority customers in the system is given recursively by

$$\tilde{b}_j^r(s) = \frac{\lambda_h}{s - \lambda_h} [\tilde{b}(\lambda_h) \frac{1 - \tilde{b}_{j-1}^r(s)}{1 - \tilde{b}_{j-1}^r(\lambda_h)} - \tilde{b}(s)], \quad j \geq 1, \quad (37)$$

where

$$\begin{aligned} \tilde{b}_0^r(s) &= \frac{\kappa_0 \lambda_h \tilde{b}(s) + \tilde{b}(s)(1 - \rho_b)(\lambda_l w_0 - \lambda)}{\kappa_0(\lambda_h - s)} \\ &+ \frac{(\pi_0 - (1 - \rho_b))\lambda v_0(\tilde{b}(s) - 1) + (1 - \rho_b)(\tilde{b}_1(s)(\lambda - s) - \lambda_l w_0)}{\kappa_0(\lambda_h - s)}. \end{aligned} \quad (38)$$

Remember that from Eq.s (24) and (25), it follows that $v_0 = \tilde{b}(\lambda_h)$ and $w_0 = \tilde{b}_1(\lambda_h)$ and ρ_b and κ_0 are given in Eq.s (27) and (A.3), respectively (where $E[A]$ is given in Theorem 7). Observe that if $\tilde{b}_1(s) = \tilde{b}(s)$, $\lambda_h = \lambda$ and $\lambda_l = 0$, Theorem 12 is identical to Corollary 2.2.1 in Kerner (2008) when setting $\lambda_n = \lambda$ for all n .

The following Algorithm gives the exceptional first service times required for the sequence of backlog queues employed for computing C_{IR} in Section 3.3.

Algorithm 1 Finding the LT of the first exceptional service times $\tilde{b}_{R_{r+1}-R_r}^r(s)$ for $r = 1, \dots, n$

[Step 0] For level R_{n+1} , set $r = n$, $\tilde{b}_1(s) := \tilde{b}(s)$ and $\lambda_h = \lambda_n^+ := \sum_{i=1}^n \lambda_i$, $\lambda_l = \lambda_n^- := 0$, and $\lambda := \sum_{i=1}^n \lambda_i$.

[Step 1] While $n \geq r \geq 1$, consider the r^{th} backlog queue:

a Use Theorem 12 to obtain and return $\tilde{b}_{R_{r+1}-R_r}^r(s)$.

b Set $\lambda_h = \lambda_{r-1}^+ := \sum_{i=1}^{r-1} \lambda_i$, $\lambda_l := \lambda_r$, and $\lambda = \lambda_r^+$.

c Set $\tilde{b}_1(s) = \tilde{b}_{R_{r+1}-R_r}^r(s)$ and $r = r - 1$ and go back to Step 1.

Algorithm 1 assumes, implicitly, that the LT of regular service times, $\tilde{b}(s)$, is known. The algorithm starts with $r = n$ at step 0, by setting the required parameters to characterize the

$(n+1)^{st}$ backlog queue: $\tilde{b}_1(s)$, λ_h , and λ_l . Then, at step 1.a., the algorithm uses Theorem 12 to find and return $\tilde{b}_{R_{r+1}-R_r}^r(s)$, the LT of the residual service time from the $(r+1)^{st}$ backlog queue. This residual service time is the exceptional first service time in the r^{th} backlog queue. At step 1.b. the algorithm sets the required arrival rates for the r^{th} backlog queue. (Note that at this stage, Theorem 9 can be used to obtain $\Pi_h(z)$ and the implied probabilities for this queue.) As step 1.c. before continuing with the same steps for the $(r-1)^{st}$ backlog queue, the algorithm updates the exceptional service time for this queue (as the residual service time resulting from the r^{th} backlog queue). The algorithm then returns to step 1 with $r = r - 1$.

Acknowledgements

The authors would like to thank the editors and referees for their helpful comments. They are also grateful to Joseph Milner for the many useful feedbacks on an early version of this paper. This work was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

References

- Abouee-Mehrzi, H., B. Balcioğlu, O. Baron. 2010. “Comparison of Policies in the Single Product Multi-Class $M/G/1$ Make-to-Stock Queue”, *Working paper, Rotman School of Management, University of Toronto*.
- Altıok, T. 1997. *Performance Analysis of Manufacturing Systems*, Springer-Verlag, NY.
- Arslan, H., S. C. Graves, T. A. Roemer. 2007. “A Single-Product Inventory Model for Multiple Demand Classes”, *Management Science*, Vol. 53, 1486–1500.
- Baron, O. 2008. “Regulated Random Walks and the LCFS Backlog Probability: Analysis and Applications”, *Operations Research*, Vol. 56, 471–486.

- Benjaafar, S., L. W. Cooper, J. S. Kim. 2005. “On the Benefits of Pooling in Production-Inventory Systems”, *Management Science*, Vol. 51, No. 4, 548–565.
- Benjaafar, S., M. ElHafsi, T. Huang. 2007. “Optimal Control of a Production-Inventory System with both Backorders and Lost Sales”, *under review*.
- Bertsimas, D., D. Nakazato. 1995. “The Distributional Little’s Law and Its Applications”, *Operations Research*, Vol. 43, No. 2, 298–310.
- Bertsimas, D. 2007. *Introduction to Queueing Systems, monograph in preparation*.
- Buzacott, J. A., J. G. Shanthikumar. 1993, *Stochastic Models of Manufacturing Systems*, Prentice Hall, New Jersey.
- Conway, R. W., W. L. Maxwell, L. W. Miller. 1967. *Theory of Scheduling*, Addison-Wesley: Reading, Mass.
- Eppen, G. D. 1979. “Effects of centralization on expected costs in a multi-location newsboy problem”, *Management Science*, Vol. 25, 498-501.
- Gayon, J., F. de Véricourt, F. Karaesmen, Y. Dallery. 2009. “Stock Rationing in an $M/E_r/1$ Multi-class Make-to-Stock Queue with Backorders”, *IIE Transactions*, Vol. 41, 1096-1109.
- Gerchak, Y., Q. M. He. 2003. “On the relation between the benefits of risk pooling and the variability of demand”, *IIE Transactions*, Vol. 35, 1027-1031.
- Gross, D., C. M., Harris. 1998. *Fundamentals of Queueing Theory*, John Wiley & Sons, New York.
- Ha, A. 1997a. “Inventory Rationing Policy in a Make-to-Stock Production System with Several Demand Classes and Lost Sales”, *Management Science* Vol. 43, 1093–1103.
- Ha, A. 1997b. “Stock-Rationing Policy for a Make-to-Stock Production System with Two Priority Classes and Backordering”, *Naval Research Logistics*, Vol. 44, 457–472.
- Ha, A. 2000. “Stock Rationing in an $M/E_k/1$ Make-to-Stock Queue”, *Management Science*, Vol. 46, 77–87.
- Haji, R., G. Newell. 1971. “A Relation Between Stationary Queue and Waiting Time

Distribution”, *Journal of Applied Probability*, Vol. 8, 617–620.

Kerner, Y. 2008. “The Conditional Distribution of the Residual Service Time in the $M_n/G/1$ Queue,” *Stochastic Models*, Vol. 24, 364–375.

Sanajian, N., B. Balcioglu. 2009. “The Impact of Production Time Variability on Make-to-Stock Queue Performance”, *European Journal of Operational Research*, Vol. 194, 847–855.

Sigman, K., U. Yechiali. 2007. “Stationary remaining service time conditional on queue length”, *OR Letters*, Vol. 35, 581–583.

Takagi, H. 1991. *Queueing Analysis*, Volume 1, Elsevier: North Holland, The Netherlands.

Veatch, M., L. M. Wein. 1996. “Scheduling a Make-to-Stock Queue: Index Policies and Hedging Points”, *Operations Research*, Vol. 44, 634–647.

de Véricourt, F., F. Karaesmen, Y. Dallery. 2001. “Assessing the Benefits of Different Stock-Allocation Policies for a Make-to-Stock Production System”, *Manufacturing & Service Operations Management*, Vol. 3, 105–121.

de Véricourt, F., F. Karaesmen, Y. Dallery. 2002. “Optimal Stock Allocation for a Capacitated Supply System”, *Management Science*, Vol. 48, 1486–1501.

Appendix A Proofs

Proof. Theorem 1.

Consider customers of classes $1, \dots, r$ as the high-priority class with an arrival rate λ_r^+ . Let $E[N_r^+]$ and $E[N_r^-]$ denote the expected number of customers of classes $1, \dots, r$ and $r + 1, \dots, n$ in the system, respectively. We call high- and low-priority classes r^+ and r^- respectively.

Using Little’s Law, we have $E[N] = -\lambda \tilde{w}'(s)|_{s=0}$ and $E[N_r^-] = -\lambda_r^- \tilde{w}'_{r^-}(s)|_{s=0}$, where $\tilde{w}(s)$ and $\tilde{w}_{r^-}(s)$ are the LT of the system times in a single class FCFS $M/G/1$ queue and customers of class r^- , respectively. Note that $\tilde{w}_{r^-}(s)$ is given in Observation 4. Since $E[N] = E[N_r^+] + E[N_r^-]$, we have

$$\begin{aligned}
\frac{E[N_r^+]}{E[N]} &= 1 - \frac{E[N_r^-]}{E[N]} = 1 - \frac{-\lambda_r^- \tilde{w}'_{r-}(s)|_{s=0}}{-\lambda \tilde{w}'(s)|_{s=0}} \\
&= 1 - \frac{\lambda_r^- \tilde{w}'(s + \lambda_r^+(1 - \theta_r^+(s)))(1 - \lambda_r^+ \theta_r^{+'}(s))|_{s=0}}{\lambda \tilde{w}'(s)|_{s=0}},
\end{aligned}$$

where due to Eq. (34) $\theta_r^+(s) = \tilde{b}(s + \lambda_r^+(1 - \theta_r^+(s)))$. Since $\theta_r^+(0) = 1$, $\tilde{w}'(0)$ cancels out and then because $\tilde{b}'(s)|_{s=0} = 1/\mu$, we have

$$\begin{aligned}
\frac{E[N_r^+]}{E[N]} &= 1 - \frac{\lambda_r^-}{\lambda} (1 - \lambda_r^+ \theta_r^{+'}(s)|_{s=0}) \\
&= 1 - \frac{\lambda_r^-}{\lambda} \left(1 - \frac{\lambda_r^+ \tilde{b}'(s)|_{s=0}}{1 + \lambda_r^+ \tilde{b}'(s)|_{s=0}}\right) = \frac{\lambda_r^+(1 - \rho)}{\lambda(1 - \rho_r^+)}, \tag{A.1}
\end{aligned}$$

where $\rho_r^+ = \lambda_r^+/\mu$ and $\rho = \lambda/\mu$.

Now consider a second system with two classes of customers where the arrival rates of high- and low-priority customers are λ_{r-1}^+ and λ_{r-1}^- , respectively. The expected number of high-priority customers in this system is $E[N_{r-1}^+]$. The expected number of customers of class r in the multi-priority class can be expressed as

$$E[N_r] = E[N_r^+] - E[N_{r-1}^+].$$

Therefore,

$$\frac{E[N_r]}{E[N]} = \frac{E[N_r^+] - E[N_{r-1}^+]}{E[N]}. \tag{A.2}$$

Substituting Eq. (A.1) applied to the the r^+ and $(r-1)^+$ customers into Eq. (A.2) and letting $\rho_r = \lambda_r/\mu$ we have

$$\begin{aligned}
\frac{E[N_r]}{E[N]} &= \frac{\lambda_r^+(1-\rho)}{\lambda(1-\rho_r^+)} - \frac{\lambda_{r-1}^+(1-\rho)}{\lambda(1-\rho_{r-1}^+)} \\
&= \frac{(1-\rho)[(\lambda_{r-1}^+ + \lambda_r)(1-\rho_{r-1}^+) - \lambda_{r-1}^+(1-\rho_r^+)]}{\lambda(1-\rho_r^+)(1-\rho_{r-1}^+)} \\
&= \frac{(1-\rho)[\lambda_{r-1}^+ + \lambda_r - \lambda_{r-1}^+\rho_{r-1}^+ - \lambda_r\rho_{r-1}^+ - \lambda_{r-1}^+ + \lambda_{r-1}^+\rho_{r-1}^+ + \lambda_{r-1}^+\rho_r^+]}{\lambda(1-\rho_r^+)(1-\rho_{r-1}^+)} \\
&= \frac{\lambda_r(1-\rho)}{\lambda(1-\rho_r^+)(1-\rho_{r-1}^+)}.
\end{aligned}$$

■

Proof. Observation 1.

$E[N]$ can be calculated using Little's Law and Eq. (33) in Corollary 2 that gives the LT of the system time in such a queue. ■

Proof. Theorem 2.

When there is a backlog in the SP system, the SPB queue is busy. Letting N^1 denote the total number of jobs in the SPB queue and N_r the number of type r jobs in this queue, the expected backlog in the SP system $E[B] = E[N^1]$ and $E[B_r] = E[N_r]$. Since the SPB queue exists $(1 - F(S - 1))$ percent of the time, Eq. (4) becomes

$$C_{SP}(S) = h \sum_{i=0}^{S-1} (S - i)p(i) + (1 - F(S - 1))E[N^1] \sum_{r=1}^n b_r \frac{E[N_r]}{E[N^1]}.$$

Substituting $E[N_r]/E[N^1]$ from Theorem 1 establishes Eq. (11). That $E[N^1]$ is given by (10) follows because the total number of customers in such queues is invariant. ■

Proof. Theorem 3.

Given the definition of b^{SP} in Eq. (12) and the optimal cost of the SP policy from Theorem 2, in order to establish Eq. (13), we only need to show that $(1 - F(S - 1))E[N^1] = \sum_{i=S}^{\infty} (i - S)p(i)$. To do this, recall that $E[N^1]$ is the expected number of the backlogs in the original system. In other words, $E[N^1] = E[N - S | N \geq S]$ where N is the total number of customers in the system with an FCFS policy (which is invariant and is the same in the SP

system). Then,

$$\begin{aligned}
(1 - F(S - 1))(E[N|N \geq S] - S) &= (1 - F(S - 1))\left(\sum_{i=S}^{\infty} ip(i|i \geq S) - S\right) \\
&= \sum_{i=S}^{\infty} ip(i) - S(1 - F(S - 1)) \\
&= \sum_{i=S}^{\infty} ip(i) - S \sum_{i=S}^{\infty} p(i) = \sum_{i=S}^{\infty} (i - S)p(i).
\end{aligned}$$

Finally, given the cost in Eq. (13), the optimal base-stock level is given in Eq. (14) as in e.g., Veatch and Wein (1996). ■

Proof. Corollary 1.

If $R_{r+1} > R_r = R_{r-1} = \dots = R_{r-k} > R_{r-k-1}$, as soon as the inventory decreases to R_r , we consider classes $r - k, r - k + 1, \dots, r$ as a single class whose demand is backlogged. The total backlog of all these classes will be $\sum_{i=r-k}^r E[N_i]$, where $E[N_i]$ is the average number of type i customers in the relevant backlog queue. This backlog results in a cost of $\sum_{i=r-k}^r b_i E[N_i]$. By aggregating these classes to a single class with a weighted backlog cost $\sum_{i=r-k}^r b_i E[N_i] / \sum_{i=r-k}^r E[N_i]$ we obtain the same cost. (Note that these ratios, $\sum_{i=r-k}^r b_i E[N_i] / \sum_{i=r-k}^r E[N_i]$, do not require the exact characterization of $b_1(\cdot)$ because they are independent of $b_1(\cdot)$ and can be obtained using Eq. (9) in Theorem 1.) ■

Proof. Theorem 4.

Consider the class r backlog queue for $r = 1, \dots, n$. From Observation 2, backlogs of customers of class r could be calculated from this backlog queue. Thus, $E[B_r] = E[N^r] \times$ (% of low-priority customers in the r^{th} modified queue). Then, Eq. (18) follows, using Theorem 1. Eq. (19) follows in a fashion similar to (10). ■

Proof. Theorem 6.

We define \mathbf{M}_n^1 as the number of low-priority customers left behind by the n th departure and consider four cases.

1. There can be at least one low-priority customer in the system at the last departure

instant; in this case, the server continues working on the next production order. If no high-priority customers arrive during this service time (with probability v_0), the next departure (a low-priority customer) leaves no high-priority customers behind. If exactly one high-priority customer arrives during this service time (with probability v_1), the next departure (a high-priority customer) leaves no high-priority customers behind. Mathematically,

$$P\{M_{n+1}^h = 0 | M_n^h = 0, M_n^l > 0\} = v_0 + v_1.$$

2. The last departure might leave the system empty. If the next customer arriving is a high-priority customer (with probability λ_h/λ) and no high-priority customers arrive during its service time (with probability w_0), the next departure (a high-priority customer) leaves no high-priority customers behind. If the next customer arriving at the idle system is a low-priority customer (with probability λ_l/λ) and, at most, one high-priority customer arrives during its service time (with probability w_0+w_1 , see item 1 for the explanation), the next departure (a high-priority customer with probability w_1 or a low-priority customer with probability w_0) leaves no high-priority customers behind. Hence,

$$P\{M_{n+1}^h = 0 | M_n^h = 0, M_n^l = 0\} = \frac{\lambda_h w_0}{\lambda} + \frac{\lambda_l (w_0 + w_1)}{\lambda} = w_0 + \frac{\lambda_l w_1}{\lambda}.$$

3. There can be at least one low-priority customer in the system at the last departure instant; in this case, the server continues working on the next production order. If $k + 1 \geq 2$ high-priority customers arrive during this service time, the next departure (a high-priority customer) leaves k high-priority customers behind. That is,

$$P\{M_{n+1}^h = k | M_n^h = 0, M_n^l > 0\} = v_{k+1}, \quad k \geq 1.$$

4. The last departure might leave the system empty. If the next customer arriving is a high-priority customer, and k additional high-priority customers arrive during its service time, or if the next customer arriving at the idle system is low-priority, and

$k + 1$ high-priority customers arrive during its service time, the next departure (a high-priority customer) leaves k high-priority customers behind. Hence,

$$P\{M_{n+1}^h = k | M_n^h = 0, M_n^l = 0\} = \frac{\lambda_h w_k}{\lambda} + \frac{\lambda_l w_{k+1}}{\lambda}, \quad k \geq 1.$$

Next, we use the above cases to construct a Markov-Chain (MC) $\widetilde{\mathbf{M}}^h$ with states $k = 0, 1, \dots$. We let its transition probabilities be p_{jk} as in Eq. (26) when $k \geq j - 1, j \geq 1$, and for $j = 0$ we let

$$\begin{aligned} p_{00} &= P\{M_{n+1}^h = 0 | M_n^h = 0, M_n^l > 0\} P\{M_n^h = 0, M_n^l > 0\} \\ &\quad + P\{M_{n+1}^h = 0 | M_n^h = 0, M_n^l = 0\} P\{M_n^h = 0, M_n^l = 0\} \\ &= P\{M_{n+1}^h = 0 | M_n^h = 0, M_n^l > 0\} \frac{\pi_0 - (1 - \rho_b)}{\pi_0} \\ &\quad + P\{M_{n+1}^h = 0 | M_n^h = 0, M_n^l = 0\} \frac{(1 - \rho_b)}{\pi_0} \\ &= \frac{1}{\pi_0} \left\{ (v_0 + v_1)(\pi_0 - (1 - \rho_b)) + (w_0 + \frac{\lambda_l w_1}{\lambda})(1 - \rho_b) \right\}, \end{aligned}$$

and for $k \geq 1$

$$\begin{aligned} p_{0k} &= P\{M_{n+1}^h = k | M_n^h = 0, M_n^l > 0\} P\{M_n^h = 0, M_n^l > 0\} \\ &\quad + P\{M_{n+1}^h = k | M_n^h = 0, M_n^l = 0\} P\{M_n^h = 0, M_n^l = 0\} \\ &= \frac{1}{\pi_0} \left\{ v_{k+1}(\pi_0 - (1 - \rho_b)) + (\lambda_h w_k + \lambda_l w_{k+1}) \frac{(1 - \rho_b)}{\lambda} \right\}. \end{aligned}$$

Note that the normalization $1/\pi_0$ on the RHS represents the time average when the system is at state $M_n^h = 0$. Finally, we observe that with the above definition

$$p_{0k} = \lim_{n \rightarrow \infty} P\{M_{n+1}^h = k | M_n^h = 0\}.$$

Thus, the Theorem follows as in Takagi (1991, p. 289).

■

Proof. Theorem 7. There are three cases:

1. With probability ρ_b , a low-priority customer finds the server busy upon its arrival. In this case, the gross processing time is identical to the one in the preemptive-repeat with re-sampling policy discussed by Conway et al. (1967, p. 171). Let A_1 denote the r.v. corresponding to this gross processing time; its LT $\tilde{a}_1(s)$ and expectation are, respectively:

$$\tilde{a}_1(s) = \frac{(s + \lambda_h)\tilde{b}(s + \lambda_h)}{s + \lambda_h\tilde{b}(s + \lambda_h)}, \quad E[A_1] = \frac{1 - \tilde{b}(\lambda_h)}{\lambda_h\tilde{b}(\lambda_h)}.$$

2. With probability $(1 - \rho_b)w_0$, a low-priority customer finds the server idle upon its arrival and no high-priority customer arrives during the first exceptional service time. Setting $z = 0$ in Eq. (24), it follows that $w_0 = \tilde{b}_1(\lambda_h)$. Let A_2 denote the r.v. corresponding to the gross processing time; its LT $\tilde{a}_2(s)$ and expectation are, respectively (see Conway et al. 1967, p. 171):

$$\tilde{a}_2(s) = \frac{\tilde{b}_1(s + \lambda_h)}{\tilde{b}_1(\lambda_h)}, \quad E[A_2] = -\frac{\tilde{b}_1'(\lambda_h)}{\tilde{b}_1(\lambda_h)}.$$

3. Finally, with probability $(1 - \rho_b)(1 - w_0)$, a low-priority customer finds the server idle upon its arrival, but during its service time at least one high-priority customer arrives. Let A_3 denote the time the low-priority customer stays on the server before a high-priority customer arrives; its LT $\tilde{a}_3(s)$ and expectation are, respectively (see Conway et al. 1967, p. 171):

$$\tilde{a}_3(s) = \frac{\lambda_h(1 - \tilde{b}_1(s + \lambda_h))}{(s + \lambda_h)(1 - \tilde{b}_1(\lambda_h))}, \quad E[A_3] = \frac{\lambda_h\tilde{b}_1'(\lambda_h) + (1 - \tilde{b}_1(\lambda_h))}{\lambda_h(1 - \tilde{b}_1(\lambda_h))}.$$

After the first high-priority customer arrives, the remaining time until the low-priority customer departs from the system will be distributed as A_1 given above. In this case, the summation of A_3 and A_1 will be the gross processing time for the low-priority customer.

Combining these three cases leads to Eq. (28). ■

Proof. Theorem 8.

Observe that $\lambda_l E[A]$ is the proportion of time the server works on orders for low-priority customers. Thus, there are no high-priority customers in the system during this time. Since ρ_b is the proportion of time the server is busy, by PASTA and departures see what arrivals do we have

$$\kappa_0 = 1 - (\rho_b - \lambda_l E[A]). \quad (\text{A.3})$$

Note that in the $M/G/1$ system, only λ_h/λ fraction of departures are high-priority customers. Thus, $\lambda_h \kappa_0/\lambda$ is the fraction -out of all departures- of high-priority customers that leave no high-priority customers in this system. Therefore, in the $M/G/1$ system, only $\lambda_h(1 - \kappa_0)/\lambda$ of departures leave high-priority customers behind, and the theorem follows.

■

Proof. Lemma 1. Based on Theorem 6, for the stochastic process $\widetilde{\mathbf{M}}^h$, the steady-state probabilities that a departure leaves behind k high-priority customers satisfy $\pi_k = \sum_{j=0}^{\infty} \pi_j p_{jk}$. Based on the discussion on the transition-probabilities presented in the proof of Theorem 6, for $k = 0$ we can write

$$\begin{aligned} \pi_0 &= \pi_0 p_{00} + \pi_1 p_{10}, \\ &= \pi_1 v_0 + (\pi_0 - (1 - \rho_b))(v_0 + v_1) + (1 - \rho_b) \left[\frac{\lambda_h}{\lambda} w_0 + \frac{\lambda_l}{\lambda} (w_0 + w_1) \right], \end{aligned}$$

and for $k \geq 1$,

$$\pi_k = \sum_{j=1}^{k+1} \pi_j v_{k-j+1} + (\pi_0 - (1 - \rho_b)) v_{k+1} + (1 - \rho_b) \left(\frac{\lambda_h}{\lambda} w_k + \frac{\lambda_l}{\lambda} w_{k+1} \right).$$

The probability generating function of the number of high-priority customers left in the two-priority class system by an arbitrary departure is

$$\begin{aligned} \Pi(z) &= \sum_{k=0}^{\infty} z^k \pi_k = (\pi_0 - (1 - \rho_b))(v_0 + v_1) + (1 - \rho_b) \left[\frac{\lambda_h}{\lambda} w_0 + \frac{\lambda_l}{\lambda} (w_0 + w_1) \right] \\ &\quad + \sum_{k=0}^{\infty} z^k \sum_{j=1}^{k+1} \pi_j v_{k-j+1} + \sum_{k=1}^{\infty} z^k \left[(\pi_0 - (1 - \rho_b)) v_{k+1} + (1 - \rho_b) \left(\frac{\lambda_h}{\lambda} w_k + \frac{\lambda_l}{\lambda} w_{k+1} \right) \right]. \end{aligned} \quad (\text{A.4})$$

Expanding the following term, which appears on the RHS of Eq. (A.4),

$$\begin{aligned}
\sum_{k=0}^{\infty} z^k \sum_{j=1}^{k+1} \pi_j v_{k-j+1} &= \pi_1 v_0 \\
&+ z\pi_1 v_1 + z\pi_2 v_0 \\
&+ z^2\pi_1 v_2 + z^2\pi_2 v_1 + z^2\pi_3 v_0 \\
&+ \dots
\end{aligned}$$

and using $V(z) = \sum_{k=0}^{\infty} z^k v_k$,

$$\begin{aligned}
\sum_{k=0}^{\infty} z^k \sum_{j=1}^{k+1} \pi_j v_{k-j+1} &= \pi_1 \sum_{k=0}^{\infty} z^k v_k + \pi_2 \sum_{k=0}^{\infty} z^{k+1} v_k \\
&+ \pi_3 \sum_{k=0}^{\infty} z^{k+2} v_k + \dots \\
&= \pi_1 V(z) + z\pi_2 V(z) + z^2\pi_3 V(z) + \dots \\
&= V(z) \sum_{k=1}^{\infty} \pi_k z^{k-1} + \frac{\pi_0 V(z)}{z} - \frac{\pi_0 V(z)}{z} \\
&= \frac{V(z) \sum_{k=0}^{\infty} \pi_k z^k}{z} - \frac{\pi_0 V(z)}{z} \\
&= \frac{\Pi(z) - \pi_0}{z} V(z).
\end{aligned} \tag{A.5}$$

Hence,

$$\begin{aligned}
\Pi(z) &= \frac{\Pi(z) - \pi_0}{z} V(z) + (\pi_0 - (1 - \rho_b))(v_0 + v_1) + (1 - \rho_b) \left[\frac{\lambda_h}{\lambda} w_0 + \frac{\lambda_l}{\lambda} (w_0 + w_1) \right] \\
&+ \sum_{k=1}^{\infty} z^k \left[(\pi_0 - (1 - \rho_b)) v_{k+1} + (1 - \rho_b) \left(\frac{\lambda_h}{\lambda} w_k + \frac{\lambda_l}{\lambda} w_{k+1} \right) \right] \tag{A.6} \\
&= \frac{\Pi(z) - \pi_0}{z} V(z) + (1 - \rho_b) \frac{\lambda_h}{\lambda} W(z) + (\pi_0 - (1 - \rho_b))(v_0 + v_1) \\
&+ (1 - \rho_b) \frac{\lambda_l}{\lambda} w_0 + \sum_{k=1}^{\infty} z^k (\pi_0 - (1 - \rho_b)) v_{k+1} + \sum_{k=0}^{\infty} z^k (1 - \rho_b) \frac{\lambda_l}{\lambda} w_{k+1} \\
&= \frac{\Pi(z) - \pi_0}{z} V(z) + (1 - \rho_b) \frac{\lambda_h}{\lambda} W(z) + (1 - \rho_b) \frac{\lambda_l}{\lambda z} (W(z) - w_0) + (1 - \rho_b) \frac{\lambda_l}{\lambda} w_0 \\
&+ (\pi_0 - (1 - \rho_b)) \left(\frac{V(z)}{z} - \frac{v_0 + z v_1}{z} \right) + (\pi_0 - (1 - \rho_b))(v_0 + v_1) \\
&= \frac{\Pi(z) V(z)}{z} + (1 - \rho_b) W(z) \frac{\lambda_h z + \lambda_l}{\lambda z} + (1 - \rho_b) \lambda_l \frac{w_0 (z - 1)}{\lambda z} \\
&- (1 - \rho_b) \frac{V(z)}{z} + (\pi_0 - (1 - \rho_b)) \frac{v_0 (z - 1)}{z}.
\end{aligned}$$

Solving for $\Pi(z)$, we obtain Eq. (30). ■

Proof. Theorem 9. If the next departing customer is a high-priority customer, there should be at least one high-priority customer present at the last departure or arriving during the current service time. Therefore, we should ignore two types of elements appearing in $\Pi(z)$: (i) the elements corresponding to departures, which leave no high-priority customers behind, and (ii) the elements corresponding to no high-priority customers arriving during the service time. We should also normalize the probabilities π_k by multiplying them by λ/λ_h so that $\Pi_h(z)$ can be obtained. A development similar to Eq. (A.5) leads to

$$\begin{aligned}
\Pi_h(z) &= \frac{\lambda}{\lambda_h} \left(\frac{\Pi(z) - \pi_0}{z} V(z) + (\pi_0 - (1 - \rho_b)) v_1 + (1 - \rho_b) \left[\frac{\lambda_h}{\lambda} w_0 + \frac{\lambda_l}{\lambda} w_1 \right] \right. \\
&\quad \left. + \sum_{k=1}^{\infty} z^k \left[(\pi_0 - (1 - \rho_b)) v_{k+1} + (1 - \rho_b) \left(\frac{\lambda_h}{\lambda} w_k + \frac{\lambda_l}{\lambda} w_{k+1} \right) \right] \right)
\end{aligned}$$

rather than Eq. (A.6) and the proof continues to be similar to Lemma 1. ■

Proof. Theorem 10. Note that a high-priority customer will leave behind n high-priority customers at its departure if there are n high-priority customers arriving during its system time. This is essentially Little's distributional law due to Haji and Newel (1971) and

Bertsimas and Nakazato (1995). Thus,

$$\Pi_h(z) = \tilde{w}_h(\lambda_h(1 - z)),$$

which, after the substitution of $s = \lambda_h(1 - z)$, gives

$$\tilde{w}_h(s) = \Pi_h\left(\frac{\lambda_h - s}{\lambda_h}\right). \quad (\text{A.7})$$

Combining Eq.s (31) and (A.7) and using Eq.s (24-25) yield Eq. (32). ■

Proof. Corollary 2. In the FCFS $M/G/1$ queue with a single class, Eq. (29) becomes $\pi_0 = 1 - \rho_b$ since without any low-priority customers $E[A] = 0$ and $\lambda_h = \lambda$. Similarly, in Eq. (32) we have $\lambda_l = 0$. These modifications reduce Eq. (32) into Eq. (33) in Corollary 2.

■

Proof. Theorem 11. We first set $\lambda_h = \lambda_r^+ = \sum_{i=1}^r \lambda_i$ and $\lambda_l = \lambda_r^- = \sum_{i=r+1}^n \lambda_i$. For a tagged customer in class $r \geq 2$, if no new arrivals come after it joins the queue, the LT of its system time density function will be $\tilde{w}_h(s)$ as given in Eq. (32). Let G be the system time in this queue. To find the actual system time of this customer, we have to include the busy periods generated by customers in classes $1, 2, \dots, r - 1$ that arrive after the tagged customer but before its service completion, namely over G . The total system time for the tagged customer is the delayed busy period, with a delay G that has a LT $\tilde{w}_h(s)$. Note that busy periods induced by customers of types $1, \dots, r - 1$ are similar to those in an $M/G/1$ queue with arrival rate λ_{r-1}^+ ; thus as in Eq. (22), their LT $\theta_{r-1}^+(s)$ is given in Eq. (35). Substituting Eq. (35) and $\tilde{w}_h(s)$ rather than $\tilde{w}(s)$ into Eq. (23) establishes Eq. (34). ■

Proof. Corollary 3. Eq. (36) is a special case of Eq. (32) of Theorem 10 obtained by substituting $\tilde{b}(s)$ for $\tilde{b}_1(s)$, i.e., without an exceptional first service time in a busy period. ■

Proof. Theorem 12. We start by considering a two-priority $M/G/1$ queue with an exceptional first service time in a busy period with density function $b_1(\cdot)$ and the other service times with a density function $b(\cdot)$. The LT of the system time of the high-priority customers in this two-priority class system is given in Eq. (32) of Theorem 10. Let $b_1^h(\cdot)$ denote the service time density function of a high-priority customer that finds no high-priority customers

in the system upon its arrival. If there are no low-priority customers in the system upon the arrival of the high-priority customer, $b_1^h(\cdot) = b_1(\cdot)$. However, if there is at least one low-priority customer in the system, $b_1^h(\cdot)$ will be distributed as the residual service time of the item currently in service. Thus, the system time of high-priority customers in this two-priority queue (with an exceptional first service time with density $b_1(\cdot)$) is identical to the one in a single class FCFS $M/G/1$ queue with an exceptional service time with density $b_1^h(\cdot)$ and an arrival rate equals to the arrival rate of the high-priority customers. Now, in the absence of low-priority customers, we can employ Eq. (33) from Corollary 2 setting $\lambda = \lambda_r^+$ and observing that $1 - \rho_b = \kappa_0$ to obtain the LT of the system time for high-priority customers

$$\tilde{w}_h(s) = \frac{\kappa_0(\lambda_h(\tilde{b}(s) - \tilde{b}_1^h(s)) + s\tilde{b}_1^h(s))}{s - \lambda_h(1 - \tilde{b}(s))}.$$

According to our construction, the $\tilde{w}_h(s)$ above equals the LT in Eq. (32). Equating these and solving for $\tilde{b}_1^h(s)$, we obtain

$$\begin{aligned} \tilde{b}_1^h(s) = & \frac{\kappa_0\lambda_h\tilde{b}(s) + \tilde{b}(s)(1 - \rho_b)(\lambda_l w_0 - \lambda)}{\kappa_0(\lambda_h - s)} \\ & + \frac{(\pi_0 - (1 - \rho_b))v_0\lambda(\tilde{b}(s) - 1) + (1 - \rho_b)(\tilde{b}_1(s)(\lambda - s) - \lambda_l w_0)}{\kappa_0(\lambda_h - s)}. \end{aligned} \quad (\text{A.8})$$

Eq. (A.8) provides the LT of the residual service time, given that there are no high-priority customers in the system, establishing Eq. (38).

To obtain the Laplace Transform of the residual service time when there is at least one customer in the system, we follow Lemma 3.1.1.1 due to Kerner (2008). To this end, we define a continuous time Markov process with states (j, η) where j is the number of high-priority customers in the system, and η denotes the remaining service time. We define $p_t(j, \eta)$ as the probability that there are j high-priority customers in the system, and remaining service time is η at time t . Furthermore, we assume the existence of limiting probabilities, i.e.,

$\lim_{t \rightarrow \infty} p_t(j, \eta) = p(j, \eta)$. Therefore, we have,

$$p_{t+dt}(1, \eta) = p_t(1, \eta + dt)(1 - \lambda_h dt) + p_t(2, 0)b(\eta)dt + p_t(0, 0)\lambda_h b_1^h(\eta)dt, \quad j = 1,$$

$$p_{t+dt}(j, \eta) = p_t(j, \eta + dt)(1 - \lambda_h dt) + p_t(j - 1, \eta + dt)\lambda_h dt + p_t(j + 1, 0)b(\eta)dt, \quad j \geq 1,$$

which, after taking the limit $t \rightarrow \infty$, and noting that $p(0, 0) = \kappa_0$ by definition, become

$$p(1, \eta) = p(1, \eta + dt)(1 - \lambda_h dt) + p(2, 0)b(\eta)dt + \kappa_0 \lambda_h b_1^h(\eta)dt, \quad j = 1$$

$$p(j, \eta) = p(j, \eta + dt)(1 - \lambda_h dt) + p(j - 1, \eta + dt)\lambda_h dt + p(j + 1, 0)b(\eta)dt, \quad j \geq 1.$$

Now, similar to the analysis in Kerner (2008) in Section 3.1.2, Lemma 3.1.3.1, and the proof of Corollary 2.2.1, we obtain Eq. (37). ■