Submitted to *Operations Research*
manuscript (Please, provide the mansucript number!)

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

Hossein Abouee-Mehrizi
Joseph L. Rotman School of Management, University of Toronto, Toronto, M5S 3E6, CANADA,
H.Abouee07@Rotman.Utoronto.Ca

Barış Balcıoğlu
Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, ON M5S 3G8, CANADA,
baris@mie.utoronto.ca

Opher Baron
Joseph L. Rotman School of Management, University of Toronto, Toronto, M5S 3E6, CANADA,
Opher.Baron@Rotman.Utoronto.Ca

Make-to-stock queues are typically investigated in the $M/M/1$ settings. For centralized single-item systems with backlogs, the Multilevel Rationing (MR) policy is established as optimal and the Strict Priority (SP) policy is a practical compromise, balancing cost and ease of implementation. However, the optimal policy is unknown when service time is general, i.e., for $M/G/1$ queues. Dynamic programming, the tool commonly used to investigate the MR policy in make-to-stock queues, is less practical when service time is general. In this paper, we focus on *customer composition*: the proportion of customers of each class to the total number of customers in the queue. We do so 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 the 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. From our results, we derive the optimal cost and control for the MR and SP policies for $M/G/1$ make-to-stock queues.

Key words: Make-to-Stock, $M/G/1$ queue, priority classes, customer composition, multilevel rationing, strict priority

1. Introduction

Market segmentation and customer differentiation are widely accepted ways to increase profitability. A common way to differentiate 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. An important research and managerial question is whether customer classes requesting the same product should be prioritized and if so how to prioritize them. In our examination of this question, we analyze inventory control strategies for a supplier using a centralized inventory to serve a single product to n classes of customers. Assuming that class 1 has the highest priority and class n has the lowest priority, we model the underlying production system as an $M/G/1$ queue.

Many policies are available to handle production and inventory control. Broadly speaking, however, inventory control policies can be characterized by whether 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 centralized inventory control policies with postponement of the allocation decision. Note that because postponing allocation provides extra information, it should result in the same or a lower total cost as not postponing.

We assume that demand that is not immediately satisfied from stock is backlogged. Similar to earlier literature, we consider a first-come-first-served (FCFS) policy analyzed by Sanajian and Balcıoğlu (2009) along with the following two centralized inventory control policies that use a base-stock level control for their production decision:

MR Policy Under a *Multilevel 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 a FCFS basis. 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 there is no stock, a finished product is allocated to the highest-priority customer backlogged (in a FCFS fashion within this class). When the inventory reaches R_{n+1}, the base-stock level, production stops.

SP Policy The *Strict Priority* policy is a special case of MR policy when $R_1 = R_2 = \cdots = R_n = 0$. That is, as long as there is stock in the centralized inventory, demand requests are satisfied on a FCFS basis. When there are backlogs, a finished product is allocated to the highest-priority customer among those with pending orders in the system.

Ha (1997a and 1997b) was the first to discuss inventory rationing problems in a centralized make-to-stock system with different classes of customers. For exponentially distributed production times, Poisson arrivals and lost sales, Ha (1997a) shows that the multilevel rationing (MR) 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 MR policy is the optimal policy for the $M/M/1$ make-to-stock queues. de Véricourt et al. (2001) introduce the strict priority (SP) policy and compare the FCFS, SP, and MR policies for an $M/M/1$ queueing system, and demonstrate that the MR policy outperforms the other two. 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. Gayon et al. (2009) propose a heuristic to approximate these levels for systems with Erlangian service times. Applications of rationing inventory have been also investigated when supply is ample; see Arslan et al. (2007) and references therein.

In this paper, we consider the SP and MR policies for a centralized single product multi-class $M/G/1$ system. While the characterization of the optimal FCFS policy in this setting is known, we are the first to consider the MR and SP policies. We focus on cases where the product allocation is postponed to the end of production when it is allocated to one customer, possibly according to the customer priority. Note that this allocation does not change the total inventory level, but may reduce costs. We ignore additional information, such as the length of time since the start of production of the current item, something which 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. While not using additional information might increase the costs of these policies relative to the optimal control policy, however, it keeps implementation simple and increases practicality.

Observe that in the MR system, the rate of change of the inventory level varies dynamically according to the rationing levels; this also changes *customer composition*, i.e., the proportion of each customer class out of the total number of customers in this queue. Note that because the total number of customers is invariant for every non-idling and non-anticipating policy (for a rigorous definition of such policies see e.g., Bertsimas, 2007), the various controls only change customer composition.

To express customer composition under MR and SP policies, we consider a series of multi-priority class $M/G/1$ queues. In these queues, the first service time in each busy period is different from other service times, i.e., these queues have exceptional first service times in their busy periods. We show that with a careful choice of the exceptional first service times, their customer composition will be the same as the original $M/G/1$ system.

We obtain closed form expressions for the optimal cost and base-stock level for an $M/G/1$ make-to-stock system under the SP policy. We also derive a computational approach to obtain the optimal cost and rationing levels for the MR policy for an $M/G/1$ system, i.e., with **general** service times. Previous work found these optimal controls using dynamic programming for exponential (or Erlang), service times, but when the service times are not exponential, dynamic programming is less practical. For example, Gayon et al. (2009) highlighted the difficulty finding the optimal controls in $M/E_k/1$ settings when the number of customer types is large. However, because the customer composition methodology employs a series of queues it allows the solution of systems with numerous customer types, as we demonstrate numerically in Section 3.4.2. We also show that the cost of the SP system is equivalent to the cost of a FCFS system with an appropriately defined backlog cost. Our theoretical and numerical results support the applicability of both the SP and MR policies for single product multi-class $M/G/1$ systems.

As discussed above, our solution for the SP and MR policies relies on (i) the exact analysis of a multi-priority $M/G/1$ queue with postponement and exceptional first service times in its busy periods, and (ii) characterizing the relevant exceptional first service times. Because the derivation of both is technical and intricate, we only present it in EC.1. In Section 2, we present the multi-class $M/G/1$ system and the terminology used in the paper. In Section 3, we derive the optimal rationing levels, base-stock levels, and costs of the FCFS, SP and MR policies. The proofs of the main results in Theorems 1 and 2 appear in Section 4 and the rest of the proofs appear in EC.2.

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

The single product multi-class $M/G/1$ system we consider has a supplier that produces a single product and caters to demand arising from n distinguishable classes. We assume that the demand of each class r (type r demand) follows a Poisson process with a rate λ_r, $r = 1, 2, \ldots, n$. We use the terms type r and class r interchangeably. We model the general production times as i.i.d with a mean $1/\mu$ and a second moment m_2. Let $b(\cdot)$ and $\tilde{b}(\cdot)$ denote the probability density function and its Laplace Transform (LT), respectively.

We assume that unsatisfied demand is backlogged. Thus, for stability, we require $\rho := \lambda/\mu < 1$, 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 $b_1 > b_2 > \cdots > b_n$ (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 MR 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 the system, and note that $I(t) < 0$ 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 MR system, we 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 (a) allocation, (b) demand, and (c) service processes that are identical to those faced by the original system. 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 the shortfall queue to match the original FCFS and SP inventory systems to a queueing model.

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 system by analyzing a multi-class $M/G/1$ queue with the same allocation, demand, and service processes as in 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, we have:

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

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

Earlier we defined *customer composition* as the proportion of each customer class in 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 MR systems when they have a backlog, we construct multi-class single-item $M/G/1$ queues with postponement of allocation and exceptional first service times in busy periods. We name these queues “backlog queues” for simplicity. 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 use the backlog queues to derive the exact cost of the SP and MR systems in Sections 3.2 and 3.3, respectively. For the sake of completeness (and better comparative analysis), in Section 3.1

we begin with the optimal control and corresponding cost for the FCFS system. We compare the performances of the three systems in Section 3.4.

The solution of multi-class $M/G/1$ queues with exceptional first service times and the derivation of the LT of the required exceptional service times are presented in EC.1.

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

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

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),

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

Note that $P(i)$ can be obtained in closed form using Eq. (12) in Kerner (2008) after setting $\lambda_i = \lambda$ as

$$P(i) = (1-\rho) \prod_{j=0}^{i-1} \frac{1-\tilde{b}_j(\lambda)}{\tilde{b}(\lambda)}, \quad i = 1, \ldots \tag{4}$$

where $\tilde{b}_j(\cdot)$ is the LT of the residual service time observed by an order arrival that sees j orders in the shortfall queue. This LT can be obtained recursively from Eq. (4) in Kerner (2008):

$$\tilde{b}_j(s) \;=\; \frac{\lambda}{s-\lambda} [\tilde{b}(\lambda) \frac{1-\tilde{b}_{j-1}(s)}{1-\tilde{b}_{j-1}(\lambda)} - \tilde{b}(s)], \quad j \geq 1,$$

where $\tilde{b}_0(s) = \tilde{b}(s)$.

3.2. The SP Policy

We next express the cost of the SP system with a base-stock level S. Let $P(B_r = i)$ denote the steady-state probability of having i backlogs from class r. The average cost for the SP system is

$$C_{SP}(S) := h\sum_{i=0}^{S}(S-i)P(i) + \sum_{r=1}^{n} b_r \sum_{i=0}^{\infty} iP(B_r = i) = h\sum_{i=0}^{S}(S-i)P(i) + \sum_{r=1}^{n} b_r E[B_r], \tag{5}$$

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 insufficient to express these costs. We 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 We construct the SPB queue to obtain the probabilistic description of the shortfall queue during periods with no inventory. To differentiate between queues, we use the terms customers in the SP system, orders in the shortfall queue, and job in the SPB queue.

We construct the SPB queue by specifying its (a) allocation, (b) arrival, and (c) service processes. As proved in Theorem 1, our construction ensures that the job composition in the SPB queue will match the customer composition in the SP system when there is no inventory in the system, i.e., when $N(t) \geq S$.

Step (a): at the end of each service completion, the SPB queue will remove the oldest job with the smallest r index, making it a priority queue with the same priorities as the SP system when it has no inventory.

Step (b): the arrival process of jobs of type r to the SPB queue will follow a Poisson process with rate λ_r, $r = 1, 2, \ldots, n$. Thus, the arrival processes for the SP system and the SPB queues are identical (in distribution).

Step (c): the first service time in each busy period of the SPB queue will be the equilibrium (steady-state) residual service time observed by an order arrival who finds S orders in the shortfall queue upon its arrival. We let $\tilde{b}_0^{SPB}(\cdot)$ denote the LT of this equilibrium residual service time. When an exceptional first service time has ended, if there are other jobs in the SPB queue, all service times until the SPB queue clears all its jobs, follow a regular service distribution, with a LT $\tilde{b}(\cdot)$.

We set the service process to include the exceptional first service time in step (c) because **every** order arrival that sees S orders in the shortfall queue creates a backlog. Thus, the service times of the first jobs in the busy periods of the SPB queue are identical in distribution to the residual service times of the customers in service in the SP system once a period with backlog starts.

To summarize: our construction in steps (a-c) indicates that the SPB queue is an $M/G/1$ priority-queue with postponement and exceptional first service times in its busy periods. These exceptional first service times have a LT $\tilde{b}_0^{SPB}(\cdot)$ 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. The LT of the other service times is that of regular service times, $\tilde{b}(\cdot)$.

Let $P_r^{SPB}(i)$ denote the steady-state probability of having i jobs of class r in the SPB queue. We next state our first main result. Its proof is given in Section 4.

THEOREM 1. *The steady-state probability of having i backlogs from class r in the SP system is*

$$P(B_r = i) = [1 - F(S-1)]P_r^{SPB}(i), \;\; r = 1, 2, ..., n, \;\; i = 1, ... \tag{6}$$

Note that Theorem 1 demonstrates that the probability of having n type r backlogs in the SP system is identical to the probability of having n type r jobs in the SPB queue given the system is out of stock. The latter depends of course on $\tilde{b}_0^{SPB}(\cdot)$. While, the theorem does not provide these probabilities, they are not required to express the cost function given in Eq. (5), all we need is the expected number of type r backlogs in the system. Given Theorem 1, this expectation is identical to the expected number of type r jobs in the SPB queue given the system is out of stock. Thus, we next characterize it.

The customer composition in the SPB queue is an essential building block in our analysis of the SP and MR systems, and is given in Theorem 2 below. The proof of the theorem requires the derivations from EC.1 and is given in Section 4.

THEOREM 2. ***Customer composition:*** *The ratio of expected number of type r customers, $E[N_r^{SPB}]$, to the expected number of total customers, $E[N^{SPB}]$, in the SPB queue is*

$$\frac{E[N_r^{SPB}]}{E[N^{SPB}]} := \frac{1-\rho}{\rho}\left(\frac{1}{1-\rho_r^+} - \frac{1}{1-\rho_{r-1}^+}\right), \tag{7}$$

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

Observe that, surprisingly, the ratio in Eq. (7) is independent of $b_0^{SPB}(\cdot)$, and this ratio only depends on the first moments of the queue's arrival and service processes.

3.2.2. Deriving the Optimal SP Policy 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 a FCFS single class $M/M/1$ queue with a specific backlog cost. Theorem 3 uses Theorems 1 and 2 to 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 a FCFS single class $M/G/1$ queue with weighted backlog cost:*

$$b^{SP} = \sum_{r=1}^{n} \frac{\lambda_r(1-\rho)b_r}{\lambda(1-\rho_r^+)(1-\rho_{r-1}^+)}. \tag{8}$$

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

$$C_{SP}(S) = h\sum_{i=0}^{S-1}(S-i)P(i) + b^{SP}\sum_{i=S}^{\infty}(i-S)P(i), \tag{9}$$

and the optimal base-stock level S^{SP^*} that minimizes Eq. (9) is

$$S^{SP^*} = \min\{i : F(i) > b^{SP}/(h+b^{SP})\}. \tag{10}$$

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. More specifically, we do not need to solve the SPB queue or characterize its exceptional first service times. 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. (8), and we express S^{SP^*} and its corresponding cost using Eq.s (10) and (9), respectively.

3.3. The Multilevel Rationing Policy

Let $C_{MR} := C(R_1 = 0, R_2, ..., R_{n+1} = S)$ be the long-run average cost of the MR system given rationing levels $R_1, R_2, ..., R_{n+1}$. In this section, we derive the closed form 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, $I(t) \in (R_i, R_{i+1}]$ for $i = 1, ..., n$ and $I(t) \leq 0$, by considering a properly defined backlog queue.

The proof of the following corollary relies on Theorem 2.

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

3.3.1. The MR Backlog Queues

Here we construct a series of backlog queues for each class $r = 1, ..., n+1$. We denote class r backlog queue by BQ_r. In Theorem 5 we show that the job composition in the backlog queues is identical to the relevant customer composition in the MR system.

We constructed the SPB queue by carefully constructing its (a) allocation, (b) arrival, and (c) service processes when $I(t) \leq 0$. We follow steps (a)-(c) below, formulating BQ_r for $I(t) \leq R_r$ as a two-priority $M/G/1$ queue with postponement and exceptional first service times in its busy periods.

Step (a): we set BQ_r as a two-priority queue in which priority is given to the jobs of classes $1, ..., r-1$ over jobs of class r.

The intuition behind step (a) is that once the inventory hits a rationing level and the customer composition changes, only the priority of a single class of customers changes; all other classes are treated as before. For example, once the inventory falls below $R_n + 1$, classes $1,...,n-1$ remain high-priority, receiving items from inventory upon arrival; and only the priority of class n customers changes from high to low. Therefore, BQ_n is a two-priority queue in which jobs of types $1,...,n-1$ are high-priority, and jobs of type n are low-priority.

Step (b): we set the arrival processes of all job types to be Poisson, and let the high- and low-priority jobs arrival rates at BQ_r be $\lambda^+_{r-1} := \sum_{i=1}^{r-1} \lambda_i$ and λ_r, respectively. This queue ignores customers of classes $r+1,...,n$.

We set the arrival rates of the low and high-priority jobs in BQ_r as defined in step (b) because:

Observation 1. For any class $r = 2,\cdots,n$, 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 is possible that all stock will be depleted and all demand 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 $j \leq r$.

Observation 1 implies that any change of class r backlog in the MR system corresponds to a change of the low-priority job in BQ_r and to a change in the high-priority job in BQ_j for $j > r$. However, this change of the class r backlog does not affect BQ_j for $j < r$. Thus, we ignore class r when considering BQ_j for $j < r$, i.e., the backlog queues of higher priority classes.

Step (c): we set the service process of the BQ_r to have exceptional first service times in busy periods and regular service times with LT of $\tilde{b}(\cdot)$ otherwise. We set the LT of the exceptional service times to be the LT of the residual service times observed by a high-priority arrival at BQ_{r+1} that sees $R_{r+1} - R_r$ jobs in the queue. We let $\Delta_r := R_{r+1} - R_r$ for $r = 1,...,n$ and denote this LT by $\tilde{b}^r_{\Delta_r}(\cdot)$. (With this notation, $\tilde{b}^n_S(\cdot)$ is identical to $\tilde{b}^{SPB}_0(\cdot)$, the LT of the exceptional first service times in the SPB queue.)

The intuition behind step (c) is as follow: for the SPB queue, we set the distribution of the exceptional first service times as the equilibrium residual service time observed by arrivals that see S orders in the shortfall queue. In the SP system, the first service times depend on all orders in the system because all arrivals reduce the inventory towards 0 (the level where the customer composition changes). However, in BQ_r only high-priority jobs in BQ_{r+1} correspond to customers that may reduce the inventory in the system to R_r. Consider high-priority job arrivals in BQ_{r+1} that see $R_{r+1} - R_r$ high-priority jobs. **Every** such arrival corresponds to a customer that decreases the inventory in the system to $R_r - 1$ or creates a class r backlog. With our construction, every such high-priority arrival corresponds to jobs that start the busy period in BQ_r. Therefore, we set the first service times in busy periods in BQ_r as the equilibrium residual service times observed by high-priority arrivals that see $R_{r+1} - R_r$ high-priority jobs in BQ_{r+1}. This choice makes the service time of the first jobs in busy periods of BQ_r identical, in distribution, to the required residual service times. As Theorem 5 below states, this construction together with steps (a) and (b) results in a job composition in BQ_r that is identical to the relevant (classes $1, ..., r$) customer composition in the MR system when $I(t) \leq R_r$.

To summarize: For $r = 2, ..., n$, BQ_r is a two-priority $M/G/1$ queue with high- and low-priority customer arrival rates $\lambda_{r-1}^{+} = \sum_{i=1}^{r-1} \lambda_i$ and λ_r, respectively, and exceptional first service times in its busy periods. The LT of the exceptional first service times is $\tilde{b}_{\Delta_r}^{r}(\cdot)$ and the LT of regular service times is $\tilde{b}(\cdot)$.

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

Note that we can calculate the backlog of class 1 customers from BQ_2 (this is $(i - R_2)^+$ where i is the number of high-priority customers in BQ_2). However, as shown in Theorem 2, finding the expected number of customers in a backlog queue can be done in closed form. Thus, to reduce the computational burden, we use BQ_1. The exceptional first service times for this queue are the residual service times seen by high-priority arrivals at BQ_2 that see Δ_1 high-priority jobs in the queue, i.e., its LT is $\tilde{b}^1_{\Delta_1}(\cdot)$.

Let ρ_b be the server utilization in BQ_r, and $1/\mu_1$ and m^1_2 be the first and second moments of the exceptional first service times in BQ_r. Both $1/\mu_1$ and m^1_2 can be obtained using $\tilde{b}^r_{\Delta_r}(s)$ that can be derived using Theorem EC.2 presented in EC.1.2. (For notational convenience and because the context is clear, we omit the superscript r from ρ_b, μ_1, and m^1_2 in BQ_r.) 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

$$\rho_b = \frac{\lambda^+_r(1-\rho_b)}{\mu_1} + \frac{\lambda^+_r\rho_b}{\mu} = \frac{\lambda^+_r\mu}{\mu_1\mu+\lambda^+_r(\mu-\mu_1)}. \tag{11}$$

Let $E[N^{BQ_r}]$ and $E[N^{BQ_r}_l]$ denote the expectation of the number of all (total) and low-priority jobs in BQ_r, respectively. Also, for $r=1,..,n+1$, let $P^{BQ_r}_h(i)$ and $P^{BQ_r}_l(i)$ denote the steady-state probability of having i high- and low-priority jobs in BQ_r, respectively. Next we derive close form expressions for $E[N^{BQ_r}]$ and $E[N^{BQ_r}_l]$, and generalize Eq. (2), which is given for the distribution of the total number of orders in a FCFS $M/G/1$ queue, to the distribution of the number of high-priority jobs in BQ_r. We define $\prod_{i=k}^{l}(\cdot) := 1$ for $k>l$.

Theorem 4. *Consider BQ_r. Then,*

1. *The expected number of type r jobs in BQ_r is*

$$E[N^{BQ_r}_l] = \sum_{j=0}^{\infty} jP^{BQ_r}_l(j) = E[N^{BQ_r}]\frac{\lambda_r}{\lambda^+_r}\frac{1-\rho^+_r}{(1-\rho^+_r)(1-\rho^+_{r-1})}, \tag{12}$$

where $\rho^+_r = \lambda^+_r/\mu$ for $r=1,\dots,n$, $\rho^+_0 := 0$, $\lambda^+_r = \sum_{i=1}^{r}\lambda_i$ as before, and

$$E[N^{BQ_r}] = (1-\rho_b)\lambda^+_r\frac{(\lambda^+_r)^2 m_2/\mu_1 + (1-\rho)(\lambda^+_r m^1_2 + 2/\mu_1)}{2(1-\rho)^2}. \tag{13}$$

2. *The probability of having i high-priority jobs in BQ_r is,*

$$P^{BQ_r}_h(i) = \frac{\lambda^+_{r-1}(1-(\rho_b-\lambda_r E[A]))}{\lambda^+_r}\prod_{j=0}^{i-1}\frac{1-\tilde{b}^{r-1}_j(\lambda^+_{r-1})}{\tilde{b}(\lambda^+_{r-1})},\ i=1,... \tag{14}$$

where ρ_b is given in Eq. (11), $E[A]$ is given in Lemma EC.2 and $\tilde{b}^{r-1}_j(\cdot)$ are given in Theorem EC.2.

The proof of Theorem 4 relies on Theorem 2 and uses a similar derivations to that in Kerner (2008). Note that $P_h^{BQ_r}(i)$ is a function of $\tilde{b}_j^{r-1}(\cdot)$ that can be obtained recursively using Algorithm 1 given in EC.1.3 starting with $\tilde{b}_0^n(\cdot) = \tilde{b}(\cdot)$.

Finally, the system's inventory and backlog probabilities can be obtained from BQ_j with $j = 2, ..., n+1$ and $j = 1, ..., n$, respectively, as given in Theorem 5 below. Although the proof of Theorem 5 is similar to the proof of Theorem 1 for the SP system, it requires more work. The proof ties BQ_r to BQ_{r+1} using induction, and then ties BQ_r to the MR system. Table 1 below summarizes the relations between these queues and the MR system.

Table 1 Relations between backlog queues and the MR system

r^{th} **backlog queue**	$(r+1)^{st}$ **backlog queue**	**The MR system**
Queue is relevant:	Once the total number of high-priority jobs in the $(r+1)^{st}$ backlog queue increases to $\triangle_r$.	when $I(t) \le R_r$.
The first service time in a busy period corresponds to:	The residual service time of a high-priority job that sees $\triangle_r$ high-priority jobs in this queue upon arrival.	The residual service time of a customer arrival of classes $1 \dots r$ that finds both $I(t) = R_r$ 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 $\triangle_r$ 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), corresponds to:	A service completion that reduces the total number of high-priority jobs in this queue to $\triangle_r$.	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, i.e., jobs of classes $1 \cdots r-1$.	Customers of classes 1 to $r-1$.

Let $F_h^{BQ_r}(i) := \sum_{j=0}^{i} P_h^{BQ_r}(j)$ and $\bar{F}_h^{BQ_r}(i) = 1 - F_h^{BQ_r}(i)$.

THEOREM 5. (i) *The steady-state probability of having i backlogs from class r in the MR system is,*

$$P(B_r = i) = \prod_{j=r+1}^{n+1} \bar{F}_h^{BQ_j}(\Delta_{j-1} - 1) P_l^{BQ_r}(i), \ r = 1, 2, ..., n, \ i = 0, 1, ... \tag{15}$$

(ii) *The steady-state probability of having $R_r - i$ inventory units in the MR system is,*

$$P(I = R_r - i) = \prod_{j=r+1}^{n+1} \bar{F}_h^{BQ_j}(\Delta_{j-1} - 1) P_h^{BQ_r}(i), \ r = 2, ..., n+1, \ i = 0, 1, ..., \Delta_{r-1} - 1. \tag{16}$$

3.3.2. The Cost of the MR Policy Here we express C_{MR} in closed form using the backlog queues defined above. Combining Theorems 4 and 5 the total cost of the MR system is (no further proof is provided):

THEOREM 6. *The long-run average cost of the MR policy is*

$$C_{MR} = h\sum_{r=2}^{n+1}\left[\prod_{j=r+1}^{n+1}\bar{F}_h^{BQ_j}(\Delta_{j-1}-1)\sum_{i=0}^{\Delta_{r-1}-1}(R_r - i)P_h^{BQ_r}(i)\right] \\ +\sum_{r=1}^{n} b_r\left[\prod_{j=r+1}^{n+1}\bar{F}_h^{BQ_j}(\Delta_{j-1}-1)E[N_l^{BQ_r}]\right]. \tag{17}$$

We remind that $E[N_l^{BQ_r}]$ and $P_h^{BQ_r}(i)$ are given in closed form in Theorem 4. To calculate Eq.s (13) and (14), we obtain the LTs of the exceptional first service times in BQ_r by recursively using Theorem EC.2. While the cost in Eq. (17) is a closed form expression, it is quite cumbersome because it uses the LT of different equilibrium residual service times.

3.3.3. Searching for the Optimal MR Policy For a given set of rationing levels $R_1,\dots,R_{n+1}$, if $R_i = R_{i+1}\cdots = R_j$, we first aggregate customers of classes $i,\dots,j$ as a single class and normalize their backlog costs using Theorem 2. Then, we calculate the cost of the MR system using Theorem 6. We start with BQ_{n+1} that is a FCFS $M/G/1$ queue with an arrival rate $\lambda = \sum_{i=1}^{n}\lambda_i$ and obtain the probabilities $P_h^{BQ_{n+1}}(i)$ using Eq. (4). To obtain the probabilities $P_h^{BQ_r}(i)$ for $BQ_2,\dots,BQ_{n+1}$, we use Eq. (14) that requires $\tilde{b}_j^{r-1}(\cdot)$. We calculate these LTs using Theorem EC.2. Finally, we obtain $E[N_l^{BQ_r}]$ using Theorem 4 without calculating $P_l^{BQ_r}(i)$. With the exact cost C_{MR} calculated using this procedure for given rationing levels, we can search over different vectors of $(R_1,R_2,...,R_{n+1})$ to find the optimal rationing levels and the corresponding cost.

3.4. Comparison of the Three Policies

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

3.4.1. Theoretical Comparison Note that the SP control is a special case of the MR 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 2 below summarizes this and provides theoretical support for the use of the MR 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 2. We have

$$C^*_{MR} \le C_{SP}(S^{SP^*}) \le C_{CF}(S^{F^*}).$$

3.4.2. Numerical Comparison Our methodology can be used to find the optimal control and cost for 2, 5, and 10 customer classes. Since $M/M/1$ make-to-stock systems have been investigated (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($\mu_1 = 1.05523, \mu_2 = 0.09477, a_1 = 0.99504$) (Altıok, 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 $\rho = 0.8, 0.9$ 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 tests. For each test we calculated the ratios

$$\Delta SP := \frac{C_{SP}(S^{SP^*}) - C^*_{MR}}{C^*_{MR}} \times 100, \quad \Delta F := \frac{C_{CF}(S^{F^*}) - C^*_{MR}}{C^*_{MR}} \times 100. \tag{18}$$

Table 2 presents the results of these numerical experiments and shows that using the MR and SP policies can significantly reduce costs, compared to the optimal FCFS policy.

Table 2 ΔSP and ΔF for multiple classes of customers

		$n=2$		$n=5$		$n=10$	
cv^2	ρ	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. Proofs of the Main Results

In this section we provide the proofs of our two main results. In Theorem 1, we show that the distribution of the number of customers in an M/G/1 queue with priorities that depend on the number of customers in the system can be deduced by investigating a multi-priority $M/G/1$ queue with an exceptional service time. In Theorem 2, we characterize the cost composition in such queues.

Proof of Theorem 1. We first prove that the steady-state distribution of number of jobs in the SPB queue is identical to the steady-state distribution of number of backlogs in the system given that the system is out of stock:

$$P(S+i)=[1-F(S-1)]P^{SPB}(i), \;\; i=0,1,..., \tag{19}$$

where $P^{SPB}(i)$ denotes the steady-state probability of having i jobs in the SPB queue. We then establish that the job composition in the SPB queue is identical to the customers backlog composition in the SP system given that the system is out of stock.

Eq. (19) states that $P^{SPB}(i)$, is identical to the steady-state probability of having $S+i$ orders in the shortfall queue given that $N(t)\geq S$.

Using Eq. (4), the steady-state probability of having $(S+i)$ orders in the shortfall queue is,

$$P(S+i)=P(0)\prod_{j=0}^{S+i-1}\frac{1-\tilde{b}_j(\lambda)}{\tilde{b}(\lambda)}=P(S)\prod_{j=0}^{i-1}\frac{1-\tilde{b}_{S+j}(\lambda)}{\tilde{b}(\lambda)}, \;\; i=0,1,... \tag{20}$$

We next obtain the steady-state probability of having i jobs in the SPB queue. The derivation is similar to the one for the $M/G/1$ queue in Kerner (2008). We define $q_t(i,\eta)$ as the probability that there are i jobs in the SPB queue, and remaining service time is η at time t. Therefore, we have,

$$q_{t+dt}(1,\eta) = q_t(1,\eta+dt)(1-\lambda dt)+q_t(2,0)b(\eta)dt+q_t(0,0)\lambda b_0^{SPB}(\eta)dt, \quad i=1, \tag{21}$$

$$q_{t+dt}(i,\eta) = q_t(i,\eta+dt)(1-\lambda dt)+q_t(i-1,\eta+dt)\lambda dt+q_t(i+1,0)b(\eta)dt, \; i>1, \tag{22}$$

where $b_0^{SPB}(\cdot)$ is the density of the first exceptional service times in the SPB queue. Then, similar to the proof of Lemma 3.1.3.1 in Kerner (2008) we have,

$$P^{SPB}(i) = P^{SPB}(0) \prod_{j=0}^{i-1} \frac{1 - \tilde{b}_j^{SPB}(\lambda)}{\tilde{b}(\lambda)}, \tag{23}$$

where $\tilde{b}_j^{SPB}(\cdot)$ is the LT of the equilibrium residual service times observed by arrivals who find j jobs in the SPB queue.

Setting $\lambda_l = 0$, $\lambda_h = \lambda$ and $\tilde{b}_0^h(s) = \tilde{b}_0^{SPB}(s) = \tilde{b}_S(s)$ (where the last equality follows by our construction in step (c)) in Theorem EC.2 we get $\tilde{b}_i^{SPB}(s) = \tilde{b}_{S+i}(s)$ for $i = 1, 2, ...$ Therefore,

$$P^{SPB}(i) = P^{SPB}(0) \prod_{j=0}^{i-1} \frac{1 - \tilde{b}_{S+j}(\lambda)}{\tilde{b}(\lambda)}, \ i = 0, 1, 2, ... \tag{24}$$

We next show that Eq. (19) holds for $i = 0$. Let $1/\mu_1$ denote the expected remaining service time of an order in service in the shortfall queue observed by an arrival who finds S orders in the shortfall queue (That is $-d\tilde{b}_S(s)/ds|_{s=0} = 1/\mu_1$). Sigman and Yechiali (2007, Eq. 1) show that

$$\frac{1}{\mu_1} = \frac{1-\rho}{\lambda P(S)}(1 - F(S)).$$

So that,

$$P(S) = \frac{1-\rho}{\frac{\lambda}{\mu_1} + 1 - \rho}(1 - F(S-1)). \tag{25}$$

Also as in Eq. (11) the utilization of the SPB queue, ρ_b, is

$$\rho_b = \frac{\lambda\mu}{\mu_1\mu + \lambda(\mu - \mu_1)} = 1 - \frac{1-\rho}{\frac{\lambda}{\mu_1} + 1 - \rho}. \tag{26}$$

Comparing Eqs. (25) and (26) we get

$$P(S) = (1 - F(S-1))(1 - \rho_b) = (1 - F(S-1))P^{SPB}(0). \tag{27}$$

Therefore, Eq. (19) holds for $i = 0$. Substituting Eq. (27) on the right hand side of Eq. (20) together with Eq. (24) establishes Eq. (19) for $i \geq 1$.

We next establish that the job composition in the SPB queue is identical to the customers backlog composition in the SP system. Intuitively, considering Eq. (19), we observe that given step

(a) of the construction of the SPB queue, the job is allocated in the SPB queue in the same way as it is allocated in the SP system while $N(t) \geq S$. Furthermore, given step (b) in the construction of the SPB queue, the job arrival process of type r in the SPB queue has the same distribution as the customer arrival process of type r in the SP system. Both observations together with Eq. (19) imply that the job composition in the SBP queue is identical to the customer composition in the SP system. This implication establishes Eq. (6).

More formally, consider the continuous time Markov chain that represents the jobs' distribution in a multi-class $M/G/1$ queue with exceptional first service times with a density of $b_0(\cdot)$. Let $\bar{\mathbf{N}} = (L_1, \cdots, L_n)$ denote the vector of the number of jobs of classes $1, \cdots, n$, $\bar{\imath}_{\bar{\mathbf{N}}} = \arg\min_r \{L_r > 0\}$ and $\bar{\mathbb{k}}_{\bar{\mathbf{N}}} = \{r : L_r > 0\}$ is the set of classes with jobs waiting in the system. Then, similar to Eqs. (21) and (22) this MC is given by:

$$\bar{h}_{t+dt}(\bar{\mathbf{N}}, \eta) = \bar{h}_t(\bar{\mathbf{N}}, \eta + dt)(1 - \lambda dt) + \sum_{k=1}^{n} \bar{h}_t(\bar{\mathbf{N}} + \mathbf{e}_k, 0) b(\eta) dt, \quad \sum_{j=1}^{n} L_j = 0 \tag{28}$$

$$\bar{h}_{t+dt}(\bar{\mathbf{N}}, \eta) = \bar{h}_t(\bar{\mathbf{N}}, \eta + dt)(1 - \lambda dt) + \sum_{k=1}^{i_{\bar{\mathbf{N}}}} \bar{h}_t(\bar{\mathbf{N}} + \mathbf{e}_k, 0) b(\eta) dt + \bar{h}_t(\bar{\mathbf{N}} - \mathbf{e}_{i_{\bar{\mathbf{N}}}}, 0) \lambda_{i_{\bar{\mathbf{N}}}} b_0(\eta) dt, \quad \sum_{j=1}^{n} L_j = 1 \tag{29}$$

$$\bar{h}_{t+dt}(\bar{\mathbf{N}}, \eta) = \bar{h}_t(\bar{\mathbf{N}}, \eta + dt)(1 - \lambda dt) + \sum_{k=1}^{i_{\bar{\mathbf{N}}}} \bar{h}_t(\bar{\mathbf{N}} + \mathbf{e}_k, 0) b(\eta) dt + \sum_{k \in \bar{\mathbb{k}}_{\bar{\mathbf{N}}}} \bar{h}_t(\bar{\mathbf{N}} - \mathbf{e}_k, \eta + dt) \lambda_k dt, \quad \sum_{j=1}^{n} L_j > 1 \tag{30}$$

where $\bar{h}_t(\bar{\mathbf{N}}, \eta)$ is the probability that there are L_r jobs of class r in the system, and remaining production time is η at time t.

Next consider the continuous time Markov chain that represents the backlogs' distribution in the SP system during the periods that the system is out of stock. Let $\mathbf{N} = (B_1, \cdots, B_n)$ denote the vector of the backlogs of classes $1, \cdots, n$, $\mathbf{e}_r$ denote the rth unit vector, $i_{\mathbf{N}} = \arg\min_r \{B_r > 0\}$ and $\mathbb{k}_{\mathbf{N}} = \{r : B_r > 0\}$ is the set of backlogged classes. Then, similar to Eq.s (21) and (22) this MC is given by:

$$h_{t+dt}(\mathbf{N}, \eta) = h_t(\mathbf{N}, \eta + dt)(1 - \lambda dt) + \sum_{k=1}^{n} h_t(\mathbf{N} + \mathbf{e}_k, 0) b(\eta) dt, \quad \sum_{j=1}^{n} B_j = 0 \tag{31}$$

$$h_{t+dt}(\mathbf{N},\eta) = h_t(\mathbf{N},\eta+dt)(1-\lambda dt) + \sum_{k=1}^{i_{\mathbf{N}}} h_t(\mathbf{N}+\mathbf{e}_k,0)b(\eta)dt + h_t(\mathbf{N}-\mathbf{e}_{i_{\mathbf{N}}},0)\lambda_{i_{\mathbf{N}}} b_0^{SPB}(\eta)dt, \quad \sum_{j=1}^{n} B_j = 1 \tag{32}$$

$$h_{t+dt}(\mathbf{N},\eta) = h_t(\mathbf{N},\eta+dt)(1-\lambda dt) + \sum_{k=1}^{i_{\mathbf{N}}} h_t(\mathbf{N}+\mathbf{e}_k,0)b(\eta)dt + \sum_{k\in\Bbbk_{\mathbf{N}}} h_t(\mathbf{N}-\mathbf{e}_k,\eta+dt)\lambda_k dt, \quad \sum_{j=1}^{n} B_j > 1 \tag{33}$$

where $h_t(\mathbf{N},\eta)$ is the probability that there are B_r backlogs of class r in the SP system given it is out of stock, and remaining production time is η at time t. Note that in this MC, $b_0^{SPB}(\cdot)$ is independent of class r backlogs because any arrival to the SP system that finds inventory level equals zero creates the first backlog and starts the backlog period.

Comparing Eqs. (31), (32) and (33) with Eqs. (28), (29) and (30), respectively, we observe that the MC representing the backlogs in the SP system given it is out of stock is identical to the MC that represents the number of jobs in an $M/G/1$ queue with exceptional first service times if $b_0(\eta) = b_0^{SPB}(\eta)$. Therefore, since the density of the first exceptional service times in the SPB queue is defined as $b_0^{SPB}(\eta)$, we observe that the MC representing the backlogs in the SP system **given it is out of stock** is identical to the MC that represents the number of items in the SPB queue, and consequently the distribution of backlogs of class r given the SP system is out of stock is identical to the distribution of jobs of class r in the SPB queue. Note that this discussion essentially establishes Eq. (19) as well. The derivation of Eq. (19) is given above as it provides the closed form expression for these probabilities.

■

We next give the proof of Theorem 2.

Proof of Theorem 2. Consider customers of classes $1,\ldots,r$ as high-priority with an arrival rate λ_r^+. Let $E[N_r^+]$ and $E[N_r^-]$ denote the expected number of customers of classes $1,\ldots,r$ and $r+1,\ldots,n$ in the SPB queue, respectively. We call high- and low-priority classes r^+ and r^- respectively.

Using Little's Law, we have $E[N^{SPB}] = -\lambda\tilde{w}'(s)|_{s=0}$ 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. 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 $\lambda_h = \lambda$. Therefore, we have a single class $M/G/1$ queue with exceptional first service times. Using Corollary EC.1 we get $\tilde{w}_h(s) = \tilde{w}_{r^-}(s) = \tilde{w}(s)$ as given in Eq. (EC.12). Since $E[N^{SPB}] = E[N^+_r] + E[N^-_r]$, we have

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

where as in Eq. (EC.11) $\theta^+_r(s) = \tilde{b}(s + \lambda^+_r(1 - \theta^+_r(s)))$. Since $\theta^+_r(0) = 1$, $\tilde{w}'(0)$ cancels out and then because $\tilde{b}'(s)|_{s=0} = 1/\mu$, we have

$$\begin{aligned}\frac{E[N^+_r]}{E[N^{SPB}]} &= 1 - \frac{\lambda^-_r}{\lambda}(1 - \lambda^+_r \theta^{+\prime}_r(s)|_{s=0}) \\ &= 1 - \frac{\lambda^-_r}{\lambda}(1 - \frac{\lambda^+_r \tilde{b}'(s)|_{s=0}}{1 + \lambda^+_r \tilde{b}'(s)|_{s=0}}) = \frac{\lambda^+_r(1 - \rho)}{\lambda(1 - \rho^+_r)},\end{aligned} \tag{34}$$

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

$$E[N^{SPB}_r] = E[N^+_r] - E[N^+_{r-1}].$$

Therefore,

$$\frac{E[N^{SPB}_r]}{E[N^{SPB}]} = \frac{E[N^+_r] - E[N^+_{r-1}]}{E[N^{SPB}]}. \tag{35}$$

Applying Eq. (34) to the r^+ and $(r-1)^+$ customers and substituting it into Eq. (35) and letting $\rho_r = \lambda_r/\mu$ we have

$$\begin{aligned}\frac{E[N^{SPB}_r]}{E[N^{SPB}]} &= \frac{\lambda^+_r(1 - \rho)}{\lambda(1 - \rho^+_r)} - \frac{\lambda^+_{r-1}(1 - \rho)}{\lambda(1 - \rho^+_{r-1})} \\ &= \frac{(1 - \rho)[(\lambda^+_{r-1} + \lambda_r)(1 - \rho^+_{r-1}) - \lambda^+_{r-1}(1 - \rho^+_r)]}{\lambda(1 - \rho^+_r)(1 - \rho^+_{r-1})}\end{aligned}$$

$$
\begin{aligned}
&= \frac{(1-\rho)[\lambda^+_{r-1} + \lambda_r - \lambda^+_{r-1}\rho^+_{r-1} - \lambda_r\rho^+_{r-1} - \lambda^+_{r-1} + \lambda^+_{r-1}\rho^+_{r-1} + \lambda^+_{r-1}\rho_r]}{\lambda(1-\rho^+_r)(1-\rho^+_{r-1})} \\
&= \frac{\lambda_r(1-\rho)}{\lambda(1-\rho^+_r)(1-\rho^+_{r-1})} = \frac{1-\rho}{\rho}\left(\frac{1}{1-\rho^+_r} - \frac{1}{1-\rho^+_{r-1}}\right).
\end{aligned}
$$

■

Acknowledgements

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

References

Altıok, T. 1997. *Performance Analysis of Manufacturing Systems,* Springer-Verlag, NY.

Baron, O. 2008. "Regulated Random Walks and the LCFS Backlog Probability: Analysis and Applications", *Operations Research*, Vol. 56, 471–486.

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.*

Conway, R. W., W. L. Maxwell, L. W. Miller. 1967. *Theory of Scheduling*, Addison-Wesley: Reading, Mass.

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.

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. Balcıoğlu. 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 the queue length” *Operations Research Letters*, Vol. 35, 581583.

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.

This page is intentionally blank. Proper e-companion title page, with INFORMS branding and exact metadata of the main paper, will be produced by the INFORMS office when the issue is being assembled.

Online Appendix

EC.1. Required Queueing Analysis

In this section, we derive the required analytical results to express the costs for the MR and SP policies. Given Theorem 2 (the proof of which requires the following derivations and theorems) expressing the cost of the SP policy only requires the solution of a FCFS $M/G/1$ queue. Expressing the cost of the MR policy requires the solution of a two-priority $M/G/1$ queue with postponement of product allocation and exceptional first service times in busy periods as well as the characterization of the first exceptional service time. In Section EC.1.1 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 exceptional first service times in its busy periods when product allocation is postponed to the end of production. In Section EC.1.2 Theorem EC.2 outputs the LT of the exceptional first service times in the busy periods for BQ_r as defined in Section 3.

EC.1.1. A Multi-Priority $M/G/1$ Queue with Exceptional First Service Times in Busy Periods

In this section, we consider a multi-priority $M/G/1$ queue with exceptional first service times in busy periods 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 the LT of the density function of the system time of class r customers, $\tilde{w}_r(s)$, in Theorem EC.1. (Because the models in Takagi and Conway et al. consider systems without postponement, their results cannot be used directly to study the MR and SP policies.) To obtain $\tilde{w}_r(s)$, we consider a system with two-priority classes in Section EC.1.1.1. In Section EC.1.1.2, 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 $\tilde{w}_r(s)$.

EC.1.1.1. A Markov-Chain Representation for the Two-Priority Class System We consider a two-priority $M/G/1$ queue with exceptional first service times where high- and low-priority customer arrival rates are λ_h and λ_l, respectively, such that $\lambda = \lambda_h + \lambda_l$. We denote the LT of the first exceptional service times in busy periods by $\tilde{b}_0(s)$. We solve this queue following Takagi (1991). We focus on the discrete stochastic process $\mathbf{M^h}$ where $\{M_n^h, n = 1, 2, ...\}$ 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 $\tilde{b}_0(s)$, respectively, we have

$$W(z) = \sum_{k=0}^{\infty} w_k z^k = \tilde{b}_0(\lambda_h(1-z)), \tag{EC.1}$$

$$V(z) = \sum_{k=0}^{\infty} v_k z^k = \tilde{b}(\lambda_h(1-z)). \tag{EC.2}$$

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

$$p_{jk} = P\{M_{n+1}^h = k | M_n^h = j\} = v_{k-j+1}, \ \ k \geq j-1, j \geq 1. \tag{EC.3}$$

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 both 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.

Lemma EC.1. *The steady-state probabilities of* $\widetilde{\mathbf{M^h}}$ *and* $\mathbf{M^h}$ *are identical:*

$$\tilde{\pi}_k = \pi_k, \ \textit{for}\ k = 0, 1, ...$$

EC.1.1.2. 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.

Lemma EC.2. *Consider a two-priority class $M/G/1$ queue with exceptional first service times in its busy periods with a LT of $\tilde{b}_0(s)$ and regular service times with LT $\tilde{b}(s)$ and allocation postponement. Then, the expected gross processing time in this queue is*

$$E[A] = \rho_b E[A_1] + (1-\rho_b)(\tilde{b}_0(\lambda_h)E[A_2] + (1-\tilde{b}_0(\lambda_h))(E[A_3]+E[A_1])), \tag{EC.4}$$

where with $\tilde{b}_0'(s) := d\tilde{b}_0(s)/ds$

$$E[A_1] = \frac{1-\tilde{b}(\lambda_h)}{\lambda_h \tilde{b}(\lambda_h)},\ E[A_2] = -\frac{\tilde{b}_0'(\lambda_h)}{\tilde{b}_0(\lambda_h)},\ E[A_3] = \frac{\lambda_h \tilde{b}_0'(\lambda_h) + (1-\tilde{b}_0(\lambda_h))}{\lambda_h(1-\tilde{b}_0(\lambda_h))}.$$

To derive the probability generating functions, we need to express π_0, which involves more work 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.

Lemma EC.3. *Consider a two-priority $M/G/1$ queue with exceptional first service times. Then,*

1. *The steady-state probability of having no high-priority customer in the system is*

$$\lambda_h/\lambda\,(1-(\rho_b-\lambda_l E[A]))\,. \tag{EC.5}$$

2. *The fraction of departures leaving no high-priority customers behind is*

$$\pi_0 = 1-\frac{\lambda_h}{\lambda}(1-\kappa_0) = 1-\frac{\lambda_h(\rho_b - E[A])}{\lambda}. \tag{EC.6}$$

Now, using π_0, the $\widetilde{\mathbf{M}}^{\mathbf{h}}$ process from Theorem EC.1, and following Takagi (1991) we show

Lemma EC.4. *The probability generating function of the number of high-priority customers left in the two-priority class system by an arbitrary departure is*

$$\Pi(z) = \frac{(1-\rho_b)V(z)}{V(z)-z} + \frac{(\lambda_h z+\lambda_l)(1-\rho_b)W(z)}{\lambda(z-V(z))} + \frac{(1-\rho_b)\lambda_l(w_0(z-1))}{\lambda(z-V(z))}$$

$$+\frac{(\pi_0-(1-\rho_b))v_0(z-1)}{z-V(z)}. \tag{EC.7}$$

Using Lemma EC.4, we can obtain the probability generating function of the number of high-priority customers in the two-priority class system with exceptional first service times in busy periods that is required to obtain the cost of the MR system:

LEMMA EC.5. *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*

$$\begin{aligned}\Pi_h(z) \;=\; & \frac{\lambda(1-\rho_b)V(z)}{\lambda_h z(V(z)-z)}[z-\frac{(\lambda_h z+\lambda_l)W(z)+\lambda_l w_0(z-1)}{\lambda}-\frac{(\pi_0-(1-\rho_b))v_0(z-1)}{1-\rho_b}]\\ & +\frac{\lambda(1-\rho_b)}{\lambda_h z}[\frac{(\lambda_h z+\lambda_l)W(z)}{\lambda}-\frac{w_0\lambda_l}{\lambda}-\frac{(\pi_0-(1-\rho_b)))v_0}{(1-\rho_b)}].\end{aligned} \tag{EC.8}$$

In Theorem 2 we used $E[N]$ and $E[N_r]$ denoting, respectively, the expected number of total and class r orders in an $M/G/1$ queue with n priority classes and exceptional first service times in busy periods. We obtain $E[N]$ and $E[N_r]$ by first characterizing the LT of the system time density function of class r customers in the system and then using Litte's Law. Let $\tilde{w_h}(s)$ denote the LT of the system time density function of the high-priority customers in a two-priority system with exceptional first service times. Then:

THEOREM EC.1. *Consider a two-priority class $M/G/1$ queue with exceptional first service times in its busy periods with a LT of $\tilde{b}_0(s)$ and regular service times with LT $\tilde{b}(s)$. Then, the LT of the system time density function of the type r customers is*

$$\tilde{w_r}(s) \;=\; \tilde{w_h}(s+\lambda^+_{r-1}(1-\theta^+_{r-1}(s))), \tag{EC.9}$$

where

$$\begin{aligned}\tilde{w_h}(s) = & \frac{\tilde{b}(s)(1-\rho_b)(\lambda_l w_0-\lambda)+(\pi_0-(1-\rho_b))v_0\lambda(\tilde{b}(s)-1)}{\lambda_h(1-\tilde{b}(s))-s}\\ & +\frac{(1-\rho_b)(\tilde{b}_0(s)(\lambda-s)-\lambda_l w_0)}{\lambda_h(1-\tilde{b}(s))-s}.\end{aligned} \tag{EC.10}$$

and

$$\theta^+_{r-1}(s)=\tilde{b}(s+\lambda^+_{r-1}(1-\theta^+_{r-1}(s))). \tag{EC.11}$$

Corollary EC.1. *Consider a single class FCFS $M/G/1$ queue with exceptional first service times in busy periods with a LT of $\tilde{b}_0(s)$ and regular service times with LT $\tilde{b}(s)$. Then, the LT of the system time density function in this queue is*

$$\tilde{w}(s) = \frac{(1-\rho_b)(\lambda(\tilde{b}(s)-\tilde{b}_0(s))+s\tilde{b}_0(s))}{s-\lambda(1-\tilde{b}(s))}. \tag{EC.12}$$

EC.1.2. Exceptional First Service Time in a Two-Priority $M/G/1$ Queue

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

The derivation of $\tilde{b}^h_j(s)$ in Theorem EC.2 is similar to the proof of part 2 in Theorem 4 that extends the approach of Kerner (2008) to the setting we require.

Theorem EC.2. *Consider a two-priority class $M/G/1$ queue with exceptional first service times in its busy periods with a LT of $\tilde{b}_0(s)$ and regular service times with LT $\tilde{b}(s)$. Then, the LT of the residual service time upon the arrival of a high-priority customer seeing j high-priority customers in the system is given recursively by*

$$\tilde{b}^h_j(s) = \frac{\lambda_h}{s-\lambda_h}[\tilde{b}(\lambda_h)\frac{1-\tilde{b}^h_{j-1}(s)}{1-\tilde{b}^h_{j-1}(\lambda_h)} - \tilde{b}(s)], \quad j \geq 1, \tag{EC.13}$$

where

$$\begin{aligned}\tilde{b}^h_0(s) &= \frac{\kappa_0\lambda_h\tilde{b}(s)+\tilde{b}(s)(1-\rho_b)(\lambda_l w_0-\lambda)}{\kappa_0(\lambda_h-s)} \\ &+ \frac{(\pi_0-(1-\rho_b))\lambda v_0(\tilde{b}(s)-1)+(1-\rho_b)(\tilde{b}_0(s)(\lambda-s)-\lambda_l w_0)}{\kappa_0(\lambda_h-s)}.\end{aligned} \tag{EC.14}$$

From Eq.s (EC.1) and (EC.2), it follows that $v_0 = \tilde{b}(\lambda_h)$ and $w_0 = \tilde{b}_0(\lambda_h)$. Also, ρ_b and κ_0 are given in Eq.s (11) and (EC.15), respectively ($E[A]$ is given in Theorem EC.2). Observe that if

$\tilde{b}_0(s) = \tilde{b}(s)$, $\lambda_h = \lambda$ and $\lambda_l = 0$, Theorem EC.2 is identical to Corollary 2.2.1 in Kerner (2008) when setting $\lambda_n = \lambda$ for all n.

Algorithm 1 in EC.1.3 below gives the LT of the residual service times observed by high-priority arrivals who find j high-priority jobs in the queue. We can obtain the exceptional first service times of BQ_r for $r = 1 \cdots n$ using this Algorithm.

EC.1.3. The residual service times observed by high-priority arrivals in $\tilde{b}_j^r(s)$

Algorithm 1. Finding the LT of the residual service times observed by high-priority arrivals, $\tilde{b}_j^r(s)$, for $j = 0, \ldots, \Delta_r$, $r = 1, \ldots, n$

[Step 0] For level R_{n+1}, set $r = n$, $\tilde{b}_0(s) := \tilde{b}(s)$ and $\lambda_h = \lambda_n^+ := \sum_{i=1}^n \lambda_i$, $\lambda_l = \lambda_n^- := 0$, $\lambda := \sum_{i=1}^n \lambda_i$, and $j = 1$. Calculate $\tilde{b}_0^h(s)$ using Eq. (EC.14).

[Step 1] While $j \leq \Delta_r$, consider the r^{th} backlog queue:

a Obtain $\tilde{b}_j^r(s) = \tilde{b}_j^h(s)$, where the latter is given in Theorem EC.2.

b Set $j = j + 1$ and go back to Step 1.

[Step 2] While $n \geq r \geq 1$, consider the r^{th} backlog queue:

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

b Set $\tilde{b}_0(s) = \tilde{b}_{\Delta_r}^r(s)$, $r = r - 1$, $j = 1$.

c Calculate $\tilde{b}_0^h(s)$ using Eq. (EC.14) and go back to Step 1.

Algorithm 1 implicitly assumes that the LT of regular service times, $\tilde{b}(s)$, is known. The algorithm starts with $r = n$ at Step 0, setting the required parameters to characterize BQ_{n+1}: $\tilde{b}_0(s)$, λ_h, and λ_l. Then, at Step 1.a., the algorithm uses Theorem EC.2 to return $\tilde{b}_j^r(s)$, the LT of the residual service times observed by high priority arrivals at BQ_{r+1} who find $j\,(= 1, \ldots, \Delta_r)$ jobs in the queue. (Note that $\tilde{b}_{\Delta_r}^r(s)$ is the exceptional first service time in BQ_r.) At Step 2.a. the algorithm sets the required arrival rates for BQ_r. (Note that at this stage, Eq. (14) can be used to obtain the implied probabilities for this queue.) In Step 2.c., before continuing with the same steps for BQ_{r-1}, the algorithm updates the exceptional service time for this queue (as the residual service time resulting from BQ_r). The algorithm then returns to Step 1 with $r = r - 1$.

EC.1.4. Proofs of the Required Queueing Analysis

Proof of Lemma EC.1. We define $\mathbf{M_n^l}$ as the number of low-priority customers left behind by the nth 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,

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

Next, we use the above cases to construct a Markov-Chain (MC) $\widetilde{\mathbf{M}^{\mathbf{h}}}$ with states $k = 0, 1, \ldots$. We let its transition probabilities be p_{jk} as in Eq. (EC.3) when $k \geq j-1, j \geq 1$, and for $j=0$ we let

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

and for $k \geq 1$

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

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

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

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

Proof of Lemma EC.2. 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 the re-sampling policy as 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. (EC.1), 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):

$$\tilde{a_2}(s) = \frac{\tilde{b_0}(s+\lambda_h)}{\tilde{b_0}(\lambda_h)}, \quad E[A_2] = -\frac{\tilde{b_1}'(\lambda_h)}{\tilde{b_1}(\lambda_h)}.$$

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):

$$\tilde{a}_3(s) = \frac{\lambda_h(1-\tilde{b}_0(s+\lambda_h))}{(s+\lambda_h)(1-\tilde{b}_0(\lambda_h))}, \quad E[A_3] = \frac{\lambda_h\tilde{b_1}'(\lambda_h)+(1-\tilde{b_1}(\lambda_h))}{\lambda_h(1-\tilde{b_1}(\lambda_h))}.$$

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. (EC.4).■

Proof of Lemma EC.3. 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

$$\kappa_0 = 1 - (\rho_b - \lambda_l E[A]). \tag{EC.15}$$

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 of high-priority customers (out of all departures) 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 of Lemma EC.4. Based on Theorem EC.1, for the stochastic process $\widetilde{\mathbf{M}}^{\mathbf{h}}$, 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 EC.1, for $k=0$ we can write

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

and for $k \geq 1$,

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

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

$$\begin{aligned}\Pi(z) = \sum_{k=0}^{\infty} z^k \pi_k &= (\pi_0 - (1-\rho_b))(v_0+v_1) + (1-\rho_b)[\frac{\lambda_h}{\lambda} w_0 + \frac{\lambda_l}{\lambda}(w_0+w_1)]\\ &+ \sum_{k=0}^{\infty} z^k \sum_{j=1}^{k+1} \pi_j v_{k-j+1} + \sum_{k=1}^{\infty} z^k [(\pi_0 - (1-\rho_b))v_{k+1} + (1-\rho_b)(\frac{\lambda_h}{\lambda} w_k + \frac{\lambda_l}{\lambda} w_{k+1})].\end{aligned} \tag{EC.16}$$

Expanding the following term, which appears on the RHS of Eq. (EC.16),

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

$$+z^2\pi_1 v_2 + z^2\pi_2 v_1 + z^2\pi_3 v_0$$

$$+...$$

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

$$\begin{aligned}\sum_{k=0}^{\infty} z^k \sum_{j=1}^{k+1} \pi_j v_{k-j+1} &= \pi_1 \sum_{k=0}^{\infty} z^k v_k + \pi_2 \sum_{k=0}^{\infty} z^{k+1} v_k \\ &+\pi_3 \sum_{k=0}^{\infty} z^{k+2} v_k + ... \\ &= \pi_1 V(z) + z\pi_2 V(z) + z^2 \pi_3 V(z) + ... \\ &= V(z)\sum_{k=1}^{\infty} \pi_k z^{k-1} + \frac{\pi_0 V(z)}{z} - \frac{\pi_0 V(z)}{z} \\ &= \frac{V(z)\sum_{k=0}^{\infty}\pi_k z^k}{z} - \frac{\pi_0 V(z)}{z} \\ &= \frac{\Pi(z)-\pi_0}{z} V(z).\end{aligned} \tag{EC.17}$$

Hence,

$$\begin{aligned}\Pi(z) &= \frac{\Pi(z)-\pi_0}{z}V(z) + (\pi_0-(1-\rho_b))(v_0+v_1) + (1-\rho_b)[\frac{\lambda_h}{\lambda}w_0 + \frac{\lambda_l}{\lambda}(w_0+w_1)] \\ &+\sum_{k=1}^{\infty} z^k[(\pi_0-(1-\rho_b))v_{k+1} + (1-\rho_b)(\frac{\lambda_h}{\lambda}w_k + \frac{\lambda_l}{\lambda}w_{k+1})] \\ &= \frac{\Pi(z)-\pi_0}{z}V(z) + (1-\rho_b)\frac{\lambda_h}{\lambda}W(z) + (\pi_0-(1-\rho_b))(v_0+v_1) \\ &+(1-\rho_b)\frac{\lambda_l}{\lambda}w_0 + \sum_{k=1}^{\infty} z^k(\pi_0-(1-\rho_b))v_{k+1} + \sum_{k=0}^{\infty} z^k(1-\rho_b)\frac{\lambda_l}{\lambda}w_{k+1} \\ &= \frac{\Pi(z)-\pi_0}{z}V(z) + (1-\rho_b)\frac{\lambda_h}{\lambda}W(z) + (1-\rho_b)\frac{\lambda_l}{\lambda z}(W(z)-w_0) + (1-\rho_b)\frac{\lambda_l}{\lambda}w_0 \\ &+(\pi_0-(1-\rho_b))(\frac{V(z)}{z} - \frac{v_0+zv_1}{z}) + (\pi_0-(1-\rho_b))(v_0+v_1) \\ &= \frac{\Pi(z)V(z)}{z} + (1-\rho_b)W(z)\frac{\lambda_h z+\lambda_l}{\lambda z} + (1-\rho_b)\lambda_l\frac{w_0(z-1)}{\lambda z} \\ &-(1-\rho_b)\frac{V(z)}{z} + (\pi_0-(1-\rho_b))\frac{v_0(z-1)}{z}.\end{aligned} \tag{EC.18}$$

Solving for $\Pi(z)$, we obtain Eq. (EC.7).■

Proof of Lemma EC.5. If the next departing customer is a high-priority customer, there should be at least one high-priority customer present at the time of the last departure or arriving during the current service time. Therefore, we should ignore two types of elements appearing in $\Pi(z)$: (i) those corresponding to departures leaving no high-priority customers behind, and (ii) those 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. (EC.17) leads to

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

rather than Eq. (EC.18) and the proof continues to be similar to Lemma EC.4.■

Proof of Theorem EC.1. We first give the LT of the system time density function of the high-priority customers in the two-priority class system, $\tilde{w}_h(s)$. 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

$$\tilde{w}_h(s) = \Pi_h(\frac{\lambda_h - s}{\lambda_h}). \tag{EC.19}$$

Combining Eq.s (EC.8) from Theorem EC.5 and (EC.19), and using Eq.s (EC.1-EC.2) yield Eq. (EC.10).

Now we obtain $\tilde{w}_r(s)$. 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 there are no new arrivals after it joins the queue, the LT of its system time density function will be $\tilde{w}_h(s)$ as given in Eq. (EC.10). 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 $1, 2, .., r-1$ arriving after the tagged customer but before its service completion, namely over G. The total system time for the tagged customer is the sum of a delay G that has a LT $\tilde{w}_h(s)$ with the delayed busy period, i.e., the busy period following this delay. Note that busy periods induced by customers of types $1,\ldots,r-1$ are similar to those in an $M/G/1$ queue with arrival rate λ^+_{r-1}; thus as Eq. (7) in Conway et al. (1967, p. 150), their LT $\theta^+_{r-1}(s)$ is as Eq. (EC.11). Eq. (9) in Conway et al. (1967, p. 151) provides the sum of such a delay and its delayed busy period as Eq. (EC.9).■

Proof of Corollary EC.1. In the FCFS $M/G/1$ queue with a single class, Eq. (EC.6) becomes $\pi_0 = 1-\rho_b$ since without any low-priority customers $E[A]=0$ and $\lambda_h=\lambda$. Similarly, in Eq. (EC.10) we have $\lambda_l=0$. These modifications reduce Eq. (EC.10) to Eq. (EC.12).■

Proof of Theorem EC.2. We start by considering a two-priority $M/G/1$ queue whose exceptional first service times in busy periods with a LT of $\tilde{b}_0(\cdot)$ and the other service times have a LT of $\tilde{b}(\cdot)$. The LT of the system time distribution of the high-priority customers in this two-priority class system is given in Eq. (EC.10) of Theorem EC.1. Let $\tilde{b}^h_0(\cdot)$ denote the LT of the service time distribution of a high-priority customer who 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, $\tilde{b}^h_0(\cdot)=\tilde{b}_0(\cdot)$. However, if there is at least one low-priority customer in the system, $\tilde{b}^h_0(\cdot)$ 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 exceptional first service times with LT of $\tilde{b}_0(\cdot)$) is identical to the one in a single class FCFS $M/G/1$ queue with an exceptional service time with LT of $\tilde{b}^h_0(\cdot)$ 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. (EC.12) from Corollary EC.1 setting $\lambda=\lambda^+_r$ and observing that $1-\rho_b=\kappa_0$ to obtain the LT of the system time for high-priority customers

$$\tilde{w}_h(s) = \frac{\kappa_0(\lambda_h(\tilde{b}(s)-\tilde{b}^h_0(s))+s\tilde{b}^h_0(s))}{s-\lambda_h(1-\tilde{b}(s))}.$$

According to our construction, the $\tilde{w}_h(s)$ above equals the LT in Eq. (EC.10). Equating these and solving for $\tilde{b}^h_0(s)$, we obtain

$$\tilde{b}_0^h(s) = \frac{\kappa_0\lambda_h\tilde{b}(s) + \tilde{b}(s)(1-\rho_b)(\lambda_l w_0 - \lambda)}{\kappa_0(\lambda_h - s)} \\ + \frac{(\pi_0 - (1-\rho_b))v_0\lambda(\tilde{b}(s)-1) + (1-\rho_b)(\tilde{b}_0(s)(\lambda - s) - \lambda_l w_0)}{\kappa_0(\lambda_h - s)}. \tag{EC.20}$$

Eq. (EC.20) provides the LT of the residual service time, given that there are no high-priority customers in the system, establishing Eq. (EC.14).

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). Similar to the proof of part 1 of Theorem 4, 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., $\lim_{t\to\infty} p_t(j,\eta) = p(j,\eta)$. Therefore, we have,

$$p_{t+dt}(1,\eta) = p_t(1,\eta+dt)(1-\lambda_h dt) + p_t(2,0)b(\eta)dt + p_t(0,0)\lambda_h b_0^h(\eta)dt, \quad j=1,$$

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

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

$$p(1,\eta) = p(1,\eta+dt)(1-\lambda_h dt) + p(2,0)b(\eta)dt + \kappa_0\lambda_h b_0^h(\eta)dt, \quad j=1$$

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

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. (EC.13).■

EC.2. Proofs

In this section we provide the proofs of Theorems 3, 4, and 5 as well as the proof of Corollary 1 that are presented in Section 3.

Proof of Theorem 3. Let N^{SPB} and N_r^{SPB} denote the total number of jobs and the number of type r jobs in the SPB queue, respectively. Using Theorem 1, the expected backlog in the SP system $E[B]=E[N^{SPB}]$ and

$$E[B_r]=(1-F(S-1))\sum_{i=0}^{\infty}P_r^{SPB}(i)=(1-F(S-1))E[N_r^{SPB}]$$

so that Eq. (5) becomes

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

We next show that $(1-F(S-1))E[N^{SPB}]=\sum_{i=S}^{\infty}(i-S)P(i)$. To do this, recall that $E[N^{SPB}]$ is the expected number of the backlogs in the original system. In other words, $E[N^{SPB}]=E[N-S|N\geq S]$ where N is the total number of orders in the shortfall queue under a FCFS policy (which is invariant and is the same in the SP system). Then,

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

Substituting $E[N_r^{SPB}]/E[N^{SPB}]$ from Theorem 2 establishes Eq. (9).

Finally, given the cost in Eq. (9), the optimal base-stock level is given in Eq. (10) as in e.g., Veatch and Wein (1996).■

Proof of Corollary 1. If $R_{r+1}>R_r=R_{r-1}=...=R_{r-k}>R_{r-k-1}$, as soon as the inventory decreases to R_r, we consider classes $r-k,r-k+1,...,r$ 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 $\sum_{i=r-k}^{r}b_iE[N_i]$. By aggregating these classes to a single class with a weighted backlog cost $\left(\sum_{i=r-k}^{r}b_iE[N_i]\right)/\left(\sum_{i=r-k}^{r}E[N_i]\right)$ we obtain the same cost. (Note that these ratios, $\sum_{i=r-k}^{r}b_iE[N_i]/\sum_{i=r-k}^{r}E[N_i]$, do not require the exact characterization of $b_1(\cdot)$ because they are independent of $b_1(\cdot)$ and can be obtained using Eq. (7) in Theorem 2.)■

Proof of Theorem 4. 1. $E[N_r] = E[N^{BQ_r}] \times$ (% of class r jobs in BQ_r). Then, Eq. (12) follows, using Theorem 2. Eq. (13) can be calculated using Little's Law and Eq. (EC.12) in Corollary EC.1 giving the LT of the system time in such a queue.

2. Consider BQ_r. To obtain $P_h^{BQ_r}(i)$, we follow Lemma 3.1.3.1 in Kerner (2008). We define a continuous time Markov process with states (j,η) where j is the number of high-priority customers in the system, while 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., $\lim_{t\to\infty} p_t(j,\eta) = p(j,\eta)$. Therefore, we have,

$$p_{t+dt}(1,\eta) = p_t(1,\eta+dt)(1-\lambda_{r-1}^{+}dt) + p_t(2,0)b(\eta)dt + p_t(0,0)\lambda_{r-1}^{+}b_0^{r-1}(\eta)dt, \quad j=1, \tag{EC.21}$$

$$p_{t+dt}(j,\eta) = p_t(j,\eta+dt)(1-\lambda_r^{+}dt) + p_t(j-1,\eta+dt)\lambda_{r-1}^{+}dt + p_t(j+1,0)b(\eta)dt,\ j\geq 1, \tag{EC.22}$$

where $b_0^{r-1}(\cdot)$ is the steady-state density function of the residual service time of a high-priority job in service in BQ_r observed by a high-priority arrival who finds 0 high-priority jobs in this queue.

Using Eq.s (EC.21) and (EC.22), and similar to the proof of Lemma 3.1.3.1 in Kerner (2008):

$$P_h^{BQ_r}(i) = P_h^{BQ_r}(0)\prod_{j=0}^{i-1}\frac{1-\tilde{b}_j^{r-1}(\lambda_{r-1}^{+})}{\tilde{b}(\lambda_{r-1}^{+})},$$

and Eq. (14) follows because $P_h^{BQ_r}(0) = \lambda_{r-1}^{+}/\lambda_r^{+}\,(1-(\rho_b-\lambda_r E[A]))$ as given in Eq. (EC.5).■

Proof of Theorem 5. The proof of Theorem 5 requires the following lemma.

Lemma EC.6. *For BQ_{r+1} we have*

$$P_h^{BQ_{r+1}}(\Delta_r) = \frac{1-\rho_r^{+}}{\frac{\lambda_r^{+}}{\mu_1^r}+(1-\rho_r^{+})}\left(1-F_h^{BQ_{r+1}}(\Delta_r-1)\right). \tag{EC.23}$$

where λ_r^{+} and $1/\mu_1^r$ are the total arrival rate and the expected first exceptional service time in BQ_r, respectively.

Proof of Lemma EC.6. For a given $\tilde{b}_0^r(s)$ (the LT of the equilibrium service times of high-priority jobs in BQ_{r+1} who observe no high-priority job in the queue upon their arrivals) the LT of the first exceptional service times in BQ_r can be obtained using Eq. (EC.13) after setting $\lambda_h = \lambda_r^+$,

$$\tilde{b}_{\Delta_r}^r(s) = \frac{\lambda_r^+}{s-\lambda_r^+}[\tilde{b}(\lambda_r^+)\frac{1-\tilde{b}_{\Delta_r-1}^r(s)}{1-\tilde{b}_{\Delta_r-1}^r(\lambda_r^+)} - \tilde{b}(s)]. \quad \Delta_r \geq 1, \tag{EC.24}$$

By taking the derivative of Eq. (EC.24) we get

$$\begin{aligned} 1/\mu_1^r = -\frac{d\tilde{b}_{\Delta_r}^r(s)}{ds}|_{s=0} &= -\frac{\lambda_r^+}{(s-\lambda_r^+)^2}\left[\tilde{b}(\lambda_r^+)\frac{1-\tilde{b}_{\Delta_r-1}^r(s)}{1-\tilde{b}_{\Delta_r-1}^r(\lambda_r^+)} - \tilde{b}(s)\right]|_{s=0} \\ &+ \frac{\lambda_r^+}{(s-\lambda_r^+)}\left[-\frac{\tilde{b}(\lambda_r^+)}{1-\tilde{b}_{\Delta_r-1}^r(\lambda_r^+)}\frac{d\tilde{b}_{\Delta_r-1}^r(s)}{ds} - \frac{d\tilde{b}(s)}{ds}\right]|_{s=0} \\ &= \frac{1}{\lambda_r^+} - \left[-\frac{\tilde{b}(\lambda_r^+)}{1-\tilde{b}_{\Delta_r-1}^r(\lambda_r^+)}\frac{d\tilde{b}_{\Delta_r-1}^r(s)}{ds}|_{s=0} + \frac{1}{\mu}\right]. \end{aligned}$$

By solving the above recursion for $-d\tilde{b}_{\Delta_r}^r(s)/ds|_{s=0}$, we get

$$1/\mu_1^r = \left(-\frac{1}{\lambda_r^+} + \frac{1}{\mu}\right)\left(1 + \sum_{j=1}^{\Delta_r-1}\prod_{k=1}^{j}\frac{\tilde{b}(\lambda_r^+)}{1-\tilde{b}_{\Delta_r-k}^r(\lambda_r^+)}\right) + \prod_{k=0}^{\Delta_r-1}\frac{\tilde{b}(\lambda_r^+)}{1-\tilde{b}_k^r(\lambda_r^+)}E\left[R_0^h\right], \tag{EC.25}$$

where $E\left[R_0^h\right] = -d\tilde{b}_0^r(s)/ds|_{s=0}$ is the expected service time of high-priority jobs in BQ_{r+1} who observe no high-priority jobs in the queue upon their arrivals. (Note that $E\left[R_0^h\right]$ is different from $1/\mu_1^r$, because the latter includes residual service times observed by low-priority job arrivals at BQ_{r+1}.)

From Eq. (14) for BQ_{r+1} we have

$$\frac{P_h^{BQ_{r+1}}(j)}{P_h^{BQ_{r+1}}(\Delta_r)} = \frac{(1-\rho_b)\prod_{k=0}^{j-1}\frac{1-\tilde{b}_k^r(\lambda_r^+)}{\tilde{b}(\lambda_r^+)}}{(1-\rho_b)\prod_{k=0}^{\Delta_r-1}\frac{1-\tilde{b}_k^r(\lambda_r^+)}{\tilde{b}(\lambda_r^+)}} = \prod_{k=1}^{\Delta_r-j}\frac{\tilde{b}(\lambda_r^+)}{1-\tilde{b}_{\Delta_r-k}^r(\lambda_r^+)}, \ j = 1,2,...,\Delta_r-1, \tag{EC.26}$$

and

$$\frac{P_h^{BQ_{r+1}}(0)}{P_h^{BQ_{r+1}}(\Delta_r)} = \frac{(1-\rho_b)}{(1-\rho_b)\prod_{k=0}^{\Delta_r-1}\frac{1-\tilde{b}_k^r(\lambda_r^+)}{\tilde{b}(\lambda_r^+)}} = \prod_{k=0}^{\Delta_r-1}\frac{\tilde{b}(\lambda_r^+)}{1-\tilde{b}_k^r(\lambda_r^+)}. \tag{EC.27}$$

By substituting Eq.s (EC.26) and (EC.27) in Eq.(EC.25) we get

$$1/\mu_1^r = \left(-\frac{1}{\lambda_r^+} + \frac{1}{\mu}\right)\left(1 + \sum_{j=1}^{\Delta_r-1}\frac{P_h^{BQ_{r+1}}(j)}{P_h^{BQ_{r+1}}(\Delta_r)}\right) + \frac{P_h^{BQ_{r+1}}(0)}{P_h^{BQ_{r+1}}(\Delta_r)}E\left[R_0^h\right]. \tag{EC.28}$$

Let $E\left[R^h\right]$ denote the expected amount of time a high-priority job actually spends on the server in BQ_{r+1}. Observe that $\lambda_r^+ E\left[R^h\right]$ denotes the proportion of time that the server works on high-priority jobs in BQ_{r+1}. Therefore,

$$E\left[R^h\right] = \frac{1-P_h^{BQ_{r+1}}(0)}{\lambda_r^+}. \tag{EC.29}$$

Also, in equilibrium and due to PASTA we have,

$$E\left[R^h\right] = \left(1-P_h^{BQ_{r+1}}(0)\right)\frac{1}{\mu} + P_h^{BQ_{r+1}}(0)E\left[R_0^h\right]. \tag{EC.30}$$

Solving for $E\left[R_0^h\right]$, from Eq.s (EC.29) and (EC.30) we get,

$$E\left[R_0^h\right] = \frac{\left(1-P_h^{BQ_{r+1}}(0)\right)}{P_h^{BQ_{r+1}}(0)}\left(\frac{1}{\lambda_r^+}-\frac{1}{\mu}\right). \tag{EC.31}$$

Substituting Eq. (EC.31) in Eq. (EC.28) we get,

$$1/\mu_1^r = \left(-\frac{1}{\lambda_r^+}+\frac{1}{\mu}\right)\frac{\left(-1+F_h^{BQ_{r+1}}\left(\Delta_r\right)\right)}{P_h^{BQ_{r+1}}(\Delta_r)} = \frac{1}{\lambda_r^+}\left(1-\rho_r^+\right)\frac{\left(1-F_h^{BQ_{r+1}}\left(\Delta_r-1\right)-P_h^{BQ_{r+1}}(\Delta_r)\right)}{P_h^{BQ_{r+1}}(\Delta_r)},$$

so that

$$P_h^{BQ_{r+1}}(\Delta_r) = \frac{1-\rho_r^+}{\frac{\lambda_r^+}{\mu_1^r}+1-\rho_r^+}\left(1-F_h^{BQ_{r+1}}\left(\Delta_r-1\right)\right).$$

■

Next, we prove Theorem 5. First consider BQ_{n+1}. Note that BQ_{n+1} is defined as the shortfall queue. Therefore, when there are $i=0,\ldots,\Delta_n-1$ jobs in BQ_{n+1}, the MR system has $(R_{n+1}-i)$ units in inventory, which establishes Eq. (16) for $r=n+1$. (Recall that, because there are no backlogs in BQ_{n+1}, Eq. (15) does not include $r=n+1$.)

We prove Eq.s (15) and (16) for $r=1,...,n$ by induction. Note that BQ_n is identical to an SPB queue with two classes of jobs (classes $1,..,n-1$ high-priority and class n low-priority) where the base-stock level of its SP system is Δ_n. Therefore, from Theorem 1, the distribution of the backlogs of class n can be calculated using BQ_n as given in Eq. (15) for $r=n$. Also, noting that all customer arrivals of class $r < n$ to the MR system who find $R_{n-1} < I(t) \leq R_n + 1$ are served immediately and each decreases the inventory level by one unit, we have

$$P(I = R_n - i) = \left[1 - F_h^{BQ_{n+1}}(\Delta_n - 1)\right] P_h^{BQ_n}(i),\ i = 0, 1, ..., \Delta_{n-1}.$$

This establishes Eq. (16) for $r = n$.

Induction hypothesis: suppose Eq.s (15) and (16) hold for $r = m + 1$.

The induction hypothesis states that the job composition in BQ_{m+1} is identical to the customer composition in the MR system (in the relevant range of inventory).

We next prove Eq.s (15) and (16) for $r = m$, i.e., the job composition in BQ_m is identical to the customer composition in the MR system (in the relevant range of inventory). The proof is similar to the proof of Theorem 1 for the SPB queue.

First assume

$$P_h^{BQ_{m+1}}(\Delta_m + i) = [1 - F_h^{BQ_{m+1}}(\Delta_m - 1)] P^{BQ_m}(i),\ i = 0, 1, ... \tag{EC.32}$$

where $P^{BQ_m}(i)$ denotes the steady-state probability of having i jobs in BQ_m. Eq. (EC.32) states that $P^{BQ_m}(i)$, is identical to the steady-state probability of having $\Delta_m + i$ high-priority jobs in BQ_{m+1} given that the number of high-priority jobs in BQ_{m+1} is greater than $\Delta_m - 1$.

Assuming Eq. (EC.32), we observe that given step (a) of the construction of BQ_m, the job is allocated in BQ_m in the same way as it is allocated in the MR system while $R_{m-1} < I(t) \leq R_m$, and type m demand is backlogged in BQ_m as it is in the MR system while $I(t) \leq R_m$. Furthermore, given step (b) of the construction of BQ_m, the job arrival process of type $1, ..., m$ in BQ_m has the same distribution of the customer arrival process of type $1, ..., m$ as in the MR system. Both observations together with Eq. (EC.32) imply:

$$P(B_m = i | I \leq R_{m+1}) = \left[1 - F_h^{BQ_{m+1}}(\Delta_m - 1)\right] P_l^{BQ_m}(i),\ i = 0, 1, ... \tag{EC.33}$$

Using Eq. (16), which holds for $m + 1$ because of the induction hypothesis, the probability of $I \leq R_{m+1}$ is

$$P(I \leq R_{m+1}) = \prod_{j=m+2}^{n+1} \bar{F}_h^{BQ_j}(\Delta_{j-1} - 1). \tag{EC.34}$$

This, together with Eq. (EC.33) establishes Eq. (15) for $r = m$.

Also, note that all customer arrivals of class $1, ..., m-1$ to the MR system who find $R_{m-1} < I(t) \leq R_m$ are served immediately and each decreases the inventory level by one unit. This implies (together with Eq. (EC.32))

$$P(I = R_m - i | I \leq R_{m+1}) = \left[1 - F_h^{BQ_{m+1}}(\Delta_m - 1)\right] P_h^{BQ_m}(i), \;\; i = 0, 1, ..., \Delta_{m-1} - 1.$$

This together with Eq. (EC.34) establishes Eq. (16) for $r = m$.

To complete the proof, we now establish Eq. (EC.32).

Using Eq. (14), the steady-state probability of having $(\Delta_m + i)$ jobs in BQ_{m+1} is,

$$P_h^{BQ_{m+1}}(\Delta_m + i) = (1 - \rho_b) \prod_{j=0}^{\Delta_m + i - 1} \frac{1 - \tilde{b}_j^m(\lambda_m^+)}{\tilde{b}(\lambda_m^+)} = P_h^{BQ_{m+1}}(\Delta_m) \prod_{j=0}^{i-1} \frac{1 - \tilde{b}_{\Delta_m + j}^m(\lambda_m^+)}{\tilde{b}(\lambda_m^+)}, \;\; i = 0, 1, ... \tag{EC.35}$$

where ρ_b denotes the utilization of BQ_{m+1}. Observe that the distribution of the total number of jobs in BQ_m is identical to the distribution of the total number of jobs in an SPB queue with exceptional first service times with a LT of $\tilde{b}_{\Delta_m}^m$. Therefore, from Eq. (24) we have

$$P^{BQ_m}(i) = P^{BQ_m}(0) \prod_{j=0}^{i-1} \frac{1 - \tilde{b}_{\Delta_m + j}^m(\lambda_m^+)}{\tilde{b}(\lambda_m^+)}, \;\; i = 0, 1, ... \tag{EC.36}$$

We next show that Eq. (EC.32) holds for $i = 0$. As in Eq. (11) the utilization of BQ_m, is

$$\frac{\lambda_m^+ \mu}{\mu_1^m \mu + \lambda_m^+(\mu - \mu_1^m)} = 1 - \frac{1 - \rho_m^+}{\frac{\lambda_m^+}{\mu_1^m} + (1 - \rho_m^+)}. \tag{EC.37}$$

By comparing Eq.s (EC.23) and (EC.37) we get

$$P_h^{BQ_{m+1}}(\Delta_m) = \left[1 - F_h^{BQ_{m+1}}(\Delta_m - 1)\right] P^{BQ_m}(0). \tag{EC.38}$$

Therefore, Eq. (EC.32) holds for $i = 0$. Substituting Eq. (EC.38) in Eq. (EC.35) together with Eq. (EC.36) establishes Eq. (EC.32) for $i \geq 0$ and completes the proof.■