OPERATIONS RESEARCH
http://dx.doi.org/10.1287/opre.1120.1062ec

e-companion

ONLY AVAILABLE IN ELECTRONIC FORM

Electronic Companion—"Strategies for a Centralized Single Product Multiclass $M/G/1$ Make-to-Stock Queue" by Hossein Abouee-Mehrizi, Barış Balcıoğlu, and Opher Baron, *Operations Research*, http://dx.doi.org/10.1287/opre.1120.1062.

Submitted to *Operations Research*
manuscript OPRE-2009-06-264.R5

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

Hossein Abouee-Mehrizi
Department of Management Sciences, University of Waterloo, Waterloo, N2L 3G1, Canada, haboueem@uwaterloo.ca

Barış Balcıoğlu
Faculty of Engineering and Natural Sciences, Sabancı University, Orhanlı-Tuzla, 34956 Istanbul, Turkey, balcioglu@sabanciuniv.edu

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

In this electronic companion, we present the required queueing analysis to express the costs of the MR and SP policies in the main text. A multi-priority $M/G/1$ queue with an exceptional first service times in busy periods is studied. Then, an algorithm is developed to obtain the residual service times observed by high-priority arrivals. In addition to the proofs of theorems of this queueing model, we also present the proofs of certain theorems of the main text here.

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

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 $\widetilde{\mathbf{M}}^{\mathbf{h}}$ that is both Markovian and tractable. We construct the transition probabilities of $\widetilde{\mathbf{M}}^{\mathbf{h}}$ 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* $\widetilde{\mathbf{M}}^{\mathbf{h}}$ *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, $\tilde{b}_j^h(s)$. 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 $\tilde{b}_j^h(s)$ 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*

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

where

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

From Eqs. (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 Eqs. (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,

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

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,

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

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

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

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 $\tilde{b}_0^h(\cdot)$ 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, $\tilde{b}_0^h(\cdot)=\tilde{b}_0(\cdot)$. However, if there is at least one low-priority customer in the system, $\tilde{b}_0^h(\cdot)$ will be distributed as the residual service time of the item currently in service. Thus, the system time of high-priority customers in this two-priority queue (with 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 $\tilde{b}_0^h(\cdot)$ and an arrival rate equals to the arrival rate of the high-priority customers. Now, in the absence of low-priority customers, we can employ Eq. (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

$$\tilde{w}_h(s) = \frac{\kappa_0(\lambda_h(\tilde{b}(s)-\tilde{b}_0^h(s))+s\tilde{b}_0^h(s))}{s-\lambda_h(1-\tilde{b}(s))}.$$

According to our construction, the $\tilde{w}_h(s)$ above equals the LT in Eq. (EC.10). Equating these and solving for $\tilde{b}_0^h(s)$, 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,

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

Substituting $E[N_r^{SPB}]/E[N^{SPB}]$ from Theorem 1 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 $\sum_{i=r-k}^{r} b_i E[N_i]$. By aggregating these classes to a single class with a weighted backlog cost $\left(\sum_{i=r-k}^{r} b_i E[N_i]\right) / \left(\sum_{i=r-k}^{r} E[N_i]\right)$ we obtain the same cost. (Note that these ratios, $\sum_{i=r-k}^{r} b_i E[N_i] / \sum_{i=r-k}^{r} E[N_i]$, do not require the exact characterization of $b_1(\cdot)$ because they are independent of $b_1(\cdot)$ and can be obtained using Eq. (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 Eqs. (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 $E[R^h]$ denote the expected amount of time a high-priority job actually spends on the server in BQ_{r+1}. Observe that $\lambda_r^+ E[R^h]$ 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 $E[R_0^h]$, 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,

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

so that

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

■

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 Eqs. (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 Δ_m+i high-priority jobs in BQ_{m+1} given that the number of high-priority jobs in BQ_{m+1} is greater than Δ_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 (Δ_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.■

References

Bertsimas, D., D. Nakazato. 1995. "The Distributional Little's Law and Its Applications", *Operations Research*, Vol. 43, No. 2, 298–310.

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

Gross, D., C. M.,Harris. 1998. *Fundamentals of Queueing Theory*, John Wiley & Sons, New York.

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.

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