Multilevel Rationing Policy for Spare Parts When Demand is State-Dependent

Pedram Sahba
University of Toronto, Department of Mechanical and Industrial Engineering
5 King's College Rd., Toronto, ON M5S 3G8, Canada,
pedram@mie.utoronto.ca
Barış Balcıoğlu
Sabancı University, Faculty of Engineering and Natural Sciences,
Orhanlı-Tuzla, 34956 Istanbul, Turkey,
balcioglu@sabanciuniv.edu
Dragan Banjevic
University of Toronto, Department of Mechanical and Industrial Engineering
5 King's College Rd., Toronto, ON M5S 3G8, Canada,
banjev@mie.utoronto.ca

Abstract

The multilevel rationing (MR) policy is the optimal inventory control policy for single-item $M/M/1$ make-to-stock queues serving different priority classes when demand rate is constant and backlogging is allowed. Make-to-repair queues serving different fleets differ from make-to-stock queues because in the setting of the former, each fleet comprises finitely many machines. This renders the characterization of the optimal control policy of the spare part inventory system difficult. In this paper, we implement the MR policy for such a repair shop/spare part inventory system. The state-dependent arrival rates of broken components at the repair shop necessitate a different queueing-based solution for applying the MR policy from that used for make-to-stock queues. We find the optimal control parameters and the cost of the MR policy; we, then compare its performance to those of the hybrid FCFS and hybrid priority policies described in the literature. We find that the MR policy performs close to the optimal policy and outperforms the hybrid policies.

Keywords and Phrases: Spare parts, multiple finite-population queueing systems, multilevel rationing policy, hybrid policies

1 Introduction

In this paper, we analyze a continuous-review control policy for an inventory system of repairable spare parts for a company with m plants/fleets of machines. The proposed model targets utility companies, airlines, manufacturing, and mining industries for whom spare part provisioning is a fundamental concern. Such companies run expensive equipment/machines in different fleets that, albeit infrequently, fail from time to time. We restrict our attention to a single type of critical component, which upon failure, is immediately sent out for repair. To sustain high production or service levels, a spare component, if available, is installed on the machine/equipment that "owns/hosts" the broken component. If there are no spare components, the machine stops production and stays down until a repaired component can be installed. Although the same type of component is used by machines in different fleets, the number of machines and the component failure rate can vary from one fleet to another. Moreover, certain fleets can be more important for the company. For instance, if each fleet is serving a different customer, the nature of the individual contracts can induce the company to assign different down time costs for different fleets, which, in return, can lead the company to prioritize its fleets. In this setting, the important questions for the company are whether there should be inventory pooling for the various fleets, and if so, what type of an allocation policy should be followed, and finally, how the destination of a repaired component should be determined. We propose employing the *multilevel rationing* (MR) policy originally considered for controlling the inventory of finished goods demanded by different priority classes of customers (e.g., Ha, 1997a, de Véricourt, Karaesmen, and Dallery, 2001).

Under the MR policy, fleets 1 to m are prioritized from highest to lowest, and there are non-decreasing threshold inventory levels L_k, $k = 1, \ldots, m+1$, with $L_1 = 0$ for a centralized inventory. If the inventory level I is at L_{m+1}, there are no broken components. If I is between L_{k+1} and L_k (i.e., $L_k < I \leq L_{k+1}$), spare components are used only when machines in fleets 1 to k fail. In other words, when $L_k < I < L_{k+1}$, even if there are down machines in fleets $k+1$ to m, the repaired component is placed in the inventory as a spare component. If there

are down machines in fleet $k+1$ when $I = L_{k+1}$ and this fleet is the only one associated with L_{k+1}, a component coming out of the repairshop is used for a down machine in this fleet. If multiple fleets have the same inventory threshold and the inventory is at this level, upon completion of its repair, a component is used for the highest priority fleet associated with that threshold level that has a down machine. Thus, when there is no positive-stock, the repaired component is allocated to the highest-priority fleet with down machines. To clarify how the policy works, consider Example No 1 in Table 4. The optimal MR threshold levels are L_4=9, $L_3 = 4$, and $L_2 = L_1 = 0$ for three fleets. If $I = 9$, there are no broken components that require fixing in the repair shop. If $I \in \{9, 8, 7, 6, 5\}$ and a machine fails in any fleet, a spare component from the inventory can be used. For $0 < I \leq 4$, if a machine breaks down in the lowest priority fleet (fleet 3), a spare from the inventory is not dispensed and that machine becomes down. If the repair of a component finishes when $I = 4$ and there is a down machine in fleet 3, the repaired component is installed on that machine. Otherwise, the repaired component is placed in the inventory, raising its level to 5. In this example, threshold inventory levels for fleets 1 and 2 are both 0. Forcing $L_3 > L_2 > L_1 = 0$, i.e., having a distinct threshold level for each fleet would make it costlier for the system. When $I = 0$, when a component is repaired, it is installed on a down machine in fleet 1 if there are any. Otherwise, it is installed on a down machine (if any) in fleet 2. If there are no down machines, the repaired component is placed in the inventory, raising its level to 1.

Instead of keeping a centralized inventory, the MR policy can be used as a transhipment policy between inventories of different fleets. In this case, when a spare is depleted by fleet k from its inventory, a component from the inventory of the lowest-priority fleet with positive stock is immediately transhipped to replenish the inventory of fleet k. If the inventories of fleets $k+1$ to m are all depleted, the inventory of fleet k decreases by 1. No spares are transhipped from inventories of higher priority fleets to lower priority fleets (or their inventories) even if the latter have down machines. When a component is repaired, it is sent to the highest priority fleet with down machines or missing spares in its inventory.

In the production/inventory systems literature on the MR policy, a system is usually

modeled as a make-to-stock queue in which a single server queue represents the production facility. Different customer classes are assumed to place orders according to homogeneous Poisson processes. Each order generates a production order for this single server queue if backlogging is permitted; hence, in a lost sales case, only demand arriving when there is stock generates a production order. In other words, in earlier research on make-to-stock queues with an inventory controlled by the MR policy, demand rates (arrival rates at the make-to-stock queue) are not state-dependent but constants. In contrast, we consider that each fleet is comprised of a finite number of machines. Although the lifetime of each component is assumed to be exponentially distributed with a constant rate, the number of functional machines in each fleet renders the rate of failing components (rate of "demand" for/broken component arrival rate at the repairshop) state-dependent. Simply stated, not having constant demand rates from each fleet prevents us from exploiting the results of earlier models. Bearing in mind this major difference in our problem, we note Ha (1997a, 1997b) as the first to study rationing policies in make-to-stock queues. More specifically, Ha (1997a) analyzes a Markovian multi-class single server system with a centralized inventory in which unsatisfied demands are lost. Ha (1997b) studies the same problem with two classes of customers when backlogging is allowed. In both cases, Ha proves that in systems with centralized inventories, the MR policy is the optimal control policy. Allowing backlogging, de Véricourt, Karaesmen, and Dallery (2001) provide an efficient algorithm to compute the optimal rationing levels and the cost of the MR policy when m classes of customers are served. In a later study, de Véricourt, Karaesmen, and Dallery (2002) prove that the MR policy is the optimal policy in $M/M/1$ systems serving m classes of customers. Without constant Poisson arrival rates and exponential service times, it is difficult to characterize the optimal policy, however. For the lost sales case, assuming Erlangian service times, Ha (2000) shows that an MR policy based on the number of exponential service stages to be completed in the make-to-stock queue is optimal. But if backlogging is permitted, Gayon et al. (2009) observe the difficulty involved in finding the optimal control policy when m is large. Abouee-Mehrizi, Balcıoğlu, and Baron (2012) obtain the optimal cost and the rationing levels of the MR policy in $M/G/1$ systems.

Since the optimal policy is unknown, they are only able to compare the performance of the MR policy with other well-known policies, such as the first-come, first-served (FCFS) policy, to demonstrate the superiority of the former. Gabor et al., (2016) employ the MR policy for two priority classes and model the production stage of the spare parts as an $M/D/\infty$ queue. After obtaining the response time distributions for both classes, they demonstrate that optimizing inventory control parameters based on response time guarantees instead of fillrate constraints decreases the stock levels.

In our problem, we model the repair facility as a single server queueing system. This follows Sahba and Balcıoğlu (2011) who demonstrate that having a centralized high capacity repair shop serving all fleets is more cost effective than having dedicated smaller capacity repair shops for each fleet. We also assume that repair times are exponentially distributed. Therefore, our model differs from that of de Véricourt, Karaesmen, and Dallery (2001) in the use of state-dependent arrival rates at the single server queue. However, this not only necessitates a completely different analysis for the underlying queueing system, but also leaves us with the fact that the optimal policy is unknown. We are only able to obtain the cost of the optimal policy numerically and this turns out to be the MR policy in most of the examples discussed in Section 5.

When we review the literature on transhipment of spares among different plants, we see that almost all authors assume demand from each plant to be homogeneous Poisson processes. Lee (1987) considers a model in which a transshipment from the inventory of a neighboring plant is requested when a plant has no stock on hand. Otherwise, if the inventory is not zero but below its base-stock level, a plant (or the plant that transhipps a spare) requests a spare from a depot. Lee models the repair shop at the depot as an infinite server queue. Axsäter (1990), noting that characterizing the optimal policy is difficult, revisits Lee's model but develops another approximation which proves to be more accurate than that of Lee's. Kukreja, Schmidt, and Miller (2001) assume that spares are consumable, thus, if a transhipment is not possible from another plant when all inventories are zero, a spare is ordered from a manufacturer. They use a queueing based approach to develop an

approximation to determine optimal inventory levels at each plant. Jung et al. (2003) model the repair facility by a multi-server queueing system. Like us, Wong, Cattrysse, and Van Oudheusden (2005) assume that each plant hosts finitely many machines, but they model the repair facility as an infinite server queue. Ignoring transportation time between the repair facility and the plants, they assume exponential transhipment times. These are assumed to be short enough that the possibility of another failure or repair completion can be safely ignored. The plant with no inventory receives a spare from the closest plant with positive stock, with mean transhipment times used to measure the distance between plants. In our model, we do not assume a transhipment delay between plants (as in Kukreja, Schmidt, and Miller, 2001) and we ignore transportation costs, simply assuming them to be much smaller than the down time costs. Unlike the examples given above, a fleet does not wait to place a transhipment request until its inventory level drops to zero. Lee (1987) suggests the transhipment be made from a plant with the maximum number of units on hand, but we stipulate that the fleets are prioritized, and the lowest priority fleet with positive stock should lend a spare to a fleet that has just used one from its own stock. van Wijk, Adan, and van Houtum (2013) consider a transhipment problem involving multiple local warehouses and a quick response warehouse which operates under a threshold policy. When a local warehouse with depleted stock faces a new demand, this can be satisfied from the quick response warehouse instead of a more expensive emergency supply. Assuming constant Poisson demand rates for each warehouse and exponential transfer times, the authors show that the quick response warehouse will follow a threshold policy; an overflow demand from a local warehouse will be satisfied only if the stock at the former is above an identified threshold.

Studies on dynamic scheduling decisions for repairs are also worth mentioning. For instance, Hausman and Scudder (1982) employ a simulation based comparative study of a sequence of work centers where multiple types of components can be fixed. Assuming constant Poisson arrival rates and constant repair times for these components, they demonstrate that scheduling decisions made dynamically based on spare part inventory level and the job's progress in the repair shop minimize the mean delay for repair completions. We refer the

reader to Sleptchenko, van der Heijden, and van Harten (2005), Tiemessen and van Houtum (2013), and the references therein for further reading on dynamic scheduling at repair shops.

Although we test the relative performance of the MR policy against the numerically computed optimal policy, we also compare it to the *hybrid FCFS* (HF) and the *hybrid priority* (HP) policies proposed by Sahba, Balcıoğlu, and Banjevic (2013a) (see Section 4). Both policies allocate inventories to each fleet but do not permit transhipment. Instead, a shared inventory immediately replenishes the inventory of a fleet when a spare part is used. The difference lies in deciding how a repaired component is to be dispatched when the shared inventory is zero. The HF policy sends the fixed component to the fleet with the longest outstanding order, whereas the HP policy, assuming fleets are prioritized, sends it to the highest-priority fleet with outstanding orders. If the optimal solution of these policies states that only a shared inventory be kept and no fleets should have its own inventory, the problem turns out to be a complete pooling policy similar to the one considered by Kukreja, Schmidt, and Miller (2001).

In Section 2, we propose a recursive method to compute the system cost under the MR policy. This algorithm makes use of single server queueing systems serving finitely many customers with an unreliable server. The analysis of such queues requires determining the distribution of the interruption period for the server to exploit the method given by Sahba, Balcıoğlu, and Banjevic (2013b) to obtain the steady-state system size distribution for each class. This proves to be difficult in our problem because of the recursive method proposed in Section 2. As a solution, in Section 3, we obtain the moments of the server interruption time distribution, and propose fitting simpler phase-type distributions to capture the first three moments of the former. Using these approximating interruption time random variables (r.v.s) as input in the exact MR algorithm developed (as we do for the numerical examples in Section 5), gives the *MR policy approximation*. In Section 4, we summarize two alternative policies, the HF and HP policies, and discuss how we numerically compute the cost of the optimal policy. In the numerical study presented in Section 5 where we test the performance of the MR policy, we also assess the accuracy of the proposed MR approximation. The

results show that the MR approximation is highly accurate; in fact, the MR policy turns out to be optimal in many cases, outperforming the HF and HP policies. Having said this, we note that the HP policy may be considered a reasonable compromise if managers find it easier to implement. All proofs appear in Appendix A.

2 The Exact Analysis of the Multilevel Rationing Policy

We consider a system of m classes/fleets of machines parameterized by $k = 1, ..., m$. Each fleet k consists of N_k machines (type k machine) that fail from time to time due to a single type of repairable critical component. When a type k machine fails, its broken component is immediately sent to a repair shop serving all fleets, modeled as a single server queueing system. We assume that repair times follow an exponential distribution with rate μ independent of the fleet from which the broken component has been sent. In addition, spare components are kept to decrease the proportion of times these fleets may have down machines because of the lack of the critical component. If there is available stock for class k, a spare component is immediately installed to replace the failed component, and the machine can stay operational without experiencing down time. Otherwise, the number of operational type k machines decreases by 1, costing the system b_k (down time cost) per unit time until a component can be installed on a failed machine. Times to failure, that is the periods between installation of a spare or repaired component on a type k machine and the next failure instant of this installed component, follow an exponential distribution with rate λ_k. This implies that each repair makes the component as good as new, and the failure rate depends only on the fleet using it. Different failure rates can be due to the type of service a fleet renders or specific operating conditions to which its machines are subject.

The system incurs two types of holding costs. The first is the capital cost tied up in additional spares (in excess of the minimum $\sum_{k=1}^{m} N_k$ components for that many machines),

and to include it in the analysis, following Louit et al. (2011) and Sahba and Balcıoğlu (2011), we assume a holding cost of h per unit spare component per unit time. The second type is the warehousing cost of h_w per unit spare per unit time during the intervals the spare component is stored in the inventory. Since we are considering slow-moving expensive components, we assume that transportation times compared to repair times, and transportation costs compared to capital holding and down time costs are negligible.

In this setting, to reduce the long-run average cost per unit time, we have to decide on a) the structure of the inventory, and b) the allocation rule for a repaired component. In broad terms, the structure of the inventory indicates whether there are reserved inventories for each fleet and/or whether inventory can be shared among fleets. The allocation rule indicates whether repaired components are dispatched on an FCFS basis or according to a priority rule among fleets needing a component. In this paper, we propose the *multilevel rationing* (MR) policy which prioritizes fleets 1 to m from highest to lowest and is applied in the following way: There are non-decreasing threshold inventory levels L_k, $k = 1, \ldots, m+1$ with $L_1 = 0$ and $L_{m+1} = S$ where S is the base-stock level of the single inventory kept for spares. If no fleets have down machines and the inventory level I is below $L_{m+1} = S$, the repaired component is placed in the inventory. When I reaches $L_{m+1} = S$, there are no more broken components in the repair shop. If $L_k < I \leq L_{k+1}$, spare components are used only if machine types 1 to k fail. In other words, when $L_k < I < L_{k+1}$, even if there are down machines in classes $k+1$ to m, the repaired component is placed in the inventory as a spare component. When $I = L_{k+1}$ and the repair of a component is finished, it is used for the highest priority fleet associated with this threshold which has a down machine; i.e., fleet $k+1$ if each threshold is associated with a single fleet. When there is no positive-stock, the repaired component is allocated to the highest-priority fleet with down machines.

In the literature, the MR policy has been modeled when demand from each customer class follows a homogeneous Poisson process. When this is the case, customers can be prioritized according to their backlogging cost (corresponding to our fleet down time cost); that is, between two customer classes, the one with the higher backlogging cost has a higher

priority (classes with the same backlogging costs are considered to be in the same class). In our problem setting, the customer arrival rate (expressed as the failed component arrival rate) is state-dependent; this varies based on the number of down machines and the vector $(L_1 = 0, L_2, \ldots, L_{m+1} = S)$. If the objective is cost minimization, the state-dependent arrival rates of the failed components prevent us from determining the priority of a fleet by simply comparing its down time cost with those of other fleets. In this case, all possible alternatives of prioritizing fleets have to considered. The same is true for the HP policy summarized in Section 4. We also note that the optimal MR and HP policies may prioritize fleets differently.

Assuming that fleets 1 to m are prioritized from highest to lowest, let $C_{MR} := C(L_1 = 0, L_2, \ldots, L_{m+1} = S)$ be the long-run average cost of the MR policy given rationing levels $L_1 = 0, L_2, \ldots, L_{m+1} = S$, stated as

$$C_{MR} = \sum_{k=1}^{m} b_k \sum_{i=0}^{N_k} (N_k - i) P_{k,i} + hS + h_w \sum_{i=0}^{S} i\pi(i), \tag{1}$$

where $\pi(i)$ and $P_{k,i}$ are the steady-state probabilities of having i spare parts in the inventory, and i machines functional in fleet k, respectively, obtained for the system under the MR policy.

We design a recursive algorithm to obtain $\pi(i)$ and $P_{k,i}$. To do so, we construct a series of *auxiliary systems* k, $k = 1, \ldots, m$, with an inventory with a base-stock level of L_{k+1}. An *auxiliary system* k serves fleets 1 to k following an MR policy with $(L_1 = 0, \ldots, L_{k+1})$ as the threshold levels. The repair rate μ and failure rate λ_j for fleet j, $j = 1, \ldots, k$ are the same as in the original system. We denote the steady-state probabilities of having i spare parts in the inventory and i functional machines in fleet j, $j = 1, \ldots, k$ in *auxiliary system* k by $\pi^k(i)$ and $P^k_{j,i}$, respectively. As will be explained below, to analyze *auxiliary system* k, we need $\pi^{k-1}(i)$ and $P^{k-1}_{j,i}$ of *auxiliary system* $k-1$. Eventually, $\pi^m(i)$ and $P^m_{k,i}$ are obtained in the last round of the algorithm for *auxiliary system* m – which is, in fact, the original system –, giving us $\pi(i) = \pi^m(i)$ and $P_{k,i} = P^m_{k,i}$ for $k = 1, \ldots, m$; thus, we can compute the cost in Eq. (1).

The algorithm starts with the special *auxiliary system* 0 explained below.

***Auxiliary system* 0:** This is a system serving a single fleet of N_1 machines for which no spares inventory is kept. Hence, $P^0_{1,i}$ can be obtained by constructing a simple birth-and-death process where the states are the number of customers (expressed as failed components) waiting in the single server queue modeling the repair shop. Obviously, the failed component arrival rate at the repair queue depends on the number of down type 1 machines.

***Auxiliary system* 1:** When we add an inventory of L_2 spares to *auxiliary system* 0, we arrive at *auxiliary system* 1. Consider the sample path of *auxiliary system* 1 given in Figure 1 where the x-axis shows the time. The positive values on the y-axis show how many spares are on hand, and the absolute value of the negative values show how many type 1 machines are down. The Markov chain (MC) superimposed on the left hand side of the figure shows the failure rate (arrival rate at the repair shop) and the repair rate based on the number of units in the inventory or the number of down machines marked on the y-axis. Note that for $\mathcal{C}_1$ proportion of the time – to be determined –, there is no inventory in *auxiliary system* 1, and during these intervals without spares, *auxiliary system* 1 reduces to *auxiliary system* 0. Thus, $P^1_{1,i} = \mathcal{C}_1 P^0_{1,i}$, $i = 0, \ldots, N_1 - 1$ gives the steady-state probability of having i functional machines when there is no inventory in *auxiliary system* 1. Then, $P^1_{1,N_1} = 1 - \sum_{i=0}^{N_1-1} P^1_{1,i}$ is the probability of having N_1 machines functional – whether or not there are spare parts in the inventory.

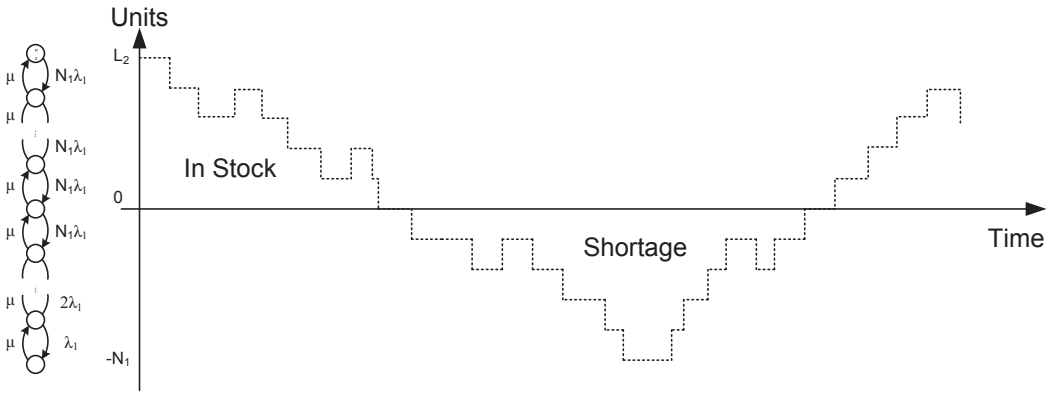

Figure 1: A Sample Path of the Single-Class Sub-system 1

Making use of the portion of the MC corresponding to the positive y values, for $1 \leq i \leq$

L_2,

$$\pi^1(i) = \left(\frac{\mu}{N_1\lambda_1}\right)^i \mathcal{C}_1 P^0_{1,N_1},$$

where P^0_{1,N_1} is the proportion of time the server is idle in *auxiliary system* 0, and

$$\mathcal{C}_1 = \left(1 + P^0_{1,N_1} \sum_{i=L_1+1}^{L_2} \left(\frac{\mu}{N_1\lambda_1}\right)^i\right)^{-1}.$$

Recall that the superscript 1 in $\pi^1(i)$ and $P^1_{1,i}$ indicates that these probabilities are found for *auxiliary system* 1.

***Auxiliary system* 2:** When we introduce fleet 2 as the low-priority class in *auxiliary system* 1, we arrive at *auxiliary system* 2. If no shared inventory is assumed ($L_3 = L_2$), any broken component from fleet 2 (i.e., class 2 customers) can be repaired only if all machines in fleet 1 are up/functional, and the inventory (reserved for high-priority fleet 1) level is at L_2. That is, periods during which the server (of the repair shop queue) is busy repairing components to reduce the number of down machines in fleet 1, or to increase the inventory level to L_2 are perceived as a server interruption by class 2 customers. Given this, each time the number of failed type 2 machines increases to 1, we will find the server idle if the inventory level is at L_2, or busy serving fleet 1 or raising the inventory level. In the latter instance, the server is perceived as interrupted by fleet 2. In the sample path given in Figure 2 where the horizontal axis shows the time, the dashed lines show the number of functional type 2 machines (via the vertical axis on the right hand side of the figure), and the solid lines show the number of spares in the inventory (via the vertical-axis on the left hand side of the figure). Here, we see that right before time instances t_A and t_B when a type 2 machine fails, leaving $N_2 - 1$ functional type 2 machines, the server is idle (with inventory level at L_2). However, two spares have already been used (for two failed type 1 machines) before time instance t_E, at which point, the number of functional type 2 machines decreases to $N_2 - 1$. At this moment, the server is trying to raise the inventory level back to L_2 but is seen as interrupted by fleet 2. At time t_C, we see that a type 1 machine fails and takes one unit from the inventory (lowering its level to $L_2 - 1$). As soon as this happens, the component

being repaired for fleet 2 is preempted until the inventory level reaches L_2 again at time t_D. This period is also seen as a server interruption by fleet 2.

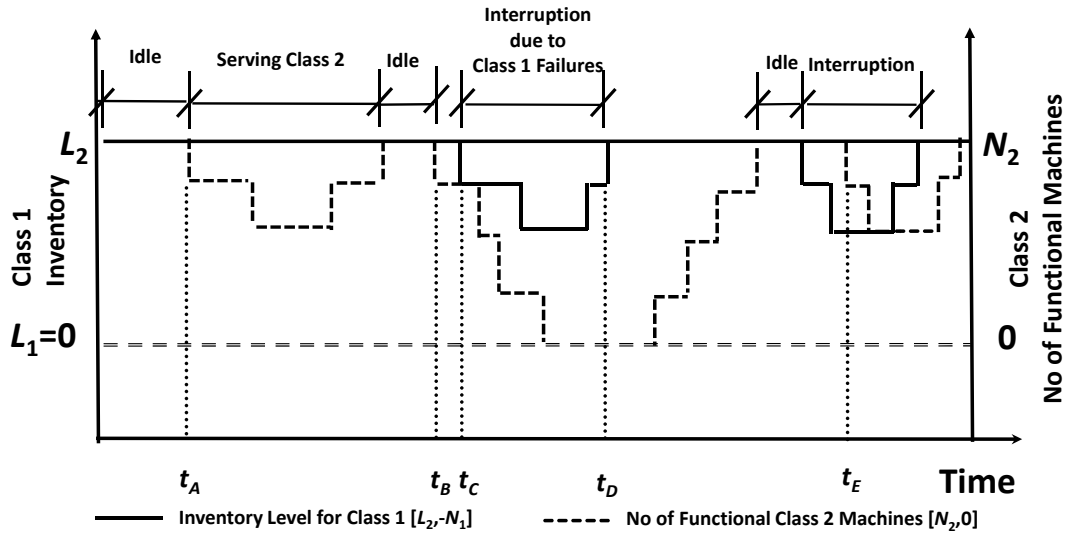

Figure 2: A Sample Path in Sub-system 2 when $L_3 = L_2$

In other words, from the standpoint of class 2 customers, the server can be interrupted when it is idle or when it is serving a type 2 customer with a failure rate of $\Lambda_1 = N_1\lambda_1$. Let D_2 denote the interruption times, starting with a class 1 arrival reducing the inventory level to $L_2 - 1$ and ending when the inventory level reaches L_2 again. Observe that D_2 is identically distributed as the first passage time from the second state at the top of the MC shown on the left side of Figure 1 (corresponding to inventory level $L_2 - 1$) to the state at the top (corresponding to inventory level L_2). The first passage times in finite state, continuous-time MC's (CTMC), and, thus, D_2, follow a phase-type distribution (PTD) (e.g., Kulkarni, 1989). Given this, when $L_3 = L_2$, *auxiliary system* 2 is an $M/M/1//N_2$ queueing system with a single unreliable server in which class 2 constitutes the only customers served, with Λ_1 as the server failure rate and D_2 modeling the server interruption periods. If the distribution

of D_2 can be characterized or accurately approximated (as in Section 3), the steady-state distribution of the number of customers out of this $M/M/1//N_2$ queue (i.e., the number of functional type 2 machines in *auxiliary system* 2) denoted by $P^{2*}_{2,i}$ (where the superscript 2* refers to $L_3 = L_2$) can be obtained. In specific,

$$P^{2*}_{2,N_2} = (1 + \Lambda_2 E[B_2])^{-1},$$

where $\Lambda_2 = N_1\lambda_1 + N_2\lambda_2$ and $E[B_2]$ is the expected length of a busy period in this queue, as found in Sahba, Balcıoğlu, and Banjevic (2013b).

If $L_3 > L_2$, the inventory is depleted at a rate of Λ_2 until it declines to L_2. Since for $\mathcal{C}_2$ proportion of the time – to be determined –, the inventory level is at or below L_2 (i.e., during these periods *auxiliary system* 2 reduces to *auxiliary system* 2 with $L_3 = L_2$ for fleet 2, and to *auxiliary system* 1 for fleet 1), for *auxiliary system* 2 with $L_3 \geq L_2$, we establish

$$\pi^2(i) = \begin{cases} \left(\frac{\mu}{\Lambda_2}\right)^{i-L_2} \mathcal{C}_2 P^{2*}_{2,N_2}, & L_2 < i \leq L_3, \\ \mathcal{C}_2 \pi^1(i), & 0 \leq i \leq L_2, \end{cases} \tag{2}$$

$$P^2_{1,i} = \mathcal{C}_2 P^1_{1,i}, \qquad i = 0, \ldots, N_1 - 1, \tag{3}$$

$$P^2_{1,N_1} = 1 - \sum_{i=0}^{N_1-1} P^2_{1,i},$$

$$P^2_{2,i} = \mathcal{C}_2 P^{2*}_{2,i}, \qquad i = 0, \ldots, N_2 - 1,$$

$$P^2_{2,N_2} = 1 - \sum_{i=0}^{N_2-1} P^2_{2,i},$$

where

$$\mathcal{C}_2 = \left(1 + P^{2*}_{2,N_2} \sum_{i=L_2+1}^{L_3} \left(\frac{\mu}{\Lambda_2}\right)^{i-L_2}\right)^{-1}.$$

Auxiliary system $k + (n-1)$**:** Consider *auxiliary system* $k-1$ with L_k as the inventory base-stock level in which fleets 1 to $k-1$ are served. Assume the steady-state probabilities of having i units in the inventory $\pi^{k-1}(i)$, $i = 0, \ldots, L_k$, and the probability of having i functional machines in fleet j, $P^{k-1}_{j,i}$, $j = 1, \ldots, k-1$, $i = 0, \ldots, N_j$. Under the MR policy, having the same threshold for some fleets instead of strictly increasing threshold levels for all

fleets may be more cost effective. To incorporate the possibility of having the same threshold for n fleets, we consider adding fleets k to $k+(n-1)$ ($1 \leq n \leq m-k+1$) at the same time to *auxiliary system* $k-1$ to arrive at *auxiliary system* $k+(n-1)$ (to allocate strictly increasing threshold level for each fleet, we merely set $n=1$ at each iteration). In this case, $L_k = L_{k+1} = \cdots = L_{k+(n-1)} \leq L_{k+n}$, and when the inventory level downcrosses L_k, the system stops serving fleets k to $k+(n-1)$ from the spares inventory. When the inventory level is at L_k, the repair shop sends the repaired component to the highest-priority fleet among fleets k to $k+(n-1)$ with down machines.

This implies that a type j customer (a broken component from fleet j), $j = k, \ldots, k+(n-1)$, can be repaired only when the inventory level is at L_k (i.e., all machines in classes 1 to $k-1$ are functional) and there are no type k to $j-1$ ($j > k$) customers in the repair shop. In other words, fleets k to $k+(n-1)$, prioritized from highest to lowest, are served under the preemptive-resume policy by an unreliable server (of the repair shop queue) becoming unavailable/interrupted at a rate of $\Lambda_{k-1} = \sum_{i=1}^{k-1} N_i\lambda_i$. If the server interruption time D_k can be characterized or well-approximated (see Section 3), following Sahba, Balcıoğlu, and Banjevic (2013b), $P_{j,i}^{k+(n-1)*}$ (where the superscript $k+(n-1)*$ refers to $L_k = L_{k+1} = \cdots = L_{k+n}$) can be obtained for each fleet $j = k, \ldots, k+(n-1)$. These are the steady-state probabilities that i type j customers are out of the multi-class priority $M/M/1//N$ queue ($N = \sum_{i=k}^{k+(n-1)} N_i$) with an unreliable server and give the number of functional type j machines in *auxiliary system* $k+(n-1)*$. Letting $E[B_{k+(n-1)}]$ denote the mean length of the busy period in this queue, as expressed in Sahba, Balcioglu and Banjevic (2013b), we have

$$P_{k+(n-1),N_{k+(n-1)}}^{k+(n-1)*} = (1+\Lambda_{k+(n-1)}E[B_{k+(n-1)}])^{-1},$$

where $\Lambda_{k+(n-1)} = \sum_{i=1}^{k+(n-1)} N_i\lambda_i$.

If an additional inventory of $L_{k+n} - L_k$ units are to be depleted by all the $k+(n-1)$ classes, we arrive at *auxiliary system* $k+(n-1)$, and the rest of the analysis mirrors that for *auxiliary system* 2. Letting $\pi^{k+(n-1)}(i)$ be the steady-state probability of having i spares stocked, the inventory is depleted at a rate of $\Lambda_{k+(n-1)}$ until it hits L_k. Note that for $\mathcal{C}_{k+(n-1)}$

– to be determined – proportion of the time, the inventory in *auxiliary system* $k+(n-1)$ is less than or equal to L_k; that is, during these periods, it reduces to *auxiliary system* $k-1$ for classes 1 to $k-1$ and *auxiliary system* $k+(n-1)*$ for classes k to $k+(n-1)$. Then, Eqs. (2)–(3) can be adjusted here as

$$\begin{aligned}
\pi^{k+(n-1)}(i) &= \begin{cases} \left(\frac{\mu}{\Lambda_{k+(n-1)}}\right)^{i-L_{k-1}} \mathcal{C}_{k+(n-1)} P^{k+(n-1)*}_{k+(n-1),N_{k+n-1}}, & L_{k-1} < i \le L_{k+n}, \\ \mathcal{C}_{k+(n-1)} \pi^{k-1}(i), & 0 \le i \le L_{k-1}, \end{cases} \\
P^{k+(n-1)}_{j,i} &= \begin{cases} \mathcal{C}_{k+(n-1)} P^{k-1}_{j,i}, & j = 1,\ldots,k-1,\ i = 0,\ldots,N_j-1, \\ \mathcal{C}_{k+(n-1)} P^{k+(n-1)*}_{j,i}, & j = k,\ldots,k+(n-1),\ i = 0,\ldots,N_j-1, \end{cases} \\
P^{k+(n-1)}_{j,N_j} &= 1 - \sum_{i=0}^{N_j-1} P^{k+(n-1)}_{j,i}, j = 1,\ldots,k+(n-1),
\end{aligned}$$

where

$$\mathcal{C}_{k+(n-1)} = \left(1 + P^{k+(n-1)*}_{k+(n-1),N_{k+(n-1)}} \sum_{i=L_{k-1}+1}^{L_{k+n}} \left(\frac{\mu}{\Lambda_{k+(n-1)}}\right)^{i-L_{k-1}} \right)^{-1}.$$

We can search different vectors of $(L_1 = 0, L_2, \ldots, L_{m+1})$ to find the optimal rationing levels and the corresponding cost given in Eq. (1).

3 Obtaining the Moments of the Server Interruption Time for Class k in *Auxiliary System* k

In this section, we derive the moments of the interruption time experienced by class k customers in *auxiliary system* k. In *auxiliary system* k with an inventory base-stock level of L_{k+1}, spares are depleted by all classes as long as the inventory is above $L_k (\le L_{k+1})$. As explained in Section 2, from the point of view of class k, server interruptions, occurring at a rate of $\Lambda_{k-1} = \sum_{j=1}^{k-1} N_j \lambda_j$, start when the inventory level decreases to $L_k - 1$ and end when it reaches L_k again. If we define the states as the number of spares in stock, the changes of the inventory level over time can be modeled as a birth-and-death process. Then, interruption

times (D_k) are the first-passage times from the state of having $L_k - 1$ units to the state of having L_k units and follow a PTD.

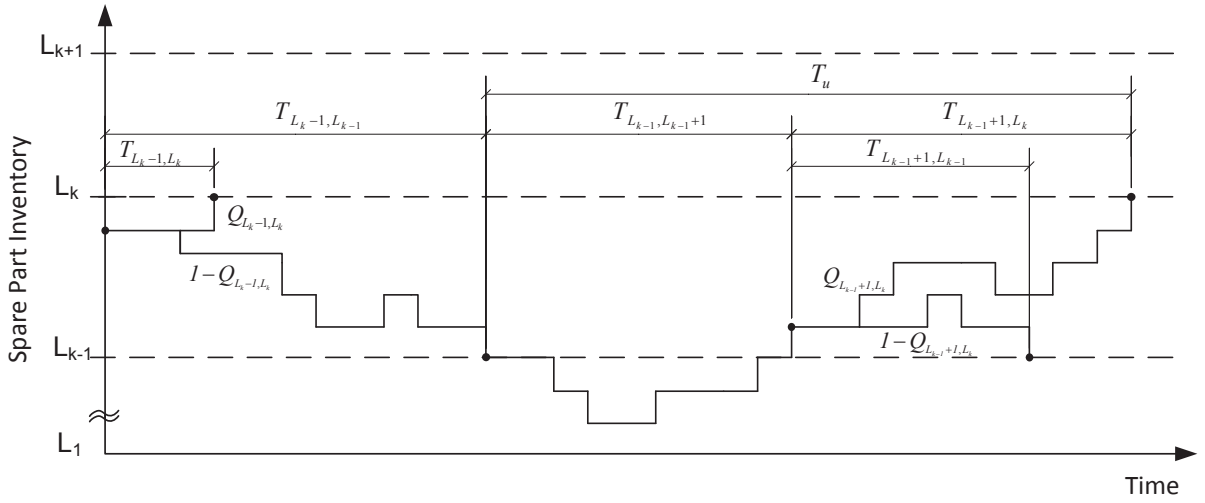

Figure 3: Breakdown of the Interruption Times of Class k

To explain how we obtain the first n moments of D_k, we use the sample path shown in Figure 3. Once the inventory level drops to $L_k - 1$ (i.e., time 0 in Figure 3), two events are possible. With probability $Q_{\mathbf{L_k-1},L_k}$, the inventory can go up to L_k – without first downcrossing L_{k-1} – in $T_{\mathbf{L_k-1},L_k}$ time units; or with probability $Q_{\mathbf{L_k-1},L_{k-1}}$, it declines to L_{k-1} in $T_{\mathbf{L_k-1},L_{k-1}}$ time units. Interpreting this as a Gambler's ruin problem, $Q_{\mathbf{L_k-1},L_{k-1}}$ ($Q_{\mathbf{L_k-1},L_k}$) is the probability of reaching (the absorbing) state L_{k-1} (L_k) from state $L_k - 1$ before reaching (the absorbing) state L_k (L_{k-1}), and

$$Q_{\mathbf{L_k-1},L_{k-1}} = 1 - Q_{\mathbf{L_k-1},L_k} = \frac{1 - \frac{\mu}{\Lambda_{k-1}}}{1 - \left(\frac{\mu}{\Lambda_{k-1}}\right)^{L_k - L_{k-1}}}. \tag{4}$$

If the inventory level reaches L_k without first downcrossing L_{k-1}, the interruption ends. Otherwise, after hitting L_{k-1} in $T_{\mathbf{L_k-1},L_{k-1}}$ time units, it takes T_u time units before the inventory level reaches L_k to end the interruption. Before presenting the next theorem, we introduce the first passage times $T_{L_{k-1},L_{k-1}+1}$ (from L_{k-1} to $L_{k-1}+1$), $T_{L_{k-1}+1,L_{k-1}}$ (from

$L_{k-1}+1$ to L_{k-1}), and $T_{L_{\mathbf{k-1}}+1,L_k}$ (from $L_{k-1}+1$ to L_k), as shown in Figure 3. After Theorem 1, we obtain their moments alongside those of $T_{\mathbf{L_k-1},L_k}$ and $T_{\mathbf{L_k-1},L_{\mathbf{k-1}}}$. Assuming that we have these moments, and noting that transition times $T_{\mathbf{L_k-1},L_{\mathbf{k-1}}}$, $T_{\mathbf{L_k-1},L_k}$, $T_{L_{\mathbf{k-1}},L_{\mathbf{k-1}}+1}$, $T_{L_{\mathbf{k-1}}+1,L_{\mathbf{k-1}}}$, and $T_{L_{\mathbf{k-1}}+1,L_k}$ are i.i.d r.v.s independent of each other, we derive the following:

Theorem 1 *The nth moment of D_k, i.e., the interruption time experienced by class k is*

$$E[D_k^n] = Q_{\mathbf{L_k-1},L_k} E[T^n_{\mathbf{L_k-1},L_k}] + Q_{\mathbf{L_k-1},L_{\mathbf{k-1}}} E\left[\left(T_{\mathbf{L_k-1},L_{\mathbf{k-1}}} + T_u\right)^n\right], \tag{5}$$

where $Q_{\mathbf{L_k-1},L_k}$ and $Q_{\mathbf{L_k-1},L_{\mathbf{k-1}}}$ are given in Eq. (4) and

$$\begin{aligned} E[T_u^n] &= Q_{L_{\mathbf{k-1}}+1,L_k} E\left[\left(T_{L_{\mathbf{k-1}},L_{\mathbf{k-1}}+1} + T_{L_{\mathbf{k-1}}+1,L_k}\right)^n\right] \\ &\quad + Q_{L_{\mathbf{k-1}}+1,L_{\mathbf{k-1}}} E\left[\left(T_{L_{\mathbf{k-1}},L_{\mathbf{k-1}}+1} + T_{L_{\mathbf{k-1}}+1,L_{\mathbf{k-1}}} + T_u\right)^n\right], \end{aligned} \tag{6}$$

where

$$Q_{L_{\mathbf{k-1}}+1,L_{\mathbf{k-1}}} = 1 - Q_{L_{\mathbf{k-1}}+1,L_k} = \frac{1-\left(\frac{\mu}{\Lambda_{k-1}}\right)^{L_k - L_{\mathbf{k-1}}-1}}{1-\left(\frac{\mu}{\Lambda_{k-1}}\right)^{L_k - L_{\mathbf{k-1}}}}. \tag{7}$$

We now show how the moments of the first passage times that appear on the right hand side of Eqs. (5) and (6) can be obtained. Assuming that we have the actual or approximate distribution of D_{k-1}, $k \geq 3$ (since there is no server interruption in *auxiliary system* 1):

Corollary 1 *The first passage time from state L_{k-1} to state $L_{k-1}+1$, $T_{L_{\mathbf{k-1}},L_{\mathbf{k-1}}+1}$, is the busy period in the $M/M/1//N_{k-1}+1$ queue with an unreliable server that fails at a rate of Λ_{k-2}, with D_{k-1} as the interruption time r.v. where each customer stays out of the queueing system for an exponentially distributed time with rate λ_{k-1}.*

To obtain D_{k-1}, we need to use Corollary 1 recursively. We start with D_2, first discussed for *auxiliary system* 2 in Section 2. The $M/M/1//N_1+1$ queue with no server failures gives the first passage time from state L_1 to state L_1+1, i.e., T_{L_1,L_1+1}. Theorem 1 is then used to obtain the moments of D_2. With Λ_{k-2} and D_{k-1}, Corollary 1 and Theorem 1 give D_k.

We use the following theorem to obtain the moments of the random variables $T_{\mathbf{L_k-1},L_k}$, $T_{\mathbf{L_k-1},L_{\mathbf{k-1}}}$, $T_{L_{\mathbf{k-1}}+1,L_k}$, and $T_{L_{\mathbf{k-1}}+1,L_{\mathbf{k-1}}}$. Before presenting the theorem, we introduce $\overline{L}_i^{(n)}$

denoting the n-th moment of the absorption time r.v. from transient state i to state 0, given that state m is avoided in a finite state continuous time Markov chain with 0 and m as the two absorbing states.

Theorem 2 *In a continuous time Markov chain with states $i \in \{0, \ldots, m\}$ and transition probabilities of $p_{i,j}$, letting states 0 and m be the absorbing states, $\overline{L}_i^{(n)}$, for $i \in \{1, \ldots, m-1\}$, is*

$$\overline{L}_i^{(n)} = E(Y_i^n) + \sum_{l=1}^{n-1} \left(\begin{array}{c} n \\ l \end{array} \right) \left(E(Y_i^l) \sum_{k \neq 0, k \neq m} \frac{Q_k}{Q_i} p_{i,k} \overline{L}_k^{(n-l)} \right) + \sum_{k \neq 0, k \neq m} \frac{Q_k}{Q_i} p_{i,k} \overline{L}_k^{(n)}, \quad (8)$$

where Y_i is the r.v. denoting the sojourn time in state i, and Q_i is the probability of reaching state 0 starting from i.

We employ Theorem 2 to obtain the moments of $T_{\mathbf{L_k - 1}, L_k}$ and $T_{L_{\mathbf{k-1}}+1, L_k}$ as given in Corollary 2, and those of $T_{L_{\mathbf{k-1}}+1, L_{\mathbf{k-1}}}$ and $T_{\mathbf{L_k - 1}, L_{\mathbf{k-1}}}$ in Corollary 3.

Corollary 2 *In auxiliary system k, we have*

$$E[T^n_{\mathbf{L_k - 1}, L_k}] = \overline{L}_1^{(n)}, \quad (9)$$

$$E[T^n_{L_{\mathbf{k-1}}+1, L_k}] = \overline{L}^{(n)}_{L_k - L_{\mathbf{k-1}} - 1}, \quad (10)$$

where $\overline{L}_1^{(n)}$ and $\overline{L}^{(n)}_{L_k - L_{\mathbf{k-1}} - 1}$ are found from Eq. (8) for a birth-and-death process with states $\{0, 1, \ldots, m = L_k - L_{k-1}\}$ by setting

$$\begin{aligned} Q_i &= \frac{1 - \left(\frac{\Lambda_{k-1}}{\mu}\right)^{m-i}}{1 - \left(\frac{\Lambda_{k-1}}{\mu}\right)^{m}}, \\ p_{i,i-1} &= 1 - p_{i,i+1} = \frac{\mu}{\mu + \Lambda_{k-1}}, \\ E[Y_i^n] &= n!(\mu + \Lambda_{k-1})^{-n}, \quad i \in \{1, \ldots, L_k - L_{k-1} - 1\}, \end{aligned} \quad (11)$$

where $m = L_k - L_{k-1}$.

Corollary 3 *In auxiliary system k, we have*

$$\begin{aligned} E[T^n_{L_{\mathbf{k-1}}+1,L_{\mathbf{k-1}}}] &= \overline{L}^{(n)}_1, \\ E[T^n_{\mathbf{L_k-1},L_{\mathbf{k-1}}}] &= \overline{L}^{(n)}_{L_k-L_{\mathbf{k-1}}-1}, \end{aligned}$$

where $\overline{L}^{(n)}_1$ and $\overline{L}^{(n)}_{L_k-L_{\mathbf{k-1}}-1}$ are found from Eq. (8) for a birth-and-death process with states $\{0,1,\ldots,m=L_k-L_{k-1}\}$ by using Eq. (11) for $E[Y^n_i]$ and setting

$$\begin{aligned} Q_i &= \frac{1-\left(\frac{\mu}{\Lambda_{k-1}}\right)^{m-i}}{1-\left(\frac{\mu}{\Lambda_{k-1}}\right)^{m}}, \\ p_{i,i-1} &= 1-p_{i,i+1}=\frac{\Lambda_{k-1}}{\mu+\Lambda_{k-1}}, \quad i\in\{1,\ldots,L_k-L_{k-1}-1\} \end{aligned}$$

where $m=L_k-L_{k-1}$.

Remark (1): If $L_k=L_{k-1}+2$, then $T_{\mathbf{L_k-1},L_k}$= $T_{\mathbf{L_k-1},L_{k-1}}$=$T_{L_{\mathbf{k-1}}+1,L_k}$= $T_{L_{\mathbf{k-1}}+1,L_{\mathbf{k-1}}}$.

Remark (2): If $L_k=L_{k-1}+1$, then $T_{\mathbf{L_k-1},L_k}$= $T_{\mathbf{L_k-1},L_{\mathbf{k-1}}}$=$T_{L_{\mathbf{k-1}}+1,L_k}$= $T_{L_{\mathbf{k-1}}+1,L_{\mathbf{k-1}}}=0$.

Remark (3): If $L_k=L_{k-1}$, we have the *auxiliary system* $k-1$.

Note that the moments of the sojourn times in each state and the state transition probabilities are not state-dependent in the birth-and-death process showing the inventory level. Therefore, labeling states $L_{k-1}, L_{k-1}+1, L_{k-1}+2,\ldots,L_k$, as states from 0 to m, Corollary 4 below presents a recursive computation method for $\overline{L}^{(n)}_i$ for $i\in\{1,\ldots,m-1\}$ (with $m=L_k-L_{k-1}$), from which we have $E[T^n_{L_{\mathbf{k-1}}+1,L_{\mathbf{k-1}}}]=\overline{L}^{(n)}_1$ and $E[T^n_{\mathbf{L_k-1},L_{\mathbf{k-1}}}]=\overline{L}^{(n)}_{m-1}$ in Corollary 3. Labeling states $L_k, L_k-1,\ldots L_{k-1}$, as states from 0 to m, Corollary 4 also provides $E[T^n_{\mathbf{L_k-1},L_k}]=\overline{L}^{(n)}_1$ and $E[T^n_{L_{\mathbf{k-1}}+1,L_k}]=\overline{L}^{(n)}_{m-1}$ given in Corollary 2.

Corollary 4 *The following recursion gives the n-th moment of the absorption time r.v. from state i, $i\in\{1,\ldots,L_k-L_{k-1}-1\}$, to state 0 given that state m is avoided as*

$$\overline{L}^{(n)}_i = b^{(n)}_i + \overline{L}^{(n)}_{i-1}, \tag{12}$$

where

$$b_{i-1}^{(n)} = \frac{C_{i-1}^{(n)} + \left(\frac{\mu}{\mu+\Lambda_{k-1}}\right) H_i^{-1} b_i^{(n)}}{1 - \left(\frac{\mu}{\mu+\Lambda_{k-1}}\right) H_i^{-1}}, \tag{13}$$

and $b_{m-1}^{(n)} = C_{m-1}^{(n)}$ *with*

$$\begin{aligned}
C_i^{(1)} &= E[Y_i], \\
C_i^{(2)} &= (E[Y_i^2] - 2E[Y_i]^2) + 2E[Y_i]\overline{L}_i^{(1)}, \\
C_i^{(3)} &= E[Y_i^3] + \left(3E[Y_i^2] - 6E[Y_i]^2\right)\left(\overline{L}_i^{(1)} - E(Y_i]\right) + 3E[Y_i]\left(\overline{L}_i^{(2)} - E[Y_i^2]\right).
\end{aligned}$$

and

$$H_i = \frac{1 - \left(\frac{\mu}{\Lambda_{k-1}}\right)^{L_k - L_{k-1} - i + 1}}{1 - \left(\frac{\mu}{\Lambda_{k-1}}\right)^{L_k - L_{k-1} - i}}.$$

Remark (4): If $\Lambda_{k-1} = \mu$, $Q_i = (m-i)/m$ in Corollaries 2 and 3 and $H_i = (m-i+1)/(m-i)$ in Corollary 4. In a similar vein, the right hand sides of Eqs. (4) and (7) become $1/m$ and $(m-1)/m$, respectively.

Recall that the exact MR model developed in Section 2 makes use of the $M/M/1//N$ queue with an unreliable server serving finitely many customers (see Sahba, Balcıoğlu, and Banjevic, 2013b) for which the server interruption distribution is required. These interruption times, namely D_k's, for the MR model we study are PTD r.v.s. The number of transient states and transition probabilities of D_k increase with the inventory levels, and the number of fleets served and their representations/structures become complex. Instead of using the original D_k's with their complex structures, we can approximate them and use the approximations as the interruption time r.v. in the $M/M/1//N$ queue analysis employed by the developed exact MR model. This will give us an MR policy, but its inventory rationing levels and cost will be an approximation of the original system. This is an approximation, not because we are exploiting a different/approximate method but because we are feeding approximate interruption time distributions into the exact model presented in Section 2.

The next question is how to approximate the original D_k's. One option is choosing a PTD r.v. with a simpler structure and the same first three moments of D_k. For instance, a

2-stage Mixture of Generalized Erlang (MGE) r.v. is an exponential r.v. with rate μ_1 (the sum of two exponential r.v.s with rates μ_1 and μ_2) with probability $1-a$ (a); it will have the same first three moments of D_k if we set (e.g., Altıok, 1997, page 52)

$$\mu_1 = \frac{X + \sqrt{X^2 - 4Y}}{2}, \text{ and } \mu_2 = X - \mu_1, \text{ and } a = \frac{\mu_2}{\mu_1}(E[D_k]\mu_1 - 1),$$

where

$$\begin{aligned} Y &= \frac{6E[D_k] - 3E[D_k^2]/E[D_k]}{(6E[D_k^2]/4E[D_k]) - E[D_k^3]}, \\ X &= \frac{1}{E[D_k]} + \frac{E[D_k^2]Y}{2E[D_k]}, \end{aligned}$$

and $E[D_k], E[D_k^2], E[D_k^3]$ are the first three moments of D_k found in Theorem 1. In Section 5, we test the accuracy of using the exact MR model with approximate interruption time distributions; we call this the MR policy approximation for ease of reference.

4 Benchmarking Policies

Alongside the MR policy proposed in Section 2, we consider two alternatives designed by Sahba, Balcıoğlu, and Banjevic (2013a): the *hybrid FCFS* (HF) and the *hybrid priority* (HP) policies. After summarizing these two policies, we close this section with a discussion of how the cost of the optimal policy can be computed numerically.

HF and HP policies have, a reserved inventory $S_k \geq 0$ for each class k and, a shared inventory $S \geq 0$ for all customers. Components from the shared inventory are expended, and only when they are depleted, are the reserved inventories used. This means that if the shared inventory is at its base-stock level S, the repair shop is idle. The dispatching decision for the repaired component comes into play when the shared inventory is empty, and some reserved inventories are below their base-stock levels, or there are some down machines. When this is the case, the repair shop has pending repair orders from fleets with down machines or fleets with missing spares in their reserved inventories. The repaired component is dispatched in an FCFS manner under the HF policy (to serve the highest

priority fleet under the HP policy with fleets 1 to m prioritized from highest to lowest) among the fleets with pending repair orders. If all machines are functional all fleets and the reserved inventories are full, the repaired component is placed in the shared inventory. HF and HP policies yield different $\pi(i)$, $\pi_k(i)$ and $P_{k,i}$, the steady-state probabilities of having i spares in the shared inventory, i spares in the reserved inventory of class k, and i machines to be functional in fleet k, respectively, defined by Sahba, Balcıoğlu, and Banjevic (2013a). These probabilities are also functions of $\mathbf{S} = (S, S_1, \ldots, S_m)$. Then, the long-run average cost, given $\mathbf{S}$ for the HF or HP policy, is

$$C_{HF/HP} = \{\sum_{k=1}^{m} b_k \sum_{i=0}^{N_k} (N_k - i) P_{k,i} + hS + h_w \left(\sum_{i=0}^{S} i\pi(i) + \sum_{k=1}^{m} \sum_{i=0}^{S_k} i\pi_k(i) \right)\}. \qquad (14)$$

We can search different vectors of $\mathbf{S} = (S, S_1, \ldots, S_m)$ to find the optimal shared and reserved inventory levels and the corresponding cost given in Eq. (14).

While the optimal policy for this problem remains unknown, the optimal cost can be numerically computed. To do so, we model the system as a semi-Markov decision process using the average cost criterion. Here, an action can be determined either when a component fails or a repair is over. The possible actions after a failure instant are either to dispatch an available spare part from the inventory or to take no action. Assuming a repaired component immediately joins the inventory, the possible actions are dispatching the component to one of the fleets with at least one down machine, or taking no action and letting the component stay in the spare parts inventory. With the assumption that a repaired component first enters the inventory, the possible actions at both decision epochs become the same. We define the state of the system as the number of down machines in each fleet and the inventory level as

$$i \ = (n_1, n_2, \ldots, n_m, l), \quad 0 \leq n_k \leq N_k, \quad k = 1, \ldots, m, \quad 0 \leq l \leq S.$$

The possible actions are

$$a \in A\,(i) = \{0, 1, \ldots, m\}\,,$$

such that if $a = 0$, no action is taken, and if $a = k$, a component is dispatched to class k. Therefore, at each decision epoch, the system may move into $m+1$ possible states as a result

of a failure or a repair completion. We assume a limited capacity of S for the inventory, i.e., when the inventory level increases to $S+1$ after the completion of a repair, taking no action is not allowed, and the component must be dispatched. Let $c_i(a)$ and $\tau_i(a)$ be the expected costs and the expected time until the next decision epoch if action a is chosen in state i. Then,

$$\tau_i(a) = \begin{cases} \frac{\sum_{k=1}^{m} n_k b_k - b_a + (l-a)h}{\mu - \lambda_a + \sum_{k=1}^{m}(N_k - n_k)\lambda_k}, & \sum_{k=1}^{m} n_k(S-l) > 0, \\ \frac{\sum_{k=1}^{m} n_k b_k - b_a + (l-a)h}{-\lambda_a + \sum_{k=1}^{m}(N_k - n_k)\lambda_k}, & \text{otherwise,} \end{cases}$$

or

$$\tau_i(a) = \begin{cases} \left(\mu - \lambda_a + \sum_{k=1}^{m}(N_k - n_k)\lambda_k\right)^{-1}, & \sum_{k=1}^{m} n_k(S-l) > 0, \\ \left(-\lambda_a + \sum_{k=1}^{m}(N_k - n_k\lambda_k\right)^{-1}, & \text{otherwise,} \end{cases}$$

where $b_0 = 0$ and $\lambda_0 = 0$.

There is a stationary deterministic average optimal policy for this finite-state semi-Markov decision process model, (see Theorem 11.4.6, page 557, Puterman, 2005). We first convert the model into a discrete-time Markov decision model and employ a version of the value-iteration algorithm (Tijms, 2003) to find a policy within ε of the optimal policy (ε-optimal policy) in the numerical examples presented in Section 5.1.

5 Numerical Experiment

In this section, we address three questions: (i) How accurate is the MR policy approximation introduced at the end of Section 3? (ii) How close is the performance of the MR policy to that of the optimal policy? Is its cost close to the optimal cost? (iii) What is the relative performance of the MR policy with respect to the HF and HP policies discussed in Section 4? Does it lead to significantly more cost savings?

To answer these questions, we consider a system in which three classes with $N_I = 5$, $N_{II} = 10$, and $N_{III} = 15$ are served. Repair times are exponentially distributed with rate

$\mu = 3$. We set the warehousing cost $h_w = 1/3$ in Eqs. (1) and (14) per unit spare per unit time in the inventory, as it is generally much smaller than the capital cost (Silver, Pyke, and Peterson, 1998, p. 45). It is assumed to be $h = 1$ for each spare part per unit time. We choose a different down time cost for each class from the set $\{10, 50, 100\}$. In addition, we set a different failure rate for each class by equating $N_k\lambda_k$, $k = I, II, III$, to a value in the set $\{0.7, 0.8, 0.9\}$, ensuring these are different from the values used for other classes. This gives a total of 36 examples presented in Table 2 and Table 3 in Appendix B.

In the rest of the discussion on numerical results, C_{MR} is the cost of the MR policy approximation. We use 2-stage MGE distributions approximating D_k for each *auxiliary system* $k \geq 2$ in the exact method developed in Section 2. These approximating MGE r.v.s have the same first three moments of D_k found by Theorem 1 (see Section 3 on the choice of MGE parameters).

Recall that down time cost is not sufficient to determine how to prioritize fleets under the MR policy without computing the system cost. Thus, for each problem, we have 6 different ways of prioritizing fleets (each is called a priority sequencing). For a given $L_4 = 0, \ldots, 12$, from 0 to L_4, L_2 can assume $L_4 + 1$ values. Given L_4 and L_2, L_3 can assume $L_4 - L_2 + 1$ values. This gives a total $(L_4/2 + 1)(L_4 + 1)$ combinations of L_3 and L_2, i.e., a total of 445 sets of inventory threshold/rationing levels for each priority sequencing. Thus, for each case, we employ the MR policy approximation 6×445 times, and the optimal cost C^*_{MR} is the one (with the corresponding priority sequencing, plus the threshold levels) that yields the minimum cost. These are presented in Table 4 and Table 5 in Appendix B. In each example, the optimal scenario turns out to have fleets that are prioritized according to their down time costs. For instance, in example 1, the highest priority class 1 is class III, and the lowest priority class 3 is class I. In all the examples, the base-stock level (L_4) is either 8, 9, or 10, and $L_2 = 0$. In each case, the MR policy allows all fleets to deplete the inventory until the inventory level hits L_3. If the inventory level is positive but less than or equal to L_3, if a machine from class 3 fails, no spare part is sent from the inventory. If there is no inventory, a repaired component is sent to the highest priority fleet with down machines.

5.1 The Accuracy of the MR Policy Approximation and its Performance Compared to the Optimal Policy

Based on the discussion of the numerical computation of the cost of the optimal policy in Section 4, we find a policy within 0.01% of the optimal policy in the numerical examples. Each resulting policy determines an action for each state, and there are between $9,504$ and $11,616$ states in each example. The algorithm run-time is around 12 hours for each example on a desk top computer with a 2.33GHz CPU. Given the priority sequencing and inventory rationing levels, it takes 0.62 seconds to compute the optimal cost of the MR policy. For each problem, the minimum cost is found for 445×6 (445 sets of inventory threshold levels and 6 priority sequencing) configurations; thus, it takes 28 minutes to arrive at optimality. In each problem, it takes 3.6 minutes to find the optimal cost of the HF policy (out of $13 \times 6 \times 6 \times 6 = 2808$ sets of S, S_1, S_2, S_3), and 6.3 minutes of the HP policy (out of 1944=324 ($12 \times 3 \times 3 \times 3$ sets of S, S_1, S_2, S_3)$\times 6$ (priority combinations) configurations).

Table 6 and Table 7 in Appendix B demonstrate a near perfect match between the MR policy and the ε-optimal policy based on the number of states with equal actions in both policies. In 22 out of 36 cases, the optimal policy is the MR policy (with 100% match), and the prioritization of classes and inventory threshold levels match those we find using the MR policy approximation. The mean/maximum absolute error of the approximate cost is 0.106%/0.13%. In other words, the MR policy approximation is extremely accurate. In all 36 cases, the mean/maximum absolute error of the approximate cost compared to the optimal cost is 0.11%/0.13%. Thus, we conclude that the MR policy performs very close to the optimal policy even when decisions differ at certain instances.

5.2 Relative Performances of the Policies

To compare the relative performances of the MR, HF and HP policies, we compute

$$\Delta_{HF}^{MR} \equiv \frac{C_{HF}^* - C_{MR}^*}{C_{HF}^*}, \ \Delta_{HP}^{MR} \equiv \frac{C_{HP}^* - C_{MR}^*}{C_{HP}^*}, \ \Delta_{HF}^{HP} \equiv \frac{C_{HF}^* - C_{HP}^*}{C_{HF}^*},$$

where C^*_{HF} and C^*_{HP} are the optimal cost of the system under the HF and HP policies, respectively, as given in Eq. (14) and found following Sahba, Balcıoğlu, and Banjevic (2013a). We present C^*_{HF} and C^*_{HP} in Table 8 and Table 9 in Appendix B along with the optimal inventory control parameters for each policy. The ratios Δ^{MR}_{HF} and Δ^{MR}_{HP} measure the cost decrease incurred by using the optimal MR policy instead of the optimal HF and HP policies, respectively. The ratio Δ^{HP}_{HF} captures how much more the HP policy reduces the cost than does the HF policy.

Table 1: Minimum, mean, median and maximum values of cost reduction of the MR policy compared to the HF and HP policies.

	Min(%)	Mean(%)	Median(%)	Max(%)
Δ^{MR}_{HF}	13.19	16.94	16.78	19.88
Δ^{MR}_{HP}	3.37	5.90	5.76	8.34
Δ^{HF}_{HP}	9.50	11.74	11.99	13.06

In Table 1, we see remarkable cost savings under the MR policy compared to the HF policy. Observe that all the three policies are flexible in the sense that they can deploy spares in different inventories or vary the threshold levels when the failure rates and down time costs are rotated among the fleets. Consequently, the optimal costs for a given policy, as listed in Tables 4-5 or Tables 8-9, do not fluctuate significantly from one problem to another one. The HP policy performs better than the HF policy. The HP policy increases the system cost by an average of 5.90% compared to the MR policy. From Tables 2-3, and Tables 8-9 in Appendix B, we see that the HF policy stores more spare parts than the other two policies. The shared inventory S is never 0, and reserved inventories are sometimes kept for one or two classes. The HP policy prioritizes the fleets based on their down time costs. The columns S_1 to S_3 show the reserved inventories for class 1 (with highest down time cost) to class 3 (lowest down time cost). The shared inventory S is never 0, and the HP policy keeps reserved inventories for classes 1 and 2. The total number of spares in the optimal HP

policy is never less than the number of spares in the optimal MR policy. In the optimal MR policy, as we recall from Table 4 and Table 5, no spare parts are reserved solely for fleet 1. Instead, fleets 1 and 2 share 3 to 4 units when the inventory level is less than or equal to L_3. As a result of this flexibility, the MR policy outperforms the HP policy in reducing the system cost. Problems 25, 26, 31, and 32 are the ones in which the savings under the MR policy are the highest – close to 20% and around 8% , respectively – when compared to the HF and HP policies. These are the problems in which the smallest fleet with 5 machines has the least down time cost while its machines have the highest failure rate. However, it is not easy for us to foresee under which scenarios the benefit of employing the MR policy can be felt more pronouncedly.

6 Conclusions

In this paper, we analyze a system of fleets, with each fleet consisting of finitely many machines which fail from time to time because of a repairable critical component. We propose employing the MR policy to control a shared inventory of spares (or as a transhipment policy between reserved inventories of fleets). The repair shop is modeled as a single server queueing system. The MR policy prioritizes classes/fleets and sets inventory threshold levels based on these priorities so that when the inventory level is below the inventory threshold identified for a class, that class is not served. We also employ MDP to obtain the cost of the ε-optimal policy for the same system. Our numerical findings indicate that the MR policy performs very close to the ε-optimal policy while outperforming the hybrid policies suggested in the literature. Although our numerical study indicates that the optimal control policy could very well be the MR policy, more research is required.

Acknowledgements

This work was supported in part by Natural Sciences and Engineering Research Council (NSERC) of Canada. The authors thank Dr. Elizabeth Thompson, for proofreading the manuscript. The authors thank the two anonymous referees and the editors for their invaluable suggestions to improve the manuscript.

References

Abouee-Mehrizi, H., B. Balcıoğlu, and O. Baron. 2012. "Strategies for a centralized single product multi-class $M/G/1$ make-to-stock queue", *Operations Research*, Vol. 60, No. 4, 803–812.

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

Axsäter, S. 1990. "Modelling Emergency Lateral Transshipments in Inventory Systems", *Management Science*, Vol. 36, No. 11, 1329–1338.

Gabor, A., L. van Vianen, G. Yang, S. Axsäter. 2016. "Enabling customer satisfaction and stock reduction through service differentiation with response time guarantees", *Erasmus School of Economics, Econometric Institute Research Papers*, No. EI2016–13.

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.

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.

Hausman, W. and G. Scudder. 1982. "Priority scheduling rules for repairable inventory systems", *Management Science*, Vol. 28, 1215–1232.

Jung, B., B. Sun, J. Kim, S. Ahn. 2003. "Modeling lateral transshipments in multiechelon repairable-item inventory systems with finite repair channels", *Computers and Operations Research*, Vol. 30, No. 9, 1401–1417.

Kukreja, A., C. P. Schmidt, and D. M. Miller. "Stocking Decisions for Low-Usage Items in a Multilocation Inventory System", *Management Science*, Vol. 47, No. 10, 1371–1383.

Kulkarni, V. G. 1989. "A new class of multivariate phase type distributions", *Operations Research*, Vol. 37, No. 1, 151–158.

Lee, H. L. 1987. "A Multi-Echelon Inventory Model for Repairable Items with Emergency Lateral Transshipments", *Management Science*, Vol. 33, No. 10, 1302–1316.

Louit, D., R. Pascual, D. Banjevic, and A.K.S. Jardine. 2011. "Optimization models for critical spare parts inventories— a reliability approach", *Journal of the Operational Research Society*, Vol. 62, 992–1004.

Puterman, M. L. 2005. *Markov Decision Processes*, John Wiley & Sons, New Jersey.

Sahba, P. and B. Balcıoğlu. 2011. "The Impact of Transportation Delays on Repairshop Capacity Pooling and Spare Part Inventories", *European Journal of Operational Research*, Vol. 214, 674–682.

Sahba, P., B. Balcıoğlu, and D. Banjevic. 2013a. "Spare Parts Provisioning for Multiple k-out-of-n : G Systems", *IIE Transactions*, Vol. 45, 953–963.

Sahba, P., B. Balcıoğlu, and D. Banjevic. 2013b. "Analysis of the Finite-source Multi-class Priority Queue with an Unreliable Server and Setup Time", *Naval Research Logistics*, Vol. 60, 331–342.

Silver, E. A., D. F. Pyke, and R. Peterson. 1998. *Inventory Management and Production Planning and Scheduling*, 3rd. Edition, John Wile & Sons.

Sleptchenko, A., M.C.van der Heijden, A. van Harten. (2005). "Using repair priorities to reduce stock investment in spare part networks", *European Journal of Operational Research*,

Vol. 163, No. 3, 733–750.

Tiemessen, H.G.H. and G.J.van Houtum. 2013. "Reducing costs of repairable inventory supply systems via dynamic scheduling", *International Journal of Production Economics*, Vol. 143, No. 2, 478–488.

Tijms, H. C. 2003. *A First Course in Stochastic Models*, John Wiley & Sons Ltd, West Sussex, England.

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

van Wijk, A.C.C., I.J.B.F. Adan, G.J. van Houtum. 2013. "Optimal allocation policy for a multi-location inventory system with a quick response warehouse", *OR Letters*, Vol. 41, 305–310.

Wong, H., D. Cattrysse, and D. Van Oudheusden. 2005. "Inventory pooling of repairable spare parts with non-zero lateral transshipment time and delayed lateral transshipments", *European Journal of Operational Research*, Vol. 165, 207–218.

Appendix A Proofs

Proof. Proof of Theorem 1. Eq. (5) is a direct result of the two possible trajectories the inventory level can follow starting from state $L_k - 1$ until reaching state L_k for the first time.

As seen in Figure 4, each time the inventory moves from L_{k-1} to $L_{k-1}+1$, with probability $Q_{L_{\mathbf{k-1}}+1,L_{\mathbf{k-1}}}$ ($Q_{L_{\mathbf{k-1}}+1,L_k}$) the inventory level, before reaching L_k, returns to state L_{k-1} in $T_{L_{\mathbf{k-1}}+1,L_{\mathbf{k-1}}}$ units, and another sub-cycle of length T_u starts (the inventory level reaches L_k ending T_u in $T_{L_{\mathbf{k-1}}+1,L_k}$ time units in a last cycle). This gives us Eq. (6). Since all states

of the underlying birth-and-death process are recurrent, the system goes through a random but a finite number of sub-cycles, each one of length T_u.

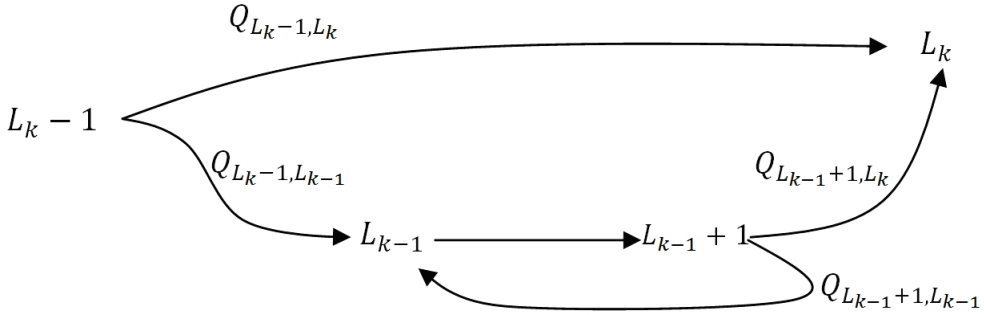

Figure 4: A Sample path of the interruption time for class k

Finally, in Eq. (7), $Q_{L_{\mathbf{k-1}}+1,L_{\mathbf{k-1}}}$ is the probability of reaching (the absorbing) state L_{k-1} from state $L_{k-1}+1$ before reaching (the absorbing) state L_k in a Gambler's ruin problem. ■

Proof. Proof of Corollary 1. We make the following analogy between the original system and the $M/M/1//N_{k-1}+1$ queue: When the inventory level hits L_{k-1} for the first time, there are N_{k-1} operational machines in the original system and the server is busy (one customer out of N_{k-1}+1 customers initiates a busy period in the $M/M/1//N_{k-1}+1$ queue). An arrival of classes 1 to $k-2$ drops the inventory level at a rate of Λ_{k-2} in the original system (the server fails in the $M/M/1//N_{k-1}+1$ queue at rate Λ_{k-2}), and it takes D_{k-1} time units before the inventory reaches L_{k-1} again (before the server interruption ends in the $M/M/1//N_{k-1}+1$ queue). During this time each type $k-1$ machine may fail at a rate of λ_{k-1} (additional customers, each with a rate of λ_{k-1}, may arrive at the $M/M/1//N_{k-1}+1$ queue). When any down machines in the original system (if there are down machines) is supplied with a fixed component while the inventory level is at L_{k-1} and one more component is fixed (corresponding to having all N_{k-1}+1 customers out of the $M/M/1//N_{k-1}+1$ queue), $T_{L_{\mathbf{k-1}},L_{\mathbf{k-1}}+1}$ (the busy period in the $M/M/1//N_{k-1}+1$ queue) ends. The moments of the busy period in the $M/M/1//N_{k-1}+1$ queue, hence those of $T_{L_{\mathbf{k-1}},L_{\mathbf{k-1}}+1}$, can be found in Sahba, Balcıoğlu, and Banjevic (2013). ■

Proof. Proof of Theorem 2. We introduce the following events and r.v.s to present the proof:

$A_{i,j}$**:** The event of reaching state j from state i in a single step of transition,

$A_{i,\circ,k}$**:** The event of eventually reaching state $k = 0, m$ after exiting state i,

X_i**:** The time to reach state 0 or m from state i ($X_k = 0$ for $k = 0, m$).

Let $I(E)$ denote the indicator function which equals 1 if event E is true and 0 otherwise. Then,

$$X_i = X_i I(A_{i,\circ,0}) + X_i I(A_{i,\circ,m}), i \neq 0, m.$$

Exiting state i, the system can be in any state after the first transition, thus implying that $\sum_k I(A_{i,k}) = 1$. Let the random variables X'_k and X_k be independent and identically distributed ($k \neq 0, m$ and $X'_0 = X'_m = 0$). Then

$$X_i = \sum_k X_i I(A_{i,k}) = \sum_k (Y_i + X'_k) I(A_{i,k}) = Y_i \sum_k I(A_{i,k}) + \sum_{k \neq 0,m} I(A_{i,k}) X'_k.$$

If the first state entered after leaving state i is either 0 or m, the remaining time to reach state 0 is zero. Otherwise it is,

$$X_i I(A_{i,\circ,0}) = Y_i I(A_{i,\circ,0}) + \sum_{k \neq 0, k \neq m} I(A_{i,k}) X'_k I(A'_{k,\circ,0}).$$

By definition, $\overline{L}_i^{(n)} = E[X_i^n | A_{i,\circ,0}] = E[X_i^n I(A_{i,\circ,0})]/Q_i$ (recall that Q_i is the probability of $A_{i,\circ,0}$ being true). Using the fact that for any random variable X_i and disjoint events B_i, $[I(B_i)]^n = I(B_i)$ and $[\sum_i X_i I(B_i)]^n = \sum_i X_i^n I(B_i)$, and that in our case, $I(A_{i,\circ,0}) I(A_{i,k}) I(A'_{k,\circ,0}) = I(A_{i,k}) I(A'_{k,\circ,0})$ for $k \neq 0, m$, we have

$$\begin{aligned}
E[(X_i^n I(A_{i,\circ,0}))] &= E[(X_i I(A_{i,\circ,0}))^n] \\
&= E\left[\left(Y_i I(A_{i,\circ,0}) + \sum_{k \neq 0, k \neq m} X'_k I(A_{i,k}) I(A'_{k,\circ,0})\right)^n\right] \\
&= E[Y_i^n] E[I(A_{i,\circ,0})] + \sum_{k \neq 0, k \neq m} E[I(A_{i,k})] E[\big((X'_k)^n I(A'_{k,\circ,0})\big)] \\
&\quad + \sum_{l=1}^{n-1} \binom{n}{l} \left(E[Y_i^l] \sum_{k \neq 0, k \neq m} E[{X'_k}^{n-l} I(A_{i,k}) I(A'_{k,\circ,0})]\right).
\end{aligned}$$

Note that $E[I(A_{i,\circ,0})] = Q_i$ and $E[I(A_{i,k})] = p_{i,k}$. Also,

$$\begin{aligned} E[{X'_k}^{n-l} I(A_{i,k}) I(A'_{k,\circ,0})] &= E[{X'_k}^{n-l} I(A'_{k,\circ,0}) | A_{i,k}] P(A_{i,k}) \\ &= E[{X_k}^{n-l} I(A_{k,\circ,0})] p_{i,k}. \end{aligned}$$

Then,

$$\begin{aligned} E[(X_i^n I(A_{i,\circ,0}))] &= E[Y_i^n] Q_i + \sum_{k\neq 0, k\neq m} p_{i,k} E[X_k^n | A_{k,\circ,0}] Q_k \\ &\quad + \sum_{l=1}^{n-1} \binom{n}{l} \left(E[Y_i^l] \sum_{k\neq 0, k\neq m} p_{i,k} E[X_k^{n-l} | A_{k,\circ,0}] Q_k \right). \end{aligned}$$

Dividing both sides by Q_i yields Eq. (8). ■

Proof. Proof of Corollary 2. Consider the birth-and-death process capturing the changes of the inventory level between levels L_k and L_{k-1}. This process has $m(= L_k - L_{k-1}) + 1$ states. If we consider the time it takes until the inventory level reaches L_k (to be interpreted as state 0) before hitting L_{k-1} (to be interpreted as state m) starting from the inventory level $L_k - 1, L_k - 2, \ldots, L_{k-1} + 1$ (to be interpreted as states $1, \ldots, m-1$, respectively), from Eq. (8), we get Eqs. (9) and (10). The probabilities Q_i and $p_{i,i-1}$ follow similarly. The duration in each state follows an exponential distribution with rate $\mu + \Lambda_{k-1}$, hence we have Eq. (11). ■

Proof. Proof of Corollary 3. The proof is similar to that for Corollary 2. We consider the time it takes until the inventory level hits L_{k-1} (to be interpreted as state 0) before reaching L_k (to be interpreted as state m), starting from the inventory level $L_{k-1}+1, L_{k-1}+2, \ldots, L_k - 1$ to be interpreted as states $1, \ldots, m-1$, respectively. ■

Proof. Proof of Corollary 4. From Eq. (8), the system of equations for the first moment of the absorption time r.v. from state i is

$$\overline{L}_i^{(1)} = E(Y_i) + \sum_{k\neq 0, k\neq m} \frac{Q_k}{Q_i} p_{i,k} \overline{L}_k^{(1)},\ 1 \leq i \leq m-1,$$

which is used alongside Eq. (8) to obtain the system of equations for the second moment as

$$\overline{L}_i^{(2)} = E(Y_i^2) + 2E(Y_i)\left(\overline{L}_i^{(1)} - E(Y_i)\right) + \sum_{k\neq 0, k\neq m} \frac{Q_k}{Q_i} p_{i,k} \overline{L}_k^{(2)},\ 1 \leq i \leq m-1.$$

Using the previous two equations together with Eq. (8), the system of equations for the third moment is

$$\overline{L}_i^{(3)} = E(Y_i^3)+\left(3E(Y_i^2) - 6E^2(Y_i)\right)\left(\overline{L}_i^{(1)} - E(Y_i)\right)+3E(Y_i)\left(\overline{L}_i^{(2)} - E(Y_i^2)\right)+\sum_{k\neq 0, k\neq m}\frac{Q_k}{Q_i}p_{i,k}\overline{L}_k^{(3)},$$

$$1 \leq i \leq m-1.$$

Let m be $L_k - L_{k-1}$ and $P_u = 1 - P_d = \mu/(\mu + \Lambda_{\mathrm{k-1}})$. Then, any of the above equations can be rewritten for $n = 1, 2, 3$ as

$$\begin{aligned}
\overline{L}_1^{(n)} &= C_1^{(n)} + P_u H_2^{-1}\overline{L}_2^{(n)}, \\
\overline{L}_i^{(n)} &= C_i^{(n)} + P_d H_i \overline{L}_{i-1}^{(n)} + P_u H_{i+1}^{-1}\overline{L}_{i+1}^{(n)}, \quad 1 < i < m-1, \\
\overline{L}_{m-1}^{(n)} &= C_{m-1}^{(n)} + P_d H_{m-1}\overline{L}_{m-2}^{(n)}.
\end{aligned}$$

Defining $b_{m-1}^{(n)} = C_{m-1}^{(n)}$ and $d_{m-1} = P_d H_{m-1}$, the equations given above become

$$\overline{L}_{m-1}^{(n)} = b_{m-1}^{(n)} + d_{m-1}\overline{L}_{m-2}^{(n)}.$$

Hence, for $i = m-1$ down to 2,

$$\overline{L}_i^{(n)} = C_i^{(n)} + P_d H_i \overline{L}_{i-1}^{(n)} + P_u H_{i+1}^{-1}\left(b_{i+1}^{(n)} + d_{i+1}\overline{L}_i^{(n)}\right),$$

or

$$\overline{L}_i^{(n)} = \frac{C_i^{(n)} + P_u H_{i+1}^{-1} b_{i+1}^{(n)}}{1 - P_u H_{i+1}^{-1} d_{i+1}} + \frac{P_d H_i}{1 - P_u H_{i+1}^{-1} d_{i+1}}\overline{L}_{i-1}^{(n)}, \tag{A.15}$$

Defining

$$b_i^{(n)} = \frac{C_i^{(n)} + P_u H_{i+1}^{-1} b_{i+1}^{(n)}}{1 - P_u H_{i+1}^{-1} d_{i+1}}, \quad d_i = \frac{P_d H_i}{1 - P_u H_{i+1}^{-1} d_{i+1}},$$

we next show that $d_i = 1$ for $1 \leq i \leq m-1$

$$d_{m-1} = P_d H_{m-1} = \frac{\Lambda_{k-1}}{\Lambda_{k-1} + \mu}\frac{1 - \left(\frac{\mu}{\Lambda_{k-1}}\right)^2}{1 - \left(\frac{\mu}{\Lambda_{k-1}}\right)} = 1,$$

and similarly, for $i = m-1$ to 2, we can show that

$$d_{i-1} = \frac{P_d H_{i-1}}{1 - P_u H_i^{-1} d_i} = 1.$$

With these, we have Eq. (13). Using it in Eq. (A.15) and noting $b_{m-1}^{(n)} = C_{m-1}^{(n)}$, we obtain Eq. (12). Moreover,

$$\overline{L}_i^{(n)} = \sum_{j=1}^{i} b_j^{(n)}, \quad i = 1, \ldots, m.$$

■

Appendix B Tables

Table 2: Parameters of the Examples-Cases 1 to 18

No	$N_I\lambda_I/$	$N_{II}\lambda_{II}$	$N_{III}\lambda_{III}$	λ_I	λ_{II}	λ_{III}	b_I	b_{II}	b_{III}
1	0.7	0.8	0.9	0.14	0.08	0.06	10	50	100
2	0.7	0.8	0.9	0.14	0.08	0.06	10	100	50
3	0.7	0.8	0.9	0.14	0.08	0.06	50	10	100
4	0.7	0.8	0.9	0.14	0.08	0.06	50	100	10
5	0.7	0.8	0.9	0.14	0.08	0.06	100	10	50
6	0.7	0.8	0.9	0.14	0.08	0.06	100	50	10
7	0.7	0.9	0.8	0.14	0.09	0.053	10	50	100
8	0.7	0.9	0.8	0.14	0.09	0.053	10	100	50
9	0.7	0.9	0.8	0.14	0.09	0.053	50	10	100
10	0.7	0.9	0.8	0.14	0.09	0.053	50	100	10
11	0.7	0.9	0.8	0.14	0.09	0.053	100	10	50
12	0.7	0.9	0.8	0.14	0.09	0.053	100	50	10
13	0.8	0.7	0.9	0.16	0.07	0.06	10	50	100
14	0.8	0.7	0.9	0.16	0.07	0.06	10	100	50
15	0.8	0.7	0.9	0.16	0.07	0.06	50	10	100
16	0.8	0.7	0.9	0.16	0.07	0.06	50	100	10
17	0.8	0.7	0.9	0.16	0.07	0.06	100	10	50
18	0.8	0.7	0.9	0.16	0.07	0.06	100	50	10

Table 3: Parameters of the Examples-Cases 19 to 36

No	$N_I\lambda_I$	$N_{II}\lambda_{II}$	$N_{III}\lambda_{III}$	λ_I	λ_{II}	λ_{III}	b_I	b_{II}	b_{III}
19	0.8	0.9	0.7	0.16	0.09	0.047	10	50	100
20	0.8	0.9	0.7	0.16	0.09	0.047	10	100	50
21	0.8	0.9	0.7	0.16	0.09	0.047	50	10	100
22	0.8	0.9	0.7	0.16	0.09	0.047	50	100	10
23	0.8	0.9	0.7	0.16	0.09	0.047	100	10	50
24	0.8	0.9	0.7	0.16	0.09	0.047	100	50	10
25	0.9	0.7	0.8	0.18	0.07	0.053	10	50	100
26	0.9	0.7	0.8	0.18	0.07	0.053	10	100	50
27	0.9	0.7	0.8	0.18	0.07	0.053	50	10	100
28	0.9	0.7	0.8	0.18	0.07	0.053	50	100	10
29	0.9	0.7	0.8	0.18	0.07	0.053	100	10	50
30	0.9	0.7	0.8	0.18	0.07	0.053	100	50	10
31	0.9	0.8	0.7	0.18	0.08	0.047	10	50	100
32	0.9	0.8	0.7	0.18	0.08	0.047	10	100	50
33	0.9	0.8	0.7	0.18	0.08	0.047	50	10	100
34	0.9	0.8	0.7	0.18	0.08	0.047	50	100	10
35	0.9	0.8	0.7	0.18	0.08	0.047	100	10	50
36	0.9	0.8	0.7	0.18	0.08	0.047	100	50	10

Table 4: The Optimal Inventory Rationing Levels and C^*_{MR} of the MR Policy-Cases 1 to 18

No	L_2	L_3	L_4	C^*_{MR}
1	0	4	9	16.174
2	0	4	9	16.12
3	0	3	9	16.476
4	0	2	9	16.458
5	0	3	9	16.371
6	0	2	9	16.398
7	0	4	9	16.071
8	0	4	9	16.19
9	0	3	9	15.932
10	0	3	9	17.024
11	0	3	9	15.908
12	0	2	9	16.859
13	0	3	8	15.551
14	0	3	8	15.344
15	0	3	9	16.949
16	0	2	9	16.29
17	0	3	9	16.977
18	0	2	9	16.467

Table 5: The Optimal Inventory Rationing Levels and C^*_{MR} of the MR Policy-Cases 19 to 36

No	L_2	L_3	L_4	C^*_{MR}
19	0	3	8	15.274
20	0	3	8	15.541
21	0	2	9	15.808
22	0	3	10	17.378
23	0	3	9	15.939
24	0	3	10	17.366
25	0	3	8	14.778
26	0	3	8	14.704
27	0	3	9	16.754
28	0	2	9	16.644
29	0	3	9	17.016
30	0	3	9	16.998
31	0	3	8	14.673
32	0	3	8	14.781
33	0	3	9	16.175
34	0	3	10	17.236
35	0	3	9	16.441
36	0	3	10	17.416

Table 6: Comparison of ε-Optimal Policy and the MR Policy-Cases 1 to 18

No	No of Iterations	No of States	% Matches with MR	Optimal Cost
1	402	10560	100.00	16.174
2	411	10560	99.99	16.110
3	479	10560	100.00	16.455
4	541	10560	99.95	16.441
5	492	10560	100.00	16.351
6	550	10560	99.91	16.382
7	406	10560	99.98	16.048
8	410	10560	100.00	16.191
9	457	10560	100.00	15.915
10	555	10560	100.00	17.001
11	469	10560	100.00	15.891
12	566	10560	99.89	16.836
13	361	9504	99.99	15.532
14	368	9504	100.00	15.326
15	499	10560	100.00	16.926
16	543	10560	99.93	16.275
17	518	10560	100.00	16.954
18	548	10560	99.91	16.449

Table 7: Comparison of ε-Optimal Policy and the MR Policy-Cases 19 to 36

No	No of Iterations	No of States	% Matches with MR	Optimal Cost
19	359	9504	100.00	15.255
20	368	9504	99.99	15.523
21	461	10560	99.92	15.780
22	612	11616	99.71	17.359
23	469	10560	100.00	15.922
24	624	11616	100.00	17.347
25	344	9504	100.00	14.764
26	348	9504	100.00	14.690
27	496	10560	100.00	16.731
28	559	10560	99.91	16.629
29	516	10560	100.00	16.993
30	563	10560	99.99	16.974
31	341	9504	100.00	14.658
32	350	9504	100.00	14.767
33	474	10560	100.00	16.155
34	609	11616	99.73	17.214
35	488	10560	100.00	16.421
36	625	11616	100.00	17.397

Table 8: The Optimal HF and HP policies-Cases 1 to 18

	HF Policy					HP Policy				
No	S	S_I	S_{II}	S_{III}	C^*_{HF}	S	S_1	S_2	S_3	C^*_{HP}
1	7	0	2	3	19.924	6	2	2	0	17.394
2	7	0	3	2	19.853	7	1	2	0	17.419
3	7	1	0	3	19.825	8	1	1	0	17.450
4	7	1	3	0	19.626	8	1	1	0	17.255
5	7	2	0	2	19.830	7	1	2	0	17.465
6	8	2	1	0	19.678	8	1	1	0	17.293
7	6	0	2	3	19.689	6	2	2	0	17.323
8	7	0	3	2	19.915	6	2	2	0	17.525
9	7	1	0	3	19.217	7	1	1	0	16.772
10	8	1	3	0	20.062	8	1	1	0	17.805
11	7	2	0	2	19.309	7	1	1	0	16.911
12	8	2	2	0	20.011	8	1	1	0	17.721
13	5	0	2	4	19.227	5	2	2	0	16.758
14	5	0	3	3	19.031	6	1	2	0	16.545
15	8	1	0	3	20.074	8	1	1	0	17.806
16	8	1	2	0	19.257	7	1	1	0	17.017
17	8	2	0	2	20.316	7	1	2	0	18.138
18	8	2	1	0	19.687	8	1	1	0	17.382

Table 9: The Optimal HF and HP policies-Cases 19 to 36

	HF Policy					HP Policy				
No	S	S_I	S_{II}	S_{III}	C^*_{HF}	S	S_1	S_2	S_3	C^*_{HP}
19	6	0	2	3	18.875	6	1	2	0	16.449
20	6	0	3	2	19.166	5	2	2	0	16.833
21	8	1	0	2	18.938	7	1	1	0	16.504
22	10	0	2	0	20.149	8	1	1	0	18.118
23	6	3	0	2	19.333	7	1	1	0	16.985
24	9	2	1	0	20.320	8	1	1	0	18.289
25	5	0	2	3	18.384	5	2	2	0	16.066
26	5	0	3	2	18.352	6	1	2	0	15.987
27	9	0	0	2	19.755	8	1	1	0	17.509
28	8	1	2	0	19.396	8	1	1	0	17.325
29	7	3	0	2	20.332	8	1	1	0	18.211
30	9	2	1	0	20.044	8	1	1	0	17.922
31	5	0	2	3	18.224	7	1	1	0	15.911
32	5	0	3	2	18.377	5	2	2	0	16.126
33	8	1	0	2	19.159	7	1	1	0	16.892
34	9	0	2	0	19.855	8	1	1	0	17.837
35	6	3	0	2	19.846	8	1	1	0	17.612
36	9	2	1	0	20.339	8	1	1	0	18.406