A queueing approach to a multi class M/G/1 make-to-stock with backlog

Opher Baron*and Yoav Kerner†

Abstract

This paper studies an $M/G/1$ production system serving several customer classes. We show that the Multilevel Rationing (MR) policy, that has been shown to be optimal in the $M/M/1$ case is not optimal in general. We propose another policy, which we call the extended MR (EMR). The EMR policy exploits the information on the number of all waiting customers at arrival epochs to assess the residual service time. We establish conditions under which the EMR policy is optimal and extend the conditions under which the MR policy is optimal.

1 Introduction

Consider a single machine that produces a single item in an $M/G/1$ environment. Customers arrive according to a Poisson process and production times are i.i.d. and independent of the arrival process. Each customer's demand is for a single unit and demand arriving when there are no items in stock is being backlogged. There are n classes of customers defer from each other only by their backlog costs. Also, there is a linear (in both time and items) inventory holding cost. A central planner faces two types of decisions: *allocation*–whether to allocate an item and to whom and *production* –whether to produce another item or to idle. The central planner's objective is to minimize the average cost per unit of time. In particular, a policy is optimal if no other policy has a smaller average cost per unit of time.

Due to the stationary nature of the problem and memoryless of the arrival process, a simple regeneration argument establishes that the optimal production policy is a stationary base stock level policy. In such policies items are make-to-stock until the base stock level, S is

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

†Department of Industrial Engineering and Management, Ben Gurion University of the Negev, P.O. 653 Beer Sheva 84105 Israel, kerneryo@bgu.ac.il

reached and production is stopped only when there are S items in stock. Of course, finding the optimal base stock level, S^*, still remains a question.

Three policies that were considered in the literature to control make-to-stock systems are the First Come First Served (FCFS), the Multi-level rationing (MR), and the strict priority policies. In the FCFS policy, when a decision to allocate an item is made, the items are allocated to the customer who is waiting the longest independent of this customer's type. The FCFS policy is not optimal because, for example, when there are backlogs and an item is allocated, it is better to allocate the item to the customer with the highest backlog cost. The MR policy is defined by a base stock level S and a sequence of rationing levels $0 = L_1 \leq L_2 \ldots \leq L_n \leq L_{n+1} = S$, such that an item is allocated to a class $j \geq i$ customer if and only if the stock level is at least L_i. Within classes $j \geq i$ items are allocated in a FCFS fashion. The strict priority policy is a special case of the MR policy where $0 = L_1 = L_2 \ldots = L_n \leq L_{n+1} = S$. That is with a strict priority policy items are allocated FCFS as long as there is inventory and are prioritized when there is backlog.

The problem of minimizing the average cost of a make-to-stock system with priorities was first investigated by Ha (1997). His work and much of the following one focused on the $M/M/1$ settings. de Vericourt et al. (2002) show that the MR policy is optimal for the $M/M/1$ make-to-stock queues. Recently, Abouee-Mehrizi et. al. (2012) provided the exact analysis of the strict priority and MR policies for the $M/G/1$ settings. They characterize the backlog cost for each customer class by considering a M/G/1 backlog queue for this class. These backlog queues were calibrated to have backlog cost identical to the one in the original system by extending the results of Kerner (2008) with respect to the distribution of the residual service time observed by arrivals. Recently, Economou and Manou (2015) provided a more intuitive derivation for these distributions. Abouee-Mehrizi et. al. (2012) note that the MR policy may not be optimal in these settings.

While in the $M/M/1$ settings information on the time spent in a specific inventory and backlog levels has no value due to memoryless of the arrival and production processes, such information may be valuable in the $M/G/1$ settings. For example, consider the case of deterministic service time of length $M = 5$ hours. Assume that we observe that an hour after production starts a low priority customer arrives and that the inventory level is such that it is optimal to backlog this customer. If there are no additional high priority arrivals after 3 more hours we know that the next production will be completed within an hour. In such a situation, we may reduce the cost by allocating an item from stock to a low priority customer (even before the next production completion). While such a control that relay on full information may be optimal, it requires a continuous review of the system and thus is hard to implement. We therefore focus on finding the optimal policy within the class of policies that only take actions at arrival and service completion times, based on the information on the inventory and backlog positions at these times. Such policies are more applicable than polices that require full information. We further note that policies within this class are static and they do not use a watch, that is the information on the time passed since the

last arrival or production completion is not kept. However, such policies could still use the information on the inventory and backlog to estimate the time to the next production completion and then use this estimation to reduce costs. Furthermore, the analysis and results below can be extended to cases with full information where a watch (i.e., information on the time elapsed since the last production starts) is available in a straight forward manner using the distribution of residual service time. Applying the resulting controls would reduce costs in these settings. This extension is straight forward and we comment to it in the paper's conclusion.

We propose the Extended MR (EMR) policy that takes allocation and production decisions only at arrivals and departures based upon the inventory and backlog positions at these times. This policy exploits the information on the residual service time given the level of backlog (queue length) to improve the control of the system. We use queue decomposition of the $M/G/1$ system with state dependent arrival rate, in the spirit of Abouee-Mehrizi and Baron (2015). We show that the EMR policy may reduce the costs of an $M/G/1$ make-to-stock system and that it is *optimal* when production times distributions are with Increasing Failure Rate (IFR). We further discuss the optimality of this policy for other service time distributions.

We describe the EMR policy in section 2. In section 3 we express the cost function for a given EMR policy, discuss cases where the EMR policy is optimal, and demonstrate numerically the benefit of the EMR policy over the MR policy. In section 4 we summarize the paper.

2 The EMR Policy

There are n classes of customers. The arrival process of class i is Poisson with rate λ_i and the total arrival rate is $\lambda = \sum_{i=1}^{n} \lambda_i$. The backlog cost of a class i customers is b_i per customer per unit of time. We assume, without a loss of generality, that $b_1 > b_2 > ... > b_n$. The holding cost is h per item per unit of time. Let $G(\cdot)$ be the cumulative distribution of the production time and let $M = \int_0^\infty x dG(x)$ be the mean production time. Assume for stability that $\rho = \lambda M < 1$. The objective is to minimize the long run average holding and backlog cost per time unit. For any given stationary policy (for convenience we omit the dependency in the policy from the notation), let B_i be a random variable having the distribution of the number of class i backlogged customers and I be a random variable having the distribution of the stock level, both under steady state. The optimization problem is to find a policy that minimizes the steady state cost

$$\min \left\{ hE\left(I\right) + \sum_{i=1}^{n} b_i E(B_i) \right\}.$$

We observe that because the problem is stationary and the only information known at any

arrival or departure time t is the inventory level $I(t)$ and the backlog levels $B_i(t)$ $i = 1, ..., n$, there is an optimal policy that is stationary with a base stock level. This observation holds because for any positive integer S, production completions epochs that leave the stock with level S are renewal epochs. Thus, if not producing at some level S is optimal once, it is always optimal. Note also that, given this objective function, once it is optimal to allocate an item when two customers are backlogged it is better to allocate the item to the customers with the higher backlog cost. However, deciding when it is optimal to allocate an item to customers is not as clear. This allocation decision captures the tradeoff between reducing costs for holding and backlog of low priority customers and the potential reduction of backlog cost for (future) high priority arrivals.

In the MR policy, this tradeoff is reflected in the sequence of rationing levels $0 = L_1 \leq \ldots \leq L_n < L_{n+1} = S$, such that an existing or arriving class i customer receives an item if and only if at her arrival time, t, the inventory level is strictly above L_i, i.e., $I(t) > L_i$. The MR policy controls the risk of backlogging future $i-$priority arrivals by only allocating inventory to classes $j \leq i$ if $I(t) > L_i$.

We define the EMR policy as follows: the production policy is according to a base stock level S, i.e., items are produced if and only if the stock level is less than S. The EMR policy is further characterized by $n-1$ rationing levels $0 \leq L_2 \leq \ldots \leq L_n < S$ as in the MR policy and, in addition, $n-1$ sets of integers $\tilde{A}_1, \ldots, \tilde{A}_{n-1}$. At production completion epochs the EMR policy behaves exactly as the MR policy – it allocates the item to the highest priority customer class that requires it in a FCFS manner within this class whenever the MR policy does. However, at arrival epochs, at time t, an item is allocated to a class i customer if and only if (i) $I(t) > L_i$, as in the MR policy, or (ii) $I(t) = L_i$ and $\sum_{j=i}^{n} B_j \in \tilde{A}_i$. An important example for the set $\tilde{A}_i$ is $\tilde{A}_i = \{q_i, q_i + 1, ...\}$, i.e., an item is allocated to class i customer if the stock level is L_i and the total number of backlogged customers is at least q_i.

We note that: when $I(t) = L_i$, we have $\sum_{j=1}^{n} B_i(t) = \sum_{j=i}^{n} B_i(t)$. The allocation in case (ii) is the extension of the MR policy. In this case, the EMR policy may reduce the inventory level, $I(t)$ to $L_i - 1$ due to an allocation to a class i arrival; this is in contrast to the MR policy that allocates items to class i only if the inventory level *after* this allocation is L_i or higher. This choice of the MR policy is equivalent to letting $\tilde{A}_i = \emptyset$, $\forall i$ in the EMR policy (alternatively, setting $L_i^{EMR} = L_i^{MR} - 1$ and $\tilde{A}_i = \{0, 1, \ldots\}$ $\forall i$, also reduces the EMR policy to an MR one).

The motivation behind the EMR policy is that while the MR policy ignores the number of backlogged low priority customers, this information is valuable when production follows a general distribution. The EMR policy uses the number of backlogged customers to better quantify the risk of backlogging future high priority customers. It can better quantify this risk because at any arrival epoch the probability of backlogging a future customer depends on the distribution of the residual production time of the current item. In the $M/G/1$ queue several authors e.g., Boxma (1983) and Kerner (2009), show that the distribution of

this residual production time depends on the total number of customers in the system. The EMR policy uses this improved information to refine the thresholds of the MR policy as above.

It is important to point out that unlike the $M/G/1$ with priorities and class dependent service times, assessing the remaining production time only depends on the *total* number of backlogged customers. That is the distribution of the residual production time depends on the total number of backlogged customer and not on their types. An intuitive justification for the latter is that our $M/G/1$ can be looked at as an $M/G/1$ with a single arrival process, where upon arrival each customer is assigned to a class with probability that is proportional to the class arrival rate. This assignment holds for the Poisson arrival case and is independent of the remaining production time.

3 Cost of the EMR Policy

In this section we derive the optimal EMR policy and its corresponding cost in two steps. First, we find the optimal MR policy as in Abouee-Mehrizi et. al. (2012). In the second step we investigate cases where the inventory level is L_i and $\sum_{j=i}^{n} B_j > 0$. In these cases, violating the MR policy and allocating items to class i customer may imply a backlog cost of future high priority arrivals, but saves holding and backlog costs. We derive the difference between the expected cost and saving implied by violating the MR policy. The set $\tilde{A}_i$ contains all the integers where this difference is negative; And the EMR policy allocates items to type i customers (assuming such exists) whenever the total backlog is in this set.

We present here the method for the case $n = 2$ and explain the generalization for $n > 2$ later. We refer to class 1 as the high priority class and to class 2 as the low priority class. For simplicity of the exposition, and to avoid trivialities, we assume that the system's parameters are such that $L_2 < L_3 = S > 0$.

Remark: The description of the two steps procedure above is helpful in conceptually understanding the idea behind the EMR: calculate the cost difference of deviating from the MR policy and change the allocation if this difference is negative. This description focuses on cases where it is beneficial to allocate items before the MR policy does. Similarly, there are cases in which it is beneficial to allocate items later than the MR policy does. Such cases may occur when the queue length indicates that the remaining production time is significantly long.

3.1 Cost Difference

Assume the stock level L_2 and a positive number of low priority customers, i.e., $B_2(t) > 0$. Also, assume that the residual production time upon a low customer arrival has a CDF $F(\cdot)$ with mean m_F. This distribution is that of the residual service time in the corresponding $M/G/1$ queue. Being more concrete, given the number of the low priority customers present $B(t)$, and the inventory level $I(t)$, the distribution F is the distribution of the residual service time in $M/G/1$ queue, given that the number of customers present is $B(t)+S-I(t)$. Explicit formulas for such distribution are given in Kerner (2008). (One might think that the fact that all backlogged customers are low priority customers should bias the distribution. Yet, due to the memoryless property of the arrival process and that each arrival can be assigned to a priority class randomly and independently, such a bias does not exist.)

Upon a low priority arrival there are two alternatives. The first is to allocate an item to the arriving customer and the second is to follow the MR policy. A key observation here is that the time until the stock level would return to L_2, under the EMR policy (when allocating the item) is identical to the time until the item is allocated to this customer under the MR policy. In other words, the impact of deviating from the MR policy on the future is limited to a time that is distributed as a busy period in an $M/G/1$, serving only class 1 customers (i.e., with arrival rate λ_1) the exceptional service level, distributed as F, and regular service times distributed as G. First, the time until we will have another unit in the inventory is a busy period of an $M/G/1$ queue, but the first service time in this busy period is distributed differently (because it is a residual production time). This is because some high priority customers may arrive before the next production completion. Second, at the end of this busy period, the inventory level will be L_2 independently of the current allocation decision: if we allocated to this low priority customer at the beginning of the busy period, the inventory went down to $L_2 - 1$ but another unit was added to the inventory at the end of the busy period (because the demand of this low priority customer was already satisfied); and if we didn't allocate to this customer at the beginning of the busy period, we will allocate this unit at the end of the busy period, that is instead of letting the inventory climbed to $L_2 + 1$ a unit is allocated to this low priority customer.

In the sequel, the term system refers to the $M/G/1$ queue with exceptional first service time described above and the term customers refers only to high priority customers. (The analysis ignores the number of low priority customers because it does not decrease before the end of the busy period under either policy.)

We denote a random variable with the distribution of such a busy period by τ_F and recall that $\mathrm{E}(\tau_F) = \frac{m_F}{1-\rho_1}$, see e.g., Takagi (1991) . Also, let $\tau_{F,L}^{-}$ be the expected amount of time during τ_F when the number of customers in the system is strictly less than L_2, i.e., the expected time that there is no backlog of high priority customers. Finally, Let Q_1 be the number of (high priority) customers in the system under the EMR policy.

Before stating and proving theorem, we provide the cost of deviating from the MR to the EMR.

Theorem 3.1. *The cost savings of deviating from the MR to the EMR-policy, i.e., the cost implied by violating the MR policy and allocating an item to a low priority customer is,*

$$C_d(F) = b_2\mathrm{E}(\tau_F) + h\tau^-_{F,L} - b_1(\mathrm{E}(\tau_F) - \tau^-_{F,L}), \tag{1}$$

where

$$\tau^-_{F,L} = \frac{m_F}{1-\rho_1}P(Q_1 < L_2). \tag{2}$$

Proof: When the controller allocates an item to a low priority customer, the direct saved cost is $b_2\mathrm{E}(\tau_F)$ from the backlog cost plus $h\tau^-_{F,L}$ from the holding cost. Yet, there is an opportunity cost–a high priority customer might be backlogged as a consequence of the inventory reduction. The expected opportunity cost is b_1 multiplied by the expected time during the busy period τ_F when the number of customers in the system is at least L; this expected time is $\mathrm{E}(\tau_F)-\tau^-_{F,L}$. Finally, (2) follows by a renewal argument, , where $P(Q_1 < L_2)$ can be obtain from e.g., Takagi (1991).⊡

Figure 1 clarifies the statement in Theorem 3.1 using a sample path argument. As long as the queue length does not exceed L_2, ($\tau^-_{F,L}$ time units in expectation), there is no backlog of high priority customer. Thus, an holding cost and a backlog cost of a single high priority customer are saved. However, when the queue length exceeds L_2 (above the horizonal line in the figure) there is a backlog of high priority customers, while a backlog of one low priority customer could be saved by allocating to her. Thus, the expected cost difference is the holding cost and backlog cost of a single low priority customer, minus a potential backlog cost of a high priority customer. In case the above cost deviation is negative, the EMR will allocate an item and its updated cost is that of the MR policy minus the cost difference in (1).

We next address the question of whether the EMR policy is optimal among the policies that only take actions at arrival and production completion times given the inventory and backlog positions at these times. It is clear from Theorem 3.1 that the optimal EMR policy leads to a lower cost than the optimal MR one. However, while the EMR's initial allocation decision is done as early as profitable (at the beginning of the relevant busy period), it is possible that additional allocation within such busy periods would further reduce costs. We note that, such cases might occur only when production times have a distribution with extremely non-smooth hazard rate function (that cause non-monotone behavior); thus we conjecture that such cases are rare. In Section 3.3 we establish that the EMR policy is optimal for a family of production time distributions.

We point out that there is no guarantee that the optimal base stock level under the MR policy is also optimal under the EMR policy. Furthermore, in the $M/M/1$ case, the optimal EMR coincides with the optimal MR policy. This is because the The EMR is based on the *conditional* distribution of the reaming production time, which in the $M/M/1$ case, is independent of the queue length.

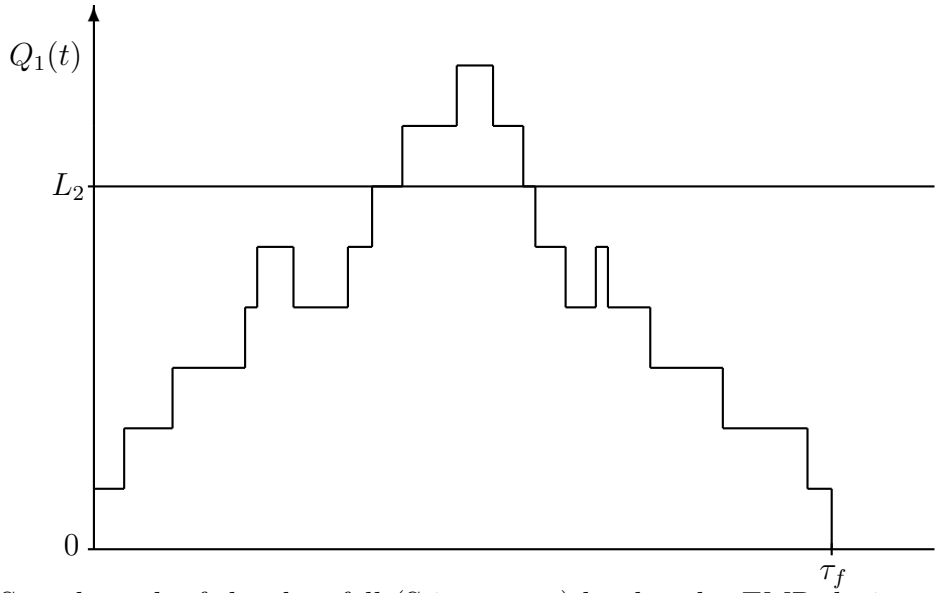

Figure 1: Sample path of the shortfall (S-inventory) level under EMR during a busy period

3.2 Comparing States

In this subsection we characterize a partial order between states in terms of their cost differences, and specify the cases of production times distribution in which this order is complete.

Theorem 3.2. *Consider two $M/G/1$ queueing systems, each with an exceptional first service time in each busy period. The two queues have the same arrival rate and the same service time distributions (beside the first, exceptional, service in each busy period). Let F_i be the CDF of the exceptional service time in queue i. If F_1 is stochastically larger*[1] *than F_2 then $C_d(F_1) > 0 \Rightarrow C_d(F_2) > 0$.*

Proof: First, dividing (1) by $\mathrm{E}(\tau_F) = \frac{\mu_F}{1-\rho_1}$, we get

$$\frac{C_d(F)}{E(\tau_F)} = (h + b_1)\frac{\bar{\tau_{F,L}}}{\mathrm{E}(\tau_F)} + b_2 - b_1.$$

Also, recall that

$$\frac{\bar{\tau_{F_i,L}}}{\mathrm{E}(\tau_{F_i})} = P(Q_i < L).$$

Thus, the statement in Theorem 3.2 is equivalent to the statement $P(Q_1 < L) < P(Q_2 < L)$.

We prove the latter by coupling the two queueing processes. Let $Q_i(t)$ be the number of customers in system i at time t. Assume that $Q_i(0-) = 0$ and $Q_i(0+) = 1$, $i = 1, 2$. Let

[1] F_1 is said to be stochastically larger than F_2 if $F_1(x) \geq F_2(x)\ \forall x$, with strict inequality at least at one point.

A_{ij}, $i = 1, 2$, $j = 1, 2, \ldots$ be the j^{th} exceptional service time of $Q_i(\cdot)$. We construct A_{1j} and A_{2j} coupled. More precisely, let F_i^{-1} be the inverse of F_i (or the pseudo inverse if the inverse does not exist[2]) and let U_j, $j = 1, 2, ...$ be a sequence of i.i.d. $U(0, 1)$ random variables. We have $A_{ij} = F_i^{-1}(U_j)$, $i = 1, 2$ and $j = 1, 2, ...$ and note that $A_{1j} \geq A_{2j}$ with probability 1. Furthermore, there are nonnegative random variables X_j such that $A_{1j} = A_{2j} + X_j$. The other service times in both queues are according to the same sample path. The processes $Q_1(t)$ and $Q_2(t)$ are not ordered. (An easy way to see this is by observing that sometime the first system may be empty while the second is not, whereas sometimes it is the other way around.) Next, we define the two processes $\tilde{Q}_1(t), \tilde{Q}_2(t)$ as follows. Take $Q_i(t)$ and change the order of busy and idle periods by first putting all busy periods one after the other and then all idle periods one after the other. That is, push all the idle periods that occurred before time t forward after all the busy periods. Now, we have $\tilde{Q}_1(t) \geq \tilde{Q}_2(t)$ with probability 1. Moreover, the fractions of the levels of the original Q_i processes and the new $\tilde{Q}_i$ processes coincide. That is,for each $L > 0$,

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T I_{\{Q_i(t)<L\}} dt = \lim_{T\to\infty} \frac{1}{T} \int_0^T I_{\{\tilde{Q}_i(t)<L\}} dt \quad , i = 1, 2, \ldots$$

Since the fraction

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T I_{\{Q_i(t)<L\}} dt$$

tends to the stationary probability $P(Q_i < L)$, our proof is completed.⊡

The application of Theorem 3.2 is as follows. Consider an arrival instant when n backlogged customers are present and let H_n be the CDF of the random variable with the distribution of a residual service time given n backlogged customers. By Theorem 3.2, if $C_d(H_n) > 0$ then for every m such that H_m is stochastically smaller than H_n we have $C_d(H_m) > 0$. We next characterize the cases in which, for a given level L, the set of numbers n such that $C_d(H_n) > 0$ is connected.

3.3 Monotone Failure Rates

In this Section we apply Theorem 3.2 to establish that threshold type EMR is optimal for production times with monotone failure rate distributions. Specifically, as in Barlow and Prochan (1996) we define:

Definition 3.1. A non-negative random variable X is said to be with Increasing (Decreasing, respectively) Failure Rate (IFR, DFR, respectively) if $P(X > t + s|X > t)$ is decreasing (increasing, respectively) with t.

[2]The pseudo inverse of F is defined to be $\inf\{x|F(x) \geq y\}$.

Proposition 3.1. *Let L_i be the rationing level of class i according to the MR policy. That is, according to the MR policy, an item is to be allocated to a class i customer if and only if the inventory level is above L_i.*

If the production time is IFR then the optimal EMR policy is to allocate an item to a class i customer if and only if: (i) $I(t) > L_i$, or (ii) $I(t) = L_i$ and $\sum_{j=1}^{N} B_j \geq q_i$, where

$$q_i = \arg\min_q C_d(H_q) > 0.$$

Proof: It has been shown in Kerner (2009) that if the service time is IFR then H_n is stochastically decreasing with n and in particular smaller than the service time distribution. Thus, by Theorem 3.2, if there exists a q such that $C_d(H_q) > 0$, then $C_d(H_{q'}) > 0$ for any $q' > q$.$\boxdot$

Observing Proposition 3.1, we can see that in the IFR case, the optimal EMR policy is globally optimal (among the policies we consider). This follows because the distribution of the time required for the inventory to reach level $L_2 + 1$ from a state $(I(t) \leq L_2, B_2(t) > 0)$ is monotone in both $I(t)$ and $B(t)$. The latter is true because the IFR property of the production time, which implies that the remaining production time, given any information is stochastically smaller than a new production time. Thus, following Theorem 3.2 and the regeneration feature of the production completion times, there is no policy that improves the optimal EMR.

One might think that in the DFR case, the optimal EMR policy would have an opposite structure to the one in the IFR case. That is, since a small number of backlogged customers indicates short elapsed production time and hence (due to the stochastic order of the DFR, see Kerner, 2009) a short residual production time, the optimal EMR is to allocate when the number of backlogged customers is small. However, a small number of backlogged should only be considered when the inventory level is above the rationing level. But above the rationing level there are no backlogs, so that the optimal policy for DFR is simply the MR policy. To demonstrate this consider two cases regarding the optimal EMR. These two cases differ at arrival epochs of a low priority customer who finds an empty system and $I(t) = L_2 + 1$. In the first case, the optimal action according to the EMR policy is not to allocate. Then, because the residual production time is stochastically increasing in the number of customers present and Theorem 3.2, the optimal action is not to allocate also when there are more than one low priority customers present. So, the EMR would be equivalent to MR with a higher rationing level (but this higher rationing level is suboptimal within the MR policy class). In the second case the optimal action according to the EMR policy is to allocate. We claim that this case is also equivalent to an MR policy (but with the same rationing level). Letting F_1 denote the residual production time at the first arrival, the optimality of this allocation decision implies that $C_d(F_1) > 0$. By the end of the busy period (serving high class customers), the inventory and backlog positions will be $I = L_2$

and $B_2 \geq 0$. Say that $B_2 = 1$, upon the next production completion another allocation to a class 2 customer is feasible or the inventory level could be raised. We observe that upon production completion the distribution of the residual production time is simply G, the distribution of a regular service time. We next claim that $C_d(G) > 0$ implies not to allocate the item. This is true because after the production completion, you can choose between not allocating the item and stay with $I = L_2 + 1$ and $B_2 \geq 1$ or to allocate the item and move to $I = L_2$ with a backlog $B_2 - 1$. However, by definition $C_d(G) > 0$ implies that allocating the item when $I = L_2 + 1$ is optimal, so that $I = L_2$ with a lower backlog is better. Finally, we recall that for DFR $F_1 \geq_{st} G$ and thus by Theorem 3.2 that $C_d(G) > 0$ so for this EMR policy it is optimal to satisfy the class 2 backlog rather than to increase inventory just as in the original MR policy.

With the discussion in this section we conclude that the exponential production case, which is both IFR and DFR and where it is established that MR is optimal has a special property. This distribution is the last IFR distribution where MR is always optimal. For any other IFR production distribution there may be some backlog cost and arrival and production rate combinations where the EMR policy may reduce cost. In contrast for any other DFR distribution the MR policy is optimal.

3.4 Numerical Examples: Erlang Service Times

To demonstrate the potential benefit of the EMR we consider several examples with IFR production times: The Erlang distribution with 10 phases. In these examples we find the optimal EMR policy for the case $N = 2$. For each example, we followed Proposition 3.1 and first derive the optimal MR policy using the method introduced in Abouee-Mehrizi et al. (2012). The system's analysis was done using Matrix Geometric technique. Then, we calculate $C_d(H_q)$ for each value of $q \geq 1$. By proposition 3.1, the optimal threshold queue length to allocate items to class 2 customers when $I(t) = L_1$, q^*, is the first with positive cost difference.

We present two sets of experiments. In the first, we examined four cases with 15 tests in each. Cases 1 and 2 are with system utility $\rho = 300/301 = 0.99668$. In case 1, 90% of the total arrival rate is high priority and in case 2 the two arrival rates are equal. In Cases 3 and 4 the system utilization is $\rho = 0.95$. In case 3, 90% of the total arrival rate is high priority and in case 4 the two arrival rates are equal. In all cases the holding cost is $h = 1$ and the low priority backlog cost is $b_2 = 5$. For each case we computed the optimal MR and EMR policies and their costs for 15 values of the high priority backlog cost $b_1 = 6, 7, ..., 20$. We provide the results S (the base stock level), L_1 (the rationing level), MR cost (the cost of the MR policy), q^*, and cost saving with $h = 1$, for $b = 6$, 10, and 20 in Table 1 below. These results are representative of all 15 tests. Figure 2 depicts the relative cost saving for all $60 = 4 * 15$ experiments. We observe that the average cost savings is 1.8%. The cost savings increase with the utilization and with the backlog cost b_1.

In the second set of experiments, we simulated the holding cost h from the $U(1,5)$ distribution, for each of the above cases. This simulation was repeated 100 times. The mean and standard deviation (SD) appear in the last two columns of Table 1 (again these results are representable of all 15 tests). We observe that the average cost savings are 1.85%,2.17%,1.02% and 1.12% for cases 1-4 respectively, and based on the standard deviation of these savings, they are substantially different than 0.

Table 1: Summary of numerical results

Case	$b1$	L	S	MR cost	q	cost saving w. $h=1$	mean w. h random	SD w. h random
1	6	1	19	336	1	0.78%	1.34%	0.03%
1	10	2	80	427	1	1.42%	2.06%	0.02%
1	20	4	163	615	2	2.64%	3.41%	0.05%
2	6	1	12	287	1	1.11%	1.61%	0.02%
2	10	1	27	355	1	1.73%	2.11%	0.04%
2	20	3	52	519	1	2.91%	3.63%	0.03%
3	6	3	9	191	1	0.49%	0.93%	0.01%
3	10	4	13	224	2	0.84%	1.25%	0.01%
3	20	6	18	298	2	1.42%	2.21%	0.02%
4	6	2	6	147	1	0.62%	1.14%	0.01%
4	10	2	8	183	1	1.02%	1.43%	0.01%
4	20	4	11	230	1	1.60%	2.32%	0.03%

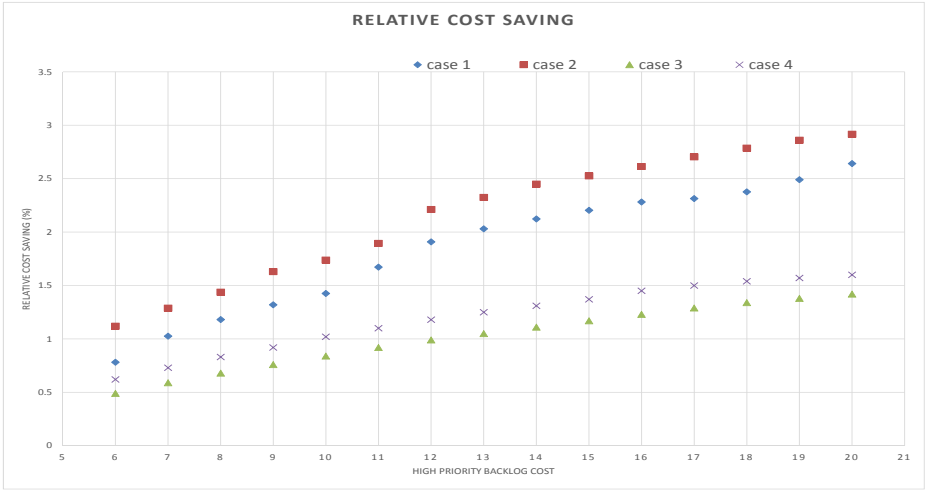

Figure 2: Relative cost saving varying with b_1

From all of our experiments we observe that the EMR usage of the extra information on the number of backlog is beneficial–it reduces the costs by an average of 1.4% and may be up to 4% in our examples. As these cost reduction are gained with little effort and because in practice most production times are close to IFR, we recommend using the EMR control policy rather than the MR one.

4 Conclusions

In this paper we introduced and analyzed the EMR policy for the control of make to stock queues with backlog and two priority classes. We establish that this policy that only takes actions at arrival and service completion times, based on the information on the inventory and backlog positions at these times–and is thus simple to implement, outperforms policies previous such policies, such as the MR policy. When the production times are IFR, we establish, numerically, that the EMR policy can reduce the average costs of such systems by 1-2%. The main contributions of this paper are threefold. We establish that the EMR control policy is a better, yet simple, control policy for make to stock queues under general production times; we show that when production times are IFR and DFR the structure of the optimal EMR policy is simple (in the DFR case it is simply MR); and we proved that the optimal EMR policy is optimal for both the DFR and IFR cases (again, in the DFR case the optimal EMR is simply MR). Extending the use and analysis of the EMR policy to more customer classes is not hard. In particular, it can be done by deriving sequentially the states for which it is fruitful to deviate from the MR policy, starting from the second-highest priority level and going downwards. However, establishing its optimality for more than two classes requires careful consideration of the different states the system can reach under the EMR policy. With more than two classes, the difference between the MR and EMR policies involves several different M/G/1 queues with first exceptional service time in each busy period (for each queue). Thus, we leave the issue of optimality in such cases as well as the extension of the EMR policy to the lost sales case for future research.

For $M/G/1$ make to stock system under more general policies-with continuous information and controls-our approach and analysis suggest that EMR policies based upon the conditional residual distributions of the service, $G(x-t)/\left(1-G(t)\right)$, as the exceptional service time in each busy period, are optimal. That is, knowing the elapsed time since the beginning of the production (instead of the queue length as an indicator for it), allows us to use the conditional distribution of the remaining production time given the elapsed time. However, establishing the optimality of this policy formally is outside the scope of this paper.

References

[1] Abouee-Mehrizi H., B. Balcioglu, and O. Baron "State-Dependent M/G/1 Queueing Systems" Queueing Systemas, Theory and Applications. Forthcoming.

[2] Abouee-Mehrizi H., B. Balcioglu, O. Baron 2011 "Strategies for a Centralized Single Product Multi-Class M/G/1 Make-to-Stock Queue," Operations Research 2012 Vol. 60, 803-812

[3] Barlow R. E., Proschan F.Mathematical Theory of Reliability (1965) (John Wiley and Sons, New York)

[4] Boxma, O.J. 1984. "Joint distribution of sojourn time and queue length in the M/G/1 queue with (in)finite capacity," European Journal of Operational Research, Vol. 16, 246-256

[5] Ha, A. 1997. "Stock-Rationing Policy for a Make-to-Stock Production System with Two Priority Classes and Backordering", Naval Research Logistics, Vol. 44, 457-472.

[6] Kerner Y. 2008. "The Conditional Distribution of the Residual Service Time in the $M_n/G/1$ Queue", Stochastic Models, Vol. 24, 364-375.

[7] Kerner Y. 2009. "Some invariance properties of monotone failure rate in the M/G/1 queue," Eurandom report 2009-018.

[8] Neuts, M.F. 1981. Matrix-geometric Solutions in Stochastic Models: An Algorithmic Approach Courier Corporation

[9] Takagi, H. 1991. Queueing Analysis, Volume 1, Elsevier: North Holland, The Netherlands.

[10] de Vericourt, F., F. Karaesmen, Y. Dallery. 2002. "Optimal Stock Allocation for a Capacitated Supply System", Management Science, Vol. 48, 1486-1501.