OPTIMAL POLICY IN A MAKE-TO-STOCK SYSTEM WITH TWO DEMAND CLASSES AND SERVICE LEVEL CONSTRIANTS

Feng Tian
College of Business and Public Administration
Governors State University
University Park, IL 60484
ftian@govst.edu

ABSTRACT

This paper studies a single-item make-to-stock production system with two demand classes. Demands follow Poisson distributions, and production time is exponentially distributed. Unsatisfied demands will be lost, and both demand classes have service level requirements. We first derive the condition of the existence of a feasible rationing policy. Then the optimality of rationing policy is shown. Numerical studies are used to compare the rationing policy with the FCFS policy, and also show how the service level constraint affects the efficiency of the rationing policy.

Keywords: production planning, make-to-stock, service level, rationing policy

1. INTRODUCTION

The practice of stock rationing means there are times when inventory is on hand but the arrival of low priority demands are denied in order to satisfy demand from high priority classes that might arrive in the future. This policy is widely applied in today's business world. One example is the inventory control problem of common components in assemble-to-order systems. Different end-products that share the same component may contribute different profit margins. The prolem seeks the optimal ordering and allocation policies for the common component. Another example is the common scenario where several customers have demands of the same product. The demands from different customers may have different prices, or different service requirements for the same product.

In this paper, we study the stock rationing problem of a single-item make-to-stock production system with two demand classes. Demand not satisfied immediately from inventory is lost. There are service level requirements for both demand classes.

(Topkis, 1968) is one of the first researchers to study the stock allocation problem. He formulates an uncapacitated, discrete time, single-item inventory system with multiple demand classes as a dynamic inventory model. Topkis shows that a base stock policy is optimal for ordering. By splitting the review period into finite subintervals, he proves that the optimal allocation policy has rationing levels for each demand class. (Ha, 1997a) studies the rationing policy of a single-item, make-to-stock production system with several demand classes and lost sales. The system is modeled as an $M/M/1$ queue. Ha characterizes the optimal policy as a sequence of monotone stock rationing levels. In a later paper, (Ha, 1997b) considers the same rationing problem with two demand classes where unmet demand is backordered. He shows that a base-stock type production policy is optimal, and the optimal allocation policy has a single monotone switching curve structure. (Vericourt et al., 2002) generalize the two demand classes problem of (Ha 1997b), and extend it to multiple-demand classes. They solve the stock allocation problem through dynamic programming, and characterize the optimal policy as a multiple threshold policy. (Deshpande et al., 2003) study a periodic review inventory system with two demand classes. They assume a (Q, r, K) policy, and provide an efficient algorithm for computing stock control and rationing parameters. However, none of these studies include a service level requirement for the demand classes.

(Nahmias and Demmy, 1981) study the demands' fill rates with and without rationing of a two-class uncapacitated inventory system under continuous review and periodic review policies. They calculate the demands' fill rates under different settings of rationing level, reorder point and order quantity. Then they construct tables that allow users to choose decision variables to meet desired fill rates. Optimizing the system cost under the service level constraint is not considered in the paper.

(Cohen et al., 1988) consider a service constrained single-location, single stage inventory system with two demand classes. They employ a (s, S) ordering policy and derive a simple priority allocation mechanism. The service level constraint is the aggregate fill-rate across both end items instead of individual fill-rate. (Bertsimas and Paschalidis, 2001) consider a single-stage make-to-stock manufacturing system with multiple products. There is a guaranteed service level for each product demand, but only one demand class exists for the each product. The work is focused on the production policy and no rationing is considered. (Benjaafar et al., 2004) study a more general demand allocation problem of multiple products and multiple production facilities with finite capacity and service level requirements. Orders are processed on a first-come-first-served (FCFS) basis in the paper.

We follow the approach and model of (Ha, 1997a). We first model the system in this paper as a single-product, $M/M/1$ make-to-stock queue plus a service level constraint. In section 2, we derive the condition of the existence of a feasible solution. Then we prove that rationing policy is optimal for traditional service level constrained model. The optimality of rationing policy for our model is also shown. In section 3, numerical studies are used to compare rationing policy and FCFS policy, and also show how the service level constraint affects the efficiency of the rationing policy. Conclusion and future work are explained in section 4.

2. THE MODEL

We consider a manufacturer who produces a single product for two different classes of customers. Finished products are stored at a common inventory. When a demand occurs, it can either be satisfied from the inventory or be denied and the sale will be lost. Once the inventory level drops to zero, all demands arriving will be lost. The lost sale costs of the two customer classes are c_1 and c_2, and we assume that $c_1 > c_2$. Demand from class i customer arrives as a Poisson process with rate λ_i, all demands will request one unit at a time. The manufacturer is modeled as a single server whose production time is exponentially distributed with mean $1/\mu$.

There are two decisions to be made in the system. The first one is what we call production decision, $u_0(t)$. At any time, the manufacturer can choose to either produce or stay idle, we denote the two actions by 1 and 0. The second one is the inventory allocation decision, $u_i(t)$. When a class i demand arrives, it may be satisfied or rejected, which is denoted as 1 and 0 respectively. If demand i is rejected, a lost sale cost c_i occurs. These two decisions can be summarized by the system control policy, $u(t)=\{u_0(t), u_i(t), i = 1, 2\}$, which specifies what action to take at time t given the current system state. We drop the time index t to let u represent the sequence of control policy. Let $X(t)$ be the inventory level at time t and $h(X(t))$ be the inventory holding cost. Denote $N_{u,i}(t)$ as the accumulated rejected units of class i demand under policy u. Let α be the discount rate. The expected system cost under policy u over an infinite horizon will be:

$$\underset{u\in U}{E}^{x}\left\{\int_0^\infty e^{-\alpha t}h(X_u(t))dt+\int_0^\infty e^{-\alpha t}c_1 dN_{u,1}(t)+\int_0^\infty e^{-\alpha t}c_2 dN_{u,2}(t)\right\} \tag{1}$$

where U is the set of all possible policies, $x = X(0)$ is the initial inventory level at time $t = 0$. Since each customer order consists of exactly one requested unit, the Type I and Type II service-level metrics equal each other. Let $D_i(t)$ be the accumulated class i demand by time t. The service level in this problem can be defined as the demand fill rate. If β_i is the service level requirement of class i demand. The service level constraint can be represented as:

$$\lim_{t\to\infty} N_{u,i}(t)/D_i(t) \le \beta_i \qquad i=1,2 \tag{2}$$

2.1 Problem Feasibility

The first problem we are facing for the service level constrained problem is whether there exists a feasible solution for the problem, i.e., given system parameters, can we find a policy that will satisfy the service level requirements. Let $\rho = (\lambda_1 + \lambda_2) / \mu$, it is obvious that if $\rho < 1$, all service level constraints can be satisfied. In fact, $\rho < 1$ means that the production rate is greater than the total demand rate. Over the long run, the system has the capability to build up the inventory level up to infinite (if we like), which means that demand fill rate can approach 100%, so is the service level. But if the total demand rate is greater than the production rate, i.e. $\rho > 1$, the service level requirement may not be able to be filled. For $\rho > 1$, we have following theorem:

Theorem 1: *If $\rho > 1$, the service level constrained problem has feasible solution if and only if $\lambda_1 \beta_1 + \lambda_2 \beta_2 < \mu$. Where β_1 and β_2 are required service levels of class 1 and class 2 demands, respectively.*

This theorem is easy to understand. For our problem setting (one unit per order), the service level requirements can be considered as the minimum percentage of orders (demand units) that should be satisfied. Then if the production rate is high enough to meet this portion of demand, then there exists a feasible solution for the problem. The theorem holds for system allows partial order filling and Type 2 service level measurement.

We also notice that the feasibility of the problem only depends on the demand rates, production rate, and the service level requirements. No cost parameters play any role in it. In the rest of the paper, we only consider problems with feasible solutions.

2.2 Service Level Constrained Model

The service level constrained model of the problem is:

$$P1 \quad \min_{u \in U} E_u^x \left\{ \int_0^\infty e^{-\alpha t} h(X_u(t))dt \right\} \tag{3}$$

$$s.t \quad \lim_{t \to \infty} N_{u,i}(t) / D_i(t) \le 1 - \beta_i \quad i = 1, 2$$

All variables are as defined early in this section. In this service level constrained model, only the inventory holding cost is considered, lost sales cost is not part of the target cost function. The major reason is that, in reality, the lost sale cost is hard to estimate, so people traditional use service level as a constraint which the policy needs to meet. The service level constrained model is considered as a substitute of the full cost model.

Before we start discussing the optimal policy of *P*1, we take a look at the full cost model first. The full cost model is:

$$P2 \quad \min_{u \in U} E_u^x \left\{ \int_0^\infty e^{-\alpha t} h(X_u(t))dt + \int_0^\infty e^{-\alpha t} c_1 dN_{u,1}(t) + \int_0^\infty e^{-\alpha t} c_2 dN_{u,2}(t) \right\} \tag{4}$$

The system parameters are (α, x, h, c_1, c_2, λ_1, λ_2, μ), which are defined earlier. The decision variable is the control policy *u*. Ha (1997a) proves that a rationing policy is optimal for *P*2, which is denoted as a (*S*, *R*) policy.

A (*S*, *R*) rationing policy means that the manufacturer will always produce until the inventory level hits level *S*. If the inventory level is above *R*, demand from both classes will be satisfied. If the inventory level is no more than *R*, then only high priority demands (class 1 demands) will be satisfied and class 2 demands will be rejected. We call *S* as the inventory target, and *R* the rationing level. If the inventory level is zero, then demands of both classes will be rejected. Given a (*S*, *R*) policy, we can easily derive the service level of each class demand as:

$$\beta_1(S,R) = 1 - \frac{(1-\rho_1)(1-\rho)\rho^{S-R}\rho_1^R}{(1-\rho_1^{R+1})(1-\rho)\rho^{S-R} + (1-\rho_1)(1-\rho^{S-R})} \tag{3}$$

$$\beta_2(S,R) = 1 - \frac{(1-\rho_1^{R+1})(1-\rho)\rho^{S-R}}{(1-\rho_1^{R+1})(1-\rho)\rho^{S-R} + (1-\rho_1)(1-\rho^{S-R})} \tag{4}$$

where $\rho = (\lambda_1 + \lambda_2) / \mu$, and $\rho_1 = \lambda_1 / \mu$.

For the (*S*, *R*) rationing policy, the stationary average inventory level can be found as:

$$I(S,R) = S - \frac{\dfrac{\rho(1-\rho_1)}{1-\rho}\left\{1-\rho^{S-R}-(1-\rho)(S-R)\rho^{S-R-1}\right\}}{(1-\rho_1)(1-\rho^{S-R})+(1-\rho)\rho^{S-R}(1-\rho_1^{R+1})} \tag{5}$$

$$-\frac{\dfrac{\rho_1(1-\rho)}{1-\rho_1}\left(\dfrac{\rho}{\rho_1}\right)^{S-R}\left\{\rho_1^{S-R}-\rho_1^{S+1}+(1-\rho)\left[(S-R)\rho^{S-R-1}-(S+1)\rho_1^S\right]\right\}}{(1-\rho_1)(1-\rho^{S-R})+(1-\rho)\rho^{S-R}(1-\rho_1^{R+1})}$$

Assuming linear inventory holding cost and the unit holding cost rate *h*, the expected system cost with (*S*, *R*) policy is:

$$C(S,R) = hI(S,R) + \lambda_1 c_1[1-\beta_1(S,R)] + \lambda_2 c_2[1-\beta_2(S,R)] \quad (6)$$

where $\beta_1(S, R)$ and $\beta_2(S, R)$ are the service levels defined in equation (3) and (4).

For the optimal policy of $P2$, we have following lemma.

Lemma 1: *Assume that S* and R* are the optimal inventory target and rationing level for a given problem P2. The corresponding $\beta_i(S^*, R^*)$ is called the optimal service level of class i demand for the same problem.*

(a) If the loss sales cost of class i demand, c_i, is zero, the optimal service level of this demand class $\beta_i(S^*, R^*) = 0$. If $c_i \to \infty$, then $\beta_i(S^*, R^*) \to 1$.

(b) If we keep h, μ, λ_1, and λ_2 fixed, then the optimal service level $\beta_i(S^*, R^*)$ is an incremental function of the loss sales cost c_i.

Lemma 1 tells us that for $P2$, if we keep h, μ, λ_1, and λ_2 fixed, and adjust the loss sales cost c_i from 0 to ∞, the optimal service level of class i will change from 0 to 1. The change is a monotonically increasing process. With Lemma 1, we can easily derive the optimal policy of $P1$ in Theorem 2.

Theorem 2: *A rationing policy is optimal for P1.*

From Lemma 1 we can see that given $P1$, we can construct a $P2$ with same parameters and adjustable loss sales costs. By adjusting the loss sales costs, we can find a $P2$ whose optimal service levels are equal to the service level constraint in $P1$. The proof of Theorem 2 shows that the optimal policy of this $P2$ is optimal for $P1$.

3. NUMERICAL STUDY

In this section, we study numerical examples to learn how the rationing policy performs under different system parameters. Through these experiments, we first look at the difference of minimum capacity requirement between the rationing policy and the First Come First Serve (FCFS) policy. We then compare the cost saving of the rationing policy against the FCFS policy. More than that, we also compare the cost savings between cases with and without the service level constraint. We want to identify how and why the service level constraint affects the cost saving. We focus on the full cost model with service level constraint on the cost saving part.

(Li, 1992) has shown that, for a single-product make-to-stock queue with FCFS policy a base-stock policy is optimal. As shown in (Ha, 1997a), the average inventory level, service level, and stationary system cost for the base-stock FCFS policy are:

$$\hat{I}(S) = \frac{S+\rho^{S+1}}{1-\rho^{S+1}} - \frac{\rho}{1-\rho}$$

$$\hat{\beta}(S) = 1 - \frac{(1-\rho)\rho^S}{1-\rho^{S+1}}$$

$$\hat{C}(S) = h\hat{I}(S) + (\lambda_1 c_1 + \lambda_2 c_2)(1-\hat{\beta}(S))$$

Notice that for the base-stock FCFS policy, both demand classes experience the same service level.

We define following experiment parameters the same as (Ha, 1997a):

1. Lost sales cost ratio: c_1/c_2

2. Relative holding cost rate: $h' = h / (\lambda_1 c_1 + \lambda_2 c_2)$
3. Demand rates ratio: λ_1 / λ_2
4. Traffic intensity of the system: $\rho = (\lambda_1 + \lambda_2) / \mu$

Then we have:

$$\rho_1 = \left(\frac{\lambda_1/\lambda_2}{1+\lambda_1/\lambda_2} \right) \rho$$

All the system decision variables can be represented by above parameters.

For the FCFS policy, the minimum cost $\hat{C}^*(S)$ is derived by searching S over $\{0, 1, \ldots S_u\}$. S_u is the minimum nonnegative integer greater than $(\rho - 1)/(h'\rho) + 1/\ln(\rho) - 1$, which Ha (1997a) shows is the upper limit of optimal base-stock level. For the (S, R) rationing policy without a service level constraint, we first find the value of R that minimize the cost $C(S, R)$ for each fixed S. The search is over $\{0, 1, \ldots, S\}$. We denote it as $R^*(S)$. Then we search S^* which minimize the cost $C(S, R^*(S))$. The search will start from $S = 0$, and end when the current cost is significantly greater than current minimum cost and the first difference of cost continues to remain positive over a large range. The cost saving is defined as $\{[\hat{C}^*(S) - C^*(S, R)] / \hat{C}^*(S)\} \times 100\%$. (The search process of rationing policy with service level will be described later.)

3.1 Capacity Comparison

Theorem 1 shows that for given service level requirements β_1 and β_2, there exists feasible rationing policy as long as $\lambda_1 \beta_1 + \lambda_2 \beta_2 < \mu$. If we define weighted service level requirement $\beta = (\lambda_1 \beta_1 + \lambda_2 \beta_2) / (\lambda_1 + \lambda_2)$, this condition is equivalent to $\beta < 1/\rho$. The FCFS policy provide the same service level to both demand, which means that the system has to meet the higher required service level. With simple mathematics, we can find out that for the FCFS policy, the condition for it to satisfy the service level requirement is $Max(\beta_1, \beta_2) < 1/\rho$, or $(\lambda_1 + \lambda_2)*Max(\beta_1, \beta_2) < \mu$. So to meet the same service level requirement, FCFS policy requires a higher minimum system capacity than the rationing policy. Following examples will show how the capacity saving, defined as $[\mu(S) - \mu(S, R)] / \mu(S)$, changes with system parameters. $\mu(S, R)$ and $\mu(S)$ are the required minimum system capacities for the rationing policy and FCFS policy, respectively. All the labels of Y-axis of this subsection are this capacity saving.

Figure 1 shows how the capacity saving changes with the demand rate ratio. Class 1 demand requires a higher service level in this experiment. So for the FCFS policy, β_1 is always the service level that system should provide to both demands, even though class 2 demand does not require such a high service level. Comparing with rationing policy, FCFS policy needs more capacity to meet this extra service level provided to class 2 demands. When λ_1 is smaller than λ_2, this excess capacity is a significant portion compared with the total capacity requirement. However, as λ_1 increases, class 1 demand will gradually dominate the demand, and the capacity saving will become less and less.

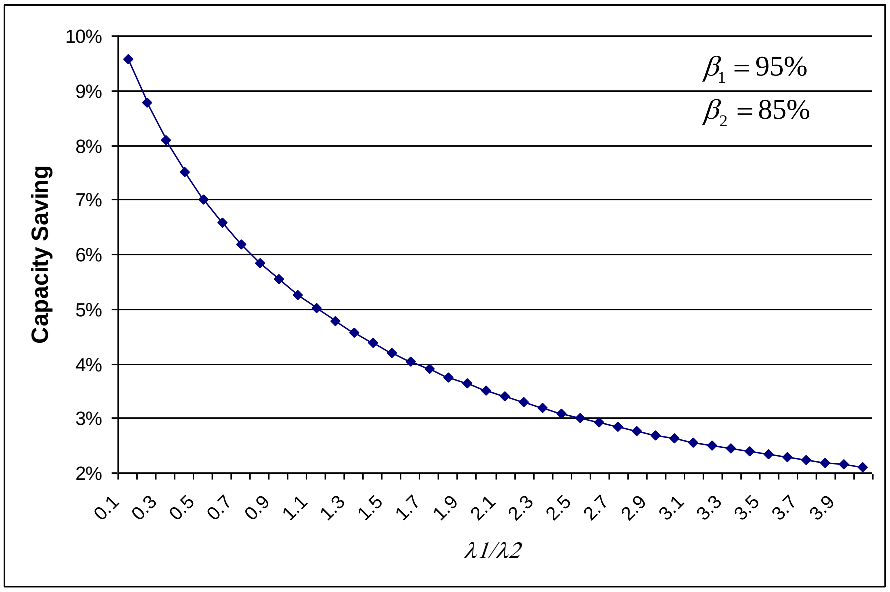


Figure 1: Capacity Saving vs. Demand Rate Ratio

Figure 2 is capacity saving changing with the required service level ratio. We fix the demand rate ratio at 1.5. It shows clearly that when the two required service levels differ from each other, the capacity saving effect exists. The capacity saving is proportional with the gap between the two required service level. However, the slopes of the curve on the two sides of the equal service level point ($\beta_1 = \beta_2$) are different. When $\beta_1 < \beta_2$, the FCFS policy will maintain the system service level at β_2, i.e. the FCFS capacity requirement is fixed for all $\beta_1 < \beta_2$. When $\beta_1 > \beta_2$, FCFS policy will meet β_1 instead. At this time, the required capacity of FCFS policy will increase with β_1, this makes the capacity saving not as salient as the case when $\beta_1 < \beta_2$.

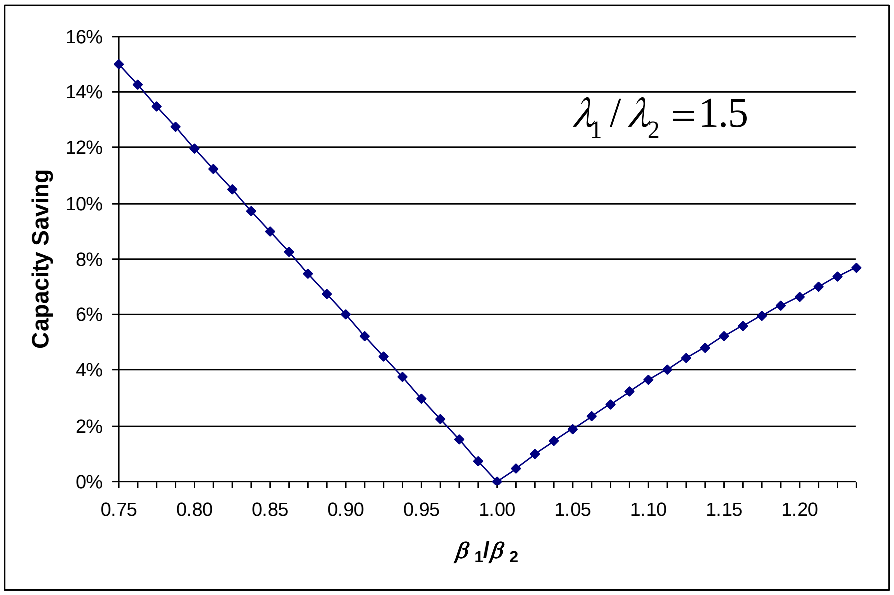
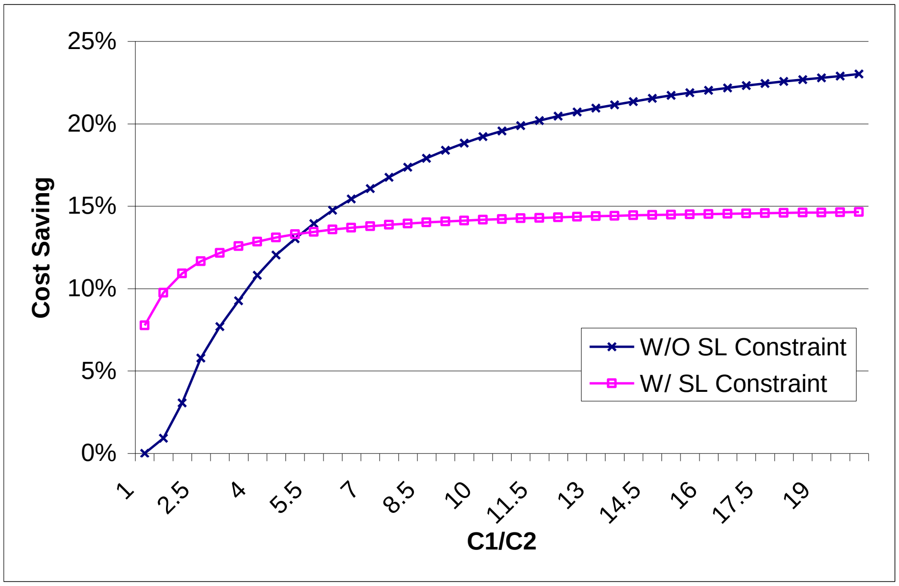


Figure 2: Capacity Saving vs. Required Service Level Ratio

3.2 Cost Comparison

Figure 3 shows the cost saving changing with the lost sales costs ratio. Without service level constraint, the cost saving of rationing policy is completely driven by the difference between lost sales costs of each demand class. If there is no difference between the two lost sale costs, then there will not be any cost saving. As the lost sales cost ratio increases, the cost saving grows rapidly. The cost saving is generated by allocating more inventory to meet the high cost demand while rejecting more low cost demand. This sacrifices the service level of the low priority demand to avoid the high priority demand stock-out. The higher the lost sale cost ration is, the more significant the cost saving is.



$$h' = \frac{h}{\lambda_1 c_1 + \lambda_2 c_2} = 0.02$$

$$\lambda_1 / \lambda_2 = 3$$

$$\rho = 0.98$$

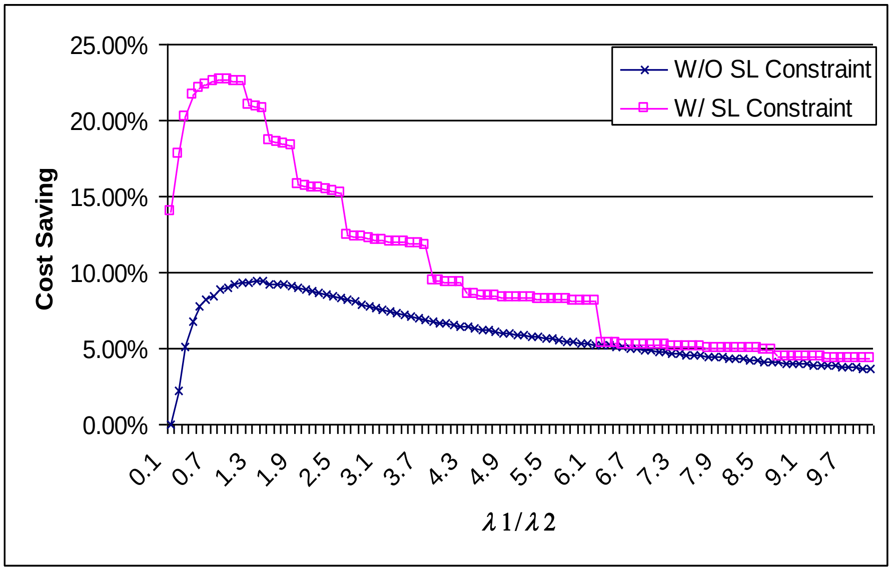
$$\beta_1 = 95\%$$

$$\beta_2 = 85\%$$

Figure 3: Cost Saving vs. Lost Sales Costs Ratio

However, once service level constraint is applied, the rationing policy works differently. The whole cost saving trend is similar, but rationing policy behaves differently when the lost sale cost ratio is close to 1. Even if the lost sales costs of the two demand classes are the same, the rationing policy still can save cost over the FCFS policy. The reason is that FCFS policy creates the same service level for both demand classes, which is the higher service level between the two. The rationing policy can provide different service levels to different demands, thereby reducing the total inventory level. When the lost sales costs ratio increases, the service level constraint will bound how the rationing policy can skew the allocation between the two demands, while no service level constraint will allow the rationing policy to completely ignore the low priority demand. Thus the cost saving is not as significant as the case without service level constraint at high lost sale costs ratio. In the figure, we can see that compared with the no service level constraint case, the cost saving of rationing policy with service level constraint is significant when the lost sale cost ratio is close to 1, and hits the upper limit more quickly as the lost sale cost ratio increases.

An extremely low or extremely high demand arrival rate ratio means that one demand dominates the other one. Rationing policy won’t be able to generate cost saving in these extreme cases because they effectively reduce the system to a single demand setting. As Figure 4 shows, both cost saving curves are a unimodal shape. Especially when the demand arrival rate ratio goes to infinity, the cost saving of both cases will eventually go to zero.

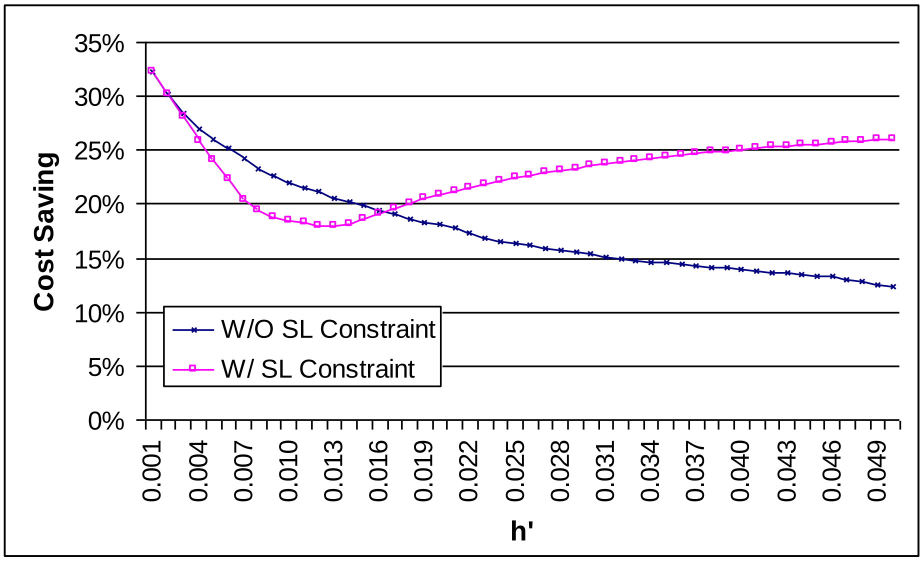


$$h' = \frac{h}{\lambda_1 c_1 + \lambda_2 c_2} = 0.02$$
$$c_1 / c_2 = 3$$
$$\rho = 0.98$$
$$\beta_1 = 95\%$$
$$\beta_2 = 85\%$$

Figure 4: Cost Saving vs. Demand Arrival Rate Ratio

However, at the low demand arrival rate ratio end, the cost saving of the two cases has a huge difference. For the no service level constraint scenario, the low demand arrival rate ratio means that the volume of higher priority demands is very limited compared with the volume of low priority demands. Thus even by rationing the inventory, there is not much to save. But once the service level constraint is applied, the cost saving will increase significantly. This is because the high priority demand has the higher service level requirement. For the FCFS policy, it has to maintain this high service level for both demands, even though only a tiny portion of the whole demand requires this high service level. The extra service level applied to low priority demand is costly, and that is why the rationing policy can generate huge cost saving in the service level constrained case. If we check figure 3, we can see that when the lost sale cost ratio is 3 (the setting of the experiment shown in Figure 4), the service level constrained case has better cost saving. This explains why the positions of the cost saving curves of the two cases in Figure 4.

When the relative holding cost rate is low, the major cost happens when the system experiences lost sales. Without surprise, the rationing policy outperforms the FCFS policy when the lost sale cost is the significant. This is shown in the left part of Figure 5. As the holding cost rate increases, holding cost gradually dominates the lost sales cost. If there is no service level constraint, both rationing policy and FCFS policy will start to focus on reducing the inventory holding cost and ignore the lost sale cost when this happens. Eventually, the two policies will converge to each other. This explains the continuously dropping of no service level constraint cost saving curve in Figure 5. The analysis is true for the service-level constrained case when holding cost is low. But the story is different for high holding costs. The service level constraint limits how much a policy can do to ignore the lost sale cost, i.e. only a specified portion of demands can be rejected. When this constraint exists, the inventory level can only be reduced to a certain level no matter how high the holding cost is. However, the rationing policy can use a much lower inventory level to meet the service level requirements, compared with FCFS policy. So as the relative holding cost rate increases, the cost saving of service level constrained case increases as well. This is different from the no service level constrained cased, as shown in Figure 5.



$\lambda_1 / \lambda_2 = 3$

$c_1 / c_2 = 6$

$\rho = 0.98$

$\beta_1 = 95\%$

$\beta_2 = 85\%$

Figure 5: Cost Saving vs. Relative Holding Cost Rate

4. CONCLUTION

In summary, this paper considers the stock rationing problem of a production system with two classes of demand and service level constraints. The condition establishes the existence of a feasible solution is derived and the optimal policy is characterized. The benefits of the rationing policy over FCFS policy can be classified into two categories. One is that the rationing policy can satisfy the system requirements with lower system capacity. The other one is the operation cost saving. These benefits differ significantly under different system parameters. When a service level constraint is required, one more factor is introduced into the system. In some cases it will amplify the saving, while in other cases, it may put a limit on the cost saving.

There are a few extensions we can do as future research. The straightforward one is to extend the conclusions we have drawn to more than two demand classes. Another possible extension is to consider the backorder case. However, we need be careful when we define the service level in the backorder case. The portion of demand satisfied immediately when they arrive may not be the best reflection of the system service level. Other chances exist in more complicated queueing and non-queueing models of production.

(References available upon request from Feng Tian at ftian@govst.edu.)