Production Control and Stock Rationing for a Make-to-Stock System with Parallel Production Channels

Önder Bulut[a], Mehmet Murat Fadıloğlu[*b]

[a]*Department of Industrial Engineering, Yaşar University, Izmir, Turkey*

onder.bulut@yasar.edu.tr

[b]*Department of Industrial Systems Engineering, Izmir University of Economics, Izmir, Turkey*

murat.fadiloglu@ieu.edu.tr

September, 2010

This paper considers the problem of production control and stock rationing in a make-to-stock production system with lost sales, multiple servers –parallel production channels--, and several customer classes. We assume independent stationary Poisson demand streams and exponential service times. At decision epochs, in conjunction with the stock allocation decision, the control specifies whether to increase the number of operational servers or not. Previously placed production orders cannot be cancelled. We model the system as an *M /M /s* make-to-stock queue and characterize properties of the optimal cost function, and of the optimal production and rationing policies. We show that the optimal production policy is a state-dependent base-stock policy, and the optimal rationing policy is of threshold type. Furthermore, we prove that the rationing levels are nonincreasing in the number of operational channels. We also show that the optimal ordering policy transforms into a bang-bang type policy when we relax the model by allowing order cancellations. Another model with partial order-cancellation flexibility is provided to fill the gap between the no-flexibility and the full-flexibility models. We quantify the additional gain that the optimal policy provides over the –suboptimal— base-stock policy proposed in the literature, along with the value of the flexibility to cancel production orders.

Keywords: Inventory/Production; Rationing; Make-to-Stock; Multiple Servers; Optimal Control.

[*] Corresponding author

1. Introduction

In this paper, we study the problem of production control and stock rationing of a single-item, make-to-stock facility with parallel production channels, several demand classes and lost sales. In a production system that keeps inventory to satisfy random demand originating from distinct customer classes, the decision maker should develop a strategy in order to efficiently use system resources and allocate inventory among different customer classes. However, in the most general setting, characterization of the optimal strategy is not analytically tractable because the decision maker should continuously adjust production and stock allocation decisions based on the current status of the production (the age information for all the outstanding production orders) and the current inventory level.

The production control issue is relatively simple when there is only a single production channel. It is optimal to produce up to a certain inventory level and then stop the production (see Ha (1997a) and Ha (1997b)). In the manufacturing systems literature, this policy is known as the production authorization mechanism (see Buzacott and Shantikumar (1993, pp 103)). A natural extension of the single production channel is parallel production channels. Buzacott and Shantikumar (1993, pp. 43) call systems with parallel production channels "single-stage systems." In these systems, "a job can be processed by any one of the machines, but only one machine is required to complete the required tasks." Zipkin (2000, pp 244) calls the same kind of systems "parallel processing systems" and provides an analysis for such systems with independent, stochastic leadtimes under the base-stock policy. The base-stock policy (in Zipkin (2000) as well as in this manuscript) is characterized by a constant order-up-to level, i.e., production target for our setting that is independent of the system state. But the base-stock policy is not optimal for this setting. Identifying optimal production policies for "parallel processing systems" and quantifying the optimality gap left by the base-stock policy are among the main issues addressed in this paper.

Another issue addressed in the paper is the problem of allocating a common stock pool among different customer classes, which is known as the stock rationing problem in the literature. It allows differentiating customer classes in order to provide different service levels and to operate the system more cost-effectively. The stock rationing policy stops serving lower priority classes when the on-hand inventory drops below a certain threshold level. Under the threshold level, only the demands from higher priority classes are satisfied. There is a threshold rationing

level for each customer class. The threshold levels could change dynamically according to the status of the production process.

The stock rationing problem frequently arises in many real life systems. A good example would be the spare parts inventory systems in which spare parts are demanded in order to repair different end products of different importance and criticality. Spare part inventory systems may experience urgent orders in case of breakdowns; and the shortage cost of such orders can be dramatically higher compared to the shortage cost of regular orders due to planned maintenance activities. Another example would be a two-echelon inventory system consisting of a warehouse and many retailers. In case of stock-outs, retailers may place urgent orders to the warehouse. Furthermore, it may be desirable to better serve certain retailers that constitute a large portion of the warehouse's business.

In this paper, we characterize structural properties of the optimal cost function, and the optimal production and stock rationing policies for a single-item, multi-server make-to-stock production system with multiple-customer-classes and lost sales. We assume that each customer class generates demand according to a stationary Poisson process independent of the other classes, and the servers – parallel production channels-- have independent exponential processing times with identical rates. In effect, we model the production system as an *M /M /s* make-to-stock queue. In the optimization model, the control determines whether to increase the number of active (operational) servers or not in conjunction with the rationing decision. Cancelling previously placed production orders is not allowed.

Another analysis for multiple replenishment channels with stochastic leadtimes is presented in Zipkin (2000) for the *M /G /s* system with a single customer class and lost sales. Zipkin provides performance analysis for the system under base-stock policy. In this paper, while characterizing the optimal policy, we show that base-stock policy is not optimal.

In order to better understand the problem, let us consider the following example. We have a company that produces spare parts for a large car manufacturer. There are two types of demand for the parts. The first one is the demand from the car manufacturer and the second one is the demand from different spare part distributors. We are obliged to provide the parts demanded by the manufacturer instantaneously or pay a hefty fine due to our contract. We do not have such an obligation with the distributors, although the sale is lost. Hence the manufacturer's demand has higher priority. Our production facility has *s* parallel channels such that each can process one

part at a time. Since the production times exhibit significant variability, we model them with exponential distribution. Given the state of the system, we would like to determine how many production channels to utilize and which types of demand to satisfy so as to operate the system optimally with respect to a predetermined cost function.

Veinott (1965) is the first to study the rationing problem. He considers a zero leadtime backordering model in the periodic review setting with exogenous supply. For both lost sales and backordering cases, Topkis (1968) shows that a time remembering rationing policy is optimal. Nahmias and Demmy (1981) derive approximate expressions for the expected number of backorders for each customer class under both periodic and continuous review. Their study is the first in the literature that considers the stock rationing problem in continuous time. They assume (Q, r) replenishment policy and static rationing levels, and provide an approximate fill rate expression for each class under the at-most-one-order-outstanding assumption.

Deshpande et al. (2003), Arslan et al. (2007) and Fadiloglu and Bulut (2007) study the problem in the same setting that Nahmias and Demmy (1981) consider. Deshpande et al. (2003) and Arslan et al. (2007) derive approximate results that are exact under the backorder clearing mechanism introduced by Deshpande et al. (2003). Their analyses yield the same results with Dekker et al. (1998) under the lot-for-lot policy. Fadiloglu and Bulut (2007) analyze the continuous-time system by constructing an embedded discrete-time Markov chain.

Dekker et al. (2002) consider the lot-per-lot replenishment policy with static rationing levels in a lost sales environment, and derive the exact service levels for all customer classes under general stochastic leadtimes. Melchiors et al. (2000) also analyze the lost sales case with static rationing levels. They assume (Q, r) replenishment policy, deterministic leadtimes, and at most one outstanding order.

There are a few studies in the literature that consider dynamic rationing policies for continuous review systems with deterministic exogenous leadtimes. For the lost sales case Melchiors (2003), and for the backordering case Teunter and Haneveld (2008) analyze time remembering rationing policies under which rationing levels are set according to the age of the outstanding order. Both studies assume at most one outstanding order. Fadiloglu and Bulut (2010) propose a dynamic rationing policy that alters the rationing levels in time, depending on the number and ages of all outstanding orders. The findings of this paper support the policy suggested in Fadiloglu and Bulut (2008). There are other works in the literature that consider settings in

which there are other sources of information such as advanced demand and assembly component inventory levels (see Benjaafar and ElHafsi (2006), Iravani et al. (2007), Gayon et al. (2009a)). For all these settings, the optimal rationing policies are state-dependent.

Ha (1997a) is the first to analyze the stock rationing problem in a production environment with capacitated replenishment channel. For a make-to-stock system with a single exponential server, no setup cost and lost sales, he shows that a base-stock policy is optimal for production control and a static threshold level policy is optimal for stock rationing. Ha (1997b) analyzes the same problem in a backordering environment with two customer classes. He shows that the rationing level is decreasing in the number of backorders of the non-critical class. Vericourt et al. (2002) extends the work of Ha (1997b) to multiple demand classes. Ha (2000) and Gayon et al. (2009b) consider Erlangian production times in the lost sales and backordering environments, respectively. They show that optimal policy is a threshold work storage level policy where work storage level is the number of completed Erlang stages. Huang and Iravani (2008) extends the findings of Ha (1997a) to batch demand.

There is also a vast literature that considers queueing control problems, which involve mechanisms such as admission control, capacity control and pricing. Queueing control problems find application in the areas of service, telecommunication, and make-to-order manufacturing systems. We direct the reader to the recent works of Cil et al. (2009) and Gans and Savin (2007) for the related literature. We would like to point out that multiple-exponential-server models are also used for service systems as in Gans and Savin (2007).

All the above mentioned studies for production environments assume a single server replenishment channel. Our work is most related to the work of Ha (1997a). We extend his study to the multi-server case using a two-dimensional state space. Along with a characterization of the optimal rationing policy, we also provide properties of the optimal production control policy and show that the optimal policy is not a base-stock policy for the multiple-servers case. Elhafsi et al. (2008) also consider a similar problem in which inventory is replenished from multiple sources of supply. They model the supply sources as exponential servers with different service rates and consider a backordering environment for a single demand class.

Section 2.1 introduces our primary model and provides the dynamic programming formulation. In Section 2.2, we present a characterization of the optimal policy. Section 3.1 discusses a variation on the main model where cancellation of any previously placed production order is

permitted, while Section 3.2 involves another variation with partial order-cancellation flexibility. In Section 4, we provide the stationary analysis of the system under both base-stock and bang-bang policies. Section 5 is devoted to a numerical study in which we quantify the benefit of the optimal policy and assess the value of order-cancellation flexibility. Section 6 consists of a discussion on the control of multiple production and replenishment channels that positions our work within the existing literature. The paper concludes with Section 7. The proofs of the lemmas and theorems presented in this work are provided in the Appendix.

2. Primary Model

2.1 Model Formulation

Consider a single-item make-to-stock production system with s identical servers having exponential production times with mean $1/\mu$. Demand is generated by $n \geq 2$ customer classes according to independent Poisson processes with rates λ_i, $i \in \{1,2,...,n\}$. We suppose that any unmet demand is lost and a lost sales cost of c_i is incurred for each unit of class i demand that is lost. Without loss of generality, we assume that $c_1 \geq c_2 ... \geq c_n$. Let h be the inventory holding cost rate, p be the production cost rate and α be the discount rate.

The state of the system is defined with two variables. Let $X(t)$ be the inventory level at time t and $Y(t)$ be the number of the operational servers at the time of the last event occurrence prior to time t. $Y(t)$ can also be considered as the number of the outstanding replenishment orders at the time of last event occurrence prior to time t. This elaborate state definition is necessary to eliminate instantaneous state transitions at decision epochs, which causes problems at the application of the uniformization technique.

At any time point t, the control specifies whether to keep the number of active servers at the same level or to increase it. We denote the production decision at time t as $u_p(t)$ where $u_p(t) \in \{Y(t), Y(t)+1, ..., s\}$. When a class i demand arrives at time t, the control specifies whether to satisfy the demand or not. We denote the rationing decision for class i at time t as $u_{r_i}(t)$ such that $u_{r_i}(t) \in \{0,1\}$, $i \in \{1,2,...,n\}$. If $u_{r_i}(t) = 0$, an arriving i^{th} class demand is rejected, otherwise it is satisfied. The complete policy for our model can be represented as

$\left\{\left(u_p(t), u_{r_1}(t), ..., u_{r_n}(t)\right) | t \geq 0\right\}$. Since the model is Markovian, optimal policies are also Markovian. Thus, it is sufficient to consider the set of admissible Markovian policies, i.e., $u_p(t) = u_p\left(X(t), Y(t)\right)$ and $u_{r_i}(t) = u_{r_i}\left(X(t), Y(t)\right), i \in \{1, ..., n\}$.

Starting at state (*x, y*), under the policy π, the infinite horizon expected discounted system cost is

$$E^{\pi}_{(x,y)}\left[\int_0^{\infty} e^{-\alpha t}\left(h(X^{\pi}(t)) + p(Y^{\pi}(t))\right)dt + \sum_{i=1}^{n}\int_0^{\infty} e^{-\alpha t} c_i dN_i^{\pi}(t)\right] \tag{1}$$

where $N_i^{\pi}(t)$ be the number of class *i* customers who have been rejected up to time *t*, $i \in \{1, 2, ..., n\}$. Given a control policy π, the process $\left\{\left(X^{\pi}(t), Y^{\pi}(t)\right) | t \geq 0\right\}$ is a continuous time Markov chain where the transition rate at state (*x, y*) is $\nu_{(x,y)} = \sum_{i=1}^{n} \lambda_i u_{r_i} + u_p \mu$. Via the uniformization technique proposed by Lippman (1975), we obtain an equivalent discrete-time problem. Let us define the uniform rate $\nu = \sum_{i=1}^{n} \lambda_i + s\mu$. Without loss of generality, we rescale the time and assume that $\alpha + \nu = 1$. Then, the optimal cost-to-go function can be expressed as

$$J(x, y) = \min_{s \geq u \geq y}\left\{hx + pu + (s-u)\mu J(x, y) + u\mu J(x+1, u-1) + T_R(x, u)\right\} \tag{2}$$

where $T_R(x, y) = \sum_{i=1}^{n} T_{R_i}(x, y)$ and for $i \in \{1, 2, ..., n\}$,

$$T_{R_i}(x, y) = \begin{cases} \lambda_i \min\left\{J(x-1, y),\ c_i + J(x, y)\right\} & , x > 0 \\ \lambda_i\left(c_i + J(0, y)\right) & , x = 0 \end{cases} \tag{3}$$

In (2), the minimization operation corresponds to the production decision, i.e., deciding the number of operational servers when there are *x* units on hand and *y* servers are operational. The term $(s-u)\mu J(x, y)$ corresponds to the fictitious self-transitions due to uniformization, while the term $u\mu J(x+1, u-1)$ corresponds to production completion at one of the *u* active production channels. The minimization operator $T_{R_i}(x, y)$ corresponds to the rationing decision for class *i*. At the boundary, when there is no stock on-hand, all the arriving demands are lost.

For notational purposes, we provide equations (4), (5) and (6) below. In (5), $u^*(x,y)$ is defined as the optimal number of operational production channels for the given state (x,y). Equation (6) defines a base-stock level for each inventory level x.

$$f(x,y,u) = hx + pu + (s-u)\mu J(x,y) + u\mu J(x+1,\, u-1) + T_R(x,\, u) \tag{4}$$

$$u^*(x,y) = \arg\min_{s \geq u \geq y} f(x,y,u) \tag{5}$$

$$S_x = x + u^*(x,0) \tag{6}$$

We also define the following operators on a function $v(x,y)$:

$$\Delta^y v(x,y) = v(x,y+1) - v(x,y)$$

$$\Delta^x v(x,y) = v(x+1,y) - v(x,y)$$

2.2 Characterization of the Optimal Production and Rationing Policies

This section provides a detailed characterization of the optimal production and rationing policies in three theorems and a corollary. The three theorems are proven via the methodology formalized by Porteus (1982). This approach is based on identifying a set of structural properties and then showing that these properties are preserved under the optimization operator. For the *M/M/s* model described in the previous section, the optimization operator is

$$T(J(x,y)) = \min_{s \geq u \geq y} f(x,y,u). \tag{7}$$

We define $\mathcal{G}$ as a set of functions on the integers such that if $v \in \mathcal{G}$, then

$$\Delta^x v(x,y+1) \geq \Delta^x v(x,y) \tag{8}$$

$$\Delta^x v(x,y) \geq \Delta^x v(x-1,y+1) \tag{9}$$

Note that, (8) can also be written as $\Delta^y v(x+1,y) \geq \Delta^y v(x,y)$.

In Theorem 1 and its corollary, we characterize the behavior of the optimal cost function and the optimal production policy with respect to the number of operational servers.

Theorem 1. If $J \in \mathcal{G}$, for given inventory level x,

i. $J(x,0) = \ldots = J\left(x, u^*(x,0)\right)$

ii. For $y \geq u^*(x,0)$, $J(x,y)$ is a convex-increasing function of y. That is,

$\Delta^y J(x, y+1) \geq \Delta^y J(x,y) > 0$.

Corollary 1. For given state (x, y),

i. $u^*(x,y) = \begin{cases} u^*(x,0), & y \leq u^*(x,0) \\ y, & y > u^*(x,0) \end{cases}$

ii. $u^*(x,y) = \min\left\{y' : J(x,y'+1) - J(x,y') > 0,\ s \geq y' \geq y\right\}$

iii. $u^*(x,y+1) = \begin{cases} u^*(x,y), & u^*(x,y) \geq y+1 \\ y+1, & u^*(x,y) = y \end{cases}$

Theorem 1 implies that the optimal cost function is constant with respect to the number of operational servers in the region where the number of operational servers is less than or equal to the optimal number of operational servers at state $(x, 0)$. On the other hand, in the complementary region, the optimal cost function is convex-increasing in the number of operational servers.

The first part of Corollary 1 indicates that if the current inventory position, $x + y$, is less than the base-stock level S_x, then it is optimal to increase the number of operational servers to $u^*(x,0)$ in order to raise the inventory position to the base-stock level. Otherwise, it is optimal not to change the number of operational production channels. It is optimal to set the number of operational servers to $u^*(x,0)$ when it is possible, i.e., the number of currently operational servers is less or equal to $u^*(x,0)$. Hence, the optimal costs for all the states in which $u^*(x,0)$ are feasible are the same as stated in the first part of Theorem 1.

As the numerical study in the next section illustrates, $u^*(x,0) = u^*(x+1,0)+1$ does not hold in general. Consequently, a single order-up-to level that is independent of the inventory position is not optimal and the optimal production policy is a state-dependent base-stock policy. As stated in the literature (Erhardt, 1984), the optimality of a simple base-stock policy cannot be guaranteed when replenishment orders cross in time. In our model, order crossing is possible

due to parallel production channels. In the single server case, order crossing does not occur since stock-units are produced one by one. In this case, base-stock policy is optimal as shown in Ha(1997a).

The second part of Corollary 1 provides an alternative definition for the optimal number of operational servers. It is optimal to increase the number of operational servers until the optimal cost function starts to increase. Finally, the last part of the corollary exhibits how the optimal number of operational channels changes with the number of currently operational servers. The optimal number of servers at state $(x, y+1)$ is either equal to the optimal number of servers at state (x, y) or one more.

In Theorem 2, we characterize the behavior of the optimal cost function, and the optimal production and rationing policies with respect to the inventory level. We also characterize the effect of the number of operational servers on the optimal rationing policy.

Theorem 2. If $J \in \mathcal{J}$, then

- **i.** $J(x, y)$ is x-convex.
- **ii.** $u^*(x+1, y) \le u^*(x, y)$.
- **iii.** $\Delta^x J(x-1, y) = J(x, y) - J(x-1, y) \ge -c_1$, and so $T_{R_1}(x, y) = \lambda_1 J(x-1, y)$. That is, it is always optimal to satisfy a class 1 demand when there is stock on hand.
- **iv.** There exists a threshold inventory level $K_x^i(y)$ for class $i \ge 2$, which is a function of operational servers, y, such that it is optimal to satisfy a class i demand above $K_x^i(y)$ and reject it otherwise. Moreover, $K_x^n(y) \ge K_x^{n-1}(y) \ge \ldots \ge K_x^2(y) \ge 0$, and $K_x^i(y+1) \le K_x^i(y)$ for $i \in \{2, \ldots, n\}$.
- **v.** There exists a threshold number of operational servers $K_y^i(x)$ for class $i \ge 2$ as a function of inventory level such that it is optimal to satisfy a class i demand above $K_y^i(x)$ and reject it otherwise. Moreover, $K_y^n(x) \ge K_y^{n-1}(x) \ge \ldots \ge K_y^2(x) \ge 0$, and $K_y^i(x+1) < K_y^i(x)$ for $i \in \{2, \ldots, n\}$.

Theorem 2 states that the optimal cost function is *x*-convex and the optimal number of production channels that should be used is non-increasing in the inventory level. The last three parts of the theorem characterize the optimal rationing policy, which is of threshold type. If there is stock on hand, it is always optimal to satisfy an arriving class 1 demand independent of the observed state. For each of the other customer classes, given the number of the operational servers, there exists a rationing inventory level, which is non-decreasing in the class index. Similarly, for each class, given the inventory level, there exists a rationing level for the number of operational servers, which is non-decreasing in the class index. Furthermore, the rationing inventory levels are non-increasing in the number of operational channels, and the rationing levels for the number of operational production channels are decreasing in the inventory level. The latter statement means that if it is optimal to satisfy an arriving class *i* demand at state $(x, y+1)$, then it is optimal to satisfy an arriving class *i* demand at state $(x+1, y)$. Moreover, a class *i* demand arriving at state $\left(x, K_y^i(x)\right)$ should be rejected, but it is optimal to satisfy an arriving class *i* demand at state $\left(x+1, K_y^i(x)\right)$.

Theorem 3 states that $J \in \mathcal{G}$, that is, the optimal cost function is an element of the function space characterized by (8) and (9). Since $J \in \mathcal{G}$ is the hypothesis of the previous two theorems, Theorem 3 is needed to ensure that the results of the previous two theorems apply to our model without any restriction.

Theorem 3. $J \in \mathcal{G}$, that is

i. $\Delta^x J\left(x, y+1\right) \geq \Delta^x J\left(x, y\right)$

ii. $\Delta^x J\left(x, y\right) \geq \Delta^x J\left(x-1, y+1\right)$

3. Variations on the Primary Model

3.1 Model with Full Order-Cancellation Flexibility

In this section, we consider a variation on the previous model in which cancellation of all previously placed production orders is permitted. The rational behind this model is twofold. Firstly, this model enables us to characterize the optimal policy for make-to-stock queues where the outstanding orders can be cancelled at a negligible cost. Secondly, this model permits us to quantify

the value of the full flexibility to cancel orders. The difference between the performances of the primary model and this model is the value of the full order-cancellation flexibility. In many systems, order cancellations are only possible at a cost. In such cases the cost of canceling orders should be compared with the value of order-cancellation flexibility.

Given that order cancellation is possible, at each decision epoch the number of operational servers can be chosen from the set $\{0,1,...,s\}$. As previously discussed, for the primary model the feasible set is $\{Y(t), Y(t)+1,...,s\}$ where $Y(t)$ is the number of operational servers at the time of the last event occurrence prior to time t. Therefore, for the model with order cancellation, there is no need to keep track of the number of operational servers and it is possible to model the system evolution with a single state variable, which is the inventory level.

For this model, the optimal cost-to-go function of this model can be expressed as

$$J(x) = \min_{s \geq u \geq 0} \{hx + pu + (s-u)\mu J(x) + u\mu J(x+1)\} + T_R(x) \tag{10}$$

where $T_R(x) = \sum_{i=1}^{n} T_{R_i}(x)$ and for $i \in \{1,2,...,n\}$,

$$T_{R_i}(x) = \begin{cases} \lambda_i \min(J(x-1), c_i + J(x)) & , x > 0 \\ \lambda_i (c_i + J(0)) & , x = 0 \end{cases} \tag{11}$$

Let $u^*(x)$ be the optimal number of operational production channels at state x. Then,

$$u^*(x) = \arg\min_{s \geq u \geq 0} \{u(p + \mu \Delta J(x))\} \tag{12}$$

It should also be noted that for $s = 1$ the model is the same with the one analyzed in Ha (1997a). Thus, the below theorem that provides the properties of the cost function and the optimal policy extends the results presented in Ha (1997a) to a multiple-servers setting.

Theorem 4.

- **i.** $u^*(x)$ is either s or 0
- **ii.** $J(x)$ is a convex function of x
- **iii.** There exists a threshold inventory level K^i for class i such that it is optimal to satisfy a class i demand above K^i and reject it otherwise.

iv. For $x > 0$, $\Delta J(x-1) \geq -c_1$, and so $T_{R_1}(x) = \lambda_1 J(x-1)$. That is, $K^1 = 0$ and it is always optimal to satisfy a class 1 demand when there is stock on hand.

v. Optimal production policy is a bang-bang type policy. That is, there exists a threshold inventory level $\bar{x}$ such that $u^*(x) = s$, for $x \in \{0,1,...,\bar{x}\}$ and $u^*(x) = 0$, for $x \in \{\bar{x}+1, \bar{x}+2, ...\}$.

Theorem 4 states that the optimal production policy when order cancellations are permitted is a bang-bang policy. Up to a certain inventory level the policy prescribes using all available production channels. Beyond that level all of the production channels are idled. A static stock-rationing policy is employed for stock allocation.

3.2 Model with Partial Order-Cancellation Flexibility

In this section, we consider a variation on the primary model in which cancellation of only a limited number of previously placed production orders is permitted. This model allows us to quantify how much value can be captured via order cancellation, when there is a limitation on the number of orders that can be cancelled. Although a manufacturer may desire to reduce its operating costs by cancelling some orders, it may not be willing to use this flexibility in an indiscriminate fashion. The optimal policy under full flexibility is a bang-bang policy that utilizes all available production channels until the inventory reaches a threshold level. When one of the orders is completed at that level, the rest of the orders have to be cancelled. For many manufacturers, this drastic cancellation practice would not be desirable. If it is possible to capture most of the value available via order cancellation while espousing a smoother production policy that involves fewer cancellations, this could be the avenue of choice for those manufacturers.

We define the cancellation flexibility index *f* as the maximum number of servers that can be shut down. At each decision epoch, the control selects the number of active servers from the set $\{(Y(t)-f)^+, ..., s\}$ where $Y(t)$ is the number of active servers at the time of the last event occurrence prior to *t*. This model has the versatility to cover both the primary model and the model with full-cancellation flexibility. The primary model and the model with full-cancellation flexibility can be obtained by setting $f = 0$, and $f = s$, respectively. The optimal cost-to-go func-

tion for this model is a straight-forward extension of the one for the primary model given in (2) and can be expressed as

$$J(x,y)=\min_{s\geq u\geq (y-f)^+}\left\{hx+pu+(s-u)\mu J(x,y)+u\mu J(x+1,\,u-1)+T_R(x,\,u)\right\}. \qquad (13)$$

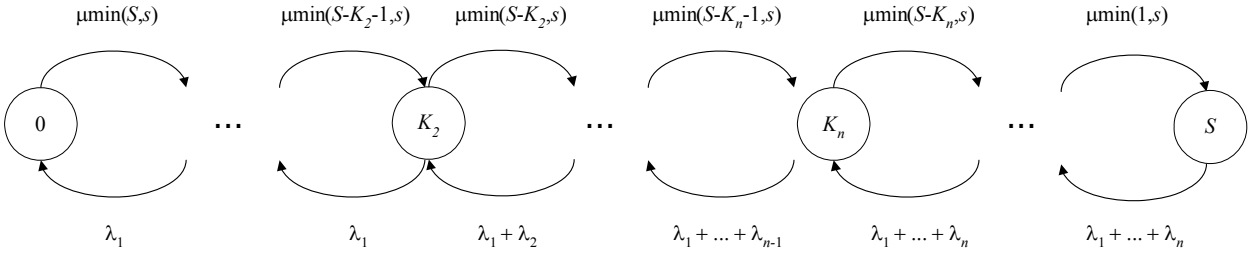
The optimal policy for the model with partial order-cancellation flexibility fully conforms to the characterization provided for the primary model.

4. Stationary Analysis

4.1 Stationary Analysis under Base-stock

In Section 2, we show that under discounted cost criterion, the optimal production and stock allocation policies are state-dependent. There is no fixed target inventory level and the rationing levels are dependent on the number of operational servers. Yet, if we operate under a simple base-stock policy with a fixed target level S, the number of operational servers is known for each inventory level, i.e., $\min\left\{(S-x)^+, s\right\}$. In this case, the stock allocation decision is solely determined by the inventory level. In this section, we provide the average cost of the system for this setting.

Figure 1 Birth-Death Process for $(S,\vec{K})$ policy



Let $(S,\vec{K})$ denote the base-stock policy with rationing whose fixed target inventory level is S and fixed threshold rationing levels are $\vec{K}=(K_1,K_2,...,K_n)$. Under $(S,\vec{K})$ policy the inven-

tory level, *x*, evolves according to a Birth-Death Process depicted in Figure 1. Births correspond to production completions whose rate is $\mu \min\left\{(S-x)^+, s\right\}$; whereas deaths correspond to inventory depletion by demand whose rate is $\sum_{i=1}^{m}\lambda_i$ if $K_m < x \le K_{m+1}$ for $m \in \{1,...,n\}$. Note that K_{n+1} is set to *S* for convenience of notation. The stationary probability that the inventory level is *j* where $K_m < j \le K_{m+1}$ for $m \in \{1,...,n\}$ can be expressed as

$$P_j = \frac{\mu^j \prod_{i=0}^{j-1} \min(S-i, s)}{\left(\sum_{i=1}^{m}\lambda_i\right)^{j-K_m} \prod_{r=1}^{m-1}\left(\sum_{i=1}^{r}\lambda_i\right)^{K_{r+1}-K_r}} P_0$$

where

$$P_0 = \left(1 + \sum_{j=1}^{S} \frac{\mu^j \prod_{i=0}^{j-1} \min(S-i, s)}{\left(\sum_{i=1}^{m}\lambda_i\right)^{j-K_m} \prod_{r=1}^{m-1}\left(\sum_{i=1}^{r}\lambda_i\right)^{K_{r+1}-K_r}}\right)^{-1}.$$

Hence, the expected inventory level, the fill rate for each class and the expected cost of the system are respectively

$$E[X] = \sum_{j=0}^{S} jP_j,$$

$$\beta_i = \sum_{j=K_i+1}^{S} P_j,\ 1 \le i \le n$$

$$C\left(S, \vec{K}\right) = hE[X] + \sum_{i=1}^{n} \lambda_i c_i (1-\beta_i).$$

In order to find the optimal policy parameters, for each fixed *S*, we can first find the optimal $\vec{K}$ vector that minimizes $C\left(S, \vec{K}\right)$ by performing an exhaustive search on each K_i (starting from $i = 2$) over $\{K_{i-1},...,S\}$ for $2 \le i \le n$. Note that Theorem 2 states it is always optimal to satisfy a class 1 demand when there is stock on hand, therefore K_1 is set to 0. Let us denote $\vec{K}^*(S)$ as the vector of optimal rationing levels for a fixed *S*. Starting from $S = 0$ we can search for the opti-

mal S, denoted as S^*, that minimizes $C\left(S,\vec{K}^*(S)\right)$. We suggest to perform an extensive search on S while keeping track of the first difference of the expected cost function, $C\left(S+1,\vec{K}(S+1)\right)-C\left(S,\vec{K}(S)\right)$, until $C\left(S,\vec{K}^*(S)\right)$ is sufficiently larger than the current minimum and the first difference continues to remain positive over a large range. Ha (1997a) also proposes a similar algorithm for the case where $s=1$ and shows that the cost function is not convex in general. The non-convexity result applies to our problem which is a generalization of Ha (1997a).

The discussion above outlines a general method for the analysis of the system under base-stock. We now consider a special case with infinitely many servers and denote the optimal base-stock for this special case as S^*_{inf}. This would provide us a bound on the number of servers beyond which the system is equivalent to a system with exogenous leadtimes, i.e., uncapacitated replenishment channel. If the number of available servers is greater or equal than S^*_{inf}, then the system never utilizes more than S^*_{inf} servers because it is the optimal base-stock level for the problem with no constraint on *s*. We can also find an upper bound on the value of S^*_{inf} by considering the $\left(S,\vec{0}\right)$ policy, i.e., the stock allocation is performed on FCFS basis. The optimal base-stock level for the $\left(S,\vec{0}\right)$ policy would obviously constitute an upper bound on S^*_{inf}, since we do not ration the inventory and the effective demand increases. By letting $\lambda=\sum_{i=1}^{n}\lambda_i$, the performances measures for an $\left(S,\vec{0}\right)$ policy can be expressed as

$$P_0=\left(\sum_{j=0}^{S}\frac{S!}{(S-j)!}\frac{\mu^j}{\lambda^j}\right)^{-1},\ \beta=1-P_0,\ E[X]=S-\frac{\lambda}{\mu}\beta$$

$$C\left(S,\vec{0}\right)=h\left(S-\frac{\lambda}{\mu}\beta\right)+(1-\beta)\sum_{i=1}^{n}\lambda_i c_i$$

Lemma 1. $C\left(S,\vec{0}\right)$ is a convex function of *S*.

Using Lemma 1, it is easy to find the optimal base-stock for the $\left(S,\vec{0}\right)$ policy, which is the smallest S that satisfies $C\left(S+1,\vec{0}\right)-C\left(S,\vec{0}\right)\geq 0$. It is also interesting that the cost function, which is not convex for finite number of servers, becomes convex when the number of servers tends to infinity. The reader may also refer to Jaarsveld and Dekker (2009) for a discussion on different algorithms proposed in the literature for finding $S^*_{\inf}$ and the corresponding optimal rationing levels. We also would like to point out that in this setting with ample servers the provided analysis is valid for general service times as well due to Palm's theorem.

4.2 Stationary Analysis under Bang-Bang Policy

In Section 3.1, we show that the optimal policy is of bang-bang type when order cancellations are allowed. In this section, we provide stationary analysis for the optimal bang-bang policy. Note that this policy is only possible when there is flexibility to cancel all outstanding orders.

The bang-bang policy is characterized with a threshold inventory level $\bar{x}$ below which all servers are utilized. Once this target level is reached, all the servers are shut down. Under this policy, the inventory level, x, corresponds to the number of customers in an $M/M/1/\bar{x}$ queue where the arrival rate (the rate of production completion) is $s\mu$ at levels and the departure rate (the rate of inventory depletion) is $\sum_{i=1}^{m}\lambda_i$ when $K_m < x \leq K_{m+1}$ for $m\in\{1,...,n\}$. Here, K_{n+1} is set to $\bar{x}$ for convenience of notation. The stationary probability of having j units of on-hand inventory where $K_m < j \leq K_{m+1}$ for $m\in\{1,...,n\}$ is

$$P_j = \frac{\left(s\mu\right)^j}{\left(\sum_{i=1}^{m}\lambda_i\right)^{j-K_m}\prod_{r=1}^{m-1}\left(\sum_{i=1}^{r}\lambda_i\right)^{K_{r+1}-K_r}}P_0$$

where

$$P_0 = \left(1+\sum_{j=1}^{\bar{x}}\frac{\left(s\mu\right)^j}{\left(\sum_{i=1}^{m}\lambda_i\right)^{j-K_m}\prod_{r=1}^{m-1}\left(\sum_{i=1}^{r}\lambda_i\right)^{K_{r+1}-K_r}}\right)^{-1}.$$

Hence, the expected inventory level, the fill rate for each class and the expected cost of the system are respectively

$$E[X] = \sum_{j=0}^{\bar{x}} jP_j ,$$

$$\beta_i = \sum_{j=K_i+1}^{\bar{x}} P_j , 1 \le i \le n$$

$$C(\bar{x}) = hE[X] + \sum_{i=1}^{n} \lambda_i c_i (1-\beta_i) .$$

In order to find the optimal policy parameters, the methodology described in Section 4.1 is directly applicable. Furthermore, the system under bang-bang policy is equivalent to the single server system of Ha (1997a) when $s\mu$ is set to be service rate. This is due to the fact that for single server systems base-stock and bang-bang policy are equivalent.

5. Numerical Study

In this section, we illustrate the results obtained in the previous sections, and quantify the impact of the system parameters on the performance of the optimal production and rationing policies. Under different scenarios, we compare the optimal production policies (for the primary model and its variations) with the base-stock policy, which is the one suggested in the literature for systems with a limited number of processing channels (see Zipkin (2000) pp. 261-263), and the optimal rationing policy with the first-come-first-served (FCFS) policy. We quantify the benefit of the optimal production and rationing policies as the percent cost reduction obtained by operating the system under the optimal policies instead of the base-stock and FCFS policies. We also present a graph illustrating the impact of the cancellation flexibility index on the performance of the optimal production policy. We obtain the numerical results presented in this section via a value iteration algorithm coded in MATLAB.

In order to illustrate the properties of the optimal policies, let us consider a two-class system with $(s, \mu, \lambda_1, \lambda_2, h, p, c_1, c_2, \alpha) = (15, 1, 5, 1, 1, 1, 4, 1, 0.6)$. Table 1 and Table 2 show the optimal production and rationing policies, respectively. In the tables, rows indicate the on-hand inventory level (0 to 4) and columns indicate the current number of operational servers (0 to 15). In

Table 1, the value corresponding to the state (*x, y*) is the optimal number of operational servers, $u^*(x,y)$, which is bounded below by *y* (the current number of operational servers) and above by *s* (the total number of available servers). Table 1 is in agreement with Corollary 1, which states, if $y \le u^*(x,0)$, then $u^*(x,y) = u^*(x,0)$, otherwise $u^*(x,y) = y$. Moreover, in parallel with Theorem 2, the optimal number of operational servers at state (*x,* 0) decreases by one or more units, for each unit increase in the on-hand inventory level, and then it remains constant at 0. It is also observed that the base-stock level for each inventory level, $S_x = x + u^*(x,0)$, varies with the inventory level.

x \ *y*	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
0	10	10	10	10	10	10	10	10	10	10	10	11	12	13	14	15
1	6	6	6	6	6	6	6	7	8	9	10	11	12	13	14	15
2	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
3	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
4	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

Table 1 Optimal Production Policy

x \ *y*	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
2	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
3	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Table 2 Optimal Rationing Policy For Class 2

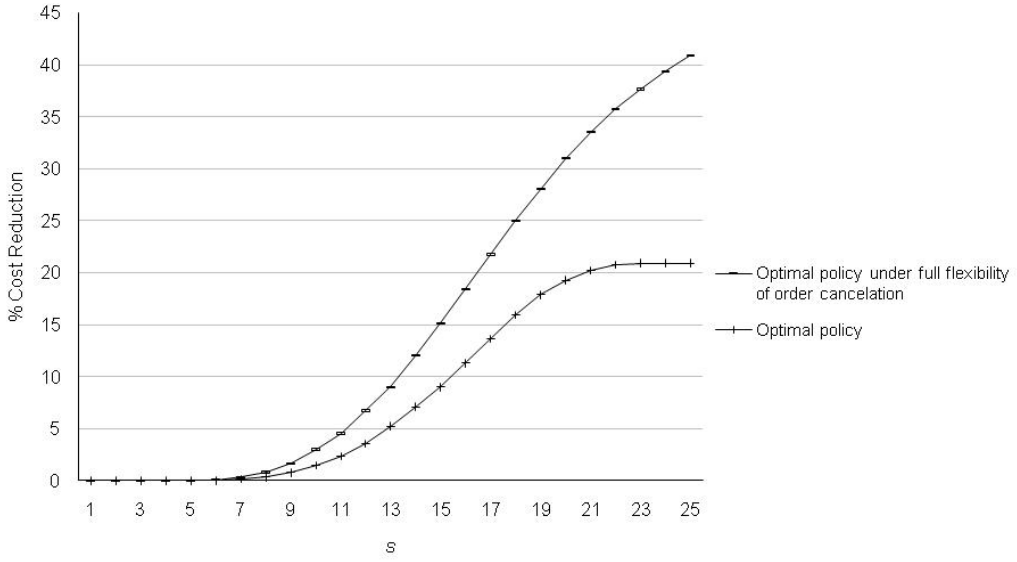
Theorem 2 states that it is always optimal to satisfy an arriving class 1 demand as long as there is inventory on-hand. Therefore, we only provide the optimal rationing policy for class 2 (Table 2). A zero in the cell corresponding to state (*x, y*) indicates that an arriving class 2 demand should be rejected at state (*x, y*), and a one indicates that the demand should be satisfied. As can be easily observed from the table, the threshold inventory rationing levels for class 2, which are

non-increasing in the number of operational servers, are: $K_x^2(0) = K_x^2(1) = 3$, $K_x^2(2) = ... = K_x^2(7) = 2$, $K_x^2(8) = ... = K_x^2(14) = 1$, and $K_x^2(15) = 0$. The rationing thresholds for the number of operational servers, which are decreasing in the on-hand inventory level until hitting -1 as shown in Theorem 2, are: $K_y^2(1) = 14, K_y^2(2) = 7, K_y^2(3) = 1$, and $K_y^2(x) = -1$ for $x \geq 4$.

Since the cost criteria in this study is the infinite-horizon discounted cost, the optimal cost depends on the initial state (x, y). Therefore, in the remaining part of this section we compare the costs of different policies for the initial state $(0,0)$.

In order to assess the value of the optimal production policy relative to the optimal base-stock policy, we suppress the effect of rationing and consider a setting with a single customer class by letting $(\lambda, \mu, h, p, c, \alpha) = (10, 2, 0.2, 0.2, 10, 0.6)$. In this setting, Figure 2 exhibits the effect of the number of available production channels *s*, on the cost reduction. As seen from the figure, the optimal policy does not provide any cost reduction for small values of *s*. This is due to the fact that when available processing channels are scarce, both of the optimal production and the optimal base-stock policies, try to use all of the limited capacity. For $s = 1$, Ha (1997a) already showed that the optimal policy is base-stock. However, for moderate values of *s*, the benefit of the optimal policy over the base-stock policy increases rapidly with *s*, because the optimal policy has flexibility to adjust the number of operational servers at each inventory level. In contrast, the base-stock policy dictates a fixed production target for all states. The percent cost reduction stabilizes after *s* exceeds the optimal number of operational processing channels at state $(0,0)$, which is 23 for the considered setting. When $s > 23$, the system is equivalent to a system with exogenous exponential leadtimes, i.e., uncapacitated replenishment channel, because it always operates with less than the available *s* processing channels and therefore additional channels do not provide further gain.

Figure 2 Optimal production policy vs. Base-stock policy: Impact of number of available servers

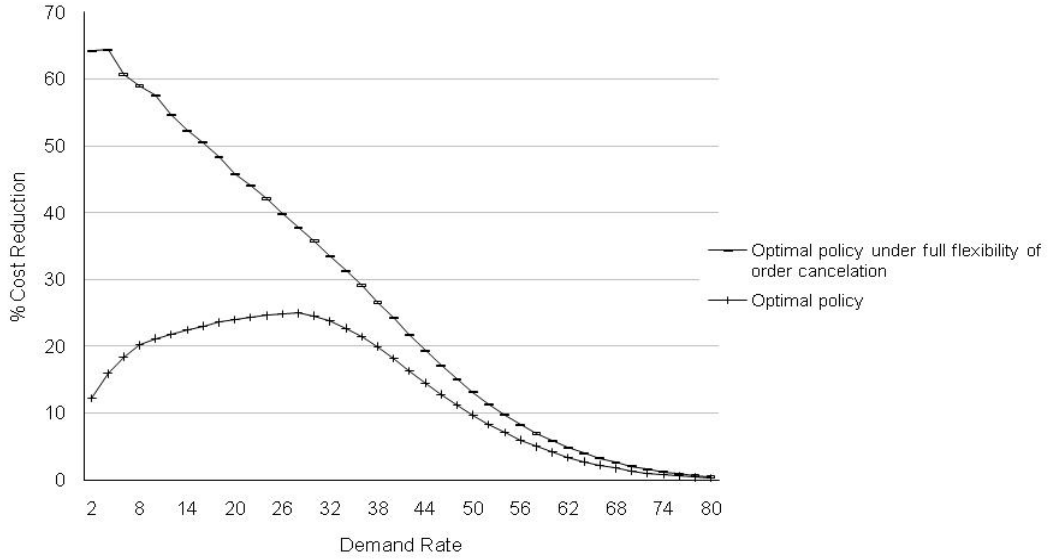


When we allow cancellation of previously placed production orders, the system operates under full flexibility. For small values of s, all optimization models (base-stock, primary, and models with order cancellation) try to use the whole capacity and, thereby, their performances are indistinguishable. For larger values of s, the savings obtained by using the optimal policy under order cancellation instead of using the best base-stock policy, or the optimal policy for the primary model, is positive and grows until hundred-percent with s. This is due to the fact that the optimal policy under order cancellation is a bang-bang type policy (as discussed in the previous section). As s tends to infinity, it is optimal to utilize all available servers at the time of each demand arrival in order to satisfy the arriving demand instantaneously (because the replenishment rate tends to infinity) and then to cancel all the other production orders after the first unit is produced.

In Figure 3, we use the same setting with Figure 2 with the exception that s is fixed to 46 and the demand rate is a variable. The figure illustrates the impact of traffic intensity on the cost reduction provided by the optimal policy. As the demand rate increases the benefit of the optimal policy over the best base-stock policy first increases and then decreases all the way down to zero. In parallel with the discussion related to Figure 2, the percent cost reduction increases until the optimal number of operational servers (at zero inventory level) hits s, which is observed when demand rate is 28. As demand rate increases beyond 28, the optimal policy is unable to open more servers at lower inventory levels, which would be needed to realize its full potential.

Therefore, when the demand rate is sufficiently large, both policies start to behave in a similar manner by utilizing all the available capacity at most of the inventory levels, and thereby the cost of the optimal policy converges to the cost of the base-stock policy. As discussed above, the optimal policy under the flexibility of the order cancellation, outperforms both the optimal policy of the primary model and the best base-stock policy for small to moderate demand rate values, and it also becomes identical with the other policies for high demand rates. For sufficiently small s values, one would not observe the region in which the cost reduction increases. Thus, for small values of s the optimal base-stock policy can be used as a good approximation of the optimal policy and this approximation performs better at high demand rates.

Figure 3 Optimal production policy vs. Base-stock policy: Impact of demand rate



Figures 2 and 3 illustrate the effect of the number of available servers and the arrival rate on cost reduction, respectively. Figure 4 investigates the joint effect of these two system parameters. For different s and traffic intensity values –traffic intensity is denoted by ρ and is equal to $\lambda/s\mu$–, the figure compares the optimal production policy with the best base-stock policy. The arrival rate and the number of servers are scaled up proportionally such that ρ remains constant while the number of servers increases. The figure exhibits the results for the case where μ is fixed to unity and $(h, p, c, \alpha) = (0.2, 0.2, 10, 0.6)$. For all ρ values, the percent cost reduction increases with s (and λ). That is, the benefit of having additional production channels (illus-

trated in Figure 2) is much more pronounced than the detriment of heavier traffic (illustrated in Figure 3) when traffic intensity is kept constant. The main advantage of the optimal policy over the base-stock policy is its flexibility in adjusting the number of operational servers at each inventory level and this flexibility increases with the number of available servers. The figure also illustrates that, for small values of *s*, the cost reduction is more significant at lower ρ values. Furthermore, there exists an *s* value for each traffic intensity beyond which the cost reduction at this intensify is higher than the cost reductions achieved at lower intensities. When the traffic intensity is kept constant, the number of servers used specifies at what proportions of the traffic intensity, the system provides service, ie., it effectively discretizes the control space. The higher the number of available servers, the finer is the discretization. This ability to adjust the instantaneous utilization more precisely is especially instrumental when the system capacity is tight.

Figure 4 Optimal production policy vs. Base -stock policy: Constant traffic intensity

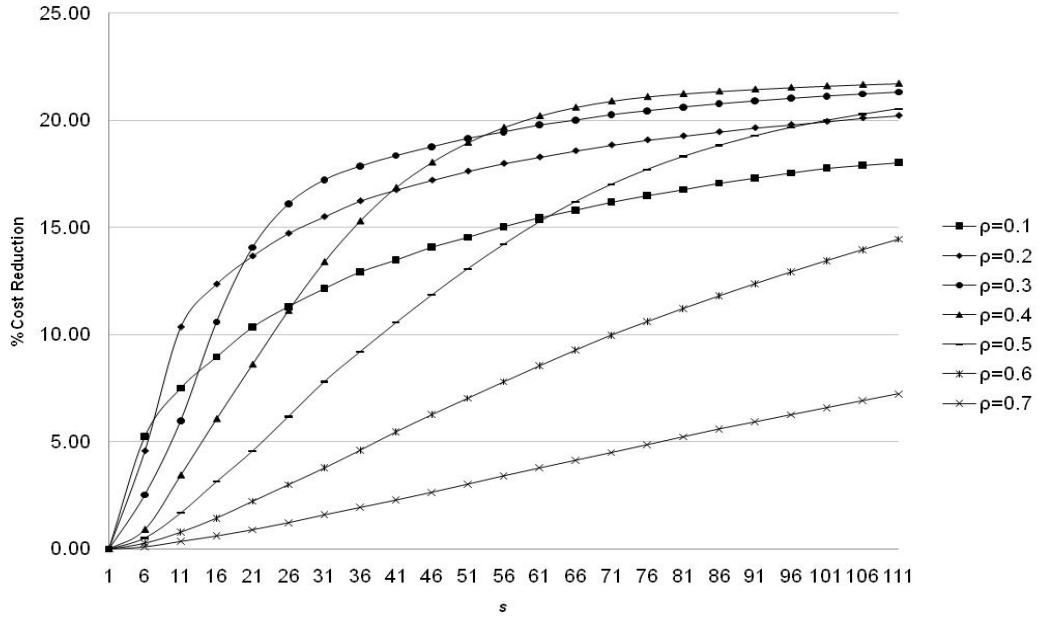
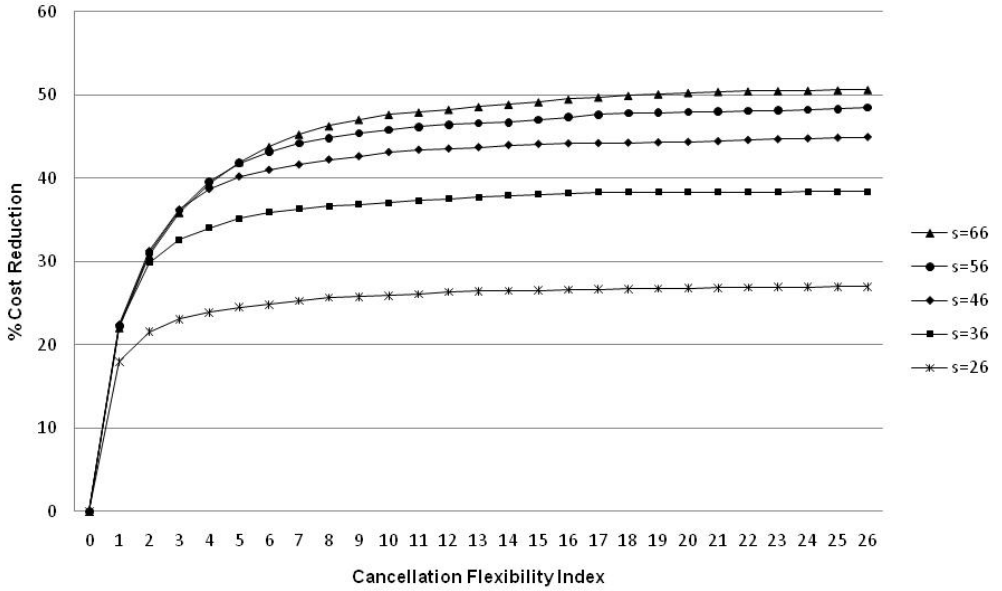


Figure 5 Value of Order Cancellation Flexibility



Figures 2 and 3 compare the performance of the optimal policies for the primary model ($f = 0$) and the model with full-cancellation flexibility ($f = s$) with the optimal base stock policy. Figure 5, which uses the same setting with Figure 2 with the exception that s is a variable, illustrates the impact of cancellation flexibility index. As stated in the discussion related to Figure 2, the benefit of order cancellation under full flexibility increases with the number of available servers. Figure 5 reveals that this is also true at any given cancellation flexibility index. It is obvious that more flexibility is better as manifested in the figure. Moreover, as the flexibility index increases, the rate of increase in the percent cost reduction obtained via order cancellation decreases sharply all the way down to zero. Thus, a little flexibility goes a long way and captures most of the value that can be realized via order cancellation. For $s = 26$, while a 27% cost reduction can be obtained with full order cancellation flexibility, having the flexibility of cancelling only one of the previously placed orders captures 67% of this potential gain. Moreover, at $f = 6$, 93% of the potential is captured. As the number of available servers increase, more flexibility is necessary to secure most of the potential gain. However, for all s values, having a little flexibility --compared to full flexibility ($f = s$)-- is sufficient to obtain a significant cost reduction as seen in the figure. As the flexibility index increases, the optimal policy becomes jitterier, i.e., it frequently shuts down operating servers. Due to this tradeoff, a manufacturer is likely to opt for

little flexibility that enables a significant reduction in operating costs while keeping the production relatively smooth.

Table 3 Optimal Rationing Policy versus FCFS policy: impact of demand mix

p_1 \ p_2	0.00	0.17	0.33	0.50	0.67	0.83	1.00
0.00	0.00	15.55	23.40	22.29	14.96	6.97	0.00
0.17	28.58	31.43	27.63	19.48	11.27	4.22	
0.33	37.38	31.60	22.98	14.85	8.05		
0.50	31.94	23.19	15.55	9.10			
0.67	19.82	12.97	7.16				
0.83	8.81	3.70					
1.00	0.00						

$s = 6$

p_1 \ p_2	0.00	0.17	0.33	0.50	0.67	0.83	1.00
0.00	0.00	3.99	10.27	13.30	13.22	9.61	0.00
0.17	13.06	15.91	17.96	16.60	11.45	0.00	
0.33	21.70	21.68	19.36	12.97	0.79		
0.50	25.10	21.68	15.04	3.31			
0.67	23.96	16.67	3.73				
0.83	16.83	2.84					
1.00	0.00						

$s = 16$

p_1 \ p_2	0.00	0.17	0.33	0.50	0.67	0.83	1.00
0.00	0.00	0.01	0.00	1.36	2.09	1.54	0.00
0.17	0.00	1.82	3.15	3.10	1.97	0.00	
0.33	4.23	4.68	3.97	2.33	0.00		
0.50	6.00	4.72	2.83	0.00			
0.67	5.37	3.35	0.00				
0.83	3.81	0.00					
1.00	0.00						

$s = 26$

p_1 \ p_2	0.00	0.17	0.33	0.50	0.67	0.83	1.00
0.00	0.00	0.00	0.00	0.00	0.00	0.25	0.00
0.17	0.00	0.00	0.00	0.52	0.55	0.00	
0.33	0.00	0.77	1.08	0.80	0.00		
0.50	1.64	1.61	1.03	0.00			
0.67	2.07	1.23	0.00				
0.83	1.44	0.00					
1.00	0.00						

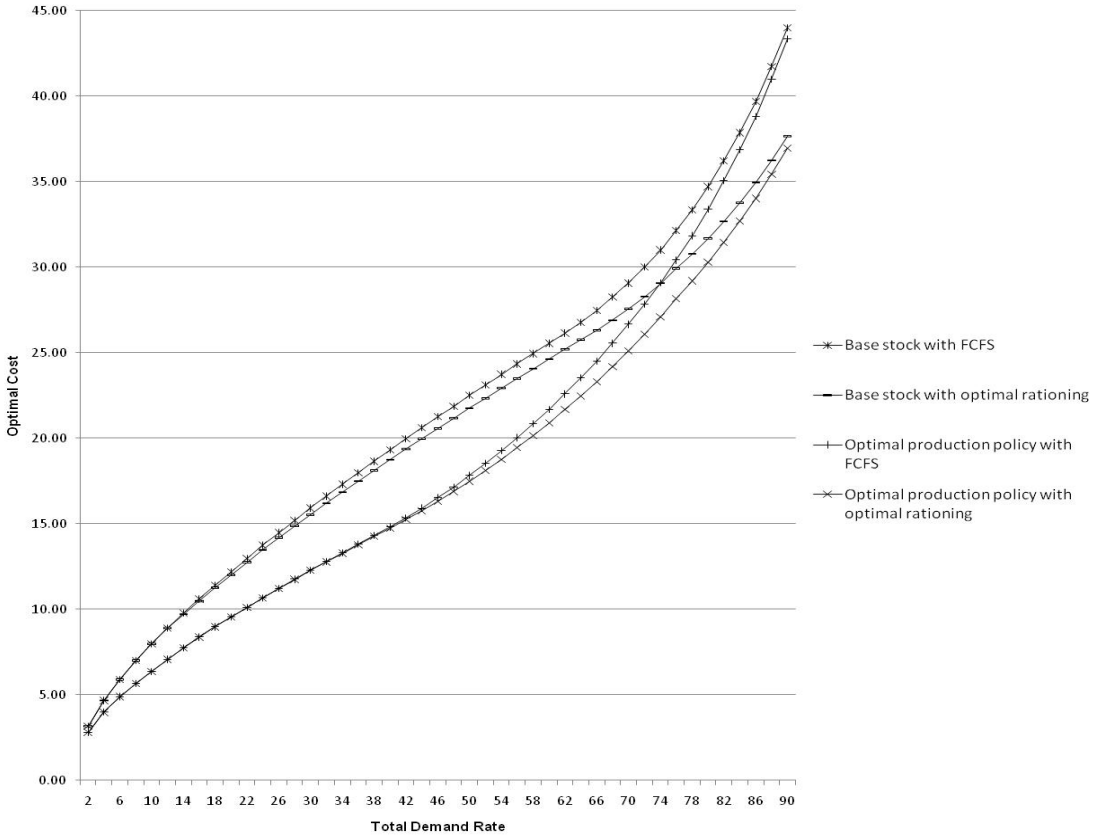
$s = 36$

For different demand mixes and number of available servers, Table 3 compares the optimal rationing policy with the FCFS policy. The comparison is performed under the respective optimal production policies. The table assumes that there are three customer classes and $(\lambda_1+\lambda_2+\lambda_3, \mu, h, p, c_1, c_2, c_3, \alpha) = (30, 2, 0.2, 0.2, 10, 6, 2, 0.6)$. The demand mix is specified using the class 1 and class 2 ratios which are $p_1 = \lambda_1/(\lambda_1+\lambda_2+\lambda_3)$ and $p_2 = \lambda_2/(\lambda_1+\lambda_2+\lambda_3)$, respectively. The table exhibits that the cost reduction obtained via the optimal rationing policy is much more pronounced at smaller values of *s*, which in contrast with our observations for the production policy. When *s* is small, base-stock policy provides a good approximation for the optimal production policy and cost reduction is achieved mainly through rationing. On the other hand, when production capacity is not scarce, the optimal production policy utilizes a large num-

ber of servers at lower inventory levels, and thereby replenishes the inventory rapidly. In such settings, the optimal rationing policy reserves limited stock for the customers from more critical classes and satisfies arriving demands on a FCFS basis at most of the inventory levels. Hence, the stock allocation policy is not very critical at larger values of s. Table 3 also shows that when the demand rate of one of the classes is zero, i.e., a two-class system is under consideration, the gap between the performance of the FCFS policy and the optimal rationing policy is maximized when the demand rates of the remaining classes are close to each other. This is an expected result since for a two-class system when one class dominates the other and the value of class differentiation diminishes. Moreover, the benefit of the optimal rationing policy over the FCFS policy is maximized when only the classes with the highest and lowest lost sales costs are present, i.e., when the medium class vanishes. When the differential between the lost sales costs is significant, there is more value to be captured via class differentiation.

Figure 6 compares the performance of the production/inventory system under four different policy combinations related to production control and stock allocation over a range of values for the total demand rate. For production control, it considers the optimal base stock and the optimal production policy derived in this paper; whereas, for stock allocation, it considers the FCFS and the optimal rationing policies. The results in the graph pertain to the case where class 1 ratio is fixed to 0.5 and $(s, \mu, h, p, c_1, c_2, \alpha) = (56, 2, 0.2, 0.2, 10, 2, 0.6)$. As expected, the combination of the optimal production policy with the optimal rationing policy provides the lowest cost at all total demand rates. The base stock policy with the FCFS policy yields the highest cost. For a given production policy, the performance gap between the FCFS and the optimal rationing policies grows with the total demand rate. Thus, the value of rationing increases with the traffic intensity. For a given stock allocation policy, as total demand rate increases the cost reduction obtained via the optimal production policy first increases and then goes down all the way to zero. Under high traffic, the optimal production policy behaves in a similar fashion to the optimal base stock policy. The figure shows that the costs of the optimal base stock and the optimal production policies converge as the traffic intensity increases irrespective of the stock allocation policy. The figure also illustrates that, for small to moderate values of the total demand rate, the optimal production policy yields more significant cost savings compared to the optimal rationing policy. However, the opposite is true for higher values of the total demand rate.

Figure 6 Optimal production and Base stock policies with or without rationing: Impact of total demand rate



6. Discussion on Control of Parallel Replenishment and Production Channels

This section aims to better position our work within the existing literature, while providing a state-of-the-art perspective on the control of parallel replenishment and production channels.

Modeling the supply process of the inventory systems with an arbitrary number of replenishment channels is instrumental in filling the gap between the two main streams of studies in the existing literature. One of these streams considers standard inventory models where supply lead-times are exogenous, i.e., the replenishment channel is uncapacitated. In this setting, the optimal policy is not fully characterized under lost sales. Recently, Zipkin (2008a) reformulates the standard periodic-review lost-sales inventory problem with a new approach based on discrete convex analysis. He shows that the optimal policy is state-dependent, i.e., the ages and the quantities of

all outstanding orders have an effect on the optimal ordering decision. On the other hand, for the backordering case, Erhardt (1984) shows that if the replenishment orders do not cross in time (e.g., deterministic leadtimes), the optimal ordering policy is independent from the status of the outstanding orders. For such settings, simple base-stock, i.e., order-up-to, policy is optimal. Most of the studies in this stream assume deterministic leadtimes and provide analyses under simple base-stock policies irrespective of the shortage dynamics. However, for the lost sales systems, numerical results of Zipkin (2008b) manifest that simple (state-independent) base-stock policies do not perform well. In many settings, it is even worse than the constant-order policy, which orders the same amount at fixed intervals.

The other stream of studies considers production-inventory systems. These systems are characterized by capacitated replenishment channels. As stated in the introduction, with the exception of the work of Elhafsi et al. (2008) and Zipkin (2000), all the works in this stream model endogenous supply leadtimes with a single server. For basic single server models (models without additional sources of information such as advance demand and assembly component inventory levels) the optimal production policy is a simple base-stock policy defined in terms of a constant produce-up-to level. This holds for both of the lost sales and the backordering cases (see Ha (1997a), Ha (1997b) and Gayon et al. (2009b)). In fact, simple base-stock is the only meaningful policy that can be considered for the single server case. On the other hand, things change when parallel servers are considered. In this setting, there is flexibility to utilize different number of servers at different inventory levels. This gives rise to the state-dependent optimal production policy, which is characterized in this work and depicted in Table 1.

Our work achieves to analyze both the exogenous and endogenous supply leadtimes within a single model. This is made possible by considering an arbitrary number of supply channels so as to cover the spectrum from the single server to the infinite servers. Our model allows analysis of single location continuous-review inventory systems with exogenous exponential leadtimes (i.e., uncapacitated replenishment channel) by letting s --the number of replenishment channels-- go to infinity. On the other hand, having $s = 1$ corresponds to the single server, capacitated production model, which is the subject of most of the literature on the control of make-to-stock queues. Furthermore, as Zipkin (2000, pp. 246) discusses, no real supply system has infinitely many processing channels. Therefore, realistic models should consider finite processing capacity. In this context, the $M/M/\infty$ model should be considered as the limiting case of the $M/M/s$ model studied

in this paper. It should also be noted that the model gives the exact solution of the $M/M/\infty$ when s is selected to be sufficiently large, since the optimal number of servers to be utilized is bounded. The existence of such a bound (beyond which the system is equivalent to a system with exogenous leadtimes) is discussed in Section 4.1 and illustrated in Figure 2. Furthermore, an algorithm is provided to calculate this bound under the average cost criterion.

In our numerical study, we investigate the performance of the base stock policy in comparison with the optimal policy. If the number servers is limited, i.e., production capacity is tight, base-stock performs well. When there is ample capacity, base-stock results in dramatic loss. Furthermore, increasing the number of servers while keeping the traffic intensity constant, undermines the base-stock's performance. In this setting, as the available number of servers increases, the control space becomes more finely discretized. Consequently, the control problem resembles to the one that Mayorga et al. (2006) consider in which the service rate of a single server is controlled over a continuous set.

7. Conclusion

This work constitutes a significant extension of the literature in the area of control of make-to-stock queues, which considers only a single server. We allow an arbitrary number of servers in our model. We show that the optimal production policy is a state-dependent base-stock policy, and the optimal rationing policy is of threshold type. We also prove that the optimal production and rationing policies are monotone in the inventory level and the number of operational servers. We compare the optimal policy with the previously suggested base-stock policy and demonstrate there are settings where the optimality gap is significant. We also provide two variations on the primary model by allowing partial and full order-cancellation flexibility. Our experiments demonstrate that a little flexibility goes a long way and captures most of the value that can be realized via order cancellation.

As the number of available servers increases, the optimal policy stops changing beyond a certain number of servers. Therefore, our work also handles the case of infinitely-many servers, i.e., exogenous supply system.

The multiple server extension provided by this work to the control of make-to-stock queues has potential to open new research avenues. There is a rich and well-established literature in the area of make-to-stock queues. It would be interesting to see how the previous findings in the lit-

erature would apply to this more general production setting that considers multiple production channels. This setting should enable us to address important issues such as the effect of pipeline inventory in rationing decisions.

The basic limitations of the models presented in this paper are that the production times are exponentially distributed and neither set-up times nor costs are considered. The assumption on exponential production times is needed for the tractability of analysis. Without this assumption, it would not be possible to use the uniformization technique. One can consider applying the semi-Markov control approach. However, this would also be problematic in our setting with multiple servers for the same reason that the analysis of $M/G/s$ queues are. Furthermore, the current literature does not even include the more tractable single server models with general production times, which was suggested as future research by Ha (1997a).

The state-of-the-art in control of make-to-stock queues does not also address the set-up times/cost issue. Extending our model to incorporate this set-up dynamics and costs would be a worthwhile effort. However, in such a setting the monotonicity results of our work will no longer hold.

References

Arslan, H., S.C. Graves, T. Roemer. 2007. A single product inventory model for multiple demand classes. *Management Science* **53** 1486-1500.

Benjaafar, S., M. Elhafsi. 2006. Production and inventory control of a single product assemble-to-order systems with multiple customer classes. *Management Science* **52** 1896-1912.

Buzacott, J.A., J.G. Shanthikumar. 1993. Stochastic models of manufacturing systems. Prentice Hall.

Cil, E.B., E.L. Ormeci, F. Karaesmen. 2009. Effects of system parameters on the optimal control structure in a class of queueing control problems. *Queueing Systems* DOI 10.1007/s11134-009-9109-x.

Dekker, R., M.J. Kleijn, P.J. de Rooij. 1998. A spare parts stocking policy based on equipment criticality. *International Journal of Production Economics* **56-57** 69-77.

Dekker, R., R.M. Hill, M.j. Kleijn, R.H. Teunter. 2002. On the (S -1, S) lost sales inventory model with priority demand classes. *Naval Research Logistics* **49** 593-610.

Deshpande, V., M.A. Cohen, K. Donohue. 2003. A threshold inventory rationing policy for service-differentiated demand classes. *Management Science* **49** 683-703.

Elhafsi, M, S. Benjaafar, Y. Yu. 2008. Production and inventory control of a system with multiple sources of supply. Working Paper.

Erhardt, R. 1984. (s,S) policies for a dynamic inventory model with stochastic lead times. *Operations Research* **32** 121-132.

Fadiloglu, M.M., O. Bulut 2007. An embedded Markov chain approach to the analysis of inventory systems with backordering under rationing. Working Paper, Department of Industrial Engineering, Bilkent University, Ankara, Turkey.

Fadiloglu, M.M., O. Bulut 2010. A dynamic rationing policy for continuous-review inventory systems. *European Journal of Operational Research* **202** 675-685.

Gans, N., S. Savin. 2007. Pricing and capacity rationing for rentals with uncertain durations. *Management Science* **53** 390-407.

Gayon, J.P, S. Benjaafar, F.D. Vericourt. 2009a. Using imperfect advance demand information in production-inventory systems with multiple customer classes. *Manufacturing&Service Operations Management* **11** 128-143.

Gayon, J.P, F.D. Vericourt, F. Karaesmen. 2009b. Stock rationing in an $M/E_r/1$ multi-class make-to-stock queue with backorders. *IIE Transactions* **41** 1096-1109.

Ha, A.Y. 1997a. Inventory rationing in a make-to-stock production system with several demand classes and lost sales. *Management Science* **43** 1093-1103.

Ha, A.Y. 1997b. Stock rationing policy for a make –to-stock production system with two priority classes and backordering. *Naval Research Logistics* **43** 458-472.

Ha, A.Y. 2000. Stock Rationing in an $M/E_k/1$ make-to-stock queue. *Management Science* **46** 77-87.

Huang, B., S.M.R Iravani. 2008. A make-to-stock system with multiple customer classes and batch ordering. *Operations Research* **56** 1312-1320.

Iravani, S.M.R, T. Liu, K.L. Luangkesorn, D.S. Levi. 2007. A produce-to-stock system with advance demand information and secondary customers. *Naval Research Logistics* **54** 331-345.

Jaarsveld, W., R. Dekker. 2009. Finding the optimal policies in the (*S*-1, *S*) lost sales inventory model with multiple demand classes. Technical Report, Econometric institute, Erasmus University Rotterdam, The Netherlands.

Lippman, S. 1975. Applying a new device in the optimization of exponential queuing systems. *Operations Research* **23** 687-710.

Mayorga, M., E., H. Ahn, J.G. Shanthikumar. 2006. Optimal control of a make-to-stock system with adjustable service rate. *Probability in the Engineering and Informational Sciences* **20** 609-634.

Melchiors, P., R. Dekker, M.J. Kleijn. 2000. Inventory rationing in an (*s*, *Q*) inventory model with lost sales and two demand classes. *Journal of Operational Research Society* **51** 111-122.

Melchiors, P. 2003. Restricted time remembering policies for the inventory rationing problem. *International Journal of Production Economics* **81** 461-468.

Nahmias, S., W. Demmy. 1981. Operating characteristics of an inventory system with rationing. *Management Science* **27** 1236-1245.

Porteus, E.L. 1982. Conditions for characterizing the structure of optimal strategies in infinite horizon dynamic programs. *Journal of Optimization Theory and Applications* **36** 419-432.

Teunter, R.H., W.K. Klein Haneveld. 2008. Dynamic inventory rationing strategies for inventory systems with two demand classes, Poisson demand and backordering. *European Journal of Operational Research* **190** 156-178.

Topkis, D.M. 1968. Optimal ordering and rationing policies in a non-stationary dynamic inventory model with n demand classes. *Management Science* **15** 160-176.

Veinott, A.F. 1965. Optimal policy in a dynamic, single product, non-stationary inventory model with several demand classes. *Operations Research* **13** 761-778.

Vericourt, F.D, F. Karaesmen, Y.Dallery. 2002. Optimal stock allocation for a capacitated supply system. *Management Science* **48** 1486-1501.

Zipkin, P. 2000. Foundations of inventory management. McGraw-Hill.

Zipkin, P. 2008a. On the structure of lost-sales inventory models. *Operations Research* **56** 937-944.

Zipkin, P. 2008a. Old and new methods for lost-sales inventory models. *Operations Research* **56** 1256-1263.

Appendix

Proof of Theorem 1: In order to prove that parts i and ii of Theorem 1 hold, it is enough to show that the optimization operator preserves these structural properties. Thence, suppose parts i and ii hold. We will first show that $u^*(x,y) = \begin{cases} u^*(x,0), & y \le u^*(x,0) \\ y, & y > u^*(x,0) \end{cases}$.

Now, for $y \le u^*(x,0)$ assume that $u^*(x,y) \neq u^*(x,0)$. Then, by the hypothesis and the assumption we have

$$f\left(x,0,u^*(x,0)\right) = f\left(x,y,u^*(x,0)\right) > f\left(x,y,u^*(x,y)\right) = f\left(x,0,u^*(x,y)\right) > f\left(x,0,u^*(x,0)\right),$$

which is a contradiction. Therefore, $u^*(x,y) = u^*(x,0)$, for $y \le u^*(x,0)$.

In order to show that $u^*(x,y) = y$, for $y > u^*(x,0)$, we will first show that

$\Delta^u f(x,u^*(x,0)+1,u^*(x,0)+1) \ge \Delta^u f(x,u^*(x,0),u^*(x,0)) > 0$ holds where

$$\begin{aligned}\Delta^u f(x,u^*(x,0)+1,u^*(x,0)+1) &= f\left(x,u^*(x,0)+1,u^*(x,0)+2\right) - f\left(x,u^*(x,0)+1,u^*(x,0)+1\right) \\ &= p + \mu\Delta^x J(x,u^*(x,0)+1) + \left(u^*(x,0)+1\right)\mu\Delta^y J(x+1,u^*(x,0)) \\ &\quad + \Delta^y T_R(x,u^*(x,0)+1)\end{aligned}$$

and

$$\begin{aligned}\Delta^u f(x,u^*(x,0),u^*(x,0)) &= f\left(x,u^*(x,0),u^*(x,0)+1\right) - f\left(x,u^*(x,0),u^*(x,0)\right) \\ &= p + \mu\Delta^x J(x,u^*(x,0)) + u^*(x,0)\mu\Delta^y J(x+1,u^*(x,0)-1) \\ &\quad + \Delta^y T_R(x,u^*(x,0))\end{aligned}$$

We can immediately say that $\Delta^u f(x,u^*(x,0),u^*(x,0)) > 0$, because $u^*(x,u^*(x,0)) = u^*(x,0)$ as shown just above. Besides, $\mu\Delta^x J(x,u^*(x,0)+1) \ge \mu\Delta^x J(x,u^*(x,0))$ holds by property (8), $\left(u^*(x,0)+1\right)\mu\Delta^y J(x+1,u^*(x,0)) \ge u^*(x,0)\mu\Delta^y J(x+1,u^*(x,0)-1)$ holds by the hypothesis. For each $i \in \{2,...,n\}$, $\Delta^y T_{R_i}(x,u^*(x,0)+1) \ge \Delta^y T_{R_i}(x,u^*(x,0))$ is shown by considering all the possible cases:

Case1. $\Delta^y T_{R_i}(x,u^*(x,0)+1) = \lambda_i \Delta^y J(x-1,u^*(x,0)+1)$, $\Delta^y T_{R_i}(x,u^*(x,0)) = \lambda_i \Delta^y J(x-1,u^*(x,0))$

Then, $\Delta^y T_{R_i}(x,u^*(x,0)+1) \ge \Delta^y T_{R_i}(x,u^*(x,0))$ holds by the hypothesis.

Case2. $\Delta^y T_{R_i}(x,u^*(x,0)+1) = \lambda_i \Delta^y J(x-1,u^*(x,0)+1)$, and

$\Delta^y T_{R_i}(x,u^*(x,0)) = \lambda_i \left(J(x-1,u^*(x,0)+1) - J(x,u^*(x,0)) - c_i \right)$.

Then, $\lambda_i \Delta^y J(x,u^*(x,0)) \geq \Delta^y T_{R_i}(x,u^*(x,0))$ by the definition of $T_{R_i}(x,u^*(x,0)+1)$, and

$\Delta^y J(x-1,u^*(x,0)+1) \geq \Delta^y J(x,u^*(x,0))$ by the property $\Delta^y v(x-1,y+1) \geq \Delta^y v(x,y)$. Thus,

$\Delta^y T_{R_i}(x,u^*(x,0)+1) \geq \Delta^y T_{R_i}(x,u^*(x,0))$.

Case3. $\Delta^y T_{R_i}(x,u^*(x,0)+1) = \lambda_i \left(J(x-1,u^*(x,0)+2) - J(x,u^*(x,0)+1) - c_i \right)$, and

$\Delta^y T_{R_i}(x,u^*(x,0)) = \lambda_i \Delta^y J(x,u^*(x,0))$.

Then, $\Delta^y T_{R_i}(x,u^*(x,0)+1) \geq \lambda_i \Delta^y J(x-1,u^*(x,0)+1)$ by the definition of $T_{R_i}(x,u^*(x,0)+1)$, and

$\Delta^y J(x-1,u^*(x,0)+1) \geq \Delta^y J(x,u^*(x,0))$ by the property $\Delta^y v(x-1,y+1) \geq \Delta^y v(x,y)$. Thus,

$\Delta^y T_{R_i}(x,u^*(x,0)+1) \geq \Delta^y T_{R_i}(x,u^*(x,0))$.

Case4. $\Delta^y T_{R_i}(x,u^*(x,0)+1) = \lambda_i \Delta^y J(x,u^*(x,0)+1)$, $\Delta^y T_{R_i}(x,u^*(x,0)) = \lambda_i \Delta^y J(x,u^*(x,0))$

Then, $\Delta^y T_{R_i}(x,u^*(x,0)+1) \geq \Delta^y T_{R_i}(x,u^*(x,0))$ holds by the hypothesis.

The results of above four cases imply that $\sum_{i=1}^{n} \Delta^y T_{R_i}(x,u^*(x,0)+1) \geq \sum_{i=1}^{n} \Delta^y T_{R_i}(x,u^*(x,0)+1)$, i.e.,

$\Delta^y T_R(x,u^*(x,0)+1) \geq \Delta^y T_R(x,u^*(x,0))$.

(Here it should be noted that following the steps of the above four cases, we can also conclude that $\Delta^y T_R(x,y+1) \geq \Delta^y T_R(x,y)$ holds for any *y*. That is, $T_R(x,y)$ is *y*-convex.)

Thus, $\Delta^u f(x,u^*(x,0)+1,u^*(x,0)+1) \geq \Delta^u f(x,u^*(x,0),u^*(x,0)) > 0$ holds. Moreover, having $J(x,u)$ is a convex-increasing function of u and $T_R(x,u)$ is u-convex implies that $f(x,y,u) = hx + pu + (s-u)\mu J(x,y) + u\mu J(x+1,u-1) + T_R(x,u)$ is a convex function of u. Therefore, for any $u \geq u^*(x,0)+2$, $\Delta^u f(x,u^*(x,0)+1,u) \geq \Delta^u f(x,u^*(x,0)+1,u^*(x,0)+1) > 0$, that is $u^*\left(u^*(x,0)+1\right) = u^*(x,0)+1$. Following the same logic used in the proof of $u^*\left(u^*(x,0)+1\right) = u^*(x,0)+1$, it is easy to show also that $u^*(x,y) = y$, for $y \geq u^*(x,0)+2$.

Hence, we conclude that $u^*(x,y)=\begin{cases}u^*(x,0), & y\le u^*(x,0)\\ y, & y>u^*(x,0)\end{cases}$ holds. Using this fact, we will show that the optimization operator preserves the properties stated in Theorem 1. Now, first consider $\Delta^y T(J(x,y))=T(J(x,y+1))-T(J(x,y))=f(x,y+1,u^*(x,y+1))-f(x,y,u^*(x,y))$. For $y<u^*(x,0)$, $u^*(x,y)=u^*(x,y+1)=u^*(x,0)$ (as shown above) and $J(x,y)=J(x,y+1)$ (by the hypothesis). Then, $\Delta^y T(J(x,y))=f(x,y,u^*(x,y))-f(x,y,u^*(x,y))=0$.

For $y\ge u^*(x,0)$, $u^*(x,y)=y$, $u^*(x,y+1)=y+1$ and $J(x,y)<J(x,y+1)$. Then, $\Delta^y T(J(x,y))=f(x,y+1,y+1)-f(x,y,y)>f(x,y,y+1)-f(x,y,y)$. Moreover, since $f(x,y,u)$ is u-convex and $u^*(x,y)=y$, $f(x,y,y+1)-f(x,y,y)\ge 0$. Hence, $\Delta^y T(J(x,y))>0$.

We will finally show that for $y\ge u^*(x,0)$, $\Delta^y T(J(x,y+1))\ge\Delta^y T(J(x,y))$, i.e., $J(x,y)$ is y-convex. For $y\ge u^*(x,0)$, $u^*(x,y)=y$, $u^*(x,y+1)=y+1$ and $u^*(x,y+2)=y+2$. Therefore,

$$\begin{aligned}&\Delta^y T(J(x,y+1))-\Delta^y T(J(x,y))\\ &\quad=(s-(y+2)\mu)(\Delta^y J(x,y+1)-\Delta^y J(x,y))\\ &\quad\quad+y\mu(\Delta^y J(x+1,y)-\Delta^y J(x+1,y-1))+2\mu(\Delta^y J(x+1,y)-\Delta^y J(x,y))\\ &\quad\quad+\Delta^y T_R(x,y+1)-\Delta^y T_R(x,y)\end{aligned}$$

where $(s-(y+2)\mu)(\Delta^y J(x,y+1)-\Delta^y J(x,y))$ and $y\mu(\Delta^y J(x+1,y)-\Delta^y J(x+1,y-1))$ are greater or equal to zero by the hypothesis, and $2\mu(\Delta^y J(x+1,y)-\Delta^y J(x,y))\ge 0$ by the property (8) . Moreover, it is shown above that $\Delta^y T_R(x,y+1)-\Delta^y T_R(x,y)\ge 0$. Thus, we conclude that $\Delta^y T(J(x,y+1))-\Delta^y T(J(x,y))\ge 0$, i.e., $J(x,y)$ is y-convex.

Proof of Corollary 1:

i. It is shown in the proof of Theorem 1 that $u^*(x,y)=\begin{cases}u^*(x,0), & y\le u^*(x,0)\\ y, & y>u^*(x,0)\end{cases}$ holds.

ii. Theorem 1 indicates that $J(x,0) = ... = J\left(x,u^*(x,0)\right) < J\left(x,u^*(x,0)+1\right) < ... < J(x,s)$. Using this result and part-i of Corollary 1, we can alternatively define the optimal number of optimal number of operational servers at state (x,y) as the number of operational servers beyond which the optimal cost function starts to increase.

iii. It is apparent from part-i of Corollary 1.

Proof of Theorem 2:

i. Immediate conclusion from properties (8) and (9).

ii. By definition of $u^*(x,y)$, $J(x,u^*(x,y)+1) - J(x,u^*(x,y)) > 0$. From property (8) we have $J(x+1,u^*(x,y)+1) - J(x+1,u^*(x,y)) \geq J(x,u^*(x,y)+1) - J(x,u^*(x,y))$. Therefore, $J(x+1,u^*(x,y)+1) - J(x+1,u^*(x,y)) > 0$. Thus, we conclude that $u^*(x+1,y) \leq u^*(x,y)$.

iii. It will be enough to show that the optimization operator preserves the property $\Delta^x J(x-1,y) \geq -c_1$. Suppose $\Delta^x J(x-1,y) \geq -c_1$. Now, for any $u \geq y$,

$$f(x,y,u) - f(x-1,y,u) = h + s\mu\Delta^x J(x-1,y) + u\mu\left(\Delta^x J(x,u-1) - \Delta^x J(x-1,y)\right) + \sum_{i=1}^{n} \Delta^x T_{R_i}(x-1,u)$$

In the above equation $h \geq 0$ and $u\mu\left(\Delta^x J(x,u-1) - \Delta^x J(x-1,y)\right) \geq 0$ by properties (8) and (9).

Moreover, by the fact that $s\mu = 1 - \sum_{i=1}^{n} \lambda_i$ and the hypothesis, $s\mu\Delta^x J(x-1,y) \geq -c_1\left(1 - \sum_{i=1}^{n} \lambda_i\right)$.

Therefore, $h + s\mu\Delta^x J(x-1,y) + u\mu\left(\Delta^x J(x,u-1) - \Delta^x J(x-1,y)\right) \geq -c_1\left(1 - \sum_{i=1}^{n} \lambda_i\right)$.

By the hypothesis, $\Delta^x T_{R_1}(x-1,u) = \lambda_1 \Delta^x J(x-2,u) \geq -\lambda_1 c_1$. For each $i \in \{2,...,n\}$, $\Delta^x T_{R_i}(x-1,u) \geq -\lambda_i c_1$ can be shown by considering all the possible cases:

Case1. $T_{R_i}(x,u) = \lambda_i J(x-1,u),\ T_{R_i}(x-1,u) = \lambda_i J(x-2,u)$

Then, $\Delta^x T_{R_i}(x-1,u) = \lambda_i (J(x-1,u) - J(x-2,u)) \geq -\lambda_i c_1$ by the hypothesis.

Case2. $T_{R_i}(x,u) = \lambda_i J(x-1,u),\ T_{R_i}(x-1,u) = \lambda_i (c_i + J(x-1,u))$

Then, $\Delta^x T_{R_i}(x-1,u) = \lambda_i (J(x-1,u) - J(x-1,u) - c_i) = -\lambda_i c_i \geq -\lambda_i c_1$ because $c_1 \geq c_i$

Case3. $T_{R_i}(x,\bar{y}) = \lambda_i (c_i + J(x,u)),\ T_{R_i}(x-1,u) = \lambda_i (c_i + J(x-1,u))$

Then, $\Delta^x T_{R_i}(x-1,u) = \lambda_i (J(x,u) - J(x-1,u)) \geq -\lambda_i c_1$ by the hypothesis.

Hence, we have $\sum_{i=1}^{n} \Delta^x T_{R_i}(x-1,u) \geq -c_1 \sum_{i=1}^{n} \lambda_i$, and so $f(x,y,u) - f(x-1,y,u) \geq -c_1$.

Having $f(x,y,u) - f(x-1,y,u) \geq -c_1$, $u \geq y$, implies that

$f(x,y,u^*(x,y)) - f(x-1,y,u^*(x,y)) \geq -c_1$.Since $f(x-1,y,u^*(x,y)) \geq f(x-1,y,u^*(x-1,y))$,

we conclude that $f(x,y,u^*(x,y)) - f(x-1,y,u^*(x-1,y)) \geq -c_1$, i.e.,

$T(J(x,y)) - T(J(x-1,y)) \geq -c_1$. Thus, $\Delta^x J(x-1,y) \geq -c_1$.

iv. For $i \in \{2,...,n\}$, let us define $K_x^i(y) = \min\{x : \Delta^x J(x,y) \geq -c_i\}$. Since $-c_n \leq -c_{n-1} ... \leq -c_2$ and $\Delta^x J(x,y)$ is non-decreasing in *x* (from part-i), $K_x^n(y) \geq K_x^{n-1}(y) \geq ... \geq K_x^2(y) \geq 0$ holds.

From property (8), $\Delta^x J(K_x^i(y), y+1) \geq \Delta^x J(K_x^i(y), y) \geq -c_i$, and since $\Delta^x J(x,y)$ is non-decreasing in *x* (from part-i) $K_x^i(y+1) \leq K_x^i(y)$ holds.

v. For $i \in \{2,...,n\}$, let us define $K_y^i(x) = \min\{y : \Delta^x J(x-1,y) \geq -c_i\} - 1$. Having $K_y^i(x) = -1$ implies that it is optimal to satisfy an arriving class *i* demand at all *y* levels when the on-hand inventory level is x. Since $-c_n \leq -c_{n-1} ... \leq -c_2$ and $\Delta^x J(x,y)$ is non-decreasing in *y*),

$K_y^n(x) \geq K_y^{n-1}(x) \geq ... \geq K_y^2(x) \geq 0$ holds.

Property (9) implies that if $\Delta^x J(x,y+1) \geq -c_i$, then $\Delta^x J(x+1,y) \geq -c_i$. That is, if

$\Delta^x J(x+1,y) \geq \Delta^x J(x,y+1) \geq -c_i$ and $\Delta^x J(x,y) < -c_i$ then $K_y^i(x) = y+1$ and $K_y^i(x+1) < y$.

Therefore, we can conclude that $K_y^i(x+1) < K_y^i(x)$.

Proof of Theorem 3:

Suppose $\Delta^x J(x,y+1) \geq \Delta^x J(x,y)$ and $\Delta^x J(x,y) \geq \Delta^x J(x-1,y+1)$. We will show that the optimization operator *T* preserves this structure.

We will first show that

$$\Delta^x f(x,y+1,y+1) \geq \Delta^x f(x,y,y) \tag{14}$$

$$\Delta^x f(x,y,y) \geq \Delta^x f(x-1,y+1,y+1) \tag{15}$$

then $\Delta^x T\big(J(x,y+1)\big) \geq \Delta^x T\big(J(x,y)\big) =$

$$f(x+1,u^*(x+1,y+1),u^*(x+1,y+1))-f(x,u^*(x,y+1),u^*(x,y+1)) \geq f(x+1,u^*(x+1,y),u^*(x+1,y))-f(x,u^*(x,y),u^*(x,y)) \quad (16)$$

and $\Delta^x T\left(J(x,y)\right) \geq \Delta^x T\left(J(x-1,y+1)\right) =$

$$f(x+1,u^*(x+1,y),u^*(x+1,y))-f(x,u^*(x,y),u^*(x,y)) \geq f(x,u^*(x,y+1),u^*(x,y+1))-f(x-1,u^*(x-1,y+1),u^*(x-1,y+1)) \quad (17)$$

From the hypothesis and part-iii of Theorem2;

$$\begin{aligned}
\Delta^x f(x,y+1,y+1) &= h+(s-y-1)\mu\Delta^x J(x,y+1)+(y+1)\mu\Delta^x J(x+1,y)+\lambda_1\Delta^x J(x-1,y+1) \\
&\quad +\sum_{i=2}^{n} T_{R_i}(x+1,y+1)-\sum_{i=2}^{n} T_{R_i}(x,y+1) \\
&\geq h+(s-y-1)\mu\Delta^x J(x,y)+y\mu\Delta^x J(x+1,y-1)+\mu\Delta^x J(x+1,y-1)+\lambda_1\Delta^x J(x-1,y) \\
&\quad +\sum_{i=2}^{n} T_{R_i}(x+1,y+1)-\sum_{i=2}^{n} T_{R_i}(x,y+1) \\
&= h+(s-y)\mu\Delta^x J(x,y)+y\mu\Delta^x J(x+1,y-1)+\mu\Delta^x J(x+1,y-1)+\lambda_1\Delta^x J(x-1,y) \\
&\quad +\sum_{i=2}^{n} T_{R_i}(x+1,y+1)-\sum_{i=2}^{n} T_{R_i}(x,y+1)+\mu\left(\Delta^x J(x+1,y-1)-\Delta^x J(x,y)\right) \\
&\geq h+(s-y)\mu\Delta^x J(x,y)+y\mu\Delta^x J(x+1,y-1)+\mu\Delta^x J(x+1,y-1)+\lambda_1\Delta^x J(x-1,y) \\
&\quad +\sum_{i=2}^{n} T_{R_i}(x+1,y+1)-\sum_{i=2}^{n} T_{R_i}(x,y+1)
\end{aligned}$$

In order to conclude that the right hand side of the above inequality is greater than $\Delta^x f(x,y,y)$, we need to show that $\forall i \in \{2,...,n\}$, $T_{R_i}(x+1,y+1)-T_{R_i}(x,y+1) \geq T_{R_i}(x+1,y)-T_{R_i}(x,y)$, i.e., $\Delta^x T_{R_i}(x,y+1) \geq \Delta^x T_{R_i}(x,y)$. We will consider three cases in order to show that the inequality holds.

Case1. $K_x^i(y+1) \leq x-1$. Then, $\Delta^x T_{R_i}(x,y+1) = \lambda_i \Delta^x J(x-1,y+1)$

i. $K_x^i(y) \leq x-1$

$\Delta^x T_{R_i}(x,y) = \lambda_i \Delta^x J(x-1,y)$. Thus, $\Delta^x T_{R_i}(x,y+1) \geq \Delta^x T_{R_i}(x,y)$.

ii. $K_x^i(y) = x$

$\Delta^x T_{R_i}(x,y) = \lambda_i\left(J(x,y)-c_i-J(x,y)\right) = -\lambda_i c_i$.

$J(x-1,y+1) \geq -c_i$ because $K_x(y+1) \leq x-1$. Thus, $\Delta^x T_{R_i}(x,y+1) \geq \Delta^x T_{R_i}(x,y)$.

iii. $K_x^i(y) \geq x+1$

This case is not possible, because the hypothesis $\Delta^x J(x,y) \geq \Delta^x J(x-1,y+1)$ implies that if $\Delta^x J(x-1,y+1) \geq -c_i$, then $\Delta^x J(x,y) \geq -c_i$. In words, if we satisfy an arriving class *i* demand when there are *x* units of inventory and (*y*+1) active servers, we should satisfy an arriving class *i* demand when there are (*x*+1) units of inventory and *y* active servers.

Therefore, $K_x^i(y) \leq x$ whenever $K_x^i(y+1) \leq x-1$.

Case2. $K_x^i(y+1) = x$. Then, $\Delta^x T_{R_i}(x,y+1) = \lambda_i\left(J(x,y+1) - c_i - J(x,y+1)\right) = -\lambda_i c_i$

i. $K_x^i(y) = x$

$\Delta^x T_{R_i}(x,y) = \lambda_i\left(J(x,y) - c_i - J(x,y)\right) = -\lambda_i c_i$. Thus,

$\Delta^x T_{R_i}(x,y+1) = \Delta^x T_{R_i}(x,y)$.

ii. $K_x^i(y) = x+1$

$\Delta^x T_{R_i}(x,y) = \lambda_i\left(c_i + J(x+1,y) - c_i - J(x,y)\right) = \lambda_i \Delta^x J(x,y)$.

And, $-\lambda_i c_i \geq \lambda_i \Delta^x J(x,y)$ because $K_x^i(y) = x+1$. Thus,

$\Delta^x T_{R_i}(x,y+1) \geq \Delta^x T_{R_i}(x,y)$.

Case3. $K_x^i(y+1) \geq x+1$.

Then, $\Delta^x T_{R_i}(x,y+1) = \lambda_i\left(c_i + J(x+1,y+1) - c_i - J(x,y+1)\right) = \lambda_i \Delta^x J(x,y+1)$.

Since we have $K_x^i(y+1) \leq K_x^i(y)$, it is true that

$\Delta^x T_{R_i}(x,y) = \lambda_i\left(c_i + J(x+1,y) - c_i - J(x,y)\right) = \lambda_i \Delta^x J(x,y)$. Then by the hypothesis

$\Delta^x T_{R_i}(x,y+1) \geq \Delta^x T_{R_i}(x,y)$.

We have just shown that (14) holds. Showing (15) is equivalent to show that

$f(x+1,y,y)-f(x,y+1,y+1)\geq f(x,y,y)-f(x-1,y+1,y+1)$, where

$$\begin{aligned}f(x+1,y,y)-f(x,y+1,y+1) &= h-p+(s-y-1)\mu\big(J(x+1,y)-J(x,y+1)\big)\\ &\quad + y\mu\big(J(x+2,y-1)-J(x+1,y)\big)+\lambda_1\big(J(x,y)-J(x-1,y+1)\big)\\ &\quad +\sum_{i=2}^{n}T_{R_i}(x+1,y)-\sum_{i=2}^{n}T_{R_i}(x,y+1)\end{aligned}$$

and

$$\begin{aligned}f(x,y,y)-f(x-1,y+1,y+1) &= h-p+(s-y-1)\mu\big(J(x,y)-J(x-1,y+1)\big)\\ &\quad + y\mu\big(J(x+1,y-1)-J(x,y)\big)+\lambda_1\big(J(x-1,y)-J(x-2,y+1)\big)\\ &\quad +\sum_{i=2}^{n}T_{R_i}(x,y)-\sum_{i=2}^{n}T_{R_i}(x-1,y+1)\end{aligned}$$

Using the hypothesis it is easy to show that the terms of the right hand side of the first equation are greater or equal to the respective terms of the right hand side of the second equation expect the terms related to the rationing decision for each class other than class 1. Therefore, we need to show that $\forall i\in\{2,...,n\}, \Delta^x T_{R_i}(x,y)\geq\Delta^x T_{R_i}(x-1,y+1)$.

Case1. $K_x^i(y+1)\leq x-2$. Then, $\Delta^x T_{R_i}(x-1,y+1)=\lambda_i\Delta^x J(x-2,y+1)$

In this case we have $\Delta^x T_{R_i}(x,y)=\lambda_i\Delta^x J(x-1,y)$. Thus, by the hypothesis

$\Delta^x T_{R_i}(x,y)\geq\Delta^x T_{R_i}(x-1,y+1)$

Case2. $K_x^i(y+1)=x-1$. Then,

$\Delta^x T_{R_i}(x-1,y+1)=\lambda_i\big(J(x-1,y+1)-c_i-J(x-1,y+1)\big)=-\lambda_i c_i$

i. $K_x^i(y)=x-1$

$\Delta^x T_{R_i}(x,y)=\lambda_i\Delta^x J(x-1,y)$, and $\Delta^x J(x-1,y)\geq -c_i$ since $K_x^i(y)=x-1$. So,

$\Delta^x T_{R_i}(x,y)\geq\Delta^x T_{R_i}(x-1,y+1)$.

ii. $K_x^i(y)=x$

$\Delta^x T_{R_i}(x,y)=\lambda_i\big(J(x,y)-c_i-J(x,y)\big)=-\lambda_i c_i=\Delta^x T_{R_i}(x-1,y+1)$

Case3. $K_x^i(y+1)\geq x$. Then,

$$\Delta^x T_{R_i}(x-1,y+1) = \lambda_i\left(c_i + J(x,y+1) - c_i - J(x-1,y+1)\right) = \lambda_i \Delta^x J(x-1,y+1)$$

i. $K_x^i(y) = x$ (possible if $K_x^i(y+1) = x$)

$\Delta^x T_{R_i}(x,y) = -\lambda_i c_i$, and $\Delta^x J(x-1,y+1) \le -c_i$ because we do not satisfy class 2 demand at state $(x,y+1)$. Thus, $\Delta^x T_{R_i}(x,y) \ge \Delta^x T_{R_i}(x-1,y+1)$.

ii. $K_x^i(y) \ge x+1$

$\Delta^x T_{R_i}(x,y) = \lambda_i \Delta^x J(x,y)$, and by the hypothesis $\Delta^x J(x,y) \ge \Delta^x J(x-1,y+1)$.

Thus, $\Delta^x T_{R_i}(x,y) \ge \Delta^x T_{R_i}(x-1,y+1)$

We have also shown (15). Now, we need to show that (16) holds in order to complete the proof of part-i of the theorem.

From Corollary1 and part-ii of Theorem2, we have

- $u^*(x,y+1) \ge u^*(x,y) \ge u^*(x+1,y)$
- $u^*(x,y+1) \ge u^*(x+1,y+1) \ge u^*(x+1,y)$
- $u^*(x,y+1) \le u^*(x,y)+1$ and $u^*(x+1,y+1) \le u^*(x+1,y)+1$

Case1. $u^*(x,y+1) \ge u^*(x+1,y+1) \ge u^*(x,y) \ge u^*(x+1,y)$

i. $u^*(x+1,y) = u^*(x,y) = u^*(x+1,y+1) = u^*(x,y+1)$

Trivial case.

ii. $u^*(x+1,y) = u^*(x,y) = u^*(x+1,y+1) = u^*(x,y+1) - 1$

Since $u^*(x+1,y) = u^*(x+1,y+1)$, we should have $u^*(x+1,0) \ge y+1$ and so $u^*(x,0) \ge y+1$. Therefore, $u^*(x,y) = u^*(x,y+1)$ should hold. Thus, this case is not possible.

iii. $u^*(x+1,y) = u^*(x,y) = u^*(x+1,y+1) - 1 = u^*(x,y+1) - 1$

Then, $u^*(x+1,0) \le u^*(x,0) \le y$. So, $u^*(x+1,y) = u^*(x,y) = y$ and $u^*(x+1,y+1) = u^*(x,y+1) = y+1$. Thus, (16) holds because (14) holds.

iv. $u^*(x+1,y) = u^*(x,y) - 1 = u^*(x+1,y+1) - 1 = u^*(x,y+1) - 1$

Then, $u^*(x+1,0) \le y$, $y^*(x+1,y) = y$. And,

$u^*(x,y) = u^*(x+1,y+1) = u^*(x,y+1) = y+1$.

Therefore, the left hand side of (16) becomes $f(x+1,y+1,y+1) - f(x,y+1,y+1)$ and the right hand side becomes $f(x+1,y,y) - f(x,y+1,y+1)$. Since $u^*(x+1,0) \le y$, due to Theorem 1 we have $f(x+1,y+1,y+1) \ge f(x+1,y,y)$ and (16) holds.

Case2. $u^*(x,y+1) \ge u^*(x,y) \ge u^*(x+1,y+1) \ge u^*(x+1,y)$

i. $u^*(x+1,y) = u^*(x+1,y+1)$ and $u^*(x,y) = u^*(x,y+1)$

Trivial case.

ii. $u^*(x+1,y) = u^*(x+1,y+1) - 1$ and $u^*(x,y) = u^*(x,y+1) - 1$

In this case (16) holds due to (14).

iii. $u^*(x+1,y) = u^*(x+1,y+1)$ and $u^*(x,y) = u^*(x,y+1) - 1$

Then, $u^*(x+1,y) = u^*(x+1,y+1)$ implies that $u^*(x,0) \ge u^*(x+1,0) \ge y+1$ and so $u^*(x,y) = u^*(x,y+1)$. Thus, this case is not possible.

iv. $u^*(x+1,y) = u^*(x+1,y+1) - 1$ and $u^*(x,y) = u^*(x,y+1)$

Then, $u^*(x+1,0) \le y$, and so $u^*(x+1,y) = y$, $u^*(x+1,y+1) = y+1$.

Therefore, the left hand side of (16) becomes $f(x+1,y+1,y+1) - f(x,u^*(x,0),u^*(x,0))$ and the right hand side becomes $f(x+1,y,y) - f(x,u^*(x,0),u^*(x,0))$. Since $u^*(x+1,0) \le y$, due to Theorem 1 we have $f(x+1,y+1,y+1) \ge f(x+1,y,y)$ and (16) holds.

Finally we will prove (17) where

$$\begin{aligned}\Delta^x T(J(x,y)) &= f(x+1,y,u^*(x+1,y)) - f(x,y,u^*(x,y)) \\ &\ge f(x+1,y,u^*(x+1,y)) - f(x,y,u^*(x+1,y)) \\ &= h + (s - u^*(x+1,y))\mu\Delta^x J(x,y) + u^*(x+1,y)\mu\Delta^x J(x+1,u^*(x+1,y)-1) \\ &\quad + \sum_{i=1}^{n} \Delta^x T_{R_i}(x,u^*(x+1,y))\end{aligned}$$

and

$$\begin{aligned}\Delta^x T\left(J(x-1,y+1)\right) &= f(x,y+1,u^*(x,y+1)) - f(x-1,y+1,u^*(x-1,y+1)) \\ &\le f(x,y+1,u^*(x-1,y+1)) - f(x-1,y+1,u^*(x-1,y+1)) \\ &= h + \left(s-u^*(x-1,y+1)\right)\mu\Delta^x J(x-1,y) + u^*(x-1,y+1)\mu\Delta^x J(x,u^*(x-1,y+1)-1) \\ &\quad + \sum_{i=1}^{n}\Delta^x T_{R_i}(x-1,u^*(x-1,y+1))\end{aligned}$$

We multiply the second inequality by -1 and add to the first one. Then,

$$\begin{aligned}&\Delta^x T\left(J(x,y)\right) - \Delta^x T\left(J(x-1,y+1)\right) \\ &\quad \ge \left(s-u^*(x+1,y)\right)\mu\Delta^x J(x,y) - \left(s-u^*(x-1,y+1)\right)\mu\Delta^x J(x-1,y) \\ &\quad + u^*(x+1,y)\mu\Delta^x J(x+1,u^*(x+1,y)-1) - u^*(x-1,y+1)\mu\Delta^x J(x,u^*(x-1,y+1)-1) \\ &\quad + \sum_{i=1}^{n}\Delta^x T_{R_i}(x,u^*(x+1,y)) - \sum_{i=1}^{n}\Delta^x T_{R_i}(x-1,u^*(x-1,y+1))\end{aligned}$$

For $i \in \{1,2,...,n\}$, we have $\Delta^x T_{R_i} J(x-1,u^*(x-1,y+1)) = \Delta^x T_{R_i} J(x-1,u^*(x+1,y+1))$ due to part-i of Theorem 1 and part-iii of Corollary 1.

Therefore, by the hypothesis

$$\begin{pmatrix} \lambda_1\Delta^x J(x-1,u^*(x+1,y)) + \sum_{i=2}^{n}\lambda_i\Delta^x T_{R_i} J(x,u^*(x+1,y)) \\ -\left(\lambda_1\Delta^x J(x-2,u^*(x-1,y+1)) + \sum_{i=2}^{n}\lambda_i\Delta^x T_{R_i} J(x-1,u^*(x-1,y+1))\right)\end{pmatrix} \ge 0 .$$

Moreover,

$\left(s-u^*(x+1,y)\right)\mu\Delta^x J(x,y) = \left(s-u^*(x-1,y+1)\right)\mu\Delta^x J(x,y) + \left(u^*(x-1,y+1)-u^*(x+1,y)\right)\mu\Delta^x J(x,y)$ and $\left(s-u^*(x-1,y+1)\right)\mu\left(\Delta^x J(x,y)-\Delta^x J(x-1,y)\right) \ge 0$. The remaining terms of the right hand side is greater or equal to the zero which can be shown by applying the operator technique once more.

Proof of Theorem 4:

i. If $p+\mu\Delta J(x) < 0$, then it is optimal to have $u^*(x) = s$ in order to minimize cost function as much as possible. Otherwise, $u^*(x)$ should be zero because having a positive number of operational servers would inflate the cost.

ii. Under maximization objective, Cil et al. (2009) show that both the production and the rationing operators are concave. It is easy to adapt their results to our case (where the value functions are cost-to-go functions) and show that $J(x)$ is a convex function of x.

iii. Since $\Delta J(x)$ is non-decreasing (from part ii), there exists a threshold rationing inventory level $K^i := \min\{x : \Delta J(x) \geq -c_i\}$

iv. For a specific value of L, Cil et al. (2009) show that $v(x-1) - v(x) \geq L$ where $v(x)$ is the reward function. Letting $J(x) = -v(x)$ and $L = -c_1$ is sufficient to show that $\Delta J(x-1) \geq -c_1$.

v. Immediately follows from parts i and ii.

Proof of Lemma 1:

We have $P_0(S) = \left(\sum_{j=0}^{S} \frac{S!}{(S-j)!} \left(\frac{\mu}{\lambda} \right)^j \right)^{-1}$,

$\Delta C(S, \vec{0}) = C(S+1, \vec{0}) - C(S, \vec{0}) = h + \left(\frac{h\lambda}{\mu} + \sum_{i=1}^{n} \lambda_i c_i \right) \Delta P_0(S)$, and

$\Delta^2 C(S, \vec{0}) = \Delta C(S+1, \vec{0}) - \Delta C(S, \vec{0}) = \left(\frac{h\lambda}{\mu} + \sum_{i=1}^{n} \lambda_i c_i \right) \Delta^2 P_0(S)$. To conclude that $\Delta^2 C(S, \vec{0}) > 0$ we first need to compute $\Delta P_0(S)$ and then $\Delta^2 P_0(S)$:

$$\Delta P_0(S) = P_0(S+1) - P_0(S) = \frac{1}{\sum_{j=0}^{S+1} \frac{(S+1)!}{(S+1-j)!} \left(\frac{\mu}{\lambda} \right)^j} - \frac{1}{\sum_{j=0}^{S} \frac{S!}{(S-j)!} \left(\frac{\mu}{\lambda} \right)^j}$$

$$= \frac{\sum_{j=0}^{S} \left(\frac{S!}{(S-j)!} - \frac{(S+1)!}{(S+1-j)!} \right) \left(\frac{\mu}{\lambda} \right)^j - (S+1)! \left(\frac{\mu}{\lambda} \right)^{S+1}}{\left(\sum_{j=0}^{S+1} \frac{(S+1)!}{(S+1-j)!} \left(\frac{\mu}{\lambda} \right)^j \right) \left(\sum_{j=0}^{S} \frac{S!}{(S-j)!} \left(\frac{\mu}{\lambda} \right)^j \right)}$$. That is,

$$\Delta P_0(S) = \frac{\sum_{j=0}^{S}\left(\frac{S!}{(S-j)!} - \frac{(S+1)(S)!}{(S+1-j)(S-j)!}\right)\left(\frac{\mu}{\lambda}\right)^j - (S+1)!\left(\frac{\mu}{\lambda}\right)^{S+1}}{\left(\sum_{j=0}^{S+1}\frac{(S+1)!}{(S+1-j)!}\left(\frac{\mu}{\lambda}\right)^j\right)\left(\sum_{j=0}^{S}\frac{S!}{(S-j)!}\left(\frac{\mu}{\lambda}\right)^j\right)}$$

$$= \frac{\sum_{j=0}^{S+1}\frac{S!}{(S-j)!}(-j)\left(\frac{\mu}{\lambda}\right)^j}{\left(\sum_{j=0}^{S+1}\frac{(S+1)!}{(S+1-j)!}\left(\frac{\mu}{\lambda}\right)^j\right)\left(\sum_{j=0}^{S}\frac{S!}{(S-j)!}\left(\frac{\mu}{\lambda}\right)^j\right)}$$

$$\Delta^2 P_0(S) = \Delta P_0(S+1) - \Delta P_0(S)$$

$$\left(\sum_{j=0}^{S}\frac{S!}{(S-j)!}\left(\frac{\mu}{\lambda}\right)^j\right)\left(\sum_{j=0}^{S+2}\frac{(S+1)!}{(S+1-j)!}(-j)\left(\frac{\mu}{\lambda}\right)^j\right) -$$

$$= \frac{\left(\sum_{j=0}^{S+2}\frac{(S+2)!}{(S+2-j)!}\left(\frac{\mu}{\lambda}\right)^j\right)\left(\sum_{j=0}^{S+1}\frac{S!}{(S-j)!}(-j)\left(\frac{\mu}{\lambda}\right)^j\right)}{\left(\sum_{j=0}^{S+2}\frac{(S+2)!}{(S+2-j)!}\left(\frac{\mu}{\lambda}\right)^j\right)\left(\sum_{j=0}^{S+1}\frac{(S+1)!}{(S+1-j)!}\left(\frac{\mu}{\lambda}\right)^j\right)\left(\sum_{j=0}^{S}\frac{S!}{(S-j)!}\left(\frac{\mu}{\lambda}\right)^j\right)}$$

$$\left(\sum_{j=0}^{S}\frac{S!}{(S-j)!}\left(\frac{\mu}{\lambda}\right)^j\right)\left(\sum_{j=0}^{S+1}\frac{(S+1)!}{(S+1-j)!}(-j)\left(\frac{\mu}{\lambda}\right)^j - (S+2)!\left(\frac{\mu}{\lambda}\right)^{S+2}\right)$$

$$= \frac{-\left(\sum_{j=0}^{S}\frac{(S+2)!}{(S+2-j)!}\left(\frac{\mu}{\lambda}\right)^j + (S+2)!\left(\frac{\mu}{\lambda}\right)^{S+1} + (S+2)!\left(\frac{\mu}{\lambda}\right)^{S+2}\right)\left(\sum_{j=0}^{S+1}\frac{S!}{(S-j)!}(-j)\left(\frac{\mu}{\lambda}\right)^j\right)}{\left(\sum_{j=0}^{S+2}\frac{(S+2)!}{(S+2-j)!}\left(\frac{\mu}{\lambda}\right)^j\right)\left(\sum_{j=0}^{S+1}\frac{(S+1)!}{(S+1-j)!}\left(\frac{\mu}{\lambda}\right)^j\right)\left(\sum_{j=0}^{S}\frac{S!}{(S-j)!}\left(\frac{\mu}{\lambda}\right)^j\right)}$$

$$\Delta^2 P_0(S) = \frac{\begin{array}{l}\left[\left(\sum_{j=0}^{S}\frac{(S+2)(S+1)S!}{(S+2-j)(S+1-j)(S-j)!}\left(\frac{\mu}{\lambda}\right)^j\right)\left(\sum_{j=0}^{S+1}\frac{(S)!}{(S-j)!}(j)\left(\frac{\mu}{\lambda}\right)^j\right)\right. \\ \left.-\left(\sum_{j=0}^{S}\frac{(S)!}{(S-j)!}\left(\frac{\mu}{\lambda}\right)^j\right)\left(\sum_{j=0}^{S+1}\frac{(S+1)S!}{(S+1-j)(S-j)!}(j)\left(\frac{\mu}{\lambda}\right)^j\right)\right] \\ +\left[(S+2)!\left(\frac{\mu}{\lambda}\right)^{S+2}\left(\sum_{j=0}^{S+1}\frac{(S)!}{(S-j)!}(j)\left(\frac{\mu}{\lambda}\right)^j-\sum_{j=0}^{S}\frac{(S)!}{(S-j)!}\left(\frac{\mu}{\lambda}\right)^j\right)\right] \\ +\left[(S+2)!\left(\frac{\mu}{\lambda}\right)^{S+1}\sum_{j=0}^{S+1}\frac{(S)!}{(S-j)!}(j)\left(\frac{\mu}{\lambda}\right)^j\right]\end{array}}{\left(\sum_{j=0}^{S+2}\frac{(S+2)!}{(S+2-j)!}\left(\frac{\mu}{\lambda}\right)^j\right)\left(\sum_{j=0}^{S+1}\frac{(S+1)!}{(S+1-j)!}\left(\frac{\mu}{\lambda}\right)^j\right)\left(\sum_{j=0}^{S}\frac{S!}{(S-j)!}\left(\frac{\mu}{\lambda}\right)^j\right)}$$

It can be easily concluded that all the terms in brackets are positive, i.e. $\Delta^2 P_0(S) > 0$. Thus, $C\left(S,\vec{0}\right)$ is a convex function of S.