Coordination of Production and Inventory for Achieving Service Quality

By Taisuke Shimamoto

A thesis submitted to
School of Information Science,
Japan Advanced Institute of Science and Technology,
in partial fulfillment of the requirements
for the degree of
Master of Information Science
Graduate Program in Information Science

Written under the direction of
Professor Milan Vlach

February 15, 1999

Contents

Chapter 1

Introduction

1.1 Background

"Customer Satisfaction" is one of the key issues in the business field today. Reflecting the abundance of products and the stagnancy of the domestic economy, consumers become more selective with purchasing. Keeping delivery promises is often considered to be the main performance measure for the quality of customer service. To respond to the customers' needs quickly, most retailers are now managing their business process with POS (Point Of Sales) system. The retailers consequently request their vendors to supply "needed items in needed amount at needed time." To meet such requests from the retailers and minimize cost, it is necessary for the vendors, i.e. the producers, to have careful policies for the production and inventory control. The producers would like to suppress the inventory cost as much as possible without making their customers wait too long. The producers promise a fixed delivery time to the customers. Sox, Thomas and McClain [5] called it *service window* (its length denoted T). The objective of the producers is to maximize the number of the orders filled within T with keeping the inventory level as low as possible. To quantify the level of the customer service regarding delivery time, *fill rate* is defined as the fraction of demand satisfied within the service window. The different kinds of finished goods are stored in the inventory which has a limited capacity. Once the total basestock level is specified, basestock allocation which maximizes the fill rate can be found using a proper method. Sox et al. proposed an algorithm of the basestock allocation for $M/M/1$ production-inventory model. $M/M/1$ represent a queueing system with the exponential distribution concerning the intervals of customer arrival and production. They also considered the production priority rules other than FIFO and showed the improvement in the fill rate.

1.2 Objectives of Research

The model proposed by Sox et al. is easily applicable but it provides optimal or approximate solutions for a rather limited number of practical situations. The main purpose of the thesis is to extend the model so that it is useful for a larger variety of real business situations and test the worthiness of the proposed extensions. To achieve these objectives, the original model of Sox et al. is extended in the following three directions.

First, the original model is restricted to exponential production time distributions. Here a larger class of production time distribution is considered, namely k-stage Erlang distributions.

Second, the original model is applicable only to situations with a single-location finished-goods inventory facility. Moreover, it does not permit limits on the individual basestock levels on subgroup of products. The proposed extension includes both the possibility of several inventory facilities and limits on the levels of basestock for subsets of products. The basestock allocation algorithm is considered for such partitioned model.

Third, Sox et al. introduced a priority rule for production, BackOrder-gets-first-Priority (BOP), and improved the fill rate over FIFO with a fixed service window. It is found that BOP is efficient production rule overall, but works negatively on the fill rate in some conditions. The proposed rule, *Modified BOP*, is expected to solve that problem and achieve higher fill rate than BOP.

1.3 Composition of Thesis

The thesis has the following composition. In Chapter 2, the basics of queueing theory, general concepts of production-inventory model and the model of Sox et al. are described for the help to understand following chapters. In Chapter 3 and Chapter 4, the extensions of production time distributions and inventory constraints are described, respectively. In Chapter 5, a new production rule Modified BOP is introduced, and its improvement over BOP and FIFO is shown. Finally in Chapter 6, the reviews of the research and the concluding remarks are stated.

Chapter 2

Preliminary

In this chapter, first the basics of the queueing theory are described, next the general concepts of the production-inventory model are outlined, then finally the model proposed by Sox et al. is explained.

2.1 Basics of Queueing Theory

The model described here is based on queueing theory. Its basic concepts related to the model are summarized as follows.

2.1.1 $M/M/1$ Model

The simplest queueing model is the "$M/M/1$" model. $M/M/1$ is known as Kendall's notation. The first "M" means that the interarrival times of customers follow exponential distribution. Customers arrive totally at random in this case. The second "M" means that service times also follow the exponential distribution. The distribution curve of the exponential is shown in Fig. 2.1. In the exponential distribution, the probability is highest when time interval is 0. The last "1" represents the number of servers. If the $M/M/1$ model is applied to the inventory model, it is reasonable to assume the exponential distribution for the customer's arrival but sometimes unreasonable [4, p.256] for the service time because a certain time lapse is necessary to give the services.

The probability distribution for this model is considered as follows. Let λ and $n(T)$ represent the average arrival rate of customers and the number of customer arrivals in the period T, respectively. Then the probability distribution of the number of customers arriving in a finite time interval $[0, T]$ has following Poisson distribution:

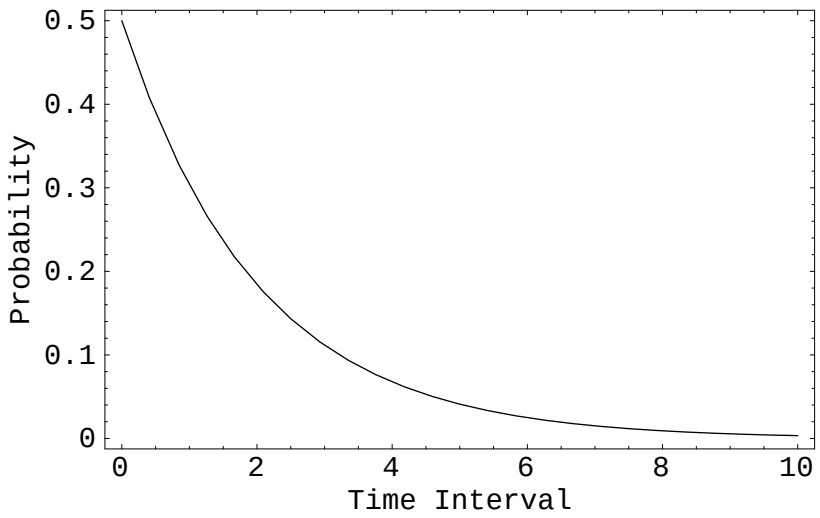


Figure 2.1: Exponential Distribution ($1/\mu = 2$)

$$P[n(T) = i] = \frac{(\lambda T)^i}{i!} e^{-\lambda T}$$

When the service rate is denoted by μ, the probability of finding n customers in the system (i.e. queue + server) is calculated as follows:

$$p_n = (1 - \rho)\rho^n$$

where

$$\rho = \frac{\lambda}{\mu}.$$

Expected number of customers $E[N(t)]$ in the systems is

$$E[N(t)] = \frac{\rho}{1 - \rho} = \frac{\lambda}{\mu - \lambda}$$

where $N(t)$ is the number of customers in the system at time t.

2.1.2 $M/E_k/1$ Model

This model has the Erlang distribution for the service time. The probability density function $f(t)$ of the k-stage Erlang distribution is

$$f(t) = \frac{(\mu k)^k}{(k-1)!} t^{k-1} e^{-k\mu t}.$$

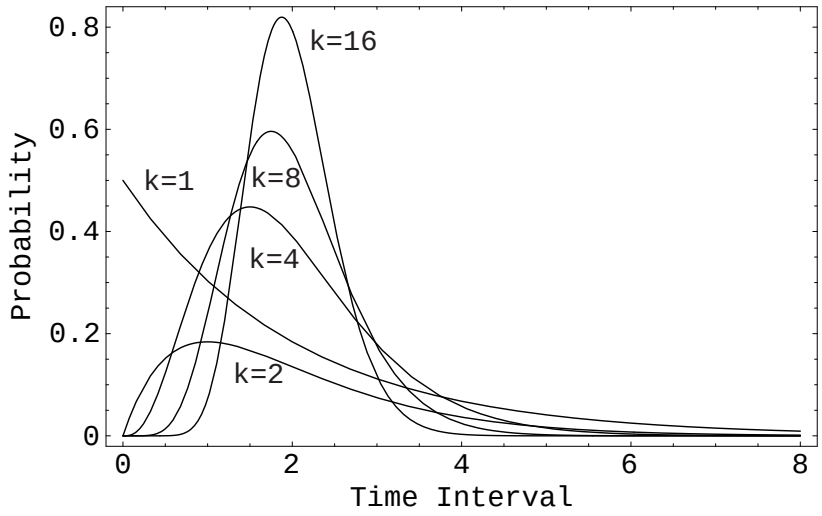


Figure 2.2: Erlang Distribution ($1/\mu = 2$)

An equivalent representation of the k-stage Erlang distribution is shown in Fig. 2.3.

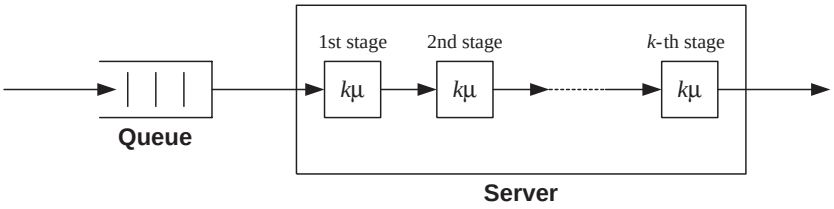


Figure 2.3: The k-stage Erlang distribution represented in terms of the exponential server

In this system, the service facility consists of k exponential servers in tandem, each of which has the service rate $k\mu$. In this representation, when a customer departs by exiting from the last (i.e., k-th) stage, a new customer may enter and proceed one stage at a time. The total time that a customer spends in the k-stage facility is distributed according to the k-stage Erlang distribution. When k is equal to 1, the distribution is exactly same to the exponential distribution. As k approaches infinity, the limit of this density function approaches a unit impulse function at the point $t = 1/\mu$. The derivation of the probability distribution function is as follows. First let consider a Poisson process with rate $k\mu$. If every k-th service is selected, then this forms a service process, in which service times have the k-stage

Erlang distribution with mean $1/\mu$. Figure 2.4 illustrates the case in which $k = 3$.

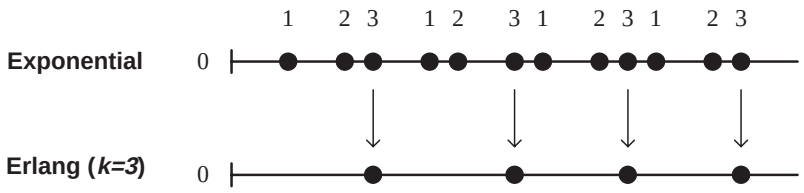


Figure 2.4: The Poisson services and the Erlang services

From Fig. 2.4 the probability that more than m customers are serviced during the interval $[0, T]$ is

$$P[n(T) \geq m] = \sum_{m=0}^{\infty} \sum_{l=km}^{k(m+1)-1} e^{-k\mu T} \frac{(k\mu T)^l}{l!}.$$

(Since this equation can be applied to the customer arrivals also by replacing μ with λ, the same notation $n(T)$ is used here.) According to Kobayashi [3, p.196], the probability distribution of finding n customers in the system is

$$p_n = (1-\rho) \sum_{j=0}^{n} (-1)^{n-j} r^{n-j-1} R^{kj} \left[\binom{kj}{n-j} r + \binom{kj}{n-j-1} \right]$$

where

$$R = 1 + \frac{\rho}{k}, \quad r = 1 - R^{-1}.$$

One can derive from the basic queueing theory [2] that the expected number of customers in the system $E[N(t)]$ and the expected waiting time W are

$$E[N(t)] = \frac{(1+1/k)\,\rho^2}{2(1-\rho)} + \rho = \frac{1+k}{2k} \frac{\lambda^2}{\mu(\mu-\lambda)} + \frac{\lambda}{\mu}$$

$$W = \frac{1+k}{2k} \frac{\lambda}{\mu(\mu-\lambda)} + \frac{1}{\mu}.$$

Now consider approximating p_n according to the method of Buzacott and Shanthikumar [1, p.76]. Because the time-average probability that no customer is found at the server is $1-\rho$, $p_0 = 1-\rho$ is obtained. Assuming that remaining probabilities have a geometric form (i.e., $p_n = a\sigma^{n-1}, n = 1, 2, \dots$)

and using the requirement that the total probability should be one, $a = \rho(1-\sigma)$ is obtained. This approximated distribution has a mean of $\rho/(1-\sigma)$. Suppose the mean number of customers in the system is approximated with $E[N(t)]$. Then requiring that this approximation be the same as the mean of the approximate distribution of the number of customers in the system, the following relations are obtained:

$$p_n \approx \tilde{p}_n = \begin{cases} 1-\rho & n = 0 \\ \rho(1-\sigma)\sigma^{n-1} & n = 1, 2, \ldots \end{cases}$$

where

$$\sigma = \frac{E[N(t)] - \rho}{E[N(t)]}.$$

Figure 2.5 shows that this approximation curve with $k = 3$ fits well to the pure Erlang distribution.

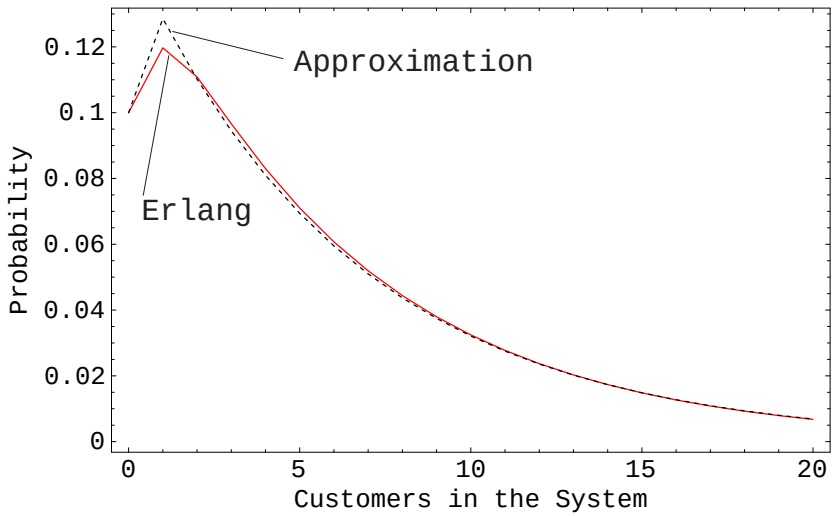


Figure 2.5: Erlang distribution and its approximation curves ($k = 3$)

2.2 Production-Inventory Model

One can find many variations in the styles of the production and inventory management in the business field. Fig. 2.6 illustrates the components of a typical production-inventory model.

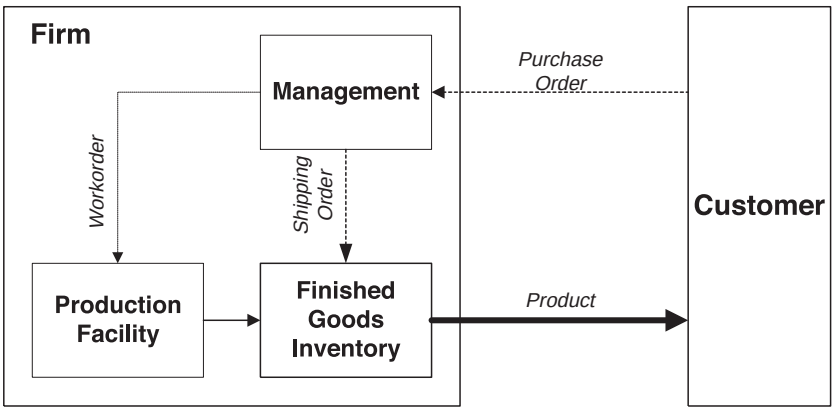


Figure 2.6: Model Components

The model consists of a single production facility and a single inventory for finished goods. There is a management department which dispatches internal orders to the production facility and the inventory. The inventory holds *basestocks* (stocks to be always kept in the inventory) to deliver products to the customers in the shorter service window. Once an unit of basestock is delivered to a customer, the management immediately issues a production order to the production facility for replenishment. This production style is sometimes called "produce-to-stock" model [1]. The model is going to be used for mathematical analyses throughout the thesis.

It is assumed that a customer arrives and orders single unit of product only at a time. No more than one customers arrive simultaneously. When a customer issues a purchase order and if there is no stock, the customer waits until his order is produced. If more than one customer wait for the same product, FIFO rule is applied.

Production facility usually has multiple independent production lines. However, a simple case with single production line like Fig. 2.7 is chosen here. Different items are produced using the same production line. The average production rate μ is same for all items. The management receives purchase orders from the customers and then issues *workorders* (production orders) to the production facility. A received purchase order is never buffered but immediately directed to the production facility as a workorder. This means

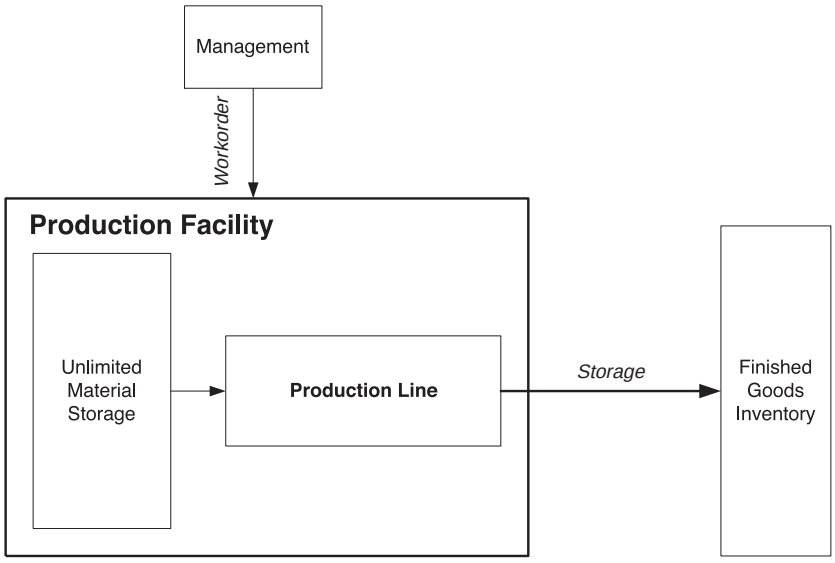


Figure 2.7: Production Facility

that no more than one workorders are issued simultaneously. Stocked items are immediately delivered to the customers when the management receives purchase orders for the item. Delivering interval is not considered.

The inventory (Fig. 2.8) holds all the products produced at the production facility.

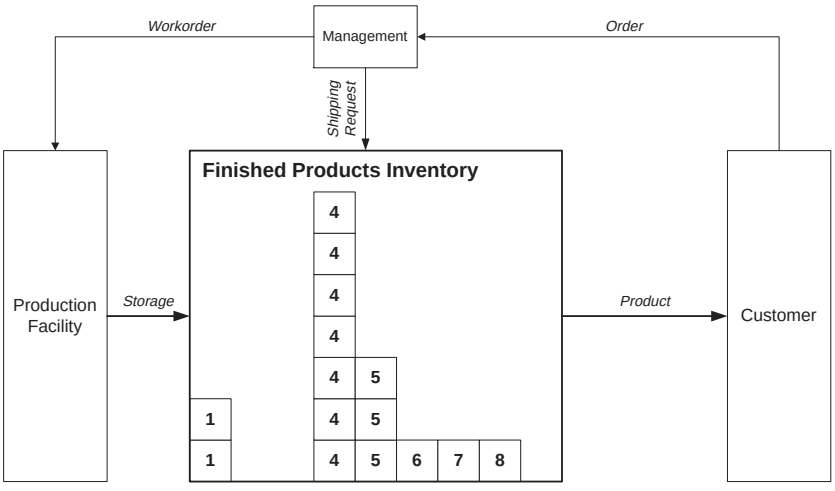


Figure 2.8: Finished Goods Inventory

The inventory has its capacity, but here it is assumed that the management can set total stock limit within the capacity because of the change of business environment. To maximize the fill rate to the customers, basestocks are filled up to the total stock limit. In Fig. 2.8, basestocks are allocated 2 for item 1, 0 for item 2 and 3, 7 for item 4, and so on. Here, the optimum basestock allocation to maximize fill rate (described next) is considered.

By the negotiation with customers, the management sets the service window T: the interval from the time of purchase order receipt to the time of delivery. The service window is a fixed value. Each customer orders an unit of products, and if the product is delivered within the service window, it is considered that the order is filled (satisfied). The fill rate can be calculated from the data of the transaction as follows:

$$(\text{Fill Rate}) = \frac{(\text{Filled Orders})}{(\text{Total Orders})}.$$

It approximates the probability that the customers are served within T.

The terminology used in general queueing theory corresponds to the one used in the production-inventory model in the following way:

Notation	Queueing Theory	The Model in the Thesis
	customer	workorder
	system	production facility and queue
λ	customer arrival	total demand
μ	service	production rate

Assumptions for the Model

The assumptions for this production-inventory model are summarized as follows:

- **There are no order cancellation from the customers**
- **The time lapse between the model components is zero**
  Time is consumed only in the production process.
- **All items are produced by same single production line with same average production rate**
- **The management doesn't accept more than one purchase orders at a time**

- **If there are more than one customers waiting for the same kind of product, their orders are processed by FIFO rule. If there are more than one customers waiting for the different kinds of products, and if the ordered items are buffered in the queue at the production line, each customer can receive the product as soon as its production is completed**

- **The average production time of every item is the same**
  Average production time is $1/\mu$ (μ = average production rate).

- **The average total demand never exceeds the average production rate**
  Using average demand rate λ and production rate μ, the relationship is written as $\lambda < \mu$. If $\lambda \geq \mu$, so that the average demand rate is greater than or equal to the average production rate, then the length of the queue would "explode" and grow without bound.

2.3 The Model of Sox et al.

Base on the $M/M/1$ production-inventory model, Sox et al. proposed a calculation method of the fill rate and an algorithm of basestock allocation which maximize the fill rate. Stepwise explanations of their method and a proof for the algorithm given by the author are shown as follows. The fill rate represents the probability that an order is satisfied within the service window T. In other words, the probability is the remainder of the probability that an order is "not" satisfied, such as

$$\text{(Fill Rate)} \quad = \quad 1 - \text{(Probability of Late-Delivery)}.$$

The probability that an order is not satisfied, i.e. late-delivery rate, is calculated first to obtain the fill rate. The events of the late-delivery occur under following conditions. Consider a production-inventory model illustrated in Fig. 2.9.

When focusing on the transactions of the item 2, for example, the two basestocks are held at the initial stage (**a**). At time $t1$, customer A arrives, orders item 2 and receives the product immediately. A workorder is issued at the same time and joined at the end of the queue (**b**). Next customer B comes in, orders the same item and receives it. Another workorder is joined at the end of the queue (**c**). The basestock for item 2 is empty when customer C arrives. His order becomes a backorder and is joined at the end of the queue (**d**). The order is filled if the number of the workorder for item 2 ahead of

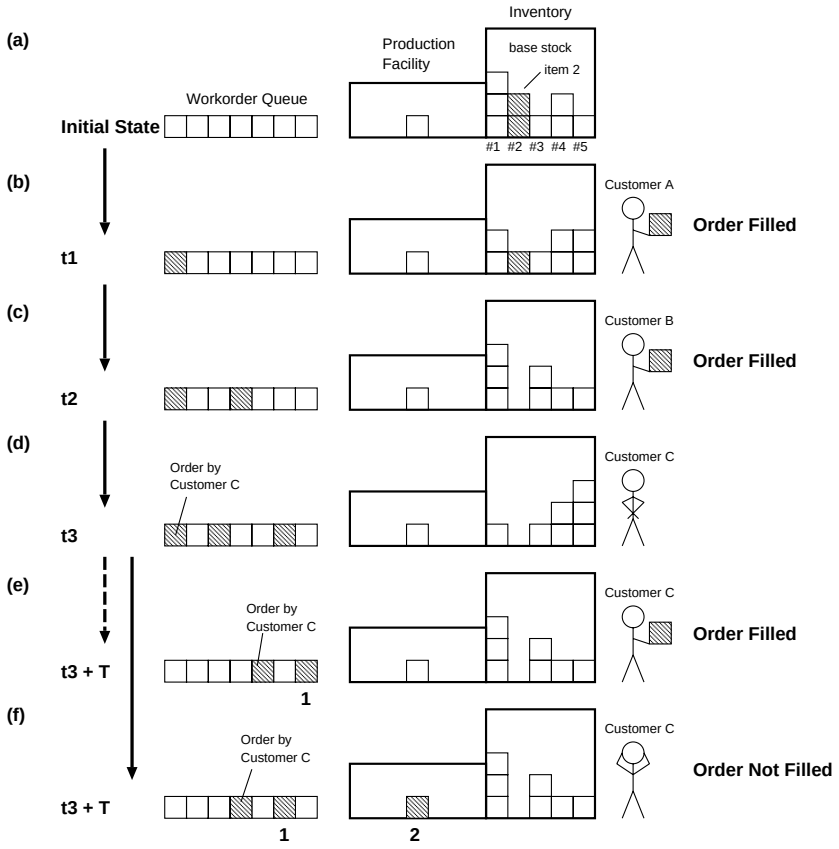


Figure 2.9: Schematic Description of Late Delivery

the order issued by customer C is smaller than the basestock level (in this case, 2) T time unit after his arrival (**e**). On the other hand, the order is not filled because the number of the workorders ahead of the order issued by customer C is greater than or equal to the basestock level (**f**). Thus, the late-delivery rate is said to be the probability that the number of workorder ahead of a backorder is greater than or equal to the basestock level T time unit after the customer of that backorder arrives.

Sox et al. showed the calculation method of the fill rate for the $M/M/1$ model. The summary of the method is shown as follows. Define $P(i, x)$ as the steady-state probability that workorder for item i has x (≥ 0) other workorders for the same item ahead of it T time unit after its arrival. The probability that n workorders exist in the system is $(1-\rho)\rho^n$. This is given by the basics of the queueing theory as described in the previous chapter. Suppose that m workorders are processed during time T. Then the probability

that there are x other workorders of item i becomes

$$\sum_{n=x+m}^{\infty} (1-\rho)\rho^n P_i^x (1-P_i)^{n-m-x} \binom{n-m}{x}$$

where $P_i = \lambda_i/\lambda$. The probability of m workorders being processed during T is given by Poisson distribution $\sum_{m=0}^{\infty} e^{-\mu T}(\mu T)^m/m!$. Then $P(i,x)$ is obtained by multiplying the two probabilities as follows:

$$P(i,x) = \sum_{m=0}^{\infty} e^{-\mu T} \frac{(\mu T)^m}{m!} \sum_{n=x+m}^{\infty} (1-\rho)\rho^n P_i^x (1-P_i)^{n-m-x} \binom{n-m}{x}.$$

The inner sum (denoted $s(m,x)$) reduces to

$$s(m,x) = (1-\gamma_i)\gamma_i^x \rho^m, \text{ where } \gamma_i = \frac{\rho P_i}{1-\rho(1-P_i)}.$$

As a result,

$$P(i,x) = (1-\gamma_i)\gamma_i^x e^{-\mu T(1-\rho)} \quad \text{for } x \geq 0$$

A late delivery occurs if $x \geq S_i$, so for item i,

$$FR_{(i,T)}(S_1,\ldots,S_K) = 1 - \sum_{x=S_i}^{\infty} (1-\gamma_i)\gamma_i^x e^{-\mu T(1-\rho)} = 1 - \gamma_i^{S_i} e^{-\mu T(1-\rho)}$$

The aggregate fill rate is the sum, weighting each item by its fraction of total demand P_i:

$$FR_T(S_1,\ldots,S_K) = 1 - e^{-\mu T(1-\rho)} \sum_{i=1}^{K} P_i \gamma_i^{S_i}.$$

Following problem is considered to allocate S units of basestock to maximize the fill rate:

Maximize

$$FR_T(S_1,S_2,\ldots,S_K) = 1 - e^{-\mu T(1-\rho)} \sum_{i=1}^{K} P_i \gamma_i^{S_i} \tag{2.1}$$

subject to

$$\sum_{i=1}^{K} S_i \leq S. \tag{2.2}$$

The increment of the fill rate when the number of basestock is increased by one is

$$FR_T(S_1,\ldots,S_i+1,\ldots,S_K) - FR_T(S_1,\ldots,S_i,\ldots,S_K) = P_i(1-\gamma_i)\gamma_i^{S_i}e^{-\mu T(1-\rho)} = \frac{(1-\rho)}{\rho}\gamma_i^{S_i+1}e^{-\mu T(1-\rho)}.$$

The fill rate when there is no stock in the inventory is

$$FR_T(0,0,\ldots,0) = 1 - e^{-\mu T(1-\rho)}.$$

Then (2.1) can be rewritten as follows:

$$FR_T(S_1,S_2,\ldots,S_K) = 1 - e^{-\mu T(1-\rho)} + \frac{(1-\rho)}{\rho}e^{-\mu T(1-\rho)}\sum_{i=1}^{K}\sum_{j=1}^{S_i}\gamma_i^j.$$

Since $\frac{(1-\rho)}{\rho}e^{-\mu T(1-\rho)}$ is positive constant, the optimization problem (2.1)-(2.2) is simplified as follows:

Maximize

$$\mathcal{FR}_T(S_1,S_2,\ldots,S_K) = \sum_{i=1}^{K}\sum_{j=1}^{S_i}\gamma_i^j \tag{2.3}$$

subject to

$$\sum_{i=1}^{K} S_i \le S. \tag{2.4}$$

Proof for the Stock Allocation Algorithm

Sox et al. stated that the optimum allocation of the basestocks may be obtained by sequentially adding 1 to S_i for the item with the largest value of $\gamma^{S_i^{(m)}+1}$. A proof for this algorithm is given as follows. First consider a generalized algorithm as follows. Let Λ be a set of non-negative real numbers $\lambda_{i,j}$ such that

$$\Lambda = \{\lambda_{i,j} |\ 1 \le i \le U,\ 1 \le j \le V\} \tag{2.5}$$

where

$$\lambda_{i,j} > \lambda_{i,j'} \quad \text{if} \quad j < j', \tag{2.6}$$

and let $\hat{\Lambda}_m$ represent a subset of Λ which consists of m largest elements of Λ. It is obvious that $\lambda_{i,j} \in \hat{\Lambda}_m$ if $\lambda_{i,j+1} \in \hat{\Lambda}_m$. Moreover, $\lambda_{i,j'} \in \hat{\Lambda}_m$ for $1 \leq j' \leq j$ if $\lambda_{i,j+1} \in \hat{\Lambda}_m$. Thus there exist U non-negative integers $V_1^{(m)}, V_2^{(m)}, \ldots, V_U^{(m)}$ which satisfy $\sum_{i=1}^{U} V_i^{(m)} = m$, and such that $\hat{\Lambda}_m$ can be represented as follows:

$$\hat{\Lambda}_m = \{\lambda_{i,j} |\ 1 \leq i \leq U,\ 1 \leq j \leq V_i^{(m)}\}. \tag{2.7}$$

A following algorithm can be considered to obtain m largest elements from Λ. First set $\hat{\Lambda}_0 = \phi$ and $V_i^{(0)} = 0$ for all i for initialization. Suppose that $\hat{\Lambda}_m$ and $V_1^{(m)}, \ldots, V_K^{(m)}$ which satisfy (2.7) are given, then $\hat{\Lambda}_{m+1}$ can be obtained by the following operation:

$$\begin{aligned} \hat{\Lambda}_{m+1} &= \hat{\Lambda}_m \cup \max_{1 \leq i \leq U} \{\lambda_{i,(V_i^{(m)}+1)}, \lambda_{i,(V_i^{(m)}+2)}, \ldots, \lambda_{i,V}\} \\ &= \hat{\Lambda}_m \cup \max_{1 \leq i \leq U} \{\lambda_{i,(V_i^{(m)}+1)}\}. \end{aligned} \tag{2.8}$$

Thus m largest elements are selected by sequentially picking up a largest element from yet unselected elements in Λ. Using this property of the generalized problem, we obtain next lemma for the stock allocation problem.

Lemma 1. *The algorithm, sequentially adding 1 to S_i for the item with the largest value of $\gamma^{S_i^{(m)}+1}$, gives the optimum set $\{S_1, S_2, \ldots, S_K\}$ which maximizes $\mathcal{FR}_T$.*

Proof. First consider a set , which is defined as

$$, = \{\gamma_i^j \mid 1 \leq i \leq K,\ 1 \leq j \leq S\}. \tag{2.9}$$

Since $0 < \gamma_i < 1$, the following relation is observed:

$$\gamma_i^j > \gamma_i^{j'} \quad \text{if} \quad j < j'. \tag{2.10}$$

The two properties (2.9) and (2.10) exactly correspond to (2.5) and (2.6), respectively. Then the subset $\hat{,}_S$ of , which consists of S largest elements of , can be represented as follows. There exist K and $S_1, S_2, \ldots, S_K$ which satisfy (2.4) and such that

$$\hat{,}_S = \{\gamma_i^j \mid 1 \leq i \leq K,\ 1 \leq j \leq S_i\}.$$

From (2.3), the maximum value of $\mathcal{FR}_T$ is obtained from the sum of S elements of γ_i^j. Then the sum of the elements of $\hat{,}_S$ is greater than or equal to $\mathcal{FR}_T$, i.e.,

$$\mathcal{FR}_T(S_1, S_2, \ldots, S_K) \leq \sum_{\gamma_i^j \in \hat{\Gamma}_S} \gamma_i^j$$

This implies that maximizing $\mathcal{FR}_T$ is equivalent to finding $\hat{,}_S$. From (2.8), $\hat{,}_S$ is obtained by sequentially selecting an element with largest value of $\gamma_i^{S_i^{(m)}+1}$. ■

The algorithm of basestock allocation is summarized as follows:

BSA : BaseStock Allocation Problem
Input: $\gamma_1, \gamma_2, \ldots, \gamma_K$, and S.
Output: $S_1, S_2, \ldots, S_K$ such that

$$\sum_{i=1}^{K} S_i \leq S,$$

and $FR_T(S_1, S_2, \ldots, S_K)$ is maximized.

```
Algorithm BSA (γ_1, γ_2, ..., γ_K, S)
  for i = 1 to K
      S_i := 0
  while (Σ_{i=1}^{K} S_i < S)
    do
      γ_max = max_i{γ_i^{S_i+1}}
      for i = 1 to K
          if (γ_i^{S_i+1} == γ_max) then j := i
      S_j := S_j + 1
    od
  return(S_1, S_2, ..., S_K)
```

2.4 Terms and Definitions

The terms and definitions used to describe the production-inventory model are summarized in the following table:

Notation	Term	Definition
	(purchase) order	A purchase order from a customer. Each customer can order one product only at a time among different kinds of products. Different customers cannot order simultaneously.
	workorder	A production order dispatched to the production facility by the management. The management can order one product only at a time. The number of workorders is equal to that of (purchase) order.
	basestock	A planned stock level to be maintained. When a product is shipped to a customer, the management immediately sends the workorder to the production facility to keep the basestock level
i	**(product) item**	Product item number ($i = 1, 2, \ldots, K$)
$N_i(t)$	**workorders in the system**	The number of workorders in the system for product i at time t. The "system" includes the waiting queue for production and the single-stage production facility. Only a workorder can enter the production facility at a time.
$N(t)$	**total workorder**	Total workorders in the system at time t. $N(t) = \sum_i^K N_i(t)$
S_i	**basestock level**	The basestock level for product i. The level is equal to the number of the product i in the inventory.
S	**limit of total basestock level**	The upper limit of total basestock level. $\sum_{i=1}^K S_i \leq S$

(continued from previous page)

Notation	Term	Definition
T	**service window**	The time interval from the point of order to the point of the delivery to customers. The service window is a fixed time and is decided by the management.
$FR_{(i,T)}$	**fill rate for** i	Fraction of the orders of the product i delivered to the customers within T.
FR_T	**fill rate**	Fraction of total orders delivered to the customers within T.
$L_{(i,T)}$	**late-delivery rate for** i	Fraction of the orders of the product i "not" delivered to the customers within T.
L_T	**late-delivery rate**	Fraction of total orders "not" delivered to the customers within T.
λ_i	**demand**	Average arriving orders per unit time for product i.
λ	**total demand**	Average arriving orders per unit time. $\lambda = \sum_{i=1}^{K} \lambda_i$
μ	**production rate**	Average produced orders per unit time. The rate is fixed and same for all kinds of products.
ρ	**utilization**	The utilization of production facility. $\rho = \lambda/\mu$
P_i	**fraction of demand**	The probability that an arriving order is for product i. $P_i = \lambda_i/\lambda$
γ_i	**effective utilization**	A parameter that indicates how much product i contributes to the utilization ρ. $\hat{\rho}_i = \lambda_i/(\mu - \lambda + \lambda_i) = \rho P_i/(1 - \rho + \rho P_i)$
$E[N(t)]$	**expected workorders in system**	Expected number of workorders in the system (queue + production facility)
σ	**approximation parameter**	The parameter used to approximate Erlang distribution. $\sigma = (E[N(t)] - \rho)/E[N(t)]$
$\hat{\sigma}_i$	**distribution parameter**	Another parameter used to approximate Erlang distribution. $\hat{\sigma}_i = \sigma P_i/(1 - \sigma + \sigma P_i)$

Chapter 3

Extension of Production Time Distributions

In this chapter, a model extension by changing production time distribution from exponential to Erlang distribution is described, and the numerical results of basestock allocation are shown.

3.1 Calculation of Fill Rate

The $M/M/1$ model is limited to an exponential distribution concerning the probability of production time. It means that production times are always random. It is unnatural to apply exponential distribution even to the case when the production times are distributed around an average value. Erlang distribution is usually used to solve this problem. However, using the formulas of Erlang distribution directly in the calculation of the fill rate makes the calculation too complicated. As described in the previous chapter, the probability distribution which is approximated for Erlang distribution will be used here, and the model will be called $M/\tilde{E}_k/1$ model in the thesis.

The procedure of the calculation is almost same as the one for the $M/M/1$ model. Different calculation methods are used according to the basestock level.

For $S_i > 0$,

$$P(i,x) = \sum_{m=0}^{\infty} \sum_{l=km}^{k(m+1)-1} e^{-k\mu T} \frac{(k\mu T)^l}{l!} \sum_{n=x+m}^{\infty} \tilde{p}_n P_i^x (1-P_i)^{n-m-x} \binom{n-m}{x} \tag{3.1}$$

Inner sum is reduced as follows:

$$\sum_{n=x+m}^{\infty} \tilde{p}_n P_i^x (1-P_i)^{n-m-x} \binom{n-m}{x} = (1-\hat{\sigma}_i)\hat{\sigma}_i^x \rho\sigma^{m-1} \qquad \text{for } x > 0$$

where

$$\hat{\sigma}_i = \frac{\sigma P_i}{1-\sigma(1-P_i)}. \tag{3.2}$$

As a result,

$$P(i,x) = (1-\hat{\sigma}_i)\hat{\sigma}_i^x e^{-k\mu T} \sum_{m=0}^{\infty} \sum_{l=km}^{k(m+1)-1} \frac{(k\mu T)^l}{l!} \rho\sigma^{m-1} \qquad \text{for } x > 0.$$

A late delivery occurs if $x \geq S_i$, so for item i,

$$\begin{aligned} L_{(i,T)}(S_1,\ldots,S_K) &= \sum_{x=S_i}^{\infty} (1-\hat{\sigma}_i)\hat{\sigma}_i^x e^{-k\mu T} \sum_{m=0}^{\infty} \sum_{l=km}^{k(m+1)-1} \frac{(k\mu T)^l}{l!} \rho\sigma^{m-1} \\ &= \hat{\sigma}_i^{S_i} e^{-k\mu T} \sum_{m=0}^{\infty} \sum_{l=km}^{k(m+1)-1} \frac{(k\mu T)^l}{l!} \rho\sigma^{m-1}. \end{aligned}$$

For $S_i = 0$, a late delivery occurs only if production during T is less than $N(t)+1$. Hence,

$$\begin{aligned} L_{(i,T)}(S_1,\ldots,S_K) &= \sum_{m=0}^{\infty} \sum_{l=km}^{k(m+1)-1} \frac{(k\mu T)^l}{l!} e^{-k\mu T} \sum_{n=m}^{\infty} \tilde{p}_n \\ &= \sum_{l=0}^{k-1} e^{-k\mu T} \frac{(k\mu T)^l}{l!} + \sum_{m=1}^{\infty} \sum_{l=km}^{k(m+1)-1} e^{-k\mu T} \frac{(k\mu T)^l}{l!} \rho\sigma^{m-1}. \end{aligned}$$

The aggregate fill rate is the sum, weighing each item by its fraction of total demand:

$$FR_T(S_1,\ldots,S_K) = 1 - \sum_{i=1}^{K} P_i L_{(i,T)}(S_1,\ldots,S_K)$$

where

$$L_{(i,T)}(S_1,\ldots,S_K) = \begin{cases} \hat{\sigma}_i^{S_i} e^{-k\mu T} \sum_{m=0}^{\infty} \sum_{l=km}^{k(m+1)-1} \frac{(k\mu T)^l}{l!} \rho\sigma^{m-1} & (S_i > 0) \\ \sum_{l=0}^{k-1} e^{-k\mu T} \frac{(k\mu T)^l}{l!} + \sum_{m=1}^{\infty} \sum_{l=km}^{k(m+1)-1} e^{-k\mu T} \frac{(k\mu T)^l}{l!} \rho\sigma^{m-1} & (S_i = 0) \end{cases}$$

3.2 Basestock Allocation Algorithm

The greedy algorithm is also applicable to the $M/\tilde{E}_k/1$ model.

The increment in the fill rate when S_i is increased by 1 is, for $S_i > 0$,

$$P_i(1-\hat{\sigma}_i)\hat{\sigma}_i^{S_i}e^{-k\mu T}\sum_{m=0}^{\infty}\sum_{l=km}^{k(m+1)-1}\frac{(k\mu T)^l}{l!}\rho\sigma^{m-1}$$
$$=\frac{1-\sigma}{\sigma}\hat{\sigma}_i^{S_i+1}e^{-k\mu T}\sum_{m=0}^{\infty}\sum_{l=km}^{k(m+1)-1}\frac{(k\mu T)^l}{l!}\rho\sigma^{m-1}, \quad (3.3)$$

for $S_i = 0$,

$$P_i(1-\hat{\sigma}_i)e^{-k\mu T}\sum_{l=0}^{k-1}\frac{(k\mu T)^l}{l!}+P_i(1-\hat{\sigma}_i)e^{-k\mu T}\sum_{m=1}^{\infty}\sum_{l=km}^{k(m+1)-1}\frac{(k\mu T)^l}{l!}\rho\sigma^{m-1}$$
$$=\frac{1-\sigma}{\sigma}\hat{\sigma}_i e^{-k\mu T}\sum_{l=0}^{k-1}\frac{(k\mu T)^l}{l!}+\frac{1-\sigma}{\sigma}\hat{\sigma}_i e^{-k\mu T}\sum_{m=1}^{\infty}\sum_{l=km}^{k(m+1)-1}\frac{(k\mu T)^l}{l!}\rho\sigma^{m-1}. \quad (3.4)$$

All the parts except $\hat{\sigma}_i$ are independent of S_i and positive. Then when $S_i > 0$, the allocation method of the basestocks can be considered in the same way to the one in the case of the $M/M/1$ model. However, that method cannot be directly applied to the case when $S_i = 0$. The relationship of the two increments, the one from $S_i = 0$ to $S_i = 1$ and another one from $S_i = 1$ to $S_i = 2$, should be examined. The increment of the fill rate when the basestock level changes from $S_i = 0$ to $S_i = 1$ is equal to (3.4), and the increase when the level changes from $S_i = 1$ to $S_i = 2$ is obtained by inputting $S_i = 1$ to (3.3) and becomes

$$\frac{1-\sigma}{\sigma}\hat{\sigma}_i^2 e^{-k\mu T}\sum_{m=0}^{\infty}\sum_{l=km}^{k(m+1)-1}\frac{(k\mu T)^l}{l!}\rho\sigma^{m-1}. \quad (3.5)$$

(3.4)−(3.5) is

$$\frac{1-\sigma}{\sigma}\hat{\sigma}_i\left(1-\frac{\rho}{\sigma}\hat{\sigma}_i\right)e^{-k\mu T}\sum_{l=0}^{k-1}\frac{(k\mu T)^l}{l!}$$
$$+\frac{1-\sigma}{\sigma}\hat{\sigma}_i(1-\hat{\sigma}_i)e^{-k\mu T}\sum_{m=1}^{\infty}\sum_{l=km}^{k(m+1)-1}\frac{(k\mu T)^l}{l!}\rho\sigma^{m-1} \quad (3.6)$$

The second term is positive because $0 < \hat{\sigma}_i < 1$. Whether the first term is positive or not depends on the part $1 - \frac{\rho}{\sigma}\hat{\sigma}_i$. When $1 - \frac{\rho}{\sigma}\hat{\sigma}_i$ is positive, $1 - \frac{\hat{\sigma}_i}{\sigma}$ is also positive because $0 \leq \rho < 1$. This term is positive because if $1 - \frac{\hat{\sigma}_i}{\sigma} < 0$, then

$$\begin{aligned} \sigma &< \hat{\sigma}_i \\ \sigma &< \frac{\sigma P_i}{1 - \sigma(1 - P_i)} \quad \text{(from (3.2))} \\ 1 &< \sigma. \end{aligned}$$

This result contradicts the property $0 < \sigma < 1$. Then $1 - \frac{\hat{\sigma}_i}{\sigma} \geq 0$. For this reason, the increment of the fill rate when the basestock level changes from $S_i = 0$ to $S_i = 1$ is greater than the increase when the level changes from $S_i = 1$ to $S_i = 2$. Thus when the increment of the fill rate is represented by $\lambda_{i,j}$ (j is the basestock level for item i), the following relation holds:

$$\lambda_{i,j} > \lambda_{i,j'} \quad \text{if} \quad j < j'.$$

Above relation is exactly same to (2.6). This implies that the optimum allocation problem for the $M/\tilde{E}_i/1$ model can be solved using the algorithm BSA (described in previous chapter) with respect to $\hat{\sigma}_i$.

From the analyses on the $M/\tilde{E}_k/1$ model, the following property is obtained.

Theorem 1. *The optimal stock levels for the $M/\tilde{E}_k/1$ models are obtained by sequentially allocating the next unit of basestock to the item with the largest value of $\hat{\sigma}_i^{S_i+1}$.*

For $k = 1$, the $M/\tilde{E}_k/1$ model is identical with $M/M/1$ model.

3.3 Numerical Results

The numerical result of the optimum basestock allocation is shown as follows.

3.3.1 Input Data

The demand rate of the products (λ_i/λ) and the utilization (ρ) as input data. The data of the 20 items is listed in Table 3.1.

Table 3.1: Demand Rate by Item

Item	Percentage of Total Demand	Item	Percentage of Total Demand
1	33.10	11	0.60
2	22.20	12	0.40
3	14.80	13	0.30
4	9.90	14	0.20
5	6.60	15	0.10
6	4.40	16	0.08
7	2.97	17	0.06
8	1.98	18	0.04
9	1.32	19	0.03
10	0.90	20	0.02

3.3.2 Result of Basestock Allocation

The results of the allocation of the 50 basestocks are shown in Table 3.2-3.4. In the table, the allocation for the exponential distribution corresponds to Erlang distribution with $k = 1$. The stage number k is changed up to 50. Only different calculation results are listed in the table. In Table 3.2 for example, the calculation results are same between $k = 1$ and $k = 2$, and between $k = 3$ and $k = 50$. Below the each table, the data is visualized with respect to the comparison of the allocation between the case with the exponential distribution ($k = 1$) and the case with the approximated Erlang distribution with the largest k in the each table. The allocation tends to shift to the lower demand items according to the increase of the value of k. Also, the allocation shifts to the higher demand items according to the increase of the utilization ρ. With the higher value of the utilization, the difference of allocation becomes greater according to the increase of k. The allocation changes more often with lower values of k than higher ones.

Table 3.2: Result of Basestock Allocation (1)

$\rho = 0.6$				
item	$k = 1$	$\cdots$	$k = 3$	$\cdots$
1	8	$\cdots$	7	$\cdots$
2	6		6	
3	5		5	
4	4		4	
5	4		4	
6	3		3	
7	3		3	
8	2		2	
9	2		2	
10	2		2	
11	2		2	
12	1		2	
13	1	$\cdots$	1	$\cdots$
$\vdots$	$\vdots$		$\vdots$	
20	1	$\cdots$	1	$\cdots$

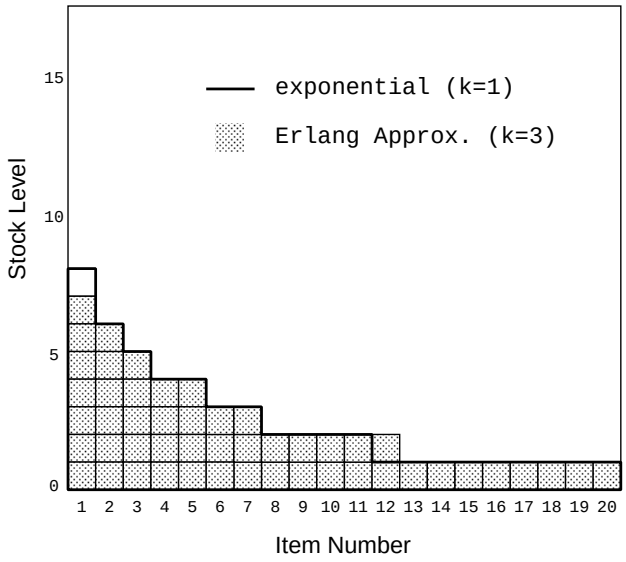


Table 3.3: Result of Basestock Allocation (2)

$\rho = 0.8$					
item	$k=1$	$k=2$	$k=3$	$k=4$	$\cdots$
1	11	10	9	9	$\cdots$
2	8	7	7	7	
3	6	6	6	5	
4	4	4	4	4	
5	4	4	3	4	
6	3	3	3	3	
7	2	2	2	2	
8	2	2	2	2	
9	2	2	2	2	
10	1	1	2	2	
11	1	1	1	1	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
17	1	1	1	1	$\cdots$
18	0	1	1	1	
19	0	1	1	1	
20	0	0	1	1	$\cdots$

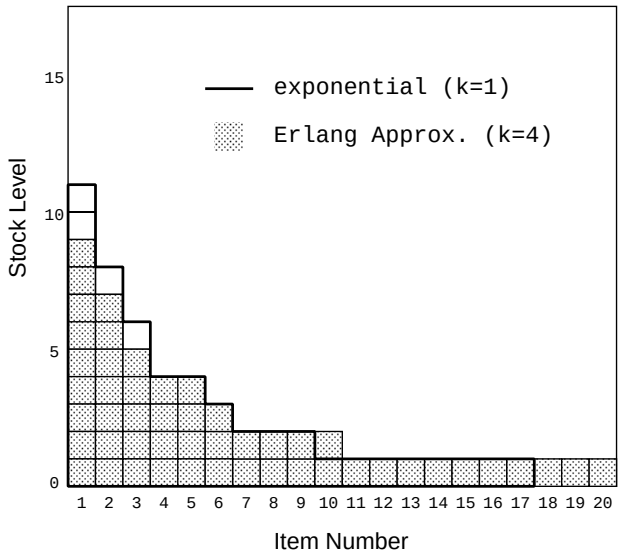
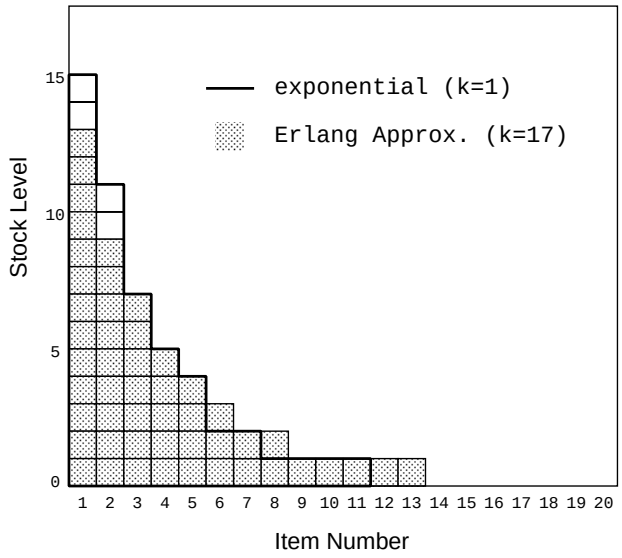


Table 3.4: Result of Basestock Allocation (3)

$\rho = 0.95$									
item	$k=1$	$k=2$	$k=3$	$\cdots$	$k=6$	$k=7$	$\cdots$	$k=17$	$\cdots$
1	15	15	14	$\cdots$	14	14	$\cdots$	13	$\cdots$
2	11	10	10		10	9		9	
3	7	7	7		7	7		7	
4	5	5	5		5	5		5	
5	4	4	4		3	3		4	
6	2	3	3		3	3		3	
7	2	2	2		2	2		2	
8	1	1	1		2	2		2	
9	1	1	1		1	1		1	
10	1	1	1		1	1		1	
11	1	1	1		1	1		1	
12	0	0	1		1	1		1	
13	0	0	0		0	1		1	
14	0	0	0	$\cdots$	0	0	$\cdots$	0	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
20	0	0	0	$\cdots$	0	0	$\cdots$	0	$\cdots$



3.3.3 Results of Fill Rate Calculation

Fig. 3.1 shows the fill rate curves with different probability distributions. The solid lines and the dot plotting represent the analytical results and the simulation results, respectively. The Erlang distribution used for the analytical result is actually the approximated one described in the previous chapter. The result of the deterministic distribution is shown just for reference. The fill rate goes up according to the increase of the stage number k. It is well known in the queueing theory that the average waiting time decreases as the service (production) time changes from random state to deterministic state. In this case, the production comes close to the deterministic state as k increases. Consequently, the decrease of the average waiting time may contribute to the increase of the fill rate.

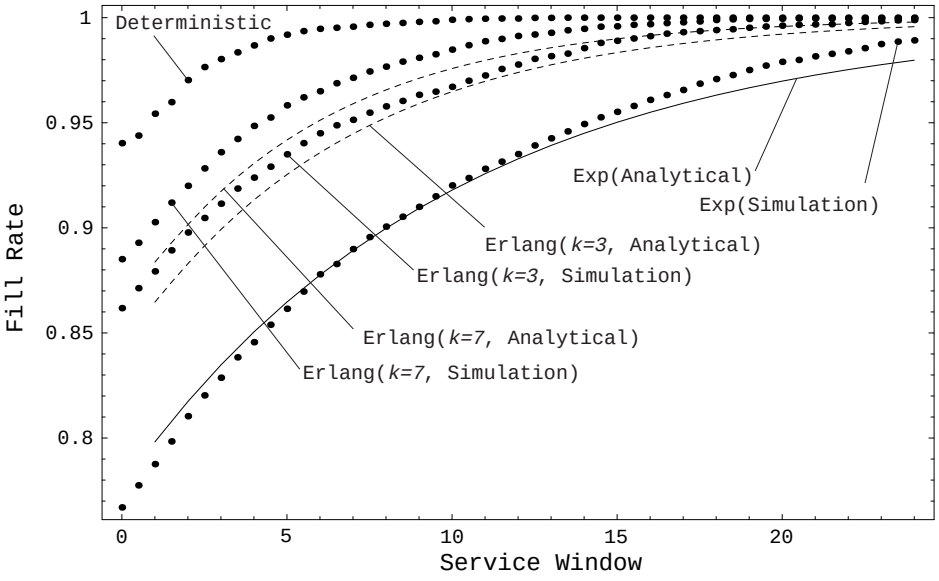


Figure 3.1: Fill Rate Curves with Different Probability Distributions

Fig. 3.2 shows the fill rate curves with changing the total basestock level from 20 to 60. To obtain the same fill rate, more basestocks are required for the exponential distribution than for the Erlang distribution. This implies that assuming exponential distribution for every type of production sometimes results in large error in the estimation of total basestocks. Since the production time tends to be periodical in the most manufacturing cases, using the Erlang distribution and adjusting k to simulate the actual distribution may be preferable in the sense of the flexibility and accuracy of the model.

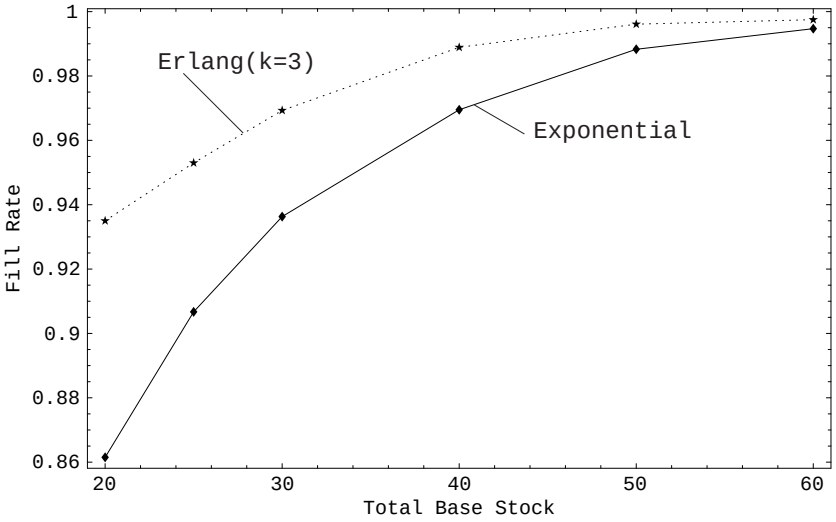


Figure 3.2: Fill Rate Curves with Different Basestock Levels

Chapter 4

Extension of Constraints

In this chapter, more general system of constraints on basestock is defined and an algorithm for the basestock allocation is presented.

4.1 Definition of Partition

Introduced general system of constraints allows to set limits of basestock for specific items or group of items. It gives the management more flexibility, such as setting the limit of stock to specific items because of the deterioration or the impulsive decrease in demand which is predicted by demand forecasting. It is also suitable for dealing with situation in which several finished-goods inventory facilities in different locations exist. The definition of the partition is given as follows. Let $\mathcal{P} = \{\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_n\}$ be a partition of $\{1, 2, \ldots, K\}$ into n subsets, that is

$$\begin{gathered} \mathcal{P}_1 \cup \mathcal{P}_2 \cup \cdots \cup \mathcal{P}_n = \{1, 2, \ldots, K\} \\ \mathcal{P}_j \cap \mathcal{P}_k = \phi \quad \text{if} \quad j \neq k \\ \mathcal{P}_j \neq \phi \quad \text{for} \quad j = 1, 2, \ldots, n \end{gathered}$$

For each $k \in \{1, 2, \ldots, n\}$, let B_k be a nonnegative integer representing the upper limit for total level of basestocks for the items in $\mathcal{P}_k$. Fig. 4.1 illustrates the concept of the partition.
The total basestock of 7 items is limited to 20. One partition member $\mathcal{P}_1$ has two elements of item 2 and item 3 and has the boundary $B_1 = 4$ for basestock allocation. Another partition member $\mathcal{P}_2$ has remaining elements and has the boundary $B_2 = 20$ for basestock allocation. These conditions

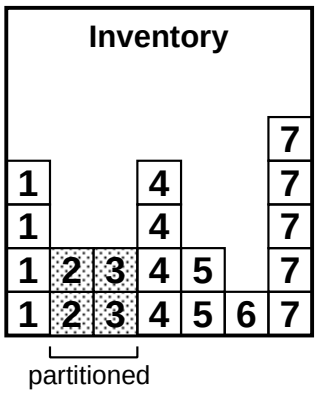


Figure 4.1: An Example of Basestock Allocation of Partitioned Model

are written in the following way:

$$\mathcal{P} = \{\mathcal{P}_1, \mathcal{P}_2\}, \quad \mathcal{P}_1 = \{2, 3\}, \quad \mathcal{P}_2 = \{1, 4, 5, 6, 7\}$$
$$S_2 + S_3 \leq B_1, \quad S_1 + S_4 + S_5 + S_6 + S_7 \leq B_2$$
$$B_1 = 4, \quad B_2 = 20.$$

4.2 Algorithm of Basestock Allocation

Next the following problem similar to (2.1)-(2.2) is considered:

Maximize

$$FR_T(S_1, S_2, \ldots, S_K) = 1 - e^{-\mu T(1-\rho)} \sum_{i=1}^{K} P_i \gamma_i^{S_i} \tag{4.1}$$

subject to

$$\sum_{i=1}^{K} S_i \leq S, \quad \sum_{i \in \mathcal{P}_k} S_i \leq B_k, \quad k = 1, 2, \ldots, n. \tag{4.2}$$

Obviously, the previous problem (2.1)-(2.2) is a special case of the problem (4.1)-(4.2). This problem is simplified also as done in (2.3)-(2.4):

Maximize

$$\mathcal{FR}_T(S_1, S_2, \ldots, S_K) = \sum_{i=1}^{K} \sum_{j=1}^{S_i} \gamma_i^j \tag{4.3}$$

subject to

$$\sum_{i=1}^{K} S_i \leq S, \quad \sum_{i \in \mathcal{P}_k} S_i \leq B_k, \quad k = 1, 2, \ldots, n. \tag{4.4}$$

An algorithm similar to BSA can be used to find the optimum allocation of the basestock even for the inventory with partition. The proof of the correctness for this is similarly shown as done in the proof of Lemma 1. A generalized algorithm is again considered as follows. The set Λ defined in (2.5),

$$\Lambda = \{\lambda_{i,j} |\ 1 \leq i \leq U,\ 1 \leq j \leq V\}$$
$$\lambda_{i,j} > \lambda_{i,j'} \quad \text{if} \quad j < j',$$

is used here also. Let $\mathcal{P}' = \{\mathcal{P}'_1, \mathcal{P}'_2, \ldots, \mathcal{P}'_n\}$ be a partition of $\{1, 2, \ldots, V\}$ into n subsets, that is

$$\mathcal{P}'_1 \cup \mathcal{P}'_2 \cup \cdots \cup \mathcal{P}'_n = \{1, 2, \ldots, V\}$$
$$\mathcal{P}'_j \cap \mathcal{P}'_k = \phi \quad \text{if} \quad j \neq k$$
$$\mathcal{P}'_j \neq \phi \quad \text{for} \quad j = 1, 2, \ldots, n$$

m largest elements are selected, and a subset

$$\hat{\Lambda}_m = \{\lambda_{i,j} |\ 1 \leq i \leq U,\ 1 \leq j \leq V_i^{(m)}\}$$

is created similarly to (2.7). Each element is sequentially selected using the greedy algorithm as done in the case without partition, but the following operation is added in each iteration for the partitioned model in order to exclude the partition which has already reached the upper limit of the partition in the m-th iteration:

if $\quad \sum_{i \in \mathcal{P}'_k} V_i^{(m)} = W_k \quad (1 \leq k \leq n) \quad$ then

$$\Lambda = \Lambda - \{\lambda_{i,(V_i^{(m)}+1)}, \lambda_{i,(V_i^{(m)}+2)}, \ldots, \lambda_{i,V}\} \quad (\text{for all } i \in \mathcal{P}'_k).$$

where W_k is a nonnegative integer representing the upper limit for total number of the elements in $\mathcal{P}'_k$. After that, a largest element is selected from yet unselected elements in Λ with taking the operation:

$$\hat{\Lambda}_{m+1} = \hat{\Lambda}_m \cup \max_{1 \leq i \leq U} \{\lambda_{i,(V_i^{(m)}+1)}\}.$$

Finally $\hat{\Lambda}_V$, a set of V largest elements, is obtained in the V-th iteration. As shown in the proof of Lemma 1, this algorithm has direct correspondence with the problem (4.3)-(4.4). Thus it is proved that the optimum allocation of basestocks is obtained by the algorithm BSA (with some modification) even for the partitioned model. The algorithm of basestock allocation is summarize as follows:

BSAP : BaseStock Allocation with Partition Problem
Input: $\gamma_1, \gamma_2, \ldots, \gamma_K,\ \mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_m,\ B_1, B_2, \ldots, B_m$, and S.
Output: $S_1, S_2, \ldots, S_K$ such that

$$\sum_{i=1}^{K} S_i \le S, \quad \sum_{i \in \mathcal{P}_k} S_i \le B_k, \quad k = 1, 2, \ldots, n,$$

and $FR_T(S_1, S_2, \ldots, S_K)$ is maximized.

```
Algorithm BSAP
  for i = 1 to K
      S_i := 0
  k := 1
  while (sum_{i=1}^{K} S_i < S) && (k > 0)
    do
      k := 0
      γ_max := 0
      for j := 1 to m
          do
            if (sum_{i ∈ P_j} S_i < B_j) then
               for i ∈ P_j
                  do
                     if (γ_max < γ_i^{S_i+1}) then
                           γ_max := γ_i^{S_i+1}
                           k := i
                  od
          od
      S_k := S_k + 1
    od
  return(S_1, S_2, ..., S_K)
```

Since this algorithm is applicable to both the $M/M/1$ and the $M/\tilde{E}_k/1$ models, the property for the extension to the partitioned model is summarized as the theorem:

Theorem 2. *The optimal stock levels for the $M/M/1$ ($M/\tilde{E}_k/1$) model with partitioned inventory are obtained by sequentially allocating the next unit of basestock to the item with the largest value of $\gamma_i^{S_i+1}$ ($\hat{\sigma}_i^{S_i+1}$), provided that if a partition reaches its limit in the course of iteration, the allocation for the items in the partition halts.*

Fig. 4.2 shows the fill rate curves with and without partition for $M/M/1$ production-inventory model. Here, the demand data in Table 3.1 is used, and the utilization is fixed to $\rho = 0.9$. The partition is given as follows:

$$\mathcal{P} = \{\mathcal{P}_1, \mathcal{P}_2\}, \quad \mathcal{P}_1 = \{1, 2, 3\}, \quad \mathcal{P}_2 = \{4, \ldots, 20\}$$
$$S_1 + S_2 + S_3 \le B_1, \quad S_4 + \cdots + S_{20} \le B_2$$
$$B_1 = 20, \quad B_2 = S.$$

It may seem unrealistic to constrain high demand items such as item 1, 2 and 3. This case is only for the experiment to see the effect of the partition on the fill rate.

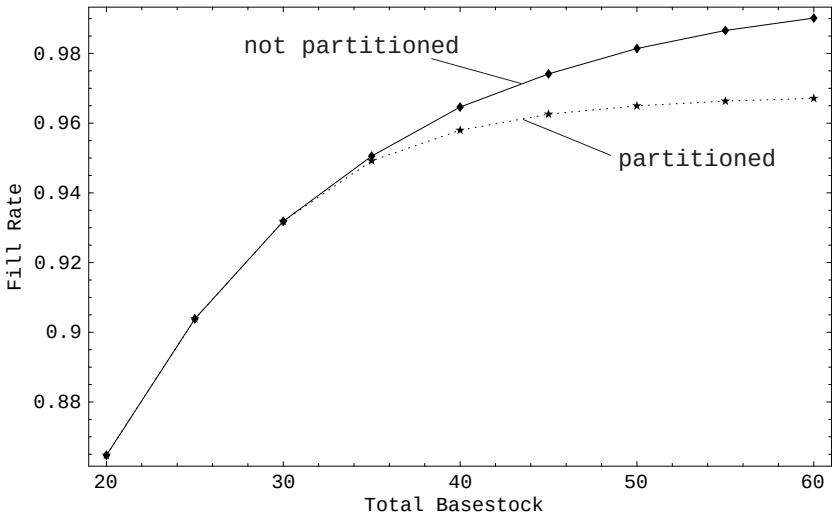


Figure 4.2: Fill Rate Curve with Partition

Chapter 5

New Production Rule

In this chapter, a new production rule Modified BOP is introduced, then its improvement over other rules (BOP, FIFO) is shown by simulation.

5.1 Previous Work

The production rule which is discussed in previous chapters is assumed FIFO (First In First Out) only. Sox et al. proposed two new rules: BOP (BackOrders get first Priority) and ANT (ANTicipate backorders). BOP is the rule that a backorder is placed at the top of the queue when it arrives. If there exist already other backorders in the queue, the last backorder is placed at the end of the backorders. ANT is not discussed in this research because there is not enough information for the algorithm in their paper so that it is hard to reproduce ANT exactly. As shown in Fig. 5.1, BOP improves the fill rate substantially in the most range of the service window but shows poor performance when $T < 1$. They explain that in such range, backordered item has already violated the criterion of delivery within T, so giving priority to such item is expending resources on a lost cause.

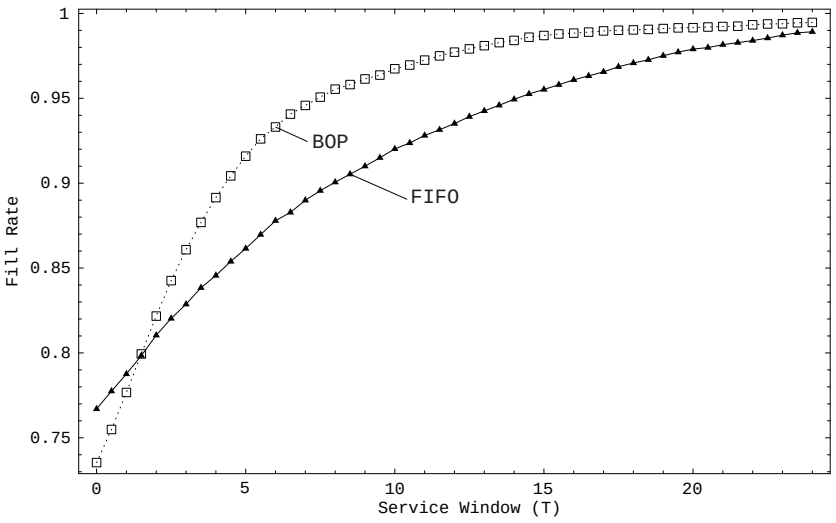


Figure 5.1: Fill Rate Curves with FIFO and BOP

5.2 Modified BOP

Modified BOP is basically same as BOP, but the number of backorders in the queue is calculated every time a new backorder arrives. Fig. 5.2 illustrates Modified BOP step by step. The workorders are processed by FIFO when there are no backorders (**a**). When a backorder 5 arrives, it gets the first priority and is placed at the top of the queue (**b**). When another backorder 6 arrives, it is placed just behind the prior backorder (**c**). Suppose that two orders are processed on average during the service window T. If next backorder arrives when two other backorders still remain in the queue, it is not attached to the prior backorder but the end of the queue because it is highly likely late in the delivery time (i.e. service window T). Other ordinary workorders will not be added extra waiting time by accepting such rule (**d**). If next backorder 9 arrives while backorder 7 still remains in the queue and there are no backorders at the top of the queue, the backorder 9, which is accompanied with 7, is placed at the top of the queue (**f**). If only 9 comes at the top of the queue (**f'**), it will not improve the fill rate because backorder 9 is given to the customer who originally issued the order 7. For that customer, the waiting time is already over T.

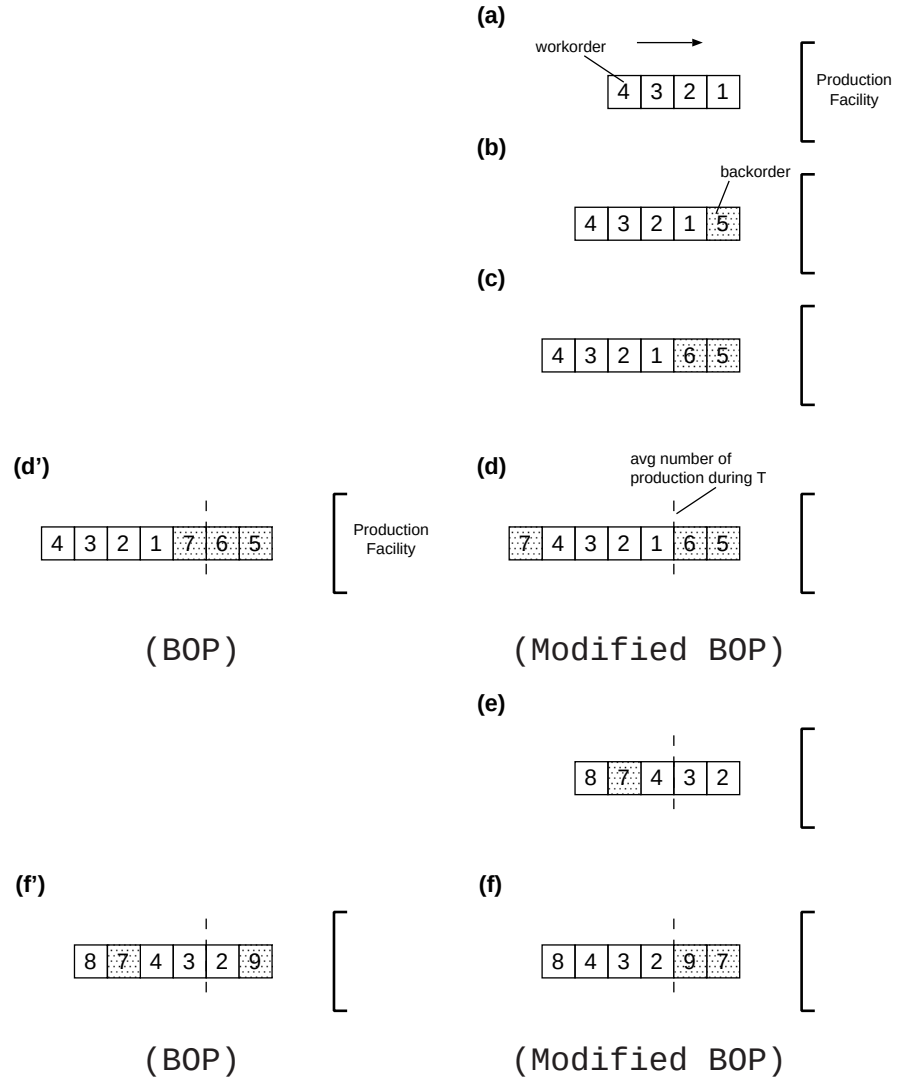


Figure 5.2: Schematic Description of Modified BOP

5.3 Simulation Result

Fig. 5.3 shows the comparison of the fill rate calculated with different production rules. BOP mostly increases the fill rate from FIFO level, but gets lower when the service window T is close to zero. This is because priority is given even to the backorders which are expected to be late in that short service window. Modified BOP solves this problem, and actually the fill rate with Modified BOP is never below the data obtained by FIFO.

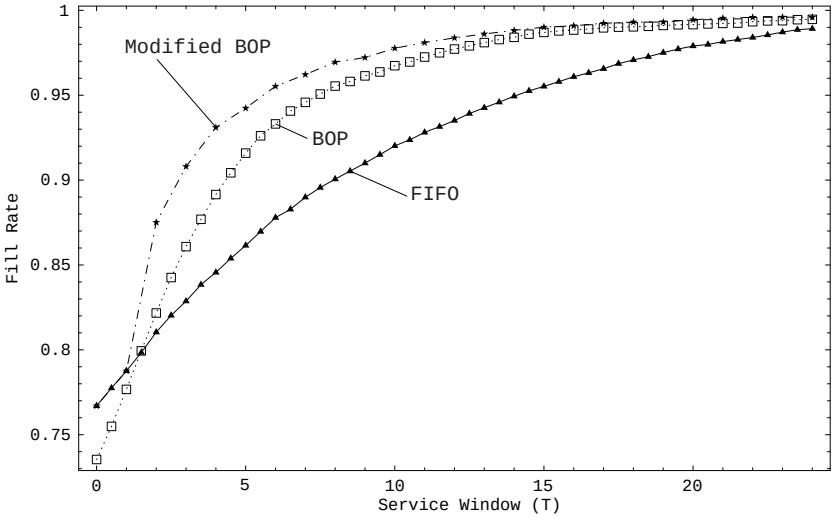
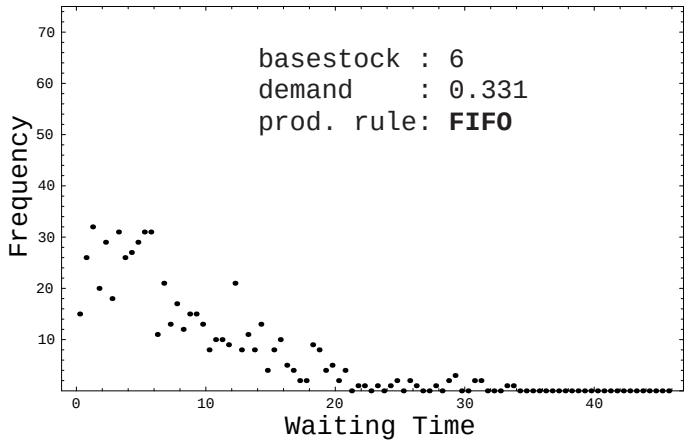
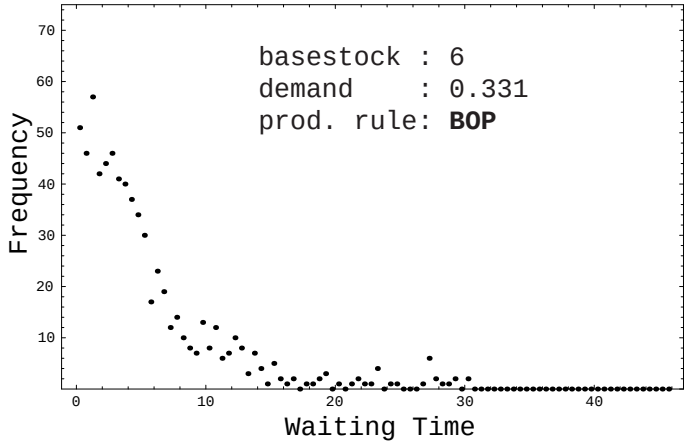


Figure 5.3: Improvement of Fill Rate by Modified BOP

Fig.5.4 and Fig.5.5 show the distribution of the waiting time calculated by simulation. Here, the demand data in Table 3.1 is used, and the utilization is fixed to $\rho = 0.9$. Item 1 and item 9 are assigned six and zero basestocks, respectively. Modified BOP generally have the customers wait for shorter time than BOP and FIFO. However, as shown in Table 5.1, Modified BOP sometimes sacrifices a few customers and makes them wait long time. This is because Modified BOP does not consider the waiting time of individual customer in the queue when it makes decision.





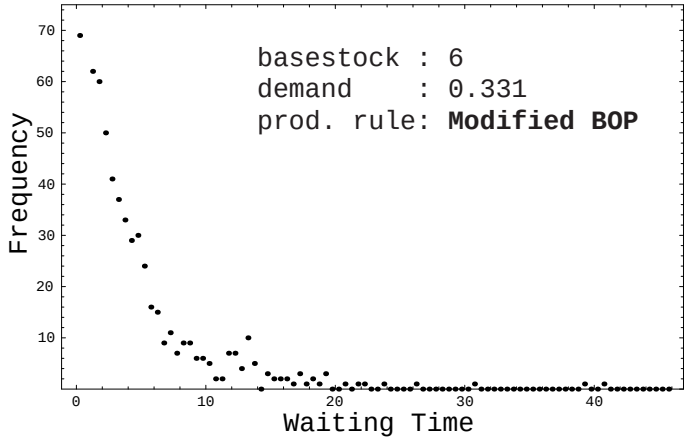
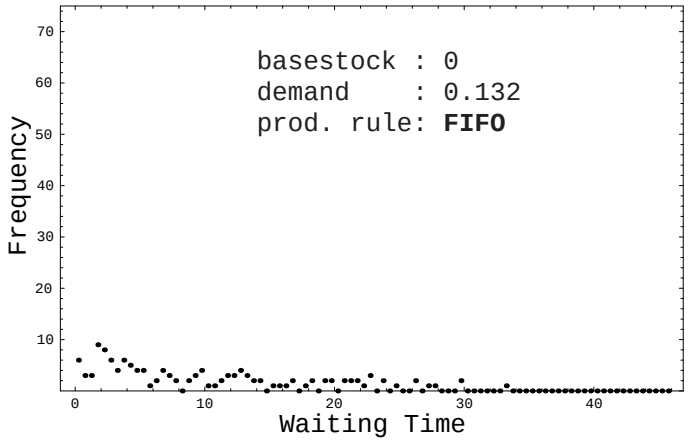
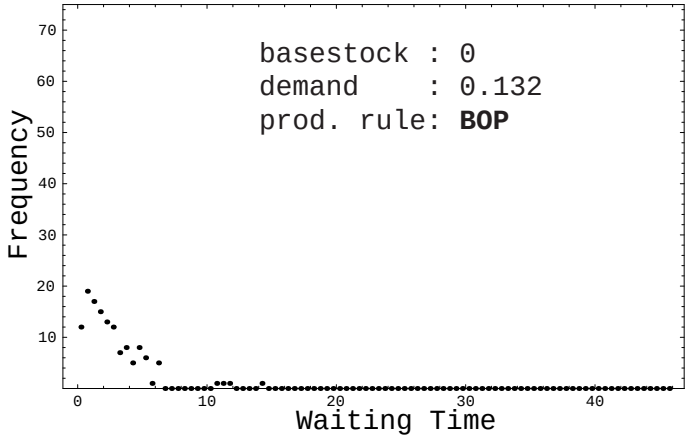


Figure 5.4: Waiting Time Distribution of Item 1





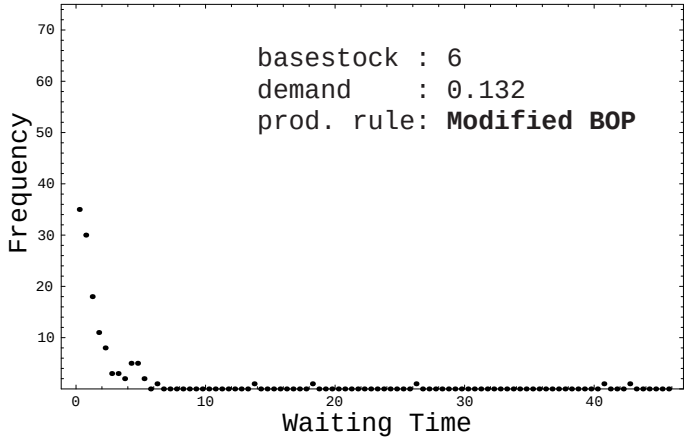


Figure 5.5: Waiting Time Distribution of Item 9

Table 5.1: Maximum Waiting Time

Production Rule	Max. Waiting Time
FIFO	45.7
BOP	30.3
Modified BOP	197.8

Chapter 6

Concluding Remarks

6.1 Review of Work

The two extensions of the $M/M/1$ model proposed by Sox et al. are introduced. One is the modification of the model using approximated Erlang distribution for production time. The allocation of basestocks changes according to the parameter k of Erlang distribution. Another is with the case where the inventory is partitioned and has extra constrains for basestock allocation. The optimum allocation of basestocks is obtained by a greedy algorithm BSAP. Modified BOP is introduced as a priority rule for production. It achieves higher fill rate than FIFO (First In First Out) or BOP (BackOrder gets first Priority) proposed by Sox et al.

6.2 Conclusions

As the conclusions of the research, the following points can be mentioned.

- The basestock allocation to maximize the fill rate is affected by the parameter k of Erlang distribution of production time. Since Erlang distribution includes exponential distribution, $M/E_k/1$ model is expected to be used for the basestock allocation of the models which have a larger class of production time distributions.
- The optimum allocation of the basestocks are obtained by a greedy algorithm for the model with partitioned inventory.
- Modified BOP rule improves conventional BOP concerning production performance and achieves higher fill rate.

6.3 Further Research

The extensions of the model are originally aimed to make the production-inventory model applicable to a larger variety of practical situations. In that sense, there is much room left for further improvement of the model. For example, the model of this research assumes single production line. Since most actual production facilities have multiple production lines for the specific groups of product items, extension from single to multiple production lines for the model is considered natural. Or the model of this research assumes the same production rate μ for all product items. The model with different μ's for the product items seems to be more realistic. Approximated Erlang distribution is used for the probability distribution of production time in this research. From theoretical point of view, using Erlang distribution would be more preferable. It was not possible to formulate mathematical model with Erlang distribution this time because of the difficulty of handling formulas. Different approach is expected to organize the model with Erlang distribution.

Acknowledgement

I am very grateful to Professor Milan Vlach for his precious and kind advice during the year. And also I would like to thank Associate Professor Kunihiko Hiraishi for sharing the time for the discussion with me. In proceeding this research I have benefited from the comments and suggestions of other members in the Foundation of System laboratory. Particular thanks go to Dr. Shao Chin Sung and Dr. Yasuhiro Takashima for giving me useful hints. Finally, I would like to thank my family for supporting my study at JAIST for two years.

Bibliography

[1] J. A. Buzacott and J. G. Shanthikumar. *Stochastic Models of Manufacturing*. Prentice-Hall, 1993.

[2] F. S. Hiller and G. J. Lieberman. *Introduction to Operation Research, sixth edition*. McGraw-Hill, 1995.

[3] H. Kobayashi. *Modeling and Analysis*. The Systems Programming Series. Addison-Wesley Publishing Company, 1981.

[4] Michael Pinedo. *Scheduling: theory, algorithms and systems*. Prentice-Hall, 1995.

[5] Sox C. R., L. J. Thomas, et al. Coordinating production and inventory to improve service. *Management Science*, Vol. 43, pp. 1189–1197, 1997.