Lead-time analysis and a control policy for production lines

Angelos A Economopoulos, Vassilis S Kouikoglou

▶ **To cite this version:**

Angelos A Economopoulos, Vassilis S Kouikoglou. Lead-time analysis and a control policy for production lines. International Journal of Production Research, 2010, pp.1. 10.1080/00207541003627845 . hal-00586984

HAL Id: hal-00586984
https://hal.science/hal-00586984

Submitted on 19 Apr 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Lead-time analysis and a control policy for production lines

Journal:	*International Journal of Production Research*
Manuscript ID:	TPRS-2009-IJPR-0791.R1
Manuscript Type:	Original Manuscript
Date Submitted by the Author:	05-Jan-2010
Complete List of Authors:	Economopoulos, Angelos; Technical University of Crete, Dept of Production Engineering and Management Kouikoglou, Vassilis; Technical University of Crete, Dept of Production Engineering and Management
Keywords:	FLOW LINES, MAKE TO STOCK PRODUCTION, QUEUEING NETWORKS, SIMULATION
Keywords (user):	CUSTOMER IMPATIENCE, LEAD-TIME ESTIMATION

For Peer Review Only

scholarONE™
Manuscript Central

Lead-time analysis and a control policy for production lines

Angelos A. Economopoulos[1], Vassilis S. Kouikoglou[*2]
Department of Production Engineering and Management
Technical University of Crete, Chania 73100, Crete, Greece
E-mail: [1]ubatou@dpem.tuc.gr, [2]kouik@dpem.tuc.gr

For Peer Review Only

Abstract

We address a control problem for a production line which produces one product to stock and faces random demand. During stockouts, the system quotes a fixed response time for each arriving order, and the customers place their orders only if the response time promised meets their deadlines. Customer orders are filled on a first come, first served basis. A penalty cost is incurred whenever a customer is served later than promised. A two-parameter admission/inventory control policy is implemented that maintains a bounded backlog and a constant inventory position (total inventory minus backlog) in the system. For production lines with exponential processing times and Poisson demand, the mean profit rate of the system is computed analytically using closed queueing network formulas. For systems with general processing or interarrival time distributions, the profit rate is estimated via simulation. Simple properties are established which ensure that the profit maximizing control parameters can be determined in finite time using exhaustive search. Numerical results show that the proposed policy performs better than the make-to-order/zero-inventory and the lost-sales/make-to-stock policies.

Keywords: Production lines, admission control, inventory control, balking, quoted lead time, closed queueing networks, simulation.

[*] Corresponding author.

1. Introduction

Production control involves decisions concerning when and how much to produce and when to accept incoming orders. Such decisions have significant effects on a number of operational indices such as throughput, mean inventory and backlog levels, and quality of service. These operational indices affect the mean profit rate of the system, which comprises profit from sales and the costs associated with inventory, backlog, delays in filling orders, possible loss of sales during stockout periods and so on (see, e.g., Porteus 1990, Section 3.7, p. 615). In the literature of production control there is a rich collection of policies that minimize the inventory and backlog costs.

A simple and often effective inventory control policy is one that specifies a target value for the stock level, called *base stock* (Zipkin 2000, Buzacott and Shanthikumar 1993); the manufacturing facility produces as long as the number of finished items is smaller than the base stock and idles otherwise.

In order to handle incoming orders during stockout periods, many systems apply a lost sales (LS - all incoming orders rejected) or a complete backordering (CB - all incoming orders accepted) policy. When applying LS, there is no cost of delay in filling orders but the loss of sales to competition during stockouts makes it mandatory to keep a high stock level. On the contrary, facilities under CB appear to have vast backlog queues and long delays in filling orders which, apart from high backlogging costs, could also lead to potential loss of sales due to customer impatience.

An admission control policy which is analogous to the base stock policy and generalizes LS and CB, balancing between loss of sales and long delays, is the *base backlog* policy (Naor 1969). This policy admits incoming orders during stockouts as long as the backlog level is less than a certain threshold called base backlog and rejects them otherwise. When a base backlog is applied to systems with zero stock level, the policy is called *make-to-order* (MTO).

In this paper, we consider a production line producing one product to meet random demand from impatient customers. We adopt a joint inventory/admission control policy called *base stock/base backlog* (BSBB) to maximize the net profit rate of the system, which includes profit from sales less the costs of inventory,

http://mc.manuscriptcentral.com/tprs Email: ijpr@lboro.ac.uk
For Peer Review Only

backordering and failure to meet the delivery deadlines promised. According to BSBB, the inventory levels are handled as in base stock and, during stockouts, a base backlog admission control policy is implemented. Joint inventory and admission control policies have been studied in the past years both for single stage systems (Caldentey 2001, Kouikoglou and Phillis 2002, Ioannidis and Kouikoglou 2008) and production networks (Ioannidis *et al*. 2008), yet without considering customer impatience and delivery due dates.

Impatience of customers is an important factor that must be taken into consideration when controlling inventory and backlog in production systems. This subject has drawn some attention in the past decades, mainly for its applicability to telecommunication systems rather than manufacturing systems. When we refer to customer's impatience we use the terms *balking* and *reneging* (see e.g. Haight 1957, 1959, 1960).

Balking occurs when a customer arrives during a stockout and decides not to place an order, rather than join the backlog queue, if he believes that he will be served later than desired. Balking causes loss of sales. It can be modeled by assigning a balking probability to the arrival process, which depends either on the state of the system (Subba Rao 1967) or on some response time estimate (Liu and Kulkarni 2008). In the former case, an arriving customer enters the backlog queue with some probability which depends on the number of outstanding orders already pending. In the latter case, each customer has a due time (deadline measured from the time of his arrival) and joins the queue only if he estimates that the facility's response time to his order is less than the due time.

Reneging, also referred as abandonment, is the process where outstanding customers decide to withdraw their orders if they have waited longer than their deadlines. Reneging not only causes loss of sales but also loss of customer goodwill. The case of queue abandonment has been studied in the past decades, but analytical results for performance measures such as mean backlog queue sizes, waiting times or abandonment rates, are limited to single-stage systems with one (Movaghar 2006, Economopoulos and Kouikoglou 2008) or several servers (Bacceli and Hebuterne 1981, Ward 1999, Brandt and Brandt 1999). For more general production systems such as production lines or networks with reneging, mathematical difficulties arise in deriving analytical results and determining

For Peer Review Only

optimal production control policies.

In this paper, we only consider balking of customers but we include a penalty cost for the system when it fails to meet a *quoted lead time* (QLT). Specifically, a customer arriving during a stockout is promised a fixed response time QLT to his order. If QLT is greater than the customer's due time, then the customer balks. Otherwise, the customer places his order, but if his waiting time exceeds QLT, then, though it is not possible to renege and abandon the backlog queue, the production facility incurs an extra delay cost as a penalty or discount offered to the customer.

Our aim is to determine optimal base stock and base backlog levels, denoted s and c respectively, for a production line so as to maximize its mean profit rate. In section 2, we describe the production line and model the balking process. In section 3, we use an equivalent closed queueing system to derive expressions for the steady-state probabilities of the system and the various components of the profit rate. In section 4, we show that the optimal values for s and c are bounded from above. This allows us to develop a simple and fast algorithm to track down the optimal solution. Numerical results are presented in section 5 and conclusions in section 6. Early results of this work are reported in Economopoulos and Kouikoglou (2009).

2. System description

Consider a production line with N machines that produce one product to meet random demand. Machine M_i along with its input buffer will be referred as node i, $i = 1, 2, \ldots, N$, and n_i will denote the number of workpieces in node i, including the one that may be processed by M_i. Raw items are first processed at M_N. Workpieces leaving M_i are sent to node $i - 1$. Machine M_1 performs the last operation and sends finished items to the output buffer. The processing times of M_i are independent exponential random variables with mean $1/\mu_i$. For simplicity, we assume that all production rates μ_i are different. Customers arrive according to a Poisson process with mean rate λ.

Each customer requests one unit of product and, if stock is available, a sale

http://mc.manuscriptcentral.com/tprs Email: ijpr@lboro.ac.uk
For Peer Review Only

occurs immediately and a new raw item is sent to the buffer of node N to replace the item sold. During stockouts, orders cannot be fulfilled immediately and customers are reluctant to place orders. To cope with order cancelations and attract reluctant customers, the firm announces a standard time frame QLT from the moment an order is placed, after which if the order is still pending, the customer would benefit from a discount. We assume that each arriving customer has a maximum waiting time tolerance upon arrival until his order is shipped. If QLT is greater than the customer's maximum waiting time he will not place the order. Otherwise, he places the order but cannot abandon the backlog queue thereafter.

Let q denote the probability that an arriving customer will place his order during a stockout period. The maximum waiting tolerance of customers is a random variable θ with known distribution function $G(t) = P(\theta \leq t)$ and the firm's QLT is constant. The probability that an arriving customer facing stockout will place his order is

$$\begin{aligned} q &= P(\text{QLT} \leq \theta) = 1 - P(\theta \leq \text{QLT}) \\ &= 1 - G(\text{QLT}). \end{aligned} \tag{1}$$

Therefore, when no stock is available, the rate at which products are requested is $q\lambda$. Every time a customer requests a product and enters the backlog queue, a new raw item is released to node N. Orders are shipped on a FCFS basis.

The number of finished items in the output buffer is denoted n_F, the number of backorders is denoted n_B and the total number of items (raw items, workpieces, products) n_H. The inventory and backlog levels of the system are controlled by implementing the joint inventory/admission control policy BSBB described in Section 1. Accordingly, the facility produces until the inventory level n_F of the output buffer reaches the base stock s and then stops. During stockouts, all incoming orders are accepted as long as the number of backorders n_B is lower than the base backlog c. A new raw item is released into the production line each time an order is accepted. It then follows that, at any time instant, the total number of items in the system equals s plus the number of backorders, i.e.,

$$n_H = s + n_B. \tag{2}$$

http://mc.manuscriptcentral.com/tprs Email: ijpr@lboro.ac.uk
For Peer Review Only

The net profit of the system comprises four types of profit or cost. Each time a product is sold, the system earns a profit p; penalties due to lost sales can be taken into account by adding a unit lost sale cost to p. The system incurs a holding cost rate h for every item held in the facility, a backlog cost rate b for every pending order and a penalty cost d for every order pending longer than QLT. Different holding costs h_1, …, h_N for the inventories can easily taken into account into the model. Yet, a uniform holding cost h is assumed for simplicity in the presentation.

The problem is to determine the optimal values of s and c so as to maximize the average net profit rate of the system. This quantity is given by

$$J(s, c) = p\mathrm{TH}_{s,c} - hH_{s,c} - bB_{s,c} - dD_{s,c}, \quad (3)$$

where $\mathrm{TH}_{s,c}$ is the throughput (mean production rate), $H_{s,c}$ is the average inventory of the total number of items in the system (raw items, workpieces, products), $B_{s,c}$ is the average backlog of pending orders and $D_{s,c}$ is mean rate of delayed orders beyond the QLT.

3. Performance evaluation

In this section, we derive analytical expressions for the performance measures involved in equation (3).

3.1 *An equivalent cyclic Jackson network*

Consider the cyclic network of queues shown in Fig. 1, in which there are $N + 1$ service nodes, $i = 0, 1, \ldots, N$, and a total population of $s + c$ jobs. Suppose that server M_0 at node 0 has a service rate μ_0 equal to the rate at which orders are placed in the production system. The servers of the queueing system have the same rates μ_i and the same contents n_i as the nodes the production facility for $i = 1, \ldots, N$.

Insert Figure 1 about here

http://mc.manuscriptcentral.com/tprs Email: ijpr@lboro.ac.uk

For Peer Review Only

Under the above assumptions, both systems evolve in the same manner. We assume that when $n_0 = 0$ in the queueing system, the production system has $n_F = 0$ finished items and $n_B = c$ orders pending, and when $n_0 = s + c$ the production system has $n_F = s$ and $n_B = 0$ (for a detailed discussion of the equivalence of the two systems see e.g. Section II.B of Ioannidis *et al.* 2008). Table 1 summarizes the relationships among the number of items in node 0 of the closed queueing system and the state variables n_F, n_B and n_H of the production system. Since the service rate of node 0 is equal to the demand rate of the production system, μ_0 is state-dependent, hence,

$$\mu_0(n_0) = \begin{cases} \lambda, & n_0 > c \\ q\lambda, & n_0 \le c. \end{cases} \tag{4}$$

Insert Table 1 about here

Next, with the use of Table 1, we derive expressions for the throughput $\mathrm{TH}_{s,c}$, average inventory $H_{s,c}$ and backlog $B_{s,c}$, and the mean rate of delayed orders $D_{s,c}$.

The throughput of the system equals the rate of accepted orders. We see from equation (4) that the latter drops from λ to λq during stockouts, i.e., when $n_F = 0$ or, in the equivalent queueing system, $n_0 \le c$. Therefore,

$$\mathrm{TH}_{s,c} = \lambda P(n_0 \ge c + 1) + q\lambda P(n_0 \le c). \tag{5}$$

The average backlog of the production line by definition is $B_{s,c} = E(n_B)$, where E is the expectation operator. By inspection of Table 1 we have $n_B = 0$ for $n_0 \le c$ and $n_0 = c - n_B$ for $n_B \ge 1$; thus,

$$B_{s,c} = \sum_{k=1}^{c} kP(n_0 = c - k)\,. \tag{6}$$

The average overall inventory of raw, semi-finished and finished items is $H_{s,c} = E(n_H)$; it follows from equation (2) that

$$H_{s,c} = s + B_{s,c}. \tag{7}$$

Finally, the mean arrival rate $D_{s,c}$ of orders delayed longer than QLT is given by

For Peer Review Only

$$D_{s,c} = \sum_{m=0}^{c-1} q\lambda P(n_0 = c - m)\Pi_m \,, \tag{8}$$

where $q\lambda P(n_0 = c - m)$ is rate at which a new order is placed when $n_B = m$ orders are already pending, and Π_m denotes the probability that the lead time (time-to-fill) of this order will exceed QLT.

Equations for the computation of Π_m will be established in the next section. All probabilities of the state variable n_0 involved in equations (5)–(8) can be calculated using standard formulas for cyclic Jackson networks (see, e.g., Section II.C of Ioannidis *et al.* 2008, Buzen 1973):

(a) Normalization constant of a cyclic Jackson network with nodes $k = i, i - 1, \ldots, j$ having constant service rates μ_k, node j feeding back to i, and a total population of n jobs,

$$G_{i,j}(n) = \sum_{n_i + \ldots + n_j = n} \prod_{k=i}^{j} \mu_k^{-n_k} \,, j \le i. \tag{9}$$

(b) Normalization constant of the original cyclic queueing system (i.e., with a total population $s + c$ and node 0 having a state-dependent service rate given by equation (4)),

$$G(s + c) = \sum_{n_0=0}^{s+c} q^{-\min(n_0, c)} \lambda^{-n_0} G_{N,1}(s + c - n_0) \,. \tag{10}$$

(c) Product-form equilibrium probabilities of the original cyclic queueing system,

$$P(n_0, \ldots, n_N) = \frac{1}{G(s+c)} q^{-\min(n_0, c)} \lambda^{-n_0} \prod_{k=1}^{N} \mu_k^{-n_k} \,. \tag{11}$$

(d) Marginal probabilities for n_0,

$$P(n_0 = k) = \frac{1}{G(s+c)} q^{-\min(k, c)} \lambda^{-k} G_{N,1}(s + c - k) \,, \tag{12}$$

$$P(n_0 \ge k) = \frac{1}{G(s+c)} \sum_{n=k}^{s+c} q^{-\min(n, c)} \lambda^{-n} G_{N,1}(s + c - n) \,. \tag{13}$$

For Peer Review Only

3.2 *Lead-time analysis*

We now derive equations for the computation of Π_m. Assume that $s \geq 1$. The case $s = 0$ will be analyzed later as a special case. We consider the system in equilibrium. Suppose that a customer arrives during a stockout interval $(t, t + dt)$ and decides to place an order while m, $m < c$, orders are already pending in the system. At time t, right before the customer places his order, there are a total of $n_H = s + m$ workpieces in the production line (a fixed inventory s plus one item per outstanding order). Since the firm adopts the FCFS policy, the lead time (time-to-fill) of the new order will be the time needed by the facility to produce a total of $m + 1$ items. Various cases have to be considered depending on the location of item $m + 1$ in the production facility. In summary, the probability that the lead time of the order exceeds QLT given that $n_B = m$ is computed in three steps: (a) we calculate the probabilities of various locations of item $m + 1$ in the production facility; (b) for each location $i = N, N - 1, \ldots, 1$, we compute the tail of the probability distribution of the total time item $m + 1$ spends in M_i, M_{i-1}, … and M_1; and (c) we add the products of (a) and (b) for all i.

3.2.1 Location of item $m + 1$

Proposition 1: Suppose that a new order arrives when m others are already pending. Then, for $s \geq 1$,

$P(\text{item } m + 1 \text{ is at } M_i \mid n_B = m \text{ at time } t \text{ and a new order arrives at } t^+) =$

$$= \frac{P(\text{item } m+1 \text{ is at } M_i \text{ and } n_B = m \text{ at time } t \text{ and a new order arrives at } t^+)}{P(n_B = m \text{ at time } t \text{ and a new order arrives at } t^+)} \quad (14)$$

$$= \frac{G_{i,1}(m)\mu_i^{-1} G_{N,i}(s-1)}{G_{N,1}(s+m)}. \quad (15)$$

Proof: First we calculate the denominator of (14). The probability that a customer will arrive during a stockout interval $(t, t + dt)$ and decide to place an order while m orders are already pending in the system, $m < c$, equals the product of

For Peer Review Only

probabilities of the following events:

(i) at time t, in steady state, the system is out of stock with m orders pending or, equivalently, a total of $s + m$ workpieces are in nodes $N, N-1, \ldots, 1$;

(ii) a new customer arrives in the interval $(t, t + dt)$ and decides to place an order.

Thus, using (4) and (12), we can calculate the denominator of (14) as follows:

$$P(n_B = m \text{ at time } t \text{ and a new customer arrives at } t^+) =$$

$$= P(n_0 = c - m) \times (\lambda q dt)$$

$$= (q\lambda dt)\frac{1}{G(s+c)}(q\lambda)^{-c+m} G_{N,1}(s+m)\,. \qquad (16)$$

The numerator of (14), is the probability of intersection of events (ii) and a subset of (i) satisfying the following conditions:

(iii) the first m workpieces are anywhere between machines $M_i, M_{i-1}, \ldots, M_1$;

(iv) one more (item $m + 1$) is in M_i;

(v) the remaining $s - 1$ workpieces are anywhere between machines M_N and M_i.

The probability of intersection of (iii)–(v) is computed as follows:

P(events (iii) and (iv) and (v)) =

$= P$(in equilibrium, item $m + 1$ is at M_i and $n_0 = c - m$) =

$$\sum_{\substack{\text{(i) } n_i'+n_{i-1}+\ldots+n_1=m \\ \text{(iii) } n_N+\ldots+n_{i+1}+n_i''=s-1}} P(n_0 = c-m, n_N, \ldots, n_i = n_i'+n_i''+1, \ldots, n_1) =$$

$=$

$$\frac{1}{G(s+c)}(q\lambda)^{-(c-m)}\left(\sum_{n_N+\ldots+n_{i+1}+n''=s-1} \mu_N^{-n_N}\ldots\mu_{i+1}^{-n_{i+1}}\mu_i^{-n_i''}\right)\mu_i^{-1}\left(\sum_{n_i'+n_{i-1}+\ldots+n_1=m} \mu_i^{-n_i'}\mu_{i-1}^{-n_{i-1}}\ldots\mu_1^{-n_1}\right)$$

$$= \frac{1}{G(s+c)}(q\lambda)^{-c+m} G_{i,1}(m)\mu_i^{-1} G_{N,i}(s-1)\,.$$

Multiplying the above by $q\lambda dt$ and dividing by (16) yields (15).

3.2.2 Remaining sojourn time distribution of item $m + 1$ from M_i on. Suppose

For Peer Review Only

that at time t item $m + 1$ is located at machine M_i awaiting to be processed by M_i, M_{i-1}, … and M_1. The distribution of its sojourn time depends on the locations of the remaining m workpieces ahead of it. The enumeration of all possibilities is fraught with combinatorial difficulties. We circumvent this using the following result.

Proposition 2: The time to produce item $m + 1$ has the same distribution as the *cycle time* of a job in a closed Jackson network with nodes i, $i - 1$, …, 1 and a total population of $m + 1$ jobs. The probability that this time is greater than QLT is given by

$$\mathrm{DP}_i(m) = \sum_{k=1}^{i} \left[\beta_{k,i}(m) \sum_{j=0}^{m} e^{-\mu_k \mathrm{QLT}} \frac{(\mu_k \mathrm{QLT})^j}{j!} \right]; \tag{17}$$

where, for distinct service rates,

$$\beta_{k,i}(m) = \frac{\alpha_{k,i} \mu_k^{-m}}{\sum_{j=1}^{i} \alpha_{j,i} \mu_j^{-m}} \quad \text{and} \quad \alpha_{k,i} = \prod_{\substack{l=1 \\ k \neq l}}^{i} \frac{\mu_l}{\mu_l - \mu_k}. \tag{18}$$

Proof: The first part of the proof is similar to the proof of Proposition 1. Consider an order arriving when the system is out of stock with m orders already pending. Suppose further that this order "sees" the system in the following state:

(iii)′ the first m workpieces are *at specific locations* between machines M_i, M_{i-1}, …, M_1;

(iv) one more (item $m + 1$) is in M_i;

(v) the remaining $s - 1$ workpieces are anywhere between machines M_N and M_i.

Event (iii)′ is a subset of (iii), which was defined in the proof of Proposition 1. The conditional probability of (iii)′ and (iv) given (iii)–(iv) is given by

$$\frac{P(\text{ arrival sees state } n_0 = c - m, n_1, \ldots, n_{i-1}, \text{ and } n_i = n_i' + n_i'' + 1 \text{ with } n_1 + \ldots + n_{i-1} + n_i' = m,)}{P(\text{ arrival sees } m \text{ orders ahead and item } m+1 \text{ at } M_i)}$$

http://mc.manuscriptcentral.com/tprs Email: ijpr@lboro.ac.uk

For Peer Review Only

$$= \frac{\frac{1}{G(s+c)}(q\lambda)^{-(c-m)}\left(\sum_{n_N+\ldots+n_{i+1}+n''=s-1}\mu_N^{-n_N}\ldots\mu_{i+1}^{-n_{i+1}}\mu_i^{-n_i''}\right)\mu_i^{-1}\left(\mu_i^{-n_i'}\mu_{i-1}^{-n_{i-1}}\ldots\mu_1^{-n_1}\right)(q\lambda dt)}{\frac{1}{G(s+c)}(q\lambda)^{-(c-m)}\left(\sum_{n_N+\ldots+n_{i+1}+n''=s-1}\mu_N^{-n_N}\ldots\mu_{i+1}^{-n_{i+1}}\mu_i^{-n_i''}\right)\mu_i^{-1}\left(\sum_{n_i'+n_{i-1}+\ldots+n_1=m}\mu_i^{-n_i'}\mu_{i-1}^{-n_{i-1}}\ldots\mu_1^{-n_1}\right)(q\lambda dt)}$$

$$= \frac{\mu_i^{-n_i''}\mu_{i-1}^{-n_{i-1}}\ldots\mu_1^{-n_1}}{G_{i,1}(m)}.$$

The above expression is the product-form equilibrium probability of a queueing network with m jobs and nodes i, $i-1$, …, and 1 in a cycle. Now the well-known *arrival theorem* for Jackson networks (see, e.g., Sevcik and Mitrani 1981) says that this probability is the same as the probability of the state seen by a job arriving at node i from node 1 in the same cyclic network but with a total population $m+1$. Therefore the distribution of the time until item $m+1$ is produced is the same as the distribution of the cycle time in the network with nodes i, $i-1$, …, and 1 and population $m+1$. Moreover, the probability that this time is greater than QLT is given by equation (17), as shown in Zazanis (2004). When all the node rates are different, $\beta_{k,i}(m)$ and $\alpha_{k,i}$ are given by equation (18). When this is not the case, the coefficients can be computed using a partial-fractions expansion of the z-transform of the normalizing constant $G_{i,1}(m+1)$ or, equivalently, the method of Lam (1977).

3.2.3 Estimation of performance indices. Combining Propositions 1 and 2 we see that the probability Π_m that the waiting time of an arriving customer will exceed QLT when m orders are already pending is given by

$$\Pi_m = \sum_{i=1}^{N} P(\text{item } m+1 \text{ is at } M_i \mid n_B = m)\ \mathrm{DP}_i(m), \quad s \geq 1. \tag{19}$$

In the special case $s = 0$ we have a make-to-order policy, and each workpiece in the system is associated with an order. Thus, when order $m+1$ is placed, the corresponding item is released into the buffer of raw items upstream of M_N. In this case, equation (16) simplifies to

$$\Pi_m = \mathrm{DP}_N(m), \quad \text{for MTO systems.}$$

Using equations (9)–(10), (12)–(13), (15), and (17)–(19) we can compute the

For Peer Review Only

performance indices given by (5)–(8) and, finally, the average profit rate of the system given by (3).

4. Optimization

We now consider the problem of determining optimal values for s and c, denoted s_{opt} and c_{opt} respectively, so as to maximize the average profit rate of the system. We have the following propositions.

Proposition 3: The optimal base stock level is bounded from above as follows

$$s_{\text{opt}} < \frac{p\lambda}{h}.$$

Proof: The throughput of the system cannot exceed the demand rate λ. Also, because $n_H = s + n_B$, we have $H_{s,c} = E(n_H) \geq s$. It follows from equation (3) that $J(s, c) \leq p\lambda - sh$. Since $J(s, c)$ must not be negative, we obtain the bound on s_{opt}.

Proposition 4: The optimal base backlog level c_{opt} is bounded from above as follows

$$c_{\text{opt}} < \frac{p\mu_1}{h+b}.$$

Proof: Consider an arriving customer who finds $c - 1$ pending orders and decides to place one more. This customer must wait until c items are produced. The best-case scenario would be that these c items are in the last machine M_1 and require only one operation. The mean time to fill this customer's order is thus greater than c/μ_1 and the associated mean holding and backlog cost is $c(h + b)/\mu_1$. The latter must be less than p to ensure that the mean profit to be earned from this customer is positive. This yields a bound on c_{opt} above which it is not profitable for the system to accept an incoming order.

For Peer Review Only

The following algorithm performs an exhaustive search to find the optimal base stock and base backlog.

Algorithm 1:

1. Set $J(0, 0) = 0$, $J_{opt} = 0$, $s_{opt} = 0$ and $c_{opt} = 0$. Initialize the base stock level and the base backlog; thus, $s = 1$ and $c = 0$.
2. Compute the profit rate $J(s, c)$. Keep the optimal control parameters and the maximum profit rate in memory; thus, if $J(s, c) > J_{opt}$, then set $J_{opt} = J(s, c)$, $s_{opt} = s$ and $s_{opt} = c$.
3. Apply Proposition 3: If $(s + 1) < p\lambda/h$, then set $s := s + 1$ and go to step 2; else go to step 4.
4. Apply Proposition 4: If $(c + 1) < p\mu_1/(h + b)$, then set $c := c + 1$, $s = 0$, and go to step 2; otherwise stop.

It follows from Propositions 3 and 4, that the above algorithm requires a finite number of iterations to find the optimal control parameters.

5. Numerical results and discussion

We now use Algorithm 1 of the previous section to illustrate the advantages of the proposed BSBB policy over two simpler, yet commonly used policies.

5.1 *Analytical results for a production line with exponential machines*

Consider a production line similar to the one described in Section 2 consisting of six machines M_i, $i = 1, 2, \ldots, 6$, with exponentially distributed service and customer interarrival times. Standard values for the system parameters are: mean demand rate $\lambda = 4.0$, mean production rates $\mu_1 = 5.25$, $\mu_2 = 6.5$, $\mu_3 = 5.5$, $\mu_4 = 5.0$, $\mu_5 = 7.0$, $\mu_6 = 6.0$, unit profit $p = 100.0$, unit costs $h = 8.0$, $b = 8.0$ and $d = 10.0$,

http://mc.manuscriptcentral.com/tprs Email: ijpr@lboro.ac.uk
For Peer Review Only

and QLT = 1.0. The maximum waiting tolerance of customers is a random variable θ with a uniform distribution function in [0, 10], i.e., $G(t) = t/10$. Upon substitution of $G(t)$ into equation (1), we obtain $q = 0.9$. This means that 90% of the customers arriving during stockouts are willing to wait one time unit and place their orders; the rest balk.

We compare BSBB, denoted (s, c), with the lost sales (LS) and make to order (MTO) policies. Because these policies are special cases of BSBB (LS is the same as $(s, 0)$ and MTO coincides with $(0, c)$), their performance will not, in general, be as good as that of BSBB. We examine the effects of varying λ, h, b, d and QLT on the net profit rate of each policy. In each experiment, we modify one of these parameters fixing the others to their standard values and compute the optimal parameters under BSBB, LS and MTO. Tables 2 and 3 show the profit rate for various values of λ and QLT. Figure 2 shows the profit rate versus the various values of the cost parameters h, b and d.

Insert Table 2 about here

From Table 2 we see that when λ is greater than $\min\mu_i = 5$, LS and BSBB perform similarly. This happens because BSBB tends to reject customers during stockouts and keeps a high stock when customers arrive frequently. However, as the demand rate decreases the probability of a stockout decreases as well, and the system does not need to maintain high stock levels.

An interesting result can be drawn from Table 3, where we observe that BSBB always performs better than but close to LS for all values of QLT we tested. When QLT is zero, all customers who arrive during stockouts place their orders with probability one. In this case, BSBB reacts to protect the system against delay penalties, so it rejects all customers beyond $c = 1$ and, therefore, it is similar to an LS policy. As QLT increases, customers become more reluctant to place orders and so a relaxed admission control policy benefits the firm more.

For Peer Review Only

Insert Table 3 about here

Insert Figure 2 about here

Finally we investigate the second-order properties of the mean profit rate for the BSBB policy with respect to s and c. Figure 3 shows that the profit rate appears to be a unimodal function in each parameter s and c when the other is fixed. Unimodality with respect to s implies that the optimal base stock for a system with fixed base backlog c is the value s_c such that $J(s_c, c) > J(s_c + 1, c)$. Unimodality with respect to c can be defined in a dual fashion. Establishing this property could reduce the computational effort for tracking down the optimal parameters for BSBB since, at step 3 or 4 of Algorithm 1, a jump to step 2 would occur whenever a local maximum for s or c, respectively, is attained regardless of validity of the condition of Proposition 3 or 4.

Insert Figure 3 about here

The profit rate also appears to be a submodular function. By submodularity, if c_s is the largest value of c at which $J(s, c)$ is maximized for some fixed s, then c_s is decreasing in s (here we take decreasing to mean nonincreasing). For example, from the second graph of Figure 3 we see that the maxima of the profit rate for $s = 0$, 6, 11, and 16 are respectively at $c_0 = 10$, $c_6 = 4$ and $c_{11} = c_{16} = 3$. Similarly, submodularity of J implies that $s_c \le s_{c-1}$. Use of submodularity can potentially reduce the search effort. For example, if $s_c \le s_{c-1}$ holds then, once we have found the optimal base stock s_{c-1} for base baclog $c - 1$, the optimal base stock s_c for c is in the subset $\{0, 1, \ldots, s_{c-1}\}$; thus in step 3 of Algorithm 1 we do not need to try values for s larger than s_{c-1} even though they may satisfy the inequality of Proposition 3.

For Peer Review Only

5.2 *Simulation results for production lines with non-exponential machines*

So far we have assumed that the processing times of all machines of the production line are exponentially distributed. This assumption permits modeling the system as a Jackson network and expressing analytically all important performance indices. However, this assumption is often violated in practice. For example, Inman (1999) studied the machine processing times in two automotive assembly plants and observed that the Erlang distribution provides more reasonable approximations of the empirical processing time distributions. To investigate the performance of the proposed joint inventory/admission control policy for more realistic systems, we examine six-machine production lines whose standard parameter values are as in the previous section. The processing times have Erlang distributions with various stages but the mean processing time of each machine M_i is equal to $1/\mu_i$, as in the exponential case. Optimal control parameters for each case are found by applying Algorithm 1 and using simulation, rather than analysis, in step 2 to compute the profit rate $J(s, c)$. Note that Propositions 3 and 4 hold regardless of the type of distribution assumed for the processing or interarrival times.

Four distributions of processing times are tested: exponential (Erlang-1), Erlang-2, Erlang-4 and deterministic (Erlang-∞). To perform a fair comparison, we used common random numbers to generate the processing times for each machine. In all experiments, the errors in estimating the mean profit rate of the system were less than 0.5% with a 95% confidence level. On an Intel Core 2 Duo 2.33GHz personal computer, the computational time for an exhaustive search of the control parameters varies from about half to three hours, depending on the parameters of the system and especially the processing time distributions.

Table 4 gives simulation results for the maximum profit rates under BSBB, LS and MTO and the corresponding optimal values for s and c. The column labeled "QLT = 1" of Table 2 and column "Exponential" of Table 4 correspond to the analytical and simulation results of the same system. We see that the simulation model gives the exact optimal control parameters, except for the MTO policy where the optimal c is overshot by 1 (yet the profit estimates $J \approx 177$ of the two models deviate by less than 0.002%). As the number of Erlang stages increases

For Peer Review Only

from 1 (exponential case) to infinity (deterministic case), the variance of processing times decreases to zero. From Table 4 we see that decreasing the variance results in an increase of the profit rate, a reduction of the safety stock s under BSBB and LS and an increase of the maximum number c of backorders allowed under BSBB and MTO.

Insert Table 4 about here

We also examine the second-order properties of the mean profit rate with respect to s and c under the BSBB policy. For this experiment we use an Erlang-2 distribution for the machine processing times and standard values for the remaining system parameters. Figure 4 shows that the profit rate is a unimodal and submodular function as in the exponential case. Establishing these properties for non-exponential production lines would substantially reduce the number of simulation runs needed to optimize the system. Because the optimal parameters for non-exponential systems are not too far from those of exponential systems, one could track down the optimal control parameters of the latter, using the fast analytical model of Section 3, and set them as initial values in a local search thereafter using simulation. Some guidelines for performing local search are given in Section 5.1.

Insert Figure 4 about here

6. Conclusion

In this paper, we have studied a control policy for production lines with exponential machines and impatient customers. During stockouts, the system informs customers about anticipated delays by quoting a fixed lead time, QLT, to each arriving order, and offers a discount per unit of delay beyond QLT. An arriving customer balks if QLT exceeds his due time, which is a random variable with known distribution.

For Peer Review Only

A joint threshold-type inventory and admission control policy is proposed to maximize the net profit rate of the system. The system accepts orders as long as the backlog level is under c and produces until the number of finished products reaches s. We derive analytical expressions for the performance indices involved in the mean profit rate of the system and develop a simple but efficient algorithm that uses bounds on the optimal parameters and performs an exhaustive search to determine their values. These bounds hold regardless of the distributions of machine processing times.

We have compared the proposed policy, BSBB, with the LS (lost sales) and MTO (make-to-order) policies. In general BSBB manages the values s and c in an optimal fashion and shows a better adaptability than LS or MTO to seasonal demand variations and changes in the cost parameters. Systems with exponentially distributed processing times are analytically tractable and the computational time required to find the optimal control parameters is very short. Systems with more general distributions are analyzed with exceptional accuracy and optimized using simulation. In all cases, BSBB performed much better than LS and MTO.

From the numerical results it appears that the mean profit rate of the system is a unimodal and submodular function with respect to the control parameters. As these properties could significantly reduce the search effort for finding the optimal control parameters, they are subjects of future work. Also, because QLT has a significant impact on system performance, considering it as an additional control parameter to be optimized is another challenging subject of research.

Acknowledgements

A. A. Economopoulos was supported in part by a doctoral scholarship from the Alexandros S. Onassis Public Benefit Foundation, Greece.

Refeences

Baccelli F. and Hebuterne, G., 1981. *On queues with impatient customers* [online]. INRIA report 94, France. Available from:

http://mc.manuscriptcentral.com/tprs Email: ijpr@lboro.ac.uk

For Peer Review Only

http://hal.inria.fr/inria-00076467/en/ [Accessed 24 August 2009].

Brandt, A. and Brandt, M., 1999. On the M(n)/M(m)/s queue with impatient calls. *Performance Evaluation*, **35** (1–2), 1–18.

Buzacott, J.A. and Shanthikumar, J.G., 1993. *Stochastic Models of Manufacturing Systems*. Englewood Cliffs, NJ: Prentice Hall.

Buzen, J.P., 1973. Computational algorithms for closed queuing networks with exponential servers. *Communications of the ACM*, **16**, 527–531.

Caldentey, R.A., 2001. *Analyzing the make-to-stock queue in the supply chain and e-business settings*. Thesis (PhD). MIT.

Economopoulos, A.A. and Kouikoglou, V.S., 2008. Analysis of a simple CONWIP system with impatient customers. *In*: Proceedings of the *4th Annual IEEE Conference on Automation Science and Engineering*, 23–26, 2008, Washington DC, USA.

Economopoulos, A.A. and Kouikoglou, V.S., 2009. Lead-time analysis and a control policy for production lines with exponential machines. *In*: Proceedings of the *7th Intrenational Conference on Stochastic Models of Manufacturing and Service Operations*, 7–12 June 2009, Ostuni, Italy

Haight, F.A., 1957. Queueing with balking. *Biometrika*, **44** (3–4), 360–369.

Haight, F.A., 1959. Queueing with reneging. *Metrika*, **2** (1), 186–197.

Haight, F.A., 1960. Queueing with balking II. *Biometrika*, **47** (3–4), 285–296.

Inman, R.R., 1999. Empirical evaluation of exponential and independence assumptions in queueing models of manufacturing systems. *Production Planning and Control*, **8** (4), 409–432.

Ioannidis, S. and Kouikoglou, V.S., 2008. Revenue management in single-stage CONWIP production systems. *International Journal of Production Research*, **46** (22), 6513–6532.

Ioannidis, S., Kouikoglou, V.S. and Phillis, Y.A., 2008. Analysis of admission and inventory control policies for production networks. *IEEE Transactions on Automation Science & Engineering*, **5** (2), 275–288.

Kouikoglou, V.S. and Phillis, Y.A., 2002. Design of product specifications and control policies in a single-stage production system. *IIE Transactions*, **34** (7), 591–600.

Lam, S.S, 1977. An extension of Moore's result for closed queueing networks. *IBM Journal of Research and Developement*, **21**, 384–387.

For Peer Review Only

Liu, L. and Kulkarni, V.G., 2008. Balking and reneging in M/G/s systems exact analysis and approximations. *Probability in the Engineering and International Sciences*, **22** (3), 355–371.

Movaghar, A., 2006. On queueing with customer impatience until the end of service. *Stochastic Models*, **22** (1), 149–173.

Naor, P., 1969. The regulation of queue size by levying tolls. *Econometrica*, **37** (1), 15–24.

Porteus, E.L., 1990. Stochastic inventory theory. *In*: D.P. Heyman and M.J. Sobel, eds. *Handbooks in Operations Research and Management Science: Stochastic Models, Vol. 2*.Amsterdam, The Netherlands: North-Holland, 605–652.

Sevcik, K.C. and Mitrani, I., 1981. The distribution of queuing network states as input and output instants. *Journal of the ACM*, **28** (2), 358–371.

Subba Rao, S., 1967. Queueing models with balking and reneging. *Annals of the Institute of Statistical Mathematics*, **19** (1), 55–71.

Ward, W., 1999. Improving service by informing customers about anticipated delays. *Management Science*, **45** (2), 192–207.

Zazanis, M.A., 2004. Cycle times in single server cyclic Jackson networks. *Operational Research Letters*, **32** (5), 422–430.

Zipkin, P.H., 2000. *Foundations of Inventory Management*. New York: McGraw-Hill.

For Peer Review Only

FOUR FIGURES IN SEPARATE PAGES

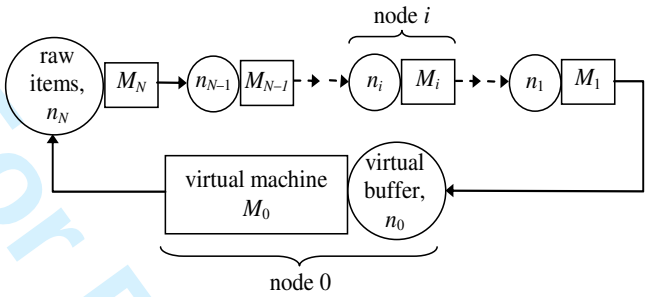

Figure 1. Equivalent cyclic network of queues.

For Peer Review Only

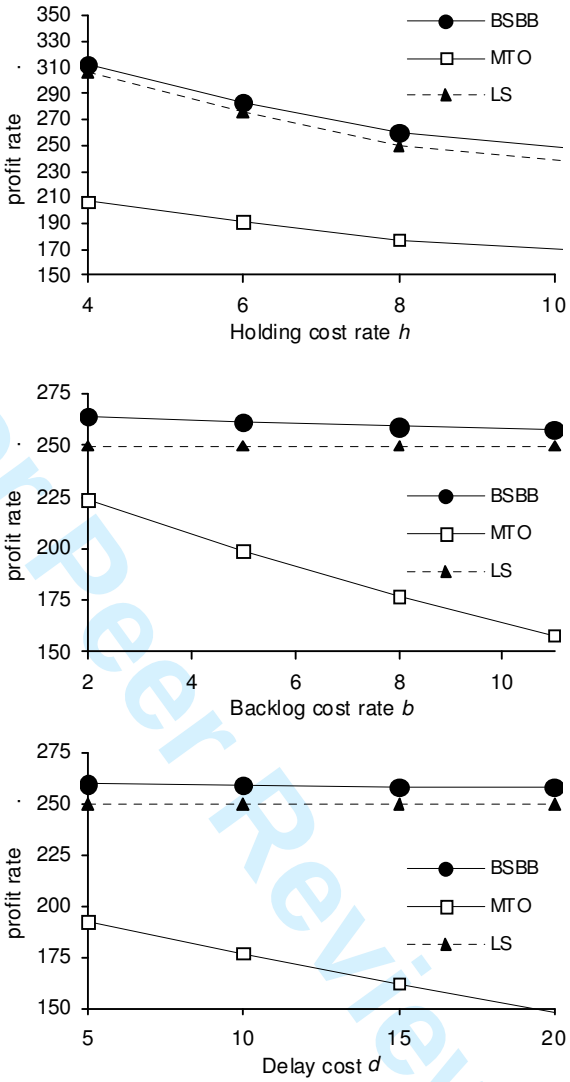

Figure 2. Profit rate versus various values of system parameters.

For Peer Review Only

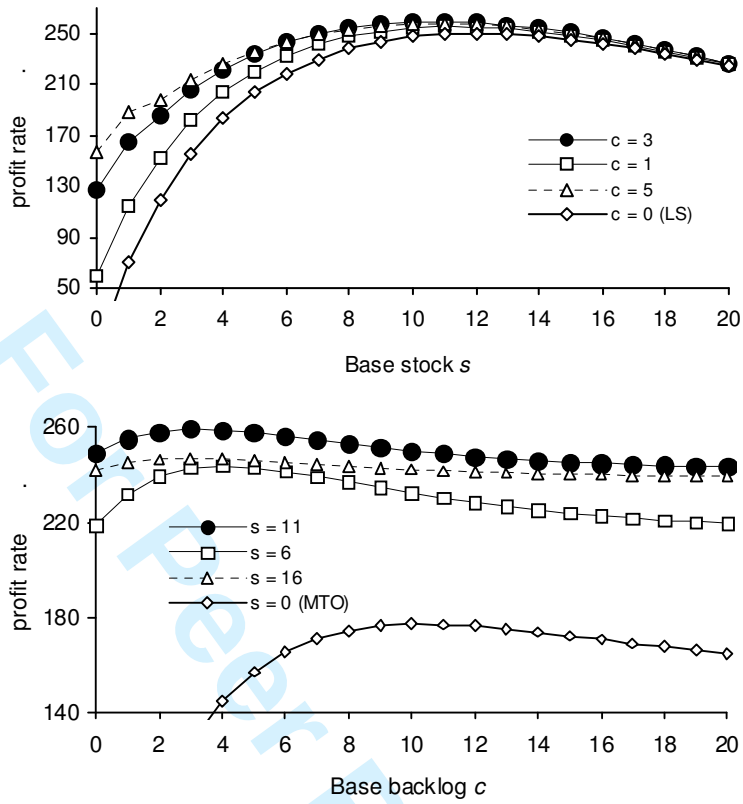

Figure 3. Profit rate versus control parameters.

For Peer Review Only

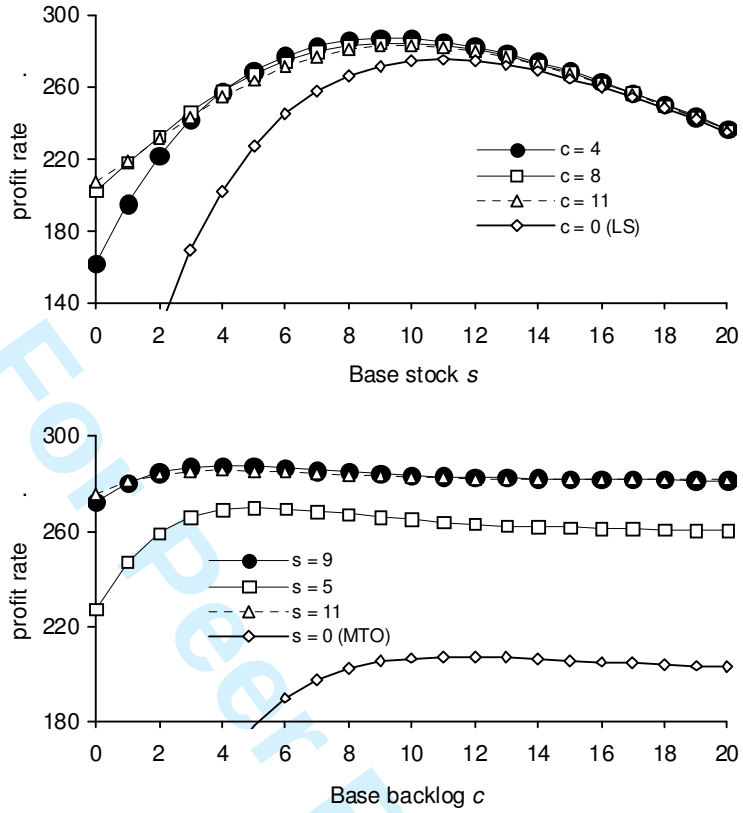

Figure 4. Profit rate versus control parameters (Erlang-2 distribution).

For Peer Review Only

FOUR TABLES IN SEPARATE PAGES

Table 1. Relationship among state variables.

Production system			Cyclic network of queues
n_F	n_B	n_H	n_0
s	0	s	$s+c$
$s-1$	0	s	$s+c-1$
⋮	⋮	⋮	⋮
1	0	s	$c+1$
0	0	s	c
0	1	$s+1$	$c-1$
⋮	⋮	⋮	⋮
0	c	$s+c$	0

For Peer Review Only

Table 2. Optimal control parameters and profit rate versus demand rate.

	Demand rate											
	$\lambda = 3$			$\lambda = 4$			$\lambda = 4.95$			$\lambda = 6.95$		
Policy	s	c	J	s	c	J	s	c	J	s	c	J
BSBB	7	6	215.97	11	3	259.21	13	2	280.39	13	1	294.17
MTO	0	13	160.29	0	10	177.06	0	9	182.81	0	8	186.51
LS	9	0	202.10	12	0	249.91	14	0	275.48	14	0	293.18

For Peer Review Only

Table 3. Optimal control parameters and profit rate versus QLT.

	Quoted lead time											
	QLT = 0			QLT = 1			QLT = 2			QLT = 5		
Policy	*s*	*c*	*J*	*s*	*c*	*J*	*s*	*c*	*J*	*s*	*c*	*J*
BSBB	12	1	250.69	11	3	259.21	11	4	259.22	11	5	254.61
MTO	0	10	179.40	0	10	177.06	0	10	185.12	0	17	148.83
LS	12	0	249.91	12	0	249.91	12	0	249.91	12	0	249.91

For Peer Review Only

Table 4. Optimal control parameters and profit rates for various distributions of machine processing times (simulation results).

	Distribution											
	Exponential			Erlang-2			Erlang-4			Deterministic		
Policy	*s*	*c*	*J*	*s*	*c*	*J*	*s*	*c*	*J*	*s*	*c*	*J*
BSBB	11	3	259.26	9	4	287.65	8	5	304.86	7	5	327.47
MTO	0	11	177.06	0	11	207.27	0	12	226.60	0	12	249.75
LS	12	0	249.96	11	0	275.73	10	0	291.62	9	0	313.46

For Peer Review Only