PRODUCTION BATCHING WITH MACHINE BREAKDOWNS AND SAFETY STOCKS

HARRY GROENEVELT
University of Rochester, Rochester, New York

LILIANE PINTELON
Katholieke Universiteit Leuven, Leuven, Belgium

ABRAHAM SEIDMANN
University of Rochester, Rochester, New York

(Received January 1990; revisions received September 1990, April, July 1991; accepted August 1991)

We study the problem of selecting the economic lot size for an unreliable manufacturing facility with a constant failure rate and general randomly distributed repair times. Safety stocks must be used to meet the managerially prescribed service level (the fraction of lost sales) because these stochastic interventions reduce the effective production capacity. We develop bounds on the range of feasible service levels and investigate the impact of several system parameters on this range. We introduce an easily implementable production control policy (but do not establish the optimality of its structure) and prove that under this policy the safety stock dynamics can be characterized fully by a renewal process analogous to the workload process of a special single server queueing system. This analogy is exploited in deriving exact and approximate expressions for the safety stock holding costs. Several operational insights are revealed by experimenting with the models developed here. We show how the results can be incorporated in a broader management framework for evaluating resource allocation decisions aimed at reducing the failure rate of the machines. A clear tradeoff is shown to exist between the overall investment in increasing the maintenance level and the resulting savings in safety stocks and repair costs.

The lot sizing decision is a fundamental management aspect of many production facilities. In spite of this, the production lot sizing decision has not been studied extensively in the context of unreliable facilities. This paper presents a new lot sizing model for an unreliable manufacturing facility operating under a constant demand rate. The random nature of the repair times reduces the effective machine capacity. Safety stocks are therefore introduced in order to maintain a managerially prescribed service level. The dynamics of the safety stock process are fully characterized. In addition, the sensitivity of the optimal lot size and safety stock levels with respect to failure rate, service level, demand rate, setup and repair times is investigated. These results are incorporated into a broader management framework for economically evaluating resource allocation decisions aimed at improving the productive reliability of the facility.

Current trends in manufacturing management have led to reduced levels of buffer inventories and increased levels of automation and equipment complexity. This trend makes effective maintenance policies more important than ever. Most of the past research on maintenance policies (for surveys see Barlow, Prochan and Hunter 1965, McCall 1965, Pierskalla and Voelker 1976, Sherif and Smith 1981), however, did not really consider the interaction effects of these policies with production planning and control systems. Furthermore, maintenance policy interactions used to get little attention in the manufacturing literature (McLeavey and Narashiman 1985, Silver and Peterson 1985, Nahmias 1989). The main focus of modeling stochastic aspects concerned uncertainty in demand, lead time or yield. Zipkin (1986) models a one-machine production facility making many products in large discrete batches, when demands and the production process are stochastic. His modeling approach integrates inventory and queueing models with a classical optimization analysis. Karmarkar (1987) examines a similar one-machine facility and studies the systematic relationships between the economic production lot size and lead times and the

Subject classifications: Inventory/production, uncertainty: impact of breakdowns on batching and safety stocks. Inventory/production, lot sizing: with breakdowns and safety stocks. Reliability, maintenance/repairs: batching and safety stocks.
Area of review: MANUFACTURING, PRODUCTION AND SCHEDULING.

0030-364X/92/4005-0959 $01.25

implications of these relationships for batching decisions and work-in-process inventories. It is proven there that the appropriate lot sizes are quite different from those associated with the fully deterministic economic manufacturing quantity (EMQ) models.

Recent research efforts attempt to incorporate the imperfections of the production process (i.e., quality and yield issues) and equipment (i.e., machine breakdowns and repairs) into the classical lot sizing decision framework. Gallego (1988a) extends the classical ELSP model (economic lot scheduling problem) by providing an algorithm for scheduling the facility after a single disruption. The later study by Gallego (1988b) presents a novel heuristic for multiple disruptions in the same context. Bielecki and Kumar (1988) analyze a one-machine single product model and show that there exists a range of parameter values describing an unreliable manufacturing system for which zero-inventory policies are exactly optimal even when the manufacturing capacity is uncertain. The unreliability of the system studied by Bielecki and Kumar is a result of stochastic breakdowns and repairs (both exponentially distributed). Lee and Rosenblatt (1987) study the joint control of production cycles and maintenance by inspection in a similar facility. The decision variables are the production lot size and the number of inspections per cycle. The inspections help to determine whether the equipment is in-control or out-of-control by looking at the current percentage of defective items produced. If the equipment is out-of-control, a maintenance intervention is required to restore it to an in-control state. Economic lot sizing implications of uncertain machine yield are extensively studied by Lee and Yano (1985), Porteus (1986), and Henig and Gerchak (1990). Groenevelt, Seidmann and Pintelon (1992) focus on the effects of machine breakdowns and corrective maintenance on economic lot sizing decisions. Assuming instantaneous setups and maintenance interventions, they show that the optimal lot sizes will always be greater than the ones needed in the deterministic EMQ case, and that significant cost savings can be realized when a threshold-like policy is used to determine the manufacturing starting point for a new production cycle.

Interest in integrated models is not only evident by recent work on production and maintenance models, but also by work in the (closely related) quality control field. Examples of this research are Peters, Schneider and Tang (1988) who combine a fixed order quantity inventory control system with a Bayesian quality control system, Tagaras (1988) who presents a model for the simultaneous optimization of the design parameters for process control and maintenance, and Tapiero (1986) who combines product quality requirements and maintenance in a linear control model. Other related studies include Pate-Cornell, Lee and Tagaras (1987), Keller and Noori (1988), and Schwaller (1988). To our knowledge, the analytical framework presented here is the first to combine lot sizing and safety stock decisions in the presence of equipment breakdowns.

The remainder of this paper is organized as follows. In Section 1 we specify the model's assumptions in detail and introduce the notation used throughout the paper. Section 2 develops the basic batching and safety stock model. Repair times can have a general distribution, while the time-to-failure is assumed to be exponentially distributed. Both the zero and the nonzero setup time cases are considered. Some computational experiments are reported in Section 3. The maintenance budget level concept is introduced in Section 4, and Section 5 presents several generic conclusions and indicates directions for further research.

1. ASSUMPTIONS AND NOTATION

The analysis presented here focuses on the lot sizing problem for a production facility with known demand and a finite processing rate. Typical examples of such manufacturing systems are: die casting, stamping and press punching in the automotive industry or those unitary cells used for printed circuit board assemblies (Seidmann and Nof 1989). The machine is subject to stochastic breakdowns and repairs, and to provide for smooth deliveries we need to establish safety stocks which are depleted over those time intervals when the machine is being repaired. Major issues in the control of these safety stocks are 1) building up the safety stocks, and 2) avoiding uncontrolled accumulation of safety stocks.

To cater to these two issues we propose the following production control scheme that results in a finite long-run average safety stock (thereby assuring ergodicity of the safety stock system): During each production run, we assume that a certain *fraction* β of the items produced is diverted into the safety stock, whereas the rest of the items are used on-schedule to meet regular customer demands. As a result, after producing a lot of q units, $q\beta$ units will be added to the safety stock, and $q(1-\beta)$ units will be added to the regular running stock to be delivered to customers during the cycle. The safety stock is used to satisfy demand when the machine undergoes repair following a breakdown. Lost sales will occur when the machine is down and *safety stocks* are depleted, regardless of the running stock level.

Figure 1 illustrates the resulting sample path for the running and safety stock levels. One distinguishes four different types of time intervals in this sample path. In periods of type I the machine is working, so both safety and running stocks increase with rates βr and $r(1 - \beta) - d$, respectively, where r is the production rate and d is the demand rate in units of product per unit time. In type II periods the machine is broken down and undergoes repair, the running stock remains constant and the safety stock decreases with a rate d. If the safety stock is not sufficient to cover the demand during the repair period (and the subsequent setup) lost sales are incurred (type III periods). During type IV periods the machine is idle, and a preventive maintenance job is carried out along with the setup for the next cycle. During this period the running stock is depleted and the safety stock level does not change. Note that the sample path for the running stock is the same as for the corresponding EMQ model if only periods of types I and IV are considered.

We assume that the (production) times between successive failures are i.i.d. exponentially distributed, and that the times to repair are i.i.d. random variables with a general distribution. A machine setup is required following each maintenance intervention (both corrective and preventive maintenance). The notation is summarized in Table I.

Table I
Summary of Notation

Item	Definition
r	Production rate (units/day)
d	Demand rate (units/day)
h	Inventory holding cost (\$/unit/day)
s	Setup cost (\$)
p	Preventive maintenance cost (\$)
c	Corrective maintenance cost (\$/day)
t_s	Setup time (days)
t_m	Preventive maintenance time (days)
q, Q	Lot size, optimal lot size, respectively (units)
α	Target service level, i.e., desired fraction of total demand satisfied
β	Fraction of the production volume diverted to safety stock accumulation
λ	Mean failure rate (1/day)
t_f	Time-to-failure (days), assumed to be exponentially distributed with rate λ
t_r	Time-to-repair, repair time (days)
f_i	Fraction of time the system is in period type i, where i = I, II, III or IV
V_i	Long-run average safety stock level in period of type i (units), $i = I, \ldots, IV$
$\underline{V}$	Long-run average overall safety stock level (units)
$E[x]$	Expected value of random variable x
Var$[x]$	Variance of random variable x

1.1. Discussion

Two comments are in order on the inventory control policy described above. First, our inventory control policy allows lost sales to occur in some cases even when cycle stock is not depleted. Second, our policy does not set an a priori maximum for the amount of safety stock that will be allowed to accumulate. These two properties make it quite likely that superior alternative inventory control rules exist. The primary motivation for considering this policy is that it permits an analysis in which the lot size decision is, to a certain extent, decoupled from the safety stock decision. This decoupling is similar in spirit to that employed in many well known inventory models, see, e.g., McLeavey and Narashiman (1985), Silver and Peterson (1985), or Nahmias (1989). It simplifies the analysis considerably and has the additional advantage that one can unambiguously allocate (the cost of carrying) inventory to the underlying causes. Alternative inventory control rules (e.g., one with upper and lower control limits, or with state-dependent lot sizes) are much harder to analyze exactly and may also be more difficult to implement, as discussed below.

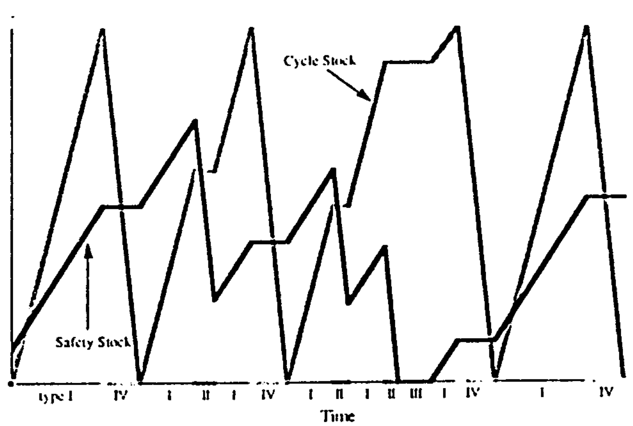

Figure 1. Safety stock model: inventory sample paths. The time-to-failure (t_f) and the time-to-repair (t_r) are random variables.

With regard to the issue of keeping the cycle stock and safety stock separate, we point out that this is a common practice in some production processes, such as automobile manufacturing, chemical processing, and circuit board assembly, to retain a certain minimal cycle stock even after the safety stock is depleted. This practice facilitates a smooth resumption of parts flow between the various production stages after the completion of a repair. In addition, the simulation results presented in Section 3 indicate that the service level and the total cost per unit time achieved by a system run according to our policy are quite close to those achieved when running and safety stocks are pulled together.

Concerning the issue of limiting the safety stock to a finite amount, we show in Section 2 that the safety stock fluctuates over time very much like the workload process of a GI/M/1 queue. Hence, the possibility exists that large amounts of safety stock accumulate from time to time. Standard numerical methods are available for estimating the probability that the safety stock level will get beyond any given threshold (Tijms 1986). A natural modification is, therefore, a policy that limits the amount of safety stock that will be allowed to accumulate. The results developed in this paper can be used to approximate the performance of that modified inventory control policy. The quality of this approximation will be best when the occupancy of the GI/M/1 queue is low, as will be the case when the required service level and the machine utilization are low. The details will be left to the interested reader.

2. BASIC BATCHING AND SAFETY STOCK MODEL

In this section, the basic model will be analyzed. In Section 2.1 the range of feasible service levels is derived, and we show how to achieve a specified level by choosing the diversion rate β appropriately. The batch size optimization problem is considered in Section 2.2, and the steady-state behavior of the safety stock is analyzed in Section 2.3.

2.1. Service Level Analysis

Let $\underset{\sim}{\alpha}(\beta)$ be the service level achieved (= the long-run average fraction of demand satisfied) when we set aside a fraction β of each lot produced as safety stock. By the renewal reward therorem (Ross 1983), the achieved service level is equal to

$$\underset{\sim}{\alpha}(\beta) = \frac{E[\text{production per cycle}]}{E[\text{demand per cycle}]}, \tag{1}$$

where a cycle is the period elapsed between the start of production of a lot of q units to the start of production of the next lot of q units, and includes production time, repair and setup time (if a failure occurs), and idle and setup time. We obviously have

$$E[\text{production per cycle}] = q,$$

and

$$E[\text{demand per cycle}]$$

$$= d \cdot E[\text{length of a cycle}]$$

$$= q(1 - \beta) + q\frac{d}{r}(E[t_r] + t_s)/E[t_f].$$

(See Appendix A for the derivation of the expected cycle length.) Hence, we get

$$\underset{\sim}{\alpha}(\beta) = \frac{r/d}{(1 - \beta)(r/d) + (E[t_r] + t_s)/E[t_f]}. \tag{2}$$

Note that the service level achieved with a given diversion rate does not depend on the lot size q! This is a consequence of the assumption that the failure rate is constant and independent of the lot size. To find the diversion rate needed to achieve a specified service level α, solve for β in the equation $\underset{\sim}{\alpha}(\beta) = \alpha$ to obtain

$$\beta = 1 - \frac{1}{\alpha} + \frac{d}{r}\frac{E[t_r] + t_s}{E[t_f]}, \tag{3}$$

which is again independent of the lot size.

This leads to the following proposition.

Proposition 1. *A service level α is feasible if and only if*

$$\frac{r/d}{r/d + (E[t_r] + t_s)/E[t_f]}$$

$$\leq \alpha \leq \min\left\{1, \frac{r/d}{1 + (E[t_r] + t_s)/E[t_f]}\right\}. \tag{4}$$

Proof. We assume that demand during production and idle periods is satisfied from running stock, hence, we require that the running stock accumulation rate exceed the demand rate, i.e., $(1 - \beta)\, r \geq d$, so $0 \leq \beta \leq (1 - d/r)$. Noting that $\underset{\sim}{\alpha}(\beta)$ is strictly increasing in β we find from (2) that

$$\frac{r/d}{r/d + (E[t_r] + t_s)/E[t_f]} \leq \alpha$$

$$\leq \frac{r/d}{1 + (E[t_r] + t_s)/E[t_f]}.$$

Since, of course, $\alpha \leq 1$, (4) follows.

It is not hard to show that a service level $\alpha = 1$ can only be achieved at the expense of an infinitely large buildup of safety stock.

2.2. Optimal Batch Size Derivation and Running Cycle Cost Analysis

Since the value of β does not depend on the lot size q (see (3)), the cost of holding safety stock does not influence the optimal lot size Q. Consequently, we can determine Q by minimizing the sum of the other components of long-run average cost: running stock, preventive maintenance, corrective maintenance, and setups. The long-run average cost of holding safety stock is derived in Section 2.3. The other cost components are:

running stock: $\frac{1}{2}hq(1-\beta-(d/r))$,

preventive maintenance:

$$\frac{p}{(q/d)(1-\beta)+(q/r)\lambda(E[t_r]+t_s)},$$

corrective maintenance:

$$\frac{\lambda(q/r)E[t_r]c}{(q/d)(1-\beta)+(q/r)\lambda(E[t_r]+t_s)},$$

and

setups: $$\frac{s(1+\lambda q/r)}{(q/d)(1-\beta)+(q/r)\lambda(E[t_r]+t_s)};$$

see Appendix A for a complete derivation of these cost expressions.

The optimal lot size Q is the maximum of Q_1 and Q_2, where Q_1 is the lot size that minimizes the average running cycle cost expression for the service level specified, and Q_2 is the minimal feasible lot size (lower bound). Lot size Q_1 is found by taking the first derivative with respect to q of the average running cycle cost and setting it equal to zero. We find

$$Q_1=\sqrt{\frac{2d(s+p)}{h[(1-\beta)-(d/r)]\cdot[(1-\beta)+\lambda(d/r)(E[t_r]+t_s)]}}. \tag{5}$$

Lot size Q_2 is the minimum for which the feasibility constraint holds. This constraint expresses the condition that the idle time per cycle must be large enough to accommodate both setup and preventive maintenance: q is a feasible lot size if and only if

$(q/r)[(1-\beta)r-d]/d \leq t_s+t_m$.

From this it is easy to see that

$$Q_2=\frac{rd(t_s+t_m)}{(1-\beta)r-d}. \tag{6}$$

It is easy to verify that the long-run average cost is convex in the lot size q. This leads to following proposition.

Proposition 2. *The lot size* $Q=\max(Q_1, Q_2)$ *minimizes the average running cycle cost.*

Whether Q will equal Q_1 or Q_2 depends on the parameter values. Figure 2 illustrates this for several values of the parameters p (preventive maintenance cost), s (setup cost), and d (demand rate). The values of the input parameters are summarized in Table II. It is clear from Figure 2 that for the high $(p+s)$ case, (i.e., $p=500$ and $s=1{,}500$), Q_1 is greater than Q_2 over all demand rates between 500 and 900. For the

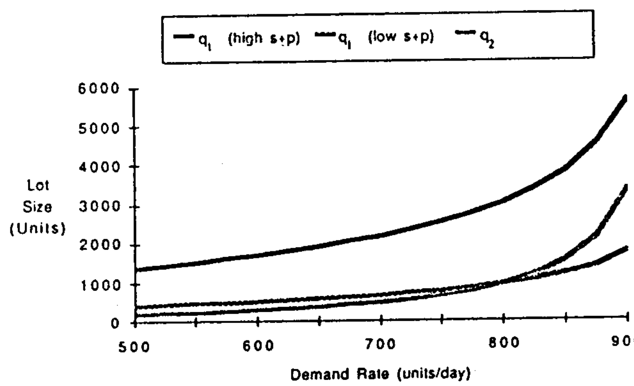

Figure 2. Safety stock model: effect of demand rate on lot sizes q_1 and q_2.

Table II
Data Used for Figures 2 to 10

Parameter	Figure 2	3	4	6, 10	7	8	9
r (units/day)	1,000	1,000	1,000	1,000	1,000	1,000	1,000
d (units/day)	*[a]	900	800, 850, 900	800	800	700, 800, 900	800
h ($/unit/day)	2	2	2	2	2	2	2
s ($)	1,500, 150	1,500	500	1,500	1,500	1,500	1,500
p ($)	500, 50	500	500	500	500	500	500
c ($/day)	1,000	1,000	1,000	1,000	1,000	1,000	1,000
t_s (days)	0.1	0.1	0	0	0	0	0
t_m (days)	0.1	0.1	0.1	0.1	0.1	0.1	0.1
λ (1/days)	0.1	0.1	0.2	*	*	*	*
$E[t_r]$ (days)	1	1	1	1	1	1	0.67, 1, 1.5
α	0.95	0.95	*	0.975	0.925, 0.95, 0.975	0.95	0.95

[a] * indicates parameters that are varied in figure.

low ($p + s$) case (i.e., $p = 50$ and $s = 150$) Q_1 is larger than Q_2 for demand rates between 500 and 800 and smaller than Q_2 for demand rates between 800 and 900. This can be explained by the structure of (5): Q_1 is proportional to $\sqrt{p + s}$, so the higher this factor the more likely it is that the correponding idle time will allow for both the setup and preventive maintenance.

2.3. Safety Stock Dynamics

Define W_q as the expected average customer waiting time in a GI/M/1 queue with a service rate of $\tilde{\mu} = d/r\beta E[t_f]$ and interarrivals distributed as $t_r + t_s$. Let ρ be the utilization ratio of this queueing system, i.e.,

$$\rho = \frac{r\beta E[t_f]}{d(E[t_r] + t_s)}.$$

With (3) it follows easily that $\rho < 1$ if and only if $\alpha < 1$. We now prove the following theorem which establishes the queueing analogy of the safety stock sample path.

Theorem 1. *The long-run average cost of holding safety stock is given by $h\underline{V}$, where the average safety stock level $\underline{V}$ is given by*

$$\underline{V} = \alpha d(W_q + E[t_r] + t_s) - (1 - \alpha)rE[t_f]. \quad (7)$$

Proof. We will calculate the average safety stock levels separately for each of the period types in the safety stock sample path (see Figure 1). The average safety stock level $\underline{V}$ is then computed by taking the weighted average of the expected inventory levels V_i associated with each period type (i = I, II, III, IV). The weights used are the fractions of time f_i the system dwells in each of these four period types.

To find the average inventory level in periods II and III, we temporarily cut out periods I and IV from the safety stock sample path. After scaling the vertical axis (d units of demand equal 1 unit of time), the resulting picture is that of the workload sample path of the GI/M/1 queueing system introduced above. We will thus interpret the resulting inventory level sample path as the workload sample path of a single server queueing system. The upward jumps in our sample path correspond to customer arrivals in the queueing system. The size of a jump in the sample path corresponds to the actual service requirement of an arriving customer in the queueing system or to the total of production volume diverted to safety stock between two consecutive failures of the machine divided by the demand rate. The down-sloping parts of the sample path correspond to service taking place between customer arrivals in the queueing system, or to safety stock depletions during corrective maintenance and setup intervals. The server's idle periods in the queueing system correspond to type III periods (lost sales) in the safety stock system. Hence, on average, a small amount of product added to the safety stock remains there for an amount of time (during periods of types II and III) equal to $W_q + 1/\tilde{\mu}$. Since $d\rho$ units of product are added to the safety stock per unit of (types II and III) time, the expected stock level in periods II and III combined, here denoted as $V_{(\mathrm{II+III})}$, is given by $d\rho(W_q + 1/\tilde{\mu})$.

Note that $V_{\mathrm{III}} = 0$ because at type III periods we have lost sales. We further note that

$$V_{(\mathrm{II+III})} = \frac{f_{\mathrm{II}}V_{\mathrm{II}} + f_{\mathrm{III}}V_{\mathrm{III}}}{f_{\mathrm{II}} + f_{\mathrm{III}}} \quad (8)$$

since the expected inventory level in period type (II + III) can be written as the weighted average of the inventory levels in periods II and III. This leads to

$$d\rho(W_q + 1/\tilde{\mu}) = V_{(\mathrm{II+III})} = \frac{f_{\mathrm{II}}}{f_{\mathrm{II}} + f_{\mathrm{III}}} V_{\mathrm{II}}. \quad (9)$$

The utilization ratio in the queueing system is, of course, equal to the fraction of time during periods II + III that safety stock is present, so

$$\rho = \frac{r\beta E[t_f]}{d(E[t_r] + t_s)} = \frac{f_{\mathrm{II}}}{f_{\mathrm{II}} + f_{\mathrm{III}}}. \quad (10)$$

From (9) and (10) we find that

$$V_{\mathrm{II}} = d(W_q + 1/\tilde{\mu}). \quad (11)$$

The following lemma is used to obtain the final expression for $\underline{V}$. Its proof is given in Appendix B.

Lemma 1. *The long-run average safety stock levels in periods of types I, II and IV are identical, i.e., $V_{\mathrm{I}} = V_{\mathrm{II}} = V_{\mathrm{IV}}$.*

It now follows that

$$\underline{V} = (f_{\mathrm{I}} + f_{\mathrm{II}} + f_{\mathrm{IV}})d(W_q + 1/\tilde{\mu})$$
$$= (1 - f_{\mathrm{III}})d(W_q + 1/\tilde{\mu}) = \alpha\, d(W_q + 1/\tilde{\mu}). \quad (12)$$

Using the definition of $\tilde{\mu}$ and (3) this is, of course, equivalent to (7).

Note that the average cost of holding safety stock can be calculated with Theorem 1 when W_q is known. For the case of exponentially distributed repairs (with rate ν) and zero setup time ($t_s = 0$), the GI/M/1 queueing system defined above reduces to an M/M/1 queueing system for which a closed formula for W_q is available (see, e.g., Gross and Harris 1985). In this case we find

$$\underline{V} = \frac{\alpha^2 \lambda d^2}{(1-\alpha)\nu^2 r} - \frac{\alpha d}{\nu}. \qquad (13)$$

When no closed form expression for W_q is available, Theorem 1 can still be employed to find the average holding cost for safety stock by finding W_q numerically. A procedure for doing this in the case of exponentially distributed repair times and positive setup time is outlined in Appendix C. An alternative approach is to use any one of the common approximations for GI/M/1 or GI/G/1 queues. For example, according to the Krämer and Langenbach-Belz approximation we have:

$$W_q \approx \frac{1}{2}(C_A^2 + 1)\frac{\rho/\bar{\mu}}{1-\rho}\exp\left\{\frac{-2(1-\rho)(1-C_A^2)^2}{3\rho(C_A^2+1)}\right\}, \qquad (14)$$

where C_A^2 is the squared coefficient of variation of the queueing system interarrival distribution (see, e.g., Tijms p. 302). Hence

$$C_A^2 = \frac{\text{var}[t_r]}{(E[t_r] + t_s)^2}.$$

Note that when $t_s = 0$ and the repair distribution is exponential, $C_A^2 = 1$, and (14) is exact.

3. NUMERICAL RESULTS

Several numerical applications using the input data of Table II are studied in this section. Figures 3 and 4 illustrate the model and several basic insights. Figure 3 shows the convex structure of the total cost per unit of time as a function of the lot size. Figure 4 depicts the optimal lot size as a function of the service level α for different production over demand (r/d) ratios. Reducing the r/d ratio increases the machine utilization and reduces the width of the feasible service level range by cutting the highest attainable service level. Note the sharp increase in the optimal lot size that takes place close to the upper limit of the feasible service level range for each r/d ratio. The optimal lot size for a given service level reduces as the r/d ratio increases. For a system with an r/d ratio of 1,000/800 (80% capacity utilization) we find the system can provide a broad range of service levels; in this case, the lot size is not very sensitive to changes in service level.

The M/M/1 analogous model for the safety stock level (zero setup time, exponential failure and repair times) was evaluated using simulation and found to be perfectly accurate. A subset of the data used for the empirical evaluation of the relative accuracy of the GI/G/1 approximations for the case of nonzero setup time, and exponential failure and repair times is presented in Table III; the results are presented in Table IV. The results in the column labeled Exact (GI/M/1) Solution were obtained with the exact

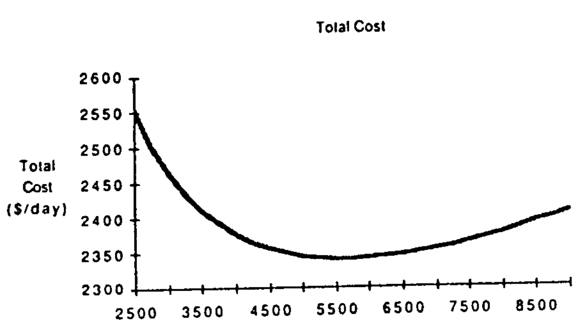

Figure 3. Safety stock model: total cost per unit time as a function of the lot size.

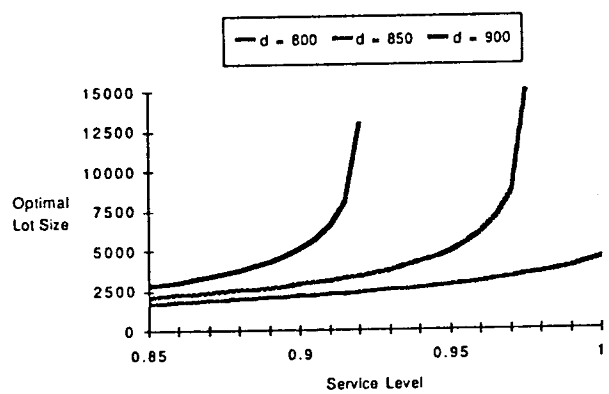

Figure 4. Safety stock model: optimal lot size as a function of the service level.

Table III
Data Used for Table IV

Parameter	Data Set 1	Data Set 2	Data Set 3
r (units/day)	1,000	1,000	1,000
d (units/day)	800	800	800
h ($/unit/day)	2	2	2
s ($)	1,500	1,500	1,500
p ($)	500	500	500
c ($/day)	1,000	1,000	1,000
t_s (days)	0.20	0.10	0.30
t_m (days)	0.20	0.20	0.20
λ (l/days)	0.10	0.10	0.10
μ (l/days)	0.50	1.5	1
α	0.95	0.95	0.95
LB, UB on α	0.8503, 1	0.9422, 1	0.9058, 1
β	0.12337	0.00870	0.05137
Q (units)	4453.67	2818.81	3197.91

Table IV
Basic Safety Stock Model, Nonzero Setup Case

Data Set	Exact (GI/M/1) Solution	K&LB Approx. Solution	Simulation I	Simulation II	% Difference Sim. I and II
Data Set 1					
Service level	0.95	0.95	0.95004 (0.00013)	0.95056 (0.00013)	0.05
Average cycle length	5.8601	5.8601	5.8598 (0.0017)	5.8566 (0.0017)	−0.05
Setup cost	369.97	369.97	369.95 (0.06)	370.15 (0.06)	0.05
PM cost	85.32	85.32	85.33 (0.01)	85.37 (0.01)	0.05
CM cost	152.00	152.00	151.97 (0.11)	152.05 (0.11)	0.05
Running stock holding cost	341.29	341.29	341.27 (0.03)	311.03 (0.07)	−8.86
Safety stock holding cost	7170.40	7333.08	7168.92 (24.58)	7011.62 (24.64)	−2.19
Total cost	8119.78	8286.66	8117.43 (24.52)	7930.23 (24.63)	−2.31
Data Set 2					
Service level	0.95	0.95	0.9499 (0.0003)	0.9640 (0.0002)	1.48
Average cycle length	3.7090	3.7090	3.7094 (0.0010)	3.6550 (0.0008)	−1.47
Setup cost	518.43	518.43	518.26 (0.25)	525.97 (0.28)	1.49
PM cost	134.81	134.81	134.79 (0.04)	136.80 (0.03)	1.49
CM cost	50.67	50.67	50.78 (0.23)	51.53 (0.24)	1.48
Running stock holding cost	539.23	539.23	539.21 (0.09)	484.17 (0.22)	−10.21
Safety stock holding cost	176.38	186.27	176.24 (1.32)	175.99 (1.34)	−0.14
Total cost	1419.52	1429.41	1419.28 (1.00)	1374.47 (1.12)	−3.16
Data Set 3					
Service level	0.95	0.95	0.9500 (0.0002)	0.9533 (0.0002)	0.35
Average cycle length	4.2078	4.2078	4.2079 (0.0010)	4.1930 (0.0010)	−0.35
Setup cost	470.48	470.48	470.48 (0.14)	472.15 (0.15)	0.35
PM cost	118.83	118.83	118.82 (0.03)	119.25 (0.03)	0.36
CM cost	76.00	76.00	76.03 (0.18)	76.30 (0.19)	0.36
Running stock holding cost	475.31	475.31	475.26 (0.07)	427.52 (0.23)	−10.05
Safety stock holding cost	1578.15	1681.78	1576.78 (8.13)	1512.75 (8.11)	−4.06
Total cost	2718.77	2822.4	2717.37 (7.91)	2607.96 (8.06)	−4.03

algorithm for GI/M/1 queues described in Appendix C. The next column gives the results obtained with the Krämer and Langenbach-Belz based GI/M/1 approximation, and shows that this approximation is fairly accurate and useful to obtain quick estimates. Of course, Theorem 1 allows the use of any convenient approximation formulas for W_q.

The results in the column Simulation I were obtained with a detailed event simulation that exactly follows the policy that we analyzed, i.e., lost sales are incurred when the safety stock is exhausted during a repair, even when running stock is still present. Since this may not be an acceptable policy in all situations, we ran a second simulation in which *running stock is used during repairs after safety stock runs out*, and safety stock is used during setups and preventive maintenance after running stock runs out. The lot size in Simulation II was kept constant, i.e., each production run ends when Q units have been produced, as opposed to when the running stock inventory reaches $Q(1-\beta) - Qd/r$ units. (In many production environments it would be impossible to extend a run on short notice due to a lack of materials or other resources.) Identical failure and repair patterns were used in both simulations. The half-widths of the confidence intervals for the simulation results are given in parentheses. The simulations were run for 20 million, 1 million, and 5 million cycles, respectively, for data sets 1, 2 and 3 to obtain tight confidence intervals. The simulations were written in Lightspeed Pascal and run on a Macintosh SE/30. The results from Simulation I confirm the analysis given in Section 2. A comparison between Simulations I and II indicates that using running stock when the safety stock runs out results in a higher service level and a lower cost. The differences are greatest for data set 2, where the service level is relatively low compared with the feasible range (and, hence, the diversion rate β is small), and the availability of the production system is largest. Conversely, when the production system availability is low and the service level and diversion rate are relatively high (cf. data set 1), the differences between Simulations I and II are small. Data set 3 lies somewhere between these two extremes. The observed differences in the total cost between the two policies followed by Simulations I and II are relatively small (2.31%–4.03%) because the holding cost for the running stock (the cost component where the largest difference occurs) is a relatively small part of the total cost. The observed differences in service level are 0.C5–1.48%, and the differences correlate closely with the observed decrease in the average cycle length. This decrease occurs because the idle time before the next cycle is reduced when running stock is used to satisfy demand during a repair in Simulation II.

4. MAINTENANCE LEVEL CONCEPT

A major management concern these days centers on the appropriate amount of effort to allocate between preventive maintenance, corrective maintenance, and safety stocks used to mitigate the effects of machine breakdowns. We investigate these tradeoffs using the same assumptions as for the basic case of Section 3. In addition, a maintenance budget level concept is introduced to capture the fact that by allocating more resources for maintenance (higher maintenance budget level), the failure rate can be reduced. This means that the failure rate $\lambda(n)$ now becomes a monotonically decreasing function of the maintenance budget level n. This budget level n can, for example, represent crew training or diagnostic equipment rental. The same approach could be followed to analyze other resource allocation decisions aimed at reducing the repair time. Details are left to the reader. We illustrate the applicability of our approach using the model of Section 2 and, for simplicity, we assume that n can only be an integer multiple of \$100. In addition, we assume that $t_s = 0$ throughout, and the repair times follow an exponential distribution. The values for the other parameters can be found in Table II.

Figure 5 illustrates three hypothetical relationships between the MTBF (mean time between failures), which is equal to $1/\lambda(n)$, and the maintenance budget level n. The first MTBF function is $\text{MTBF}_1 = 10 + 20(1 - e^{-0.005n})$. This function implies that, without any additional investments in maintenance ($n = 0$), the MTBF will be 10 days; additional investments in maintenance result in diminishing marginal returns, and the MTBF cannot exceed a value of 30. The total cost per unit of time (\$/day) as a function of the maintenance budget level n is illustrated in Figure 6, where the maintenance budget varies from 0 to 2,500 \$/day and the corresponding MTBF varies from 10 to almost 30. This total cost curve shows a sharp decline due to the initial investments in reducing the MTBF and increases in almost a linear fashion beyond the optimal operating point. The optimal maintenance budget level is \$400, and the additional maintenance investments constitute 25% of the total cost at the optimum. This level of the maintenance budget also corresponds to the point where the MTBF curve starts to flatten out; see Figure 5.

To study the sensitivity of the optimal maintenance level to some of the parameters we used another

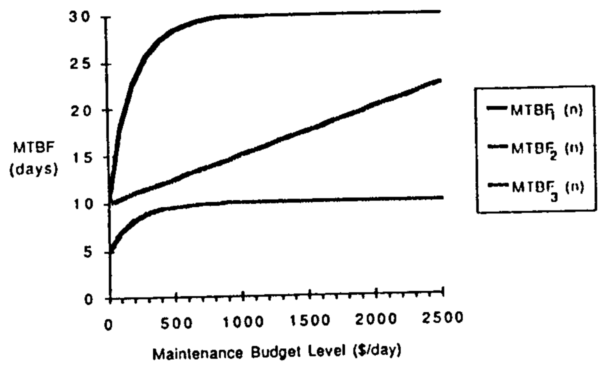

Figure 5. Safety stock model (maintenance budget level concept): MTBF (mean-time-between failures) as a function of the maintenance budget level.

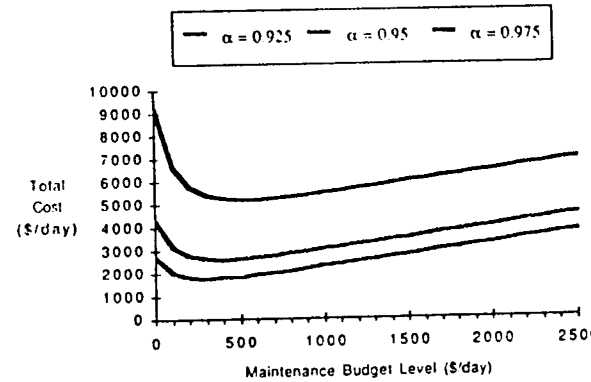

Figure 7. Safety stock model (maintenance budget level concept): total cost per unit of time as a function of the maintenance budget level for different service levels (using $MTBF_2$).

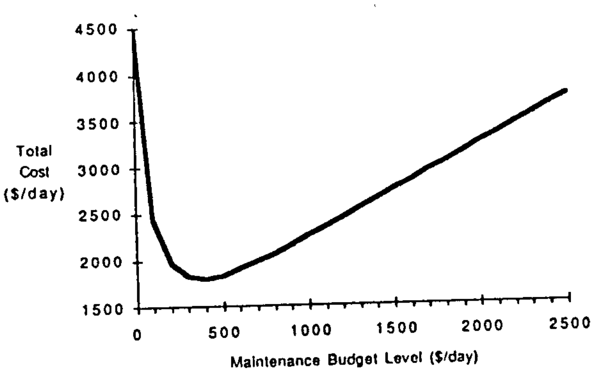

Figure 6. Safety stock model (maintenance budget level concept): total cost per unit of time as a function of the maintenance budget level (using $MTBF_1$).

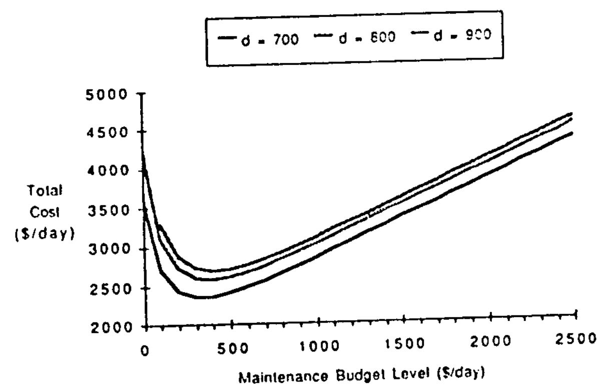

Figure 8. Safety stock model (maintenance budget level concept): total cost per unit of time as a function of the maintenance budget level for different demand rates (using $MTBF_2$).

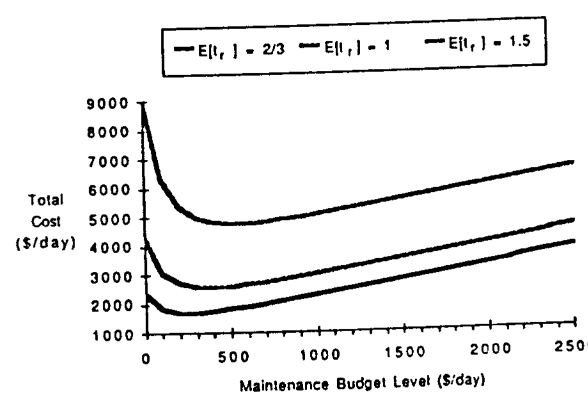

Figure 9. Safety stock model (maintenance budget level concept): total cost per unit of time as a function of the maintenance budget level for different repair rates (using $MTBF_2$).

MTBF function: $MTBF_2 = 5 + 5(1 - e^{-0.005n})$, cf. Figure 5. The maintenance budget was again varied between 0 and 2,500 $/day. The total cost functions per unit of time ($/day) for three different service levels are presented in Figure 7. Clearly, the total cost increases sharply with increasing service level requirements. The optimal maintenance budget levels are $300/day, $400/day and $500/day, and the optimal lot sizes are 2,783, 3,018 and 3,312 for service levels 0.925, 0.95 and 0.975, respectively. Figures 8 and 9 illustrate the effect of changes in demand rate and mean repair times on total cost and optimal maintenance buuget levels. As expected, the total cost is an increasing function of both the demand rate and the mean repair time. On the other hand, the optimal maintenance level appears to be less sensitive to changes in the demand rate than to changes in the mean repair time for the range of parameters considered.

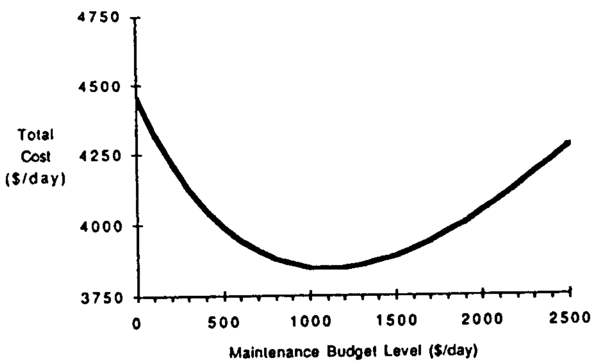

Figure 10. Safety stock model (maintenance budget level concept): total cost per unit of time as a function of the maintenance budget level (using $MTBF_3$).

Both $MTBF_1$ and $MTBF_2$ implicitly assume that there is some technological constraint that prevents the MTBF to be made higher than 30, respectively, 10. The third MTBF function used, $MTBF_3 = 10 + 0.005n$, assumes that the MTBF can be made as large as desired provided that enough additional maintenance investments are made, cf. Figure 5. The corresponding total cost per unit of time curve of Figure 10 shows that even in this case there exists an optimal maintenance budget level beyond which further investments are not justified.

5. CONCLUSIONS

This paper is one of the first studies to address the problem of determining the economic lot size for an unreliable manufacturing facility. It recognizes the fact that both breakdown and repair times are randomly distributed and therefore reduce the effective production capacity. Safety stocks are used in this context in order to meet a managerially prescribed service level. Produced items are drawn from these safety stocks whenever the machine is broken and undergoes repair. We developed bounds on the range of feasible service levels and showed that this range gets smaller with increasing failure rate and increasing machine utilization. We propose an easily implementable production control policy (but do not display its optimality), and develop a closed form expression for deriving the optimal lot size under this policy. We also show that the safety stock dynamics can be described by a special renewal process. The structure of this renewal process is proven to be analogous to the workload process of a single server queueing system with an arrival rate proportional to the mean repair plus setup times and a service rate proportional to the ratio of the failure rate over the expected machine utilization. The same approach can be used to approximately analyze more general cases assuming general failure and repair distributions.

Extensive numerical experimentations along with detailed simulation were used to validate the results of the analytical and approximate models developed. Several key management insights are also revealed by experimenting with the models developed here. We show that the optimal lot size and the expected safety stock increase with the failure rate, service level, demand rate, and setup and repair times. We also show how these results can be incorporated in a broader framework for evaluating resource allocation decisions aimed at reducing the failure rate of the machines. A clear tradeoff is shown to exist between the overall investment in increasing the maintenance level and the resulting savings in safety stocks and repair costs. Also, the total cost function is asymmetric: Slight overinvestments in machine maintenance are significantly less expensive than similarly sized underinvestments.

Possible future extensions of the modeling approach developed in this paper include the analysis of multimachine and multiproduct production systems. Other possible extensions can be based on exploiting the structure of the queueing approach presented here for dealing with general time to failure distributions, finite storage space, and state dependent safety stock control schemes.

APPENDIX A

Running Cycle Cost Components for the Safety Stock Model With Exponentially Distributed Failure and Repair Times

A cycle is the period elapsed between the start of production of a lot of q units to the start of the following lot of q units. It includes production plus idle time, and repair time (if failures occur). As in the ordinary EMQ model, production plus idle time in a cycle is equal to the total running stock production per cycle divided by the demand rate, i.e., $q(1-\beta)/d$. The expected total repair time in a cycle equals the expected number of breakdowns ($\lambda q/r$) times the expected repair time plus setup time ($E[t_r] + t_s$). We thus find

$$E[\text{cycle length}] = \frac{q(1-\beta)}{d} + \frac{q}{r}\lambda(E[t_r] + t_s). \qquad \text{(A.1)}$$

The long-run average inventory holding cost for running stock (RSHC) is equal to the holding cost rate h multiplied with the long-run average running inventory. On average, during periods I and IV of a cycle, running stock is half the maximum running stock as in the ordinary EMQ model. Long-run average inventory during machine breakdowns (type II and III periods) is also half the maximum running stock because the failure distribution is exponential. Hence

$$\text{RSHC} = \frac{1}{2} hq(1 - \beta - (d/r)). \tag{A.2}$$

The other long-run average cost components are easy to derive with the renewal reward theorem (see, e.g., Ross) as the expected cost incurred per cycle divided by the expected cycle length. Since there is exactly one preventive maintenance per cycle, the long-run average cost of preventive maintenance cost is given by

$$\text{PMC} = \frac{p}{E[\text{cycle length}]}. \tag{A.3}$$

Since the expected number of failures during a cycle equals $\lambda(q/r)$, and the average cost per repair is given by $E[t_r]c$, the long-run average cost of corrective maintenance equals

$$\text{CMC} = \frac{\lambda(q/r)E[t_r]c}{E[\text{cycle length}]}. \tag{A.4}$$

The number of setups per cycle is given by the number of failures + 1, so the long-run average cost of setups is

$$\text{SUC} = \frac{s + s\lambda\,(q/r)}{E[\text{cycle length}]}. \tag{A.5}$$

APPENDIX B

Proof of Lemma 1

Consider the safety stock process on periods of types I (increasing safety stock level) and II (decreasing safety stock level) only. This process regenerates each time the safety stock level reaches zero. In each regeneration cycle the fraction of period I time equals $f_I/(f_I + f_{II})$, and the fraction of period II time equals $f_{II}/(f_I + f_{II})$. During type I periods, the safety stock level increases at a constant rate $r\beta$, while during type II periods, the safety stock level decreases with a constant rate d. Let $G_i(x)$ = the fraction of time during type i periods that the safety stock level is at or below x, i = I, II. Without loss of generality we assume that the safety stock is managed according to a LIFO (last in, first out) system. Then any small amount δ of product that was added to safety stock while the safety stock level was less than x can only be withdrawn when the safety stock level is less than x. It follows that in each regeneration cycle the amount of safety stock accumulation while the safety stock level is less than x equals the amount of withdrawals while the safety stock level is less than x, and hence,

$$\frac{f_I}{f_I + f_{II}} G_I(x) \cdot r\beta = \frac{f_{II}}{f_I + f_{II}} G_{II}(x) \cdot d, \tag{B.1}$$

and since $d \cdot f_{II} = r\beta f_I$ this implies

$$G_I(x) = G_{II}(x). \tag{B.2}$$

It then follows that

$$V_I = \int_0^\infty x\, dG_I(x) = \int_0^\infty x\, dG_{II}(x) = V_{II}.$$

To show that $V_I = V_{IV}$, consider the safety stock sample path $X(t)$ during type I periods only. It is easy to see that the distribution of the safety stock during type IV periods is obtained by sampling $X(t)$ every $Q(1 - \beta)/r$ time periods. Let $\tilde{X}$ be the limiting distribution of $X(t)$. Then

$$X\left(\frac{Q(1-\beta)}{r} k\right) \to \tilde{X} \text{ in distribution as } k \to \infty,$$

and hence,

$$\frac{1}{n}\sum_{k=1}^{n} X\left(\frac{Q(1-\beta)}{r} k\right) \to \tilde{X} \text{ in distribution as } n \to \infty.$$

We conclude that

$$\begin{aligned} V_{IV} &= \lim_{n\to\infty} E\left[\frac{1}{n}\sum_{k=1}^{n} X\left(\frac{Q(1-\beta)}{r} k\right)\right] \\ &= E[\tilde{X}] \\ &= \lim_{t\to\infty} E\left[\frac{1}{t}\int_0^t X(s)\, ds\right] = V_I. \end{aligned}$$

APPENDIX C

Computing W_q in GI/M/1 Queue With G = Delayed Exponential

Notation

- μ = the service rate;
- $\mathbf{I}$ = the interarrival time random variable;
- $\mathbf{N}_i$ = the number of customers in the system seen by arrival i;
- $p_k = \lim_{i\to\infty} \Pr\{\mathbf{N}_i = k\}$;
- $q_m = \Pr\{m$ service times are completed during an interarrival time$\}$.

The average waiting time in queue for the GI/M/1 system can be found by studying the discrete time Markov chain $\{\mathbf{N}_i\}$ (see, e.g., Tijms). In our case we have $\mathbf{I} = C + \mathbf{U}$, where $\mathbf{U} \sim \text{Exp}(\lambda)$ and $C \geq 0$. It follows that

$$q_n = \sum_{i=0}^{n} \frac{\lambda}{\lambda + \mu}\left(\frac{\mu}{\lambda + \mu}\right)^{n-i} \frac{(\mu C)^i}{i!} e^{-\mu C}. \tag{C.1}$$

The limiting distribution of $\{N_i\}$ can be found by solving the system of linear equations

$$p_j = \sum_{k=j-1}^{\infty} p_k q_{k+1-j}, \quad j = 1, 2, \ldots, \tag{C.2a}$$

$$p_0 = \sum_{k=0}^{\infty} p_k \left(1 - \sum_{m=0}^{k} q_m\right), \tag{C.2b}$$

$$\sum_{k=0}^{\infty} p_k = 1. \tag{C.2c}$$

We solved system C.2 with the Gauss-Seidel method with overrelaxation. Finally, the average waiting time in queue can now be computed as

$$W_q = \frac{1}{\mu} \sum_{n=1}^{\infty} n p_n. \tag{C.3}$$

ACKNOWLEDGMENT

We thank Professor Uday S. Karmarkar from the University of Rochester for stimulating discussions and useful comments, and Professor Awi Federgruen from Columbia University for pointing out a simplification in the proof of Lemma 1 in Appendix B. We also thank two anonymous referees and the associate editor for their excellent suggestions. Finally, we are indebted to one of the referees for pointing out that it is common practice to retain a minimal cycle stock after the safety stock has been depleted.

REFERENCES

BARLOW, R. E., F. PROSCHAN AND L. C. HUNTER. 1965. *Mathematical Theory of Reliability*. John Wiley, New York.

BIELECKI, T., AND P. R. KUMAR. 1988. Optimality of Zero-Inventory Policies for Unreliable Manufacturing Systems. *Opns. Res.* **36**, 532–541.

GALLEGO, G. 1988a. Linear Control Policies for Scheduling a Single Facility After an Initial Disruption. Technical Report 770, School of OR and IE, College of Engineering, Cornell University, Ithaca, N.Y.

GALLEGO, G. 1988b. Produce-Up-To Policies for Scheduling a Single Facility After an Initial Disruption. Technical Report 771, School of OR and IE, College of Engineering, Cornell University, Ithaca, N.Y.

GROENEVELT, H., A. SEIDMANN AND L. PINTELON. 1992. Production Lotsizing With Machine Breakdowns. *Mgmt. Sci.* **38**, 104–123.

GROSS, D., AND C. HARRIS. 1985. *Fundamentals of Queueing Theory*. John Wiley, New York.

HENIG, M., AND Y. GERCHAK, 1990. The Structure of Periodic Review Policies in the Presence of Random Yield. *Opns. Res.* **38**, 634–643.

KARMARKAR, U. S. 1987. Lot Sizes, Lead Times and In-Process Inventories. *Mgmt. Sci.* **33**, 409–418.

KELLER, G., AND H. NOORI. 1988. Impact of Investing in Quality Improvement on the Lot Size Model. *OMEGA* **16**, 595–601.

LEE, H. L., AND M. J. ROSENBLATT. 1987. Simultaneous Determination of Production Cycle and Inspection Schedules in a Production System. *Mgmt. Sci.* **33**, 1125–1136.

LEE, H. L., AND C. A. YANO. 1985. Production Control in Systems With Variable Yield Losses. Technical Report 85-32, Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, Mich.

MCCALL, J. J. 1965. Maintenance Policies for Stochastically Failing Equipment: A Survey. *Mgmt. Sci.* **11**, 493–524.

MCLEAVEY, D. W., AND S. L. NARASHIMAN. 1985. *Production and Inventory Control*. Allyn & Bacon, Boston.

NAHMIAS, S. 1989. *Production and Operations Analysis*. Irwin, Homewood, Ill.

PATE-CORNELL, M. E., H. L. LEE AND G. TAGARAS. 1987. Warnings of Malfunction: The Decision to Inspect and Maintain Production Processes on Schedule and on Demand. *Mgmt. Sci.* **33**, 1277–1290.

PETERS, M. H., H. SCHNEIDER AND K. TANG. 1988. Joint Determination of Optimal Inventory and Quality Control Policy. *Mgmt. Sci.* **34**, 991–1004.

PIERSKALLA, W. J., AND J. A. VOELKER, 1976. A Survey of Maintenance Models: The Control and Surveillance of Deteriorating Systems. *Naval Res. Logist. Quart.* **23**, 353–388.

PORTEUS, E. L. 1986. Optimal Lot Sizing, Process Quality Improvement and Setup Cost Reduction. *Opns. Res.* **34**, 137–144.

ROSS, S. M. 1983. *Stochastic Processes*. John Wiley, New York.

SCHWALLER, R. L. 1988. EOQ Under Inspection Costs. *Prod. Inv. Mgmt.* **3rd Q.**, 22–24.

SEIDMANN, A., AND S. Y. NOF. 1989. Operational Analysis of an Autonomous Assembly Robotic Station. *IEEE Trans. Robot. Automat.* **5**, 4–15.

SHERIF, Y. S., AND M. L. SMITH. 1981. Optimal Maintenance Models for Systems Subject to Failure: A Review. *Naval Res. Logist. Quart.* **28**, 47–74.

SILVER, E. A., AND R. PETERSON. 1985. *Decision Systems for Inventory Management and Production Planning*. John Wiley, New York.

TAGARAS, G. 1988. An Integrated Cost Model for the Joint Optimization of Process Control and Maintenance. *J. Opns. Res. Soc.* **39**, 757–766.

TAPIERO, C. S. 1986. Continuous Quality Production and Machine Maintenance. *Naval Res. Logis. Quart.* **33**, 489–499.

TIJMS, H. C. 1986. *Stochastic Modelling and Analysis*. John Wiley, New York.

ZIPKIN, P. 1986. Models for Design and Control of Stochastic, Multi-Item Batch Production Systems. *Opns. Res.* **34**, 91–104.