Recent developments in single product, discrete-time, capacitated production-inventory systems

SRIDHAR TAYUR

239, Graduate School of Industrial Administration, Carnegie Mellon University, Pittsburgh, PA 15213, USA

e-mail: stayur@grobner.gsia.cmu.edu

Abstract. We present here some recent results in single-product, capacitated production-inventory systems in discrete time. The key results are: (1) structure of optimal policy for single stage systems; (2) analysing via a shortfall process; (3) using simulation to optimize; (4) an approximation using tail probabilities. We consider periodic demand, and multiple stages – serial, distribution and assembly. Related topics of re-entrant flow shops, lead time quotation and value of information are also discussed briefly.

Keywords. Multi-echelon systems; inventory; optimal policies; simulation; approximations.

1. Introduction

The motivation for this stream of research has come from problems faced by a diverse set of companies, such as IBM, Sematech, AMD, Allegheny Ludlum, GE, Proctor and Gamble, Westinghouse, Intel, American Standard and McDonald's. Smaller local (to Pittsburgh) companies such as Sintermet, Blazer Diamond, ASKO and Northside Packing have also provided several interesting issues to pursue. At the heart of many of the problems is the interaction between demand variability and non-stationarity, available production capacity, holding costs of inventory (at different locations), lead times and desired service levels. The central goal of this research stream is to understand the interactions in simple single and multiple stage settings and to provide insights and implementable solutions for managing inventories in a cost-effective manner for complex systems. The goal of this paper is to introduce in a systematic manner some recent advances in 'Discrete-time, Capacitated Production-Inventory Systems facing Stochastic Demands' and we limit ourselves to single produc! t ! setting.

1.1 *Quantitative models for supply chain management*

A modern manufacturing network, consisting of multiple manufacturing facilities and several external vendors, can be modelled as a multi-stage, capacitated, assembly system;
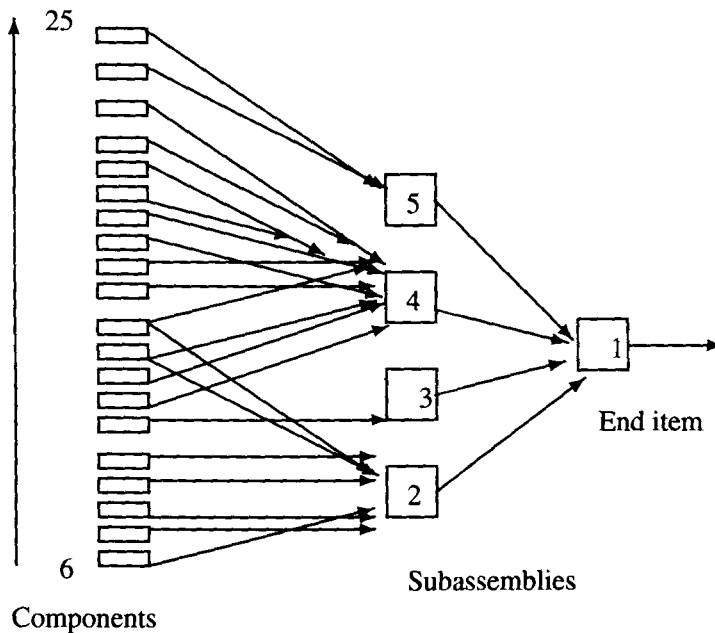
Figure 1. A typical supply chain: IBM assembly.

see figure 1 for a representation of an IBM supply chain. Until 1991, the only major result that was available in *capacitated systems* was the structure of the optimal policy for a single product, single stage system facing a stationary demand process (Federgruen & Zipkin 1986a). Even for this case ('a simple model'), no computational method was available to compute the optimal parameters for a given instance. Since then, significant progress has been made in this area. In some sense, these models form the backbone of quantitative modelling for *supply chain management*.

Among the many papers that are now available (since 1991), five papers on this topic – single-product, capacitated systems in discrete time – make the following contributions: (1) Develop a method to compute the parameters for this simple model; (2) find the optimal policy and provide a computational procedure for the case when this system faces a non-stationary (periodic) demand process (this paper generalises) (Karlin 1960; Morton 1978; Federgruen & Zipkin 1986; Zipkin 1989; Ciarallo *et al* 1994; Morton & Pentico 1995); (3) study the stability of a multi-stage capacitated system operated by a base-stock policy; (4) develop a computational method to compute good parameters for a multi-stage system operated by a base-stock policy; and (5) develop a very quick and accurate approximation method for the same problem as above. Re-entrant flow hops, multi-product systems, component commonality (and delayed differentiation) and other topics have been studied in greater detail since that time. Similarly, several papers and research! t! hemes in continuous time models are available as well.

No attempt has been made to provide a comprehensive literature review; however, most references may be obtained from the papers mentioned here. For a thorough survey of results (mainly in uncapacitated systems), see Van Houtum *et al* (1995). Other useful surveys and books include Graves *et al* (1992) and Buzzacott & Shantikumar (1993).

1.2 Basics

In our models, we will assume *time buckets*. Each bucket can be a day or a week long (or a month long) depending on the situation. These models will be called *discrete* time models. Each unit of time will be called a *period*. This approach is appropriate at the plant and system levels. *Continuous* time models are appropriate at shop-floor and machine levels.

A *base stock* policy with order up to level z means that we produce in any period just enough to reach this target. If we cannot reach it, due to capacity limitations or lack of raw material, we do the best we can. An *echelon* base stock policy is exactly the same, except that all quantities (mainly inventories) considered are cumulative in order to include this stage and all stages downstream (near the end product, close to the customer).

We will study these systems via a combination of analytical methods and simulation. Models are good for simple situations and to grasp concepts. To compute numbers for real world situations, simulation is preferred. The assumptions made are more realistic and the solutions obtained are more believable. Furthermore, certain flexibility that decision makers would prefer to have is better handled by simulation. In terms of acceptance by end users, a validated simulation has had better luck than complicated mathematical models. A shortcoming of simulation as compared to mathematical models is that it takes a much longer time to find answers. What we do then in reality is use models to get rough estimates and to provide intuition to fellow team members; then choose an alternative that shows most promise; finally, we simulate to get accurate solutions.

1.3 Literature survey: Papers before 1991

Clark & Scarf (1960) developed a periodic review inventory control model for a serial system without setup costs. By using a discounted cost framework, they established that an *order up to* policy at each node is indeed optimal. Federgruen & Zipkin (1984) extended these results further. Muckstadt *et al* (1984) conducted a computational study using the Clark & Scarf (1960) model. A continuous review version of the Clark and Scarf model is studied by Debodt & Graves (1985). An in-depth analysis of an assembly structure with only two inputs, again by using the discounted cost framework is presented by Schmidt & Nahmias (1985). Rosling (1989) showed that under some initial conditions, an assembly system can be reduced to a serial system with modified lead times so that the results of Clark & Scarf (1960) may be applied to this equivalent serial system.

A model of a production and distribution network in which manufacturing is modelled by a single node is described by Cohen & Lee (1990). It also differs from much of the earlier work in an important way in that decentralized control is assumed and the model itself is a framework that combines separate models of production and distribution. A supply chain planning model that can be used to study production scale and scope economics is presented by Cohen & Moon (1990).

A model for supply chain management which assumes decentralized control at each node in the manufacturing network, controlled by periodic review order up to inventory policies is presented by Lee *et al* (1991). Once the service levels are set for each node, the overall relationships between cost and service can be obtained by applying this model. Although

capacity considerations are not addressed, they allow for uncertainty in the supplier lead times.

The literature on inventory control systems and production-distribution systems is extensive and hence we limit our review to the work that is closely related to the theme of this paper. Similarly, there is a vast body of literature on single location production-inventory systems (addressing many aspects of interest) that is not reviewed here.

The Clark & Scarf (1960) model and many of its extensions, including the one by Rosling (1989) analyse the model within a discounted cost framework. These results are fairly involved and further, the computational procedures are not easy to describe or program. The assembly system inventory control problem in an average cost framework is studied by Langenhoff & Zijm (1989) and Kamesam & Tayur (1993). This analysis leads to an exact decomposition of the assembly system into several single location problems. Even this decomposition is not easy to handle, but Van Houtum & Zijm (1990) describe computational approximations that lead to a simplified computational procedure.

Except for that by Federgruen & Zipkin (1986), not much work was done in capacitated systems in discrete time. In this survey paper, we will begin with their model and then describe recent developments of several extensions of this basic model. The notation we use may change between sections to remain consistent with the papers that are being summarized.

2.   Single-stage, single-product models

The first progress since 1991 was the introduction of *shortfall* [1], and the connection that was made between the capacitated inventory model operated under a base stock policy and a dam model that has been studied extensively by applied probabilists.

2.1   *Basic model*

Tayur (1993) provides a method to compute the optimal policy for a basic inventory problem addressed previously by Federgruen & Zipkin (1986a). We are to determine the base stock level of a single item at a single location under periodic review, when

- the unit variable cost of purchase is c per item and there are no fixed costs;

- the holding cost (h) and stockout cost ($p > c$) are per period and per item;

- demands in successive periods are non-negative i.i.d. random variables with known distributions (labelled by d);

- there is an infinite horizon and the costs are not discounted;

- all demands that are not satisfied by stock on hand are backordered;

- there is a finite production capacity, C, in every period;

- the cost is computed on the amount of inventory or backorder at the end of each period;

---

[1] Our thanks to a referee for pointing out that this shortfall type connection was known to queueing theorists before 1991.

• we are to minimize the expected cost of holding plus penalty per period.

Recall that *inventory position* is defined as (stock on hand) + (stock on order) − (back-orders). An *order up-to* (or a base stock) policy with a critical number z is one in which the inventory position (x) is raised to z if $x < z$, and no production is done if $x \geq z$.

Unlike previous approaches, we provide a different construction of the sequence of problems that converge to the problem of interest. In particular, we do not consider finite horizon problems of the capacitated problem and then take the limit as the number of periods go to infinity. Rather, we have a sequence of uncapacitated, multi-stage, serial infinite horizon versions that converge to the desired system. We use results from uncapacitated multi-stage serial systems coupled with results in storage stochastic processes.

Specifically, our steps are the following. *First*, we show that the inventory model of interest is equivalent to a problem in dams. This suggests an analysis based, not on the evolution of (inventory on hand minus backorder) process but, rather, by considering a shortfall process. The shortfall is defined as the *amount on order that has not yet been produced* because of the capacity constraint. If X_n is the shortfall at the end of period n, then $X_n = \max(0, X_{n-1} + d_n - C)$, where d_n is the demand in period n.

It is important to differentiate between backorders and shortfalls: the former represents what the customer did not obtain, while the latter represents what the manufacturer could not produce because of the capacity constraint. Thus, the backorder at the end of period n equals $\max(0, X_n - z)$ where z is the order up to level. The penalty cost p is on the backorder; there is no direct penalty on shortfall. Similarly, the amount of inventory at the end of period n is $\max(0, z - X_n)$. The cost in period n, therefore, equals $p\max(0, X_n - z) + h\max(0, z - X_n)$.

Second, we show that we can replace the single-stage capacitated inventory model by constructing a specially structured uncapacitated infinite-stage inventory model: This is simply a mathematical artifact. The sequence of multi-stage problems alluded to above will converge to this infinite stage system.

2.2 *Connection with a dam model*

Figure 2 shows the sample path of a typical single-stage capacitated inventory system under periodic review that is operated by a base stock policy where excess demand is backlogged. The capacity (C) is 30, the order-up-to level (z) is 45, and the inventory at time 0 (I_0) is 10. Let $d_1 = 15, d_2 = 9, d_3 = 37, d_4 = 21$ be the demands in the first four periods. Figure 3 shows the sample path of a dam (see Prabhu 1965, 1980) that has an infinite height, a release capability of at most C, and an initial water level of 35. Let the rainfall in the first four periods be 15, 9, 37, and 21. The dam releases as much water as it can, and if the water level is less than C, the dam becomes empty. The equivalence of the two sample paths is obvious. Let $(a)_+$ stand for $\max(a, 0)$. If Z_n is the content of the dam in period n just after release, then it satisfies $Z_n = (Z_{n-1} + d_{n-1} - C)_+$ and if X_n is the amount on order in period n that has not yet been produced, it satisfies $X_n = (X_{n-1} + d_{n-1} - C)_+$ (a similar recursion arises in the study of a D/G/1 queue also). Note that $\{X_n, n = 1, 2, \ldots\}$ is a Markov chain. This motivates us to study the capacitated inventory system in terms of the process X_n, and provide results in terms of the steady
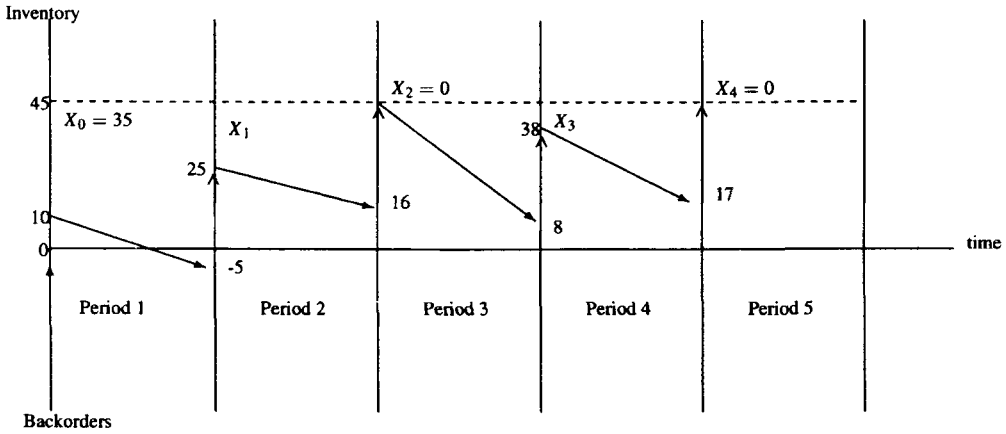
The order-up-to level is 45

The capacity is 30

Negative inventories represent backorders

X_t is the amount in period t that is on order but not yet produced

Downward sloping arrows represent demands; upward arrows denote production

Figure 2. Sample path of inventory model.

state distribution of $X = \lim_{n \to \infty} X_n$. Table 1 summarizes the equivalence between the capacitated inventory model and the dam model.

### 2.3 Computing the optimal base-stock level

Let $K(x)$ be the distribution of the input to the dam in any period. If $F(x)$ is the steady state distribution of the water content in the dam just after release, then the optimal value of the order up to level, z, in the capacitated inventory system satisfies

$$(F * K)(z) = p/(p + h), \tag{1}$$

where $*$ represents convolution. $F(x)$ is known for all discrete distributions of the demand, and for Erlang distributions. Intuitively, we are adding the following two independent random variables: (1) demand in a period and (2) the amount on order at the beginning of the period that has not yet been produced. The necessary and sufficient condition for the distribution $F(x)$ to exist is that the expected demand (input to the dam) in a period be less than the capacity, C. Intuitively, the result states that it is the sum of two independent random variables (demand in a period and the amount on order not yet produced at the beginning of that period) that adds up to z. Penalty p is incurred if this sum crosses (at the end of the period) z and is proportional to the excess, and a holding cost (h) is imposed if the sum is less than z and is proportional to the amount on hand at the end of the period.

Example 1. If the demand is exponentially distributed with mean rate λ ($K(u) = 1 - e^{-\lambda u}$, $u \geq 0$; $K(u) = 0$, otherwise) and the capacity is C, then the steady state distribution of the water content in the dam just after release is given by

$$F(x) = 1 - e^{-\mu(x+C)} \quad (x \geq 0, \ \lambda C > 1)$$

WATER CONTENT
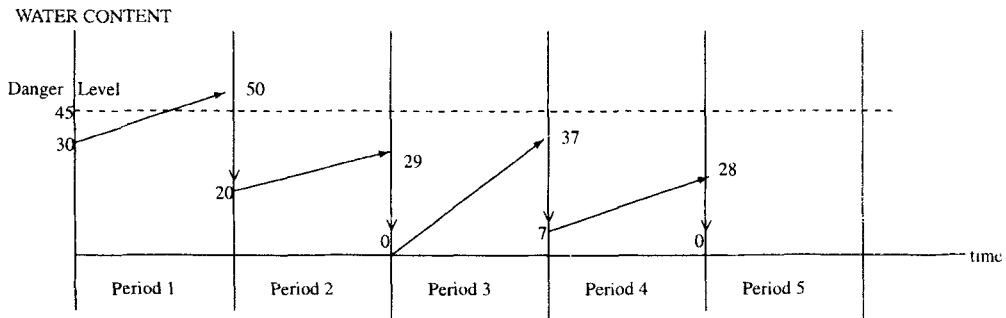
Upward arrows represent rainfall

Downward arrows represent release

One cannot release more than the water level

The danger level is simply shown to connect with the inventory model

Figure 3. Sample path of dam model.

where μ is the largest positive root of the equation

$$\mu = \lambda - \lambda e^{-(\mu + C)}.$$

Thus, the optimal order upto policy has a critical number z^*, and is obtained by solving $(F * K)(z^*) = p/(p + h)$.

Independently, the similarity of the basic inventory model operated via a base stock policy to a D/G/1 queue is recognized by Van Houtum & Zijm (1994).

2.4 *Optimal policy for an extended model*

In the previous subsection, we computed the optimal base stock level. We now want to show that among all policies, a base stock policy is optimal. Federgruen & Zipkin (1986a) show this for the stationary case above. Kapuscinski & Tayur (1996a) provide a proof for an extension that allows for periodic demand as well as capacities by considering the following variant of the basic single-stage, single-item, discrete-time production–inventory model. The demands (stochastic) follow a periodic pattern with a period K. As before, there is a maximum production capacity (C) in any period; demands not satisfied in a period are backlogged to the next; there is holding cost (h) per unit of inventory per period and a cost of penalty (p) per unit of backlog per period. We want to find policies that minimize the

Table 1. Comparison of dam and inventory models.

Dam model	Inventory model
Maximum release	Capacity
Water content	Amount not yet produced (shortfall)
Empty dam	Order-up-to level achieved
Rainfall	Demand
Danger level crossed	Backorders

finite-horizon costs, the infinite-horizon discounted cost and the infinite-horizon average cost (respectively) of operating this stage.

For all the three cases – finite-horizon cost, discounted infinite-horizon cost and infinite-horizon average cost – we show that an order up-to (or base-stock, critical-number) policy is optimal. This extends the results of Karlin (1960b) and Zipkin (1989) for uncapacitated, non-stationary models and Federgruen & Zipkin (1986) for capacitated, stationary models. Our proof for the finite-horizon case follows standard steps. The proof of optimality for the infinite-horizon discounted case is simpler than that provided by Federgruen & Zipkin (1986) because we are able to use more recent results from Bertsekas (1987). To provide the optimality proof in the average cost case, we use the framework of Federgruen *et al* (1983), but our approach is different from that used by Federgruen & Zipkin (1986a) for the stationary case. All proofs can be found in Kapuscinski & Tayur (1996a).

2.5   *Sequence of results leading to optimal policy structure*

The following is the sequence of events at the beginning of a period: (1) some inventory or backlog exists; (2) a decision to increase the inventory is taken (limited by the production capacity); and (3) demand arrives. Holding or penalty costs are charged on the inventory after demand arrives. The notation is mostly standard. We have suppressed the time subscript in x, y and d below and these are assumed to be reals unless mentioned explicitly as integers. We will write them when necessary. We define

- x: inventory at the beginning of a period;

- y: inventory after ordering, but before demand arrives.

We assume that $Ed_i < \infty$ for all period types $i = 1, \ldots, K$.

2.5a   *Finite horizon*:   As is standard practice, we begin with the finite-horizon case, which is not only the simplest situation to consider but also sets the framework for the infinite horizon cases.

Let $0 < \beta \le 1$ be the discount factor. Define (recursively) $v_n(x)$ = minimum total expected discounted cost with a time horizon of n periods. Note (in this subsection only) that *we start in period n and count downward towards period 1*, the end of the horizon. The demand in period n is one of the K period types ($i = 1, \ldots K$). We may assume that period 1 has type 1 demand and period $n = K + 1$ has type 1 demand again and so on. Thus:

$$v_0(\cdot) = 0,$$
$$v_n(x) = \inf_{(x,y) \in A} \{c(y - x) + L_n(y) + \beta E_{d_n} v_{n-1}(y - d_n)\},$$
$$\text{for } x \in R \text{ and } n \ge 1,$$

where

$$A = \{(x, y) \in R^2 | x \le y \le x + C\} = \bigcup_{x \in R} Y(x)$$

is the feasible set. For $x \in R$,

$$Y(x) = [x, x + C],$$

and for $y \in R$,

$$L_n(y) = E_{d_n}(h(y - d_n)^+ + p(y - d_n)^-).$$

We can express v_n using additional functions J_n and I_n as follows.

$$v_n(x) = -cx + I_n(x), \quad x \in R, \quad n \geq 1,$$
$$I_n(x) = \min\{J_n(y) : y \in Y(x)\}, \quad x \in R,$$
$$J_n(y) = cy + L_n(y) + \beta E v_{n-1}(y - d_n), \quad y \in R,$$

with $v_0(\cdot) = J_0(\cdot) = I_0(\cdot) = 0$.

Our first lemma generalizes theorem 2 of Federgruen & Zipkin (1986b) to the cyclic case. The proof is a direct adoption of theirs.

Lemma 1. The set of feasible pairs (x, y), A (as defined above), is convex. For all $n \in N$:

(a) *The expected sum of holding and penalty costs, $L_n(y)$, is convex;*

(b) *J_n, v_n, and I_n are convex;*

(c) *$v_n \geq 0$; and*

(d) *For $n \geq 1$: $v_n(x_n) \to \infty$ when $|x_n| \to \infty$, and if $p > c$ then $J_n(y_n) \to \infty$ when $|y_n| \to \infty$.*

Theorem 1. *Let y_n^* be the smallest value minimizing J_n. The optimal policy in period n is order up-to y_n^*. As $\lim_{|y_n| \to \infty} J_n(y_n) = \infty$, we have $y_n^* < \infty$.*

As a first property, we have the following.

Property 1. For any $x \in R$. $v_{nK+i}(x)$ is increasing in n.

Proof. We show by induction that $v_{m+K}(x) \geq v_m(x)$ for all x and for all m. First, $v_K(x) \geq 0 = v_0(x)$. Let $v_{K+m}(x) \geq v_m(x)$ for a certain m and all x. Then,

$$v_{K+m+1}(x) = \min_{y \in Y(x)} \{c(y - x) + L_{K+m+1}(y) + \beta E v_{K+m}(y - d_{K+m+1})\}$$
$$\geq \min_{y \in Y(x)} \{c(y - x) + L_{m+1}(y) + \beta E v_m(y - d_{m+1})\}$$
$$= v_{m+1}(x). \qquad \square$$

Note that the convexity of functions J_n, I_n, v_n implies continuity of these functions. One-sided derivatives exist at all points. Also, two-sided derivatives exist with the exception of a denumerable set of points. Points where two-sided derivatives do not exist are generated by mass points of demands. Although derivatives do not have to be continuous, they are monotonic and bounded on any compact set. Therefore, in this paper we will define them as right-hand-side limits. We will use a prime $(')$ to denote these derivatives.

We define the *myopic* solution to period i, $\bar{y}_i$, as the one that satisfies $c + L_i'(\bar{y}_i) = 0$. Property 2 provides a simple lower bound as in the uncapacitated case; see Zipkin (1989).

Property 2. Assume that period 1 has the minimum myopic solution; $\bar{y}_1 = \min\{\bar{y}_i: i = 1, \ldots, K\}$ where $\bar{y}_i$ is the myopic solution for period i. Then: $y_n^* \geq \bar{y}_1$, $\forall n$.

Strictly speaking, our next result is not a finite-horizon result. However, this appears to be the most appropriate point to state it. Part (c) of the technical lemma (lemma 2) is used for the next result. We will use this property in proving property 6 in § 2.5a below and when analysing the infinite horizon average cost case in § 2.3.

Property 3. For a given $0 < \beta \leq 1$, the sequence y_n^* is bounded. That is, $\lim \sup\{y_n^*: n \in N\} < \infty$.

Property 4. For any finite-horizon (n periods) problem consider a policy that produces up-to z_i in period $i = 1, \ldots, n$. For all i, the cost of such a policy is convex in z_i.

2.5b *Infinite horizon: Discounted model*: We now move to the infinite-horizon discounted case. It was natural in the finite-horizon case to label periods as time to the end of the horizon. In the infinite-horizon case, we typically start the process at some point and continue indefinitely. To make the notation more intuitive, starting from this subsection, *we will number periods in increasing order*: following period n, we have period $n + 1$.

Federgruen & Zipkin (1986b) showed the next result for the stationary, capacitated case. We provide a simpler proof as we are able to use results from Bertsekas (1987).

Theorem 2. *Let* $0 < \beta < 1$ *(discounted case). The optimal policy for the infinite-horizon is cyclic up-to level policy.*

Since $E(d_i) < \infty$ for $i = 1, \ldots, K$, we have the following.

Property 5. Let $w_n(x, i) := v_{nK+i}(x)$ for $i = 1, \ldots, K$, $n \in N_0 := N \cup \{0\}$, and $x \in R$. The limit, $\lim_{n \to \infty} w_n(x, i) = w(x, i)$ (where values of w are in $R \cup \{\infty\}$), exists and $w < \infty$.

$w < \infty$ does not imply that stationary up-to levels are finite. However, it is possible to prove the following (since $E(d_l) < \infty$ for $l = 1, \ldots, K$).

Property 6. The up-to levels are finite (i.e. $z_l < \infty$).

2.5c *Average cost criterion*: This case is the most difficult one to analyse. We need some technical results before the optimality of base-stock policies can be proved.

Property 7. (Convexity) Consider the infinite-horizon case (with cyclic demands) and the class of up-to level policies, where levels z_i for $= 1, \ldots, K$ are period-type specific. For both the discounted cost and average cost criteria, the average cost is finite and convex in each of z_i's.

Let $z_{\max}$ and $z_{\min}$ be the maximum and minimum respectively among the levels for a given vector $(z_1, \ldots, z_K)$.

Property 8. (Coupling) If $z_{\max}$ is strongly regenerative, then for any two processes with starting points $x_{i_0}^{(1)}$ and $x_{i_0}^{(2)}$ respectively in period i_0, there exists (with probability 1) a period n such that starting from this period the two processes coincide.

Property 9. (Shortfall stability) Consider an up-to policy with a vector $z_1, \ldots, z_K$. If $\sum_{i=1}^{K} E d_i < KC$ then $E S_i < \infty$ for all i's, where $S_i = z_i - y_i$ is the shortfall in period type i.

Average cost criterion is easy to analyse when either the number of states the process can take is finite or the one-period cost function is bounded (see Bertsekas 1987). Conditions when an optimal policy exists for semi-Markovian process with average cost objective function with denumerable state space and unbounded one-period cost function are derived by Federgruen *et al* (1983). These conditions were used by Federgruen & Zipkin (1986a) to derive optimality of up-to policy for a capacitated stationary model. We extend it to a cyclical model as follows.

We first show that any policy can be dominated by a policy that requires reducing (any) backlog and not exceeding some stationary level A^*. Then we show that among such policies, the up-to policy is optimal. The main structure of our proof is based on results of Federgruen *et al* (1983), but the proof that the required conditions are satisfied is shown by a different method as compared to that by Federgruen & Zipkin (1986a).

Fact 1. If an optimal policy for the problem exists (including possibility of randomized policies), then it has the following form:

(a) for $x \le -C$, $y = x + C$.

(b) there exists $A^* < \infty$ such that for $y \ge A^*$ for all period types it is better to produce nothing rather than take any other action.

Theorem 3. *Consider a capacitated system with cyclic discrete demands and linear ordering, penalty, and holding costs. For the average cost criterion, the cyclic up-to policy is optimal.*

Lemma 2. Consider the policy $\delta[0]$ (produce up-to 0). Let $\sum_{i=1}^{K} E d_i < KC$ and $E(d_i)^{2k+2} < \infty$ for a certain $k \ge 1$ and for all $i = 1, \ldots, K$. Consider a point process defined by points i, for which $y_i = 0$ (no backlog of previous demands). Let N be a random variable equal to time between two consecutive points of this point process. Then for any starting period-type:

(a) $E(N^k) < \infty$

(b) $E(|y_i|^k) < \infty$ and $E(|x_i|^k) < \infty$

(c) If $0 < \beta < 1$ then $E(\sum_{n=0}^{\infty} |y_n|^k \beta^n)$, $E(\sum_{n=0}^{\infty} |x_n|^k \beta^n) < \infty$ for any x_0.

(d) For $0 < \beta \le 1$, there exists $A \in R$, such that $J_n(0) \le A_n$.

2.6   *Basic properties*

We show several properties of the optimal policy including the following: (1) capacity smooths the base-stock levels in a manner that is different from that due to holding costs; (2) the limit of finite horizon order up-to levels are bounded; (3) the optimal levels are higher than the minimum of the K myopic levels; (4) in an infinite-horizon average cost case, the optimal levels are lower than the maximum of the K stationary optimal levels and higher than the minimum of the K stationary optimal levels; (5) if demands are stochastically larger or the capacity is lower, the base-stock levels are higher; and (6) for $K = 2$, as the penalty cost is increased, the difference between the maximum and minimum levels is bounded by C under fairly general assumptions on demand distributions.

2.7   *Computational technique*

Exact computation of optimal levels by analytical formulas appears difficult. We provide a simulation based method using infinitesimal perturbation analysis (IPA) to find these levels. The basic idea is simple: instead of using the derivative of the expected cost in a gradient search method, we use the expected value of the sample path derivative (obtained via simulation). To validate this approach and prove the optimality result for average cost case, we derive several technical properties of base-stock policies – convexity, regeneration, coupling and stability. See Glasserman (1991) for an excellent reference and a later subsection for details of a multi-stage system analysed in this manner. A numerical study indicates that our IPA method is robust and finds solutions within a few minutes on a workstation. The steps are similar to those described later in § 3 for a serial system, and so we do not detail them here. See Kapuscinski & Tayur (1996a).

2.8   *Insights into some complex issues*

We also numerically study several issues that provide insight into the behaviour of optimal solutions. Examples of issues studied include the following: (1) what is the increase in cost if all periods are forced to have the same base-stock level? (2) what are the benefits of changing capacities in each period based on the demand type? (3) how many periods are affected by smoothing of the levels? (4) how are the above results affected by high penalty cost or high utilizations? (5) what is the relationship between service level and costs? Several of the qualitative and technical properties in our capacity setting differ from the uncapacitated non-stationary model.

   See Kapuscinski & Tayur (1996a) for extensive computational testing. The basic insights are as follows.

(1) *Basic insights.* Several basic properties of the system confirm what we expect. All other things being equal:

   - with an increased mean demand, the up-to levels increase,
   - with decreased capacity, the up-to levels increase, and
   - with increased variance of demand, the up-to levels increase.

(2) *Effect of capacity and range of demands on 'smoothing'.* By smoothing, we mean that values z_i are affected by demands and the capacity of period $j \neq i$. The difference between $z_{\max}$ and $z_{\min}$ is lower as compared to the difference between the maximum and the minimum of K independent stationary capacitated models; so some higher levels are brought down and some lower levels are lifted up. We have two types of smoothing working in opposite directions: (a) in anticipation of high demand, some levels are lifted up, and (b) in anticipation of low demand, other levels are decreased. The second type was described by Karlin (1960b) and Zipkin (1989) in the uncapacitated setting. The first one is induced by finite capacity, and the second by the holding cost.

At times there are changes in the ordering of the levels as capacity decreases. For example, the period with the minimum level may be different depending on the capacity. Furthermore, although the differences between levels do not disappear as capacity is reduced, the levels increase and the ratio between the maximum and the minimum of the levels gets closer to 1.

(3) *Constant base-stock versus optimal (period wise) base-stock.* We compare the cost of the optimal *constant* base-stock policy to the optimal cost. (Note that the cost is a convex function of this constant level and the simulation based method remains valid.) Not surprisingly, the optimal constant z lies between the minimum and the maximum optimal base stock levels When capacity is tight, the cost ratio (say R) becomes close to 1. We note that in some cases the difference $z_{\max} - z_{\min}$ first increases and then decreases as capacity decreases, while in other cases it only increases with capacity. Even in cases where $z_{\max} - z_{\min}$ of optimal policy increases as capacity decreases, R monotonically decreases to 1. One explanation is the fact that the cost function is relatively flat around the optimum (see item 8 below), and more so as the capacity becomes tighter.

The effect of variance on R is as follows. R is close to 1 in most cases. A case when the cost ratio is large has two small but different variances (example: $N(70, 1)$, $N(70, 10^2)$ in a $K = 2$ situation). With increase in variance the levels increase and the ratio of costs goes to 1. We can also show that even with significantly different means, the ratio of costs goes to 1 when at least one of the variances is high.

(4) *Changing capacity with periods according to period type.* Rather than remain constant, capacity can vary according to period type. (The cost is still a convex function of $(z_1, \ldots, z_K)$ and the optimal policy is still order up-to. The simulation-based method remains valid and recursions can be easily adapted.) Several capacity allocations for four situations each with $K = 2$ were tested: (i) both period types have low variance; (ii) period with lower mean has a high variance; (iii) period with higher mean has a high variance; (iv) both periods have high variance. We are able to test if two natural alternatives to constant capacity (with the same total capacity KC in a cycle) – (a) proportional to mean demand and (b) proportional to (mean + constant $*$ standard deviation) of demand – do well.

We note from our experiments that the optimal spread of total capacity among periods is not a simple function of mean demand alone and quite often not one based on variance

either. Typically the spread proportional to mean demands did well only in small-variance cases, but in these cases nearly all spreads did well. In case of high variance, a proportional (based on mean and variance) spread did much better than one based on mean alone, but surprisingly it was possible to find cases when it was optimal to allocate more to a period type with a demand having both a smaller mean and a smaller variance. We think that in these cases there is typically sufficient capacity in low demand period to raise the inventory to optimal level in higher demand period, but the effect of a surge in demand in high demand period can be repaired faster when a little bit more capacity is assigned to low demand period. This is seen very clearly in the following example: let $K = 4$ with demands wi! th! means and variances equal to $(20, 1)$, $(20, 1)$, $(90, 50)$, $(20, 1)$ respectively. We find that the optimal capacity allocation is $(22, 22, 93, 143)$. So the use of the term 'recovery capacity' seems appropriate. In all cases, allocating equal capacity to periods did very well.

(5) *Service level.* In all the experiments – constant capacity, equal base-stock levels, period-wise capacity allocation, different K's, different demand distributions, different holding and penalty costs – the type-1 service, defined as P (inventory after demand > 0), at optimality, was $p/p + h$. See Tayur (1993) for a proof of this connection between service level and costs in the single-stage capacitated stationary case, and Glasserman & Tayur (1995) for the multistage stationary capacitated case.

(6) *Increasing penalty cost.* With increased penalty the up-to levels increased. Obviously larger differences in means and variances caused relatively larger differences in up-to levels, but this effect was small over a range of penalty costs. For most part, the levels rise in parallel. Furthermore, in all our experiments, the ordering of the levels did not change as a function of p (unlike item 2 above). We also noticed that the difference between the maximum and the minimum levels at optimality was bounded by C for $K = 2$ as we increased p. Property 12 below provides an explanation.

(7) *Mean utilization vs. variance.* As penalty cost is increased, the optimal up-to levels of low variance system with high utilization increase faster than in a system with high variance and lower utilization. This indicates that the rate of increase of base-stock levels depends on both the variance and excess capacity.

(8) *Cost sensitivity.* Cost is not very sensitive to the up-to levels around the optimum. This was noticed in the stationary multi-stage system studied by Glasserman (1991) also.

3.   Single-product, serial system

It has been shown that for a serial capacitated system, the optimal policy (under the cost criteria discussed above) is not base-stock in general; see Speck & Van der Waal (1991).

3.1   *A simulation-based optimization procedure: Single-stage simulation*

Let s be any base stock level; the optimal value for s is what we eventually need. The notation used is consistent with the papers referenced on this topic where detailed proofs are available.
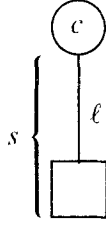
Figure 4. Illustration of a single stage.

We need notation.

I_n = inventory – backlog in period n;

R_n = production in period n;

ℓ = leadtime from production to inventory;

$T_n = R_{n-1} + \cdots + R_{n-\ell}$

= pipeline inventory;

D_n = demand in period n;

s = base-stock level;

c = production capacity.

We always assume that $c > 0$ and $D_n \geq 0$ for all n. Under a (modified) base-stock policy, the production level in each period is set to try to restore the inventory position $I_n + T_n - D_n$ to s. If production were uncapacitated, this would be achieved by setting $R_n = D_n$. Since, however, R_n cannot exceed c, it may take multiple periods of production to offset demand in a single period.

The system evolves as follows:

$$I_{n+1} = I_n - D_n + R_{n-\ell}, \tag{2}$$

$$R_n = c \wedge [s + D_n - (I_n + T_n)]^+, \tag{3}$$

$$T_{n+1} = T_n + R_n - R_{n-\ell}. \tag{4}$$

x^+ denotes $\max(0, x)$, x^- denotes $\max(0, -x)$ and $a \wedge b$ denotes $\min(a, b)$.

What about costs? In period n the cost will be $C_n = hI_n^+ + pI^-$. The average cost over the long term will be $(\sum_{n=1}^{N} C_n)/N$ for a large N. N is the number of periods simulated.

3.2   *A simulation-based optimization procedure: Single-stage gradient*

We can certainly find the gradient of expected cost by simulating one more time and changing (say, increasing) only the order up to level by 1: the difference in cost between the two simulations is the gradient. This would be too much work especially if we had several stages whose order up to levels we want to optimize. Can we get the gradient by not doing any more simulations? Yes; and here goes the basic trick.

(1) If we had inventory in period n, having started at a higher base stock level means an *increase* in holding by h.

(2) If we had backlog in period n, we will *decrease* penalty by p.

Thus, the changes are either h or $-p$. We will just average these changes. The point is we are claiming that average of changes is the change in the average. In this situation, and in several others, this claim is good.

How do we make the computer do this? In principle, we are just differentiating the recursions. The easiest way is to write the following two lines in the code to be executed in every period:

$$\text{dCost} = \text{dCost} + h, \quad \text{if } I_n > 0,$$
$$\text{dCost} = \text{dCost} - p, \quad \text{if } I_n < 0.$$

Then dividing dCost by N provides the derivative. Note that at the beginning of the simulation dCost is set to zero.

3.3 A simulation-based optimization procedure: Serial system

3.3a *Operation*: We now link multiple stages in series. There are m stages: stage 1 supplies external demands and stage i supplies components for stage $i - 1, i = 2, \ldots, m$. Stage m draws raw materials from an unlimited supply. To specify an echelon-inventory base-stock policy for the system, we let

$$s^i = \text{echelon base-stock level for stage } i.$$

Naturally, we require $s^1 \leq s^2 \leq \cdots \leq s^m$. Let the variables R_n^i, T_n^i, ℓ^i, and c^i have the same meaning as before, applied to stage i. For $i = 2, \ldots, m$, let I_n^i be the installation inventory at stage i, and let $I_n^1 = I_n$, with I_n as in § 2.1. In period n, stage i sets production to try to restore the echelon inventory position,

$$\sum_{j=1}^{i}(I_n^j + T_n^j) - D_n,$$

to its base-stock level s^i.

Two features distinguish the multi-echelon system from a single stage: production at stage i, $i < m$, is constrained by available component inventory I_n^{i+1}, as well as by the capacity limit c^i; and for $i > 1$ the amount removed in period n from the store at stage i is the downstream production level R_n^{i-1}, rather than the external demand. Thus, for stage $i = 2, \ldots, m - 1$ we have

$$I_{n+1}^i = I_n^i - R_n^{i-1} + R_{n-\ell^i}^i, \tag{5}$$

$$R_n^i = c^i \wedge \left[s^i + D_n - \sum_{j=1}^{i}(I_n^j + T_n^j) \right]^+ \wedge I_n^{i+1}, \tag{6}$$

$$T_{n+1}^i = T_n^i + R_n^i - R_{n-\ell^i}^i. \tag{7}$$

At stage m, raw materials are unlimited so the last term in (5) is absent. To subsume these special cases in (5–7), we take $R_n^0 \equiv D_n$ and $I_n^{m+1} = \infty$ for all n. To complete our specification of the model, we need initial conditions; for simplicity, we take $I_1^1 = s^1$,

$I^i_1 = s^i - s^{i-1}$, $i = 2, \ldots, m$, and all other variables zero. In other words, the system starts with full inventory. For details see Glasserman & Tayur (1995).

Similar to the single stage case, the derivatives with respect to the base stocks can be computed.

3.4 Validation of technique

Validation of finite horizon derivatives – inventory and costs – is quite straightforward. We show that (right-side) derivatives exist with probability one at a given value of s, then appeal to Lipschitz continuity and finish by applying the dominated convergence theorem. See Glasserman & Tayur (1995) for details.

3.5 Stability and recurrence

For IPA to work in the infinite horizon, several conditions have to be satisfied by the underlying stochastic process. These are derived by Glasserman & Tayur (1994). When capacity limits are introduced, an ineffective policy may lead to increasingly large order backlogs: the *stability* of the system becomes an issue. In this paper, we examine the stability of a multi-echelon system in which each node has limited production capacity and operates under a *base-stock* policy. We show that if the mean demand per period is smaller than the capacity at every node, then inventories and backlogs are stable, having a unique stationary distribution to which they converge from all initial states. Under i.i.d. demands we show that the system is a Harris ergodic Markov chain and is thus wide-sense regenerative. Under slightly stronger conditions, inventories return to their target levels infinitely often, with probability one. We discuss cost implications of these results, and give extensions to systems with random leadtimes and periodic demands.

3.6 Extensions

Let us discuss two obvious extensions.

3.6a *Assembly system*: In an assembly system, each node i requires components from a set of predecessor nodes. These are assembled into stage-i finished goods. By changing units, if necessary, we may assume that components from predecessor stages are assembled in equal quantities.

To keep the notation simple, we consider a representative example, rather than the general case. Figure 5 depicts a three-node system in which node 1 assembles components supplied by nodes 2 and 3. Node 1 feeds external demands; the other nodes draw raw materials from infinite sources. The evolution of inventory at node 1 is characterized by

$$\begin{aligned}
I^1_{n+1} &= I^1_n - D_n + R^1_{n-\ell^1}, \\
R^1_n &= c^1 \wedge [s^1 + D_n - (I^1_n + T^1_n)]^+ \wedge I^2_n \wedge I^3_n, \\
T^1_{n+1} &= T^1_n + R^1_n - R^1_{n-\ell^1}.
\end{aligned} \tag{8}$$

The assembly feature is reflected in the dependence of R^1_n on I^2_n and I^3_n in (8). Nodes 2 and 3 are characterized by the basic recursions (5–6), with obvious modifications to the

indexing. The only notable difference is that now I_n^2 and i_n^3 are decreased by the same production level R_n^1 each period.

3.6b   *Distribution system*   Another variant of the serial system allows intermediate stages to supply multiple lower-echelon stages, typically in a tree topology. Our results extend without difficulties to such models. For ease of exposition we describe a less general setting – a serial system in which each stage faces external demands for components in addition to internal demands from the downstream stage. A manufacturer of electronic equipment, for example, may face demands for integrated circuits, and for circuits assembled into circuit packs, along with demands for finished goods.

To characterize the operation of such a system, we need to specify how each stage allocates inventory to internal and external demand. Rather than restrict ourselves to any one policy, we describe a class of policies consistent with our results. Let I_{0n}^i be the stage-i inventory reserved for external demands at stage i in period n, and let $I_{1n}^i = I_n^i - I_{0n}^i$ be the inventory available for downstream production. Suppose stage i has base-stock levels s_0^i and s_1^i to supply external and internal demands, and let $s^i = s_0^i + s_1^i$. Denote by D_n^i the external demand at stage i in period n.

The operation of stage 1 is unchanged. At stage i, production is now set according to

$$R_n^i = c^i \wedge \left[s^i + \sum_{j=1}^{i} (D_n^j - I_n^j - T_n^j) \right]^+ \wedge I_{1n}^{i+1}.$$

Also,

$$I_{n+1}^i = I_n^i - R_n^{i-1} - [D_n^i \wedge I_{0n}^i]^+ + R_{n-\ell^i}^i.$$

Pipeline inventory follows (5), just as before. It only remains to specify how I_{0n}^i and I_{1n}^i are determined.

A broad class of policies sets

$$I_{1n}^i = g^i \left(I_n^i, s_1^i + D_n^1 - I_{1n}^i - \left(\sum_{j=1}^{i-1} I_n^j + T_n^j \right), s_0^i + D_n^i - I_{0n}^i \right)$$
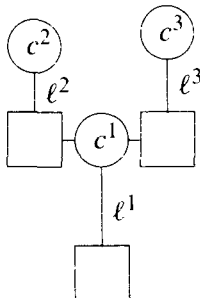

Figure 5.   An assembly system. Super-scripts represent node numbers.

and $I_{0n}^i = I_n^i - I_{1n}^i$, for some function g^i; in other words, the inventory reserved for internal production is a function of the available inventory and of the shortfalls in meeting internal and external demands. The choice of g^i determines the particular policy.

We briefly describe two types of policies. The first type gives strict priority to internal or external demands, allocating as much inventory as needed to the high-priority demand and using any excess for the other. The second type attempts to balance shortfalls, allocating inventory to minimize the resulting difference between inventories and base-stock levels. This type of policy implements relative priorities through the choice of s_0^i and s_1^i. The assumption that the system can, in fact, be balanced in each period is essential to many analytical approaches; our setting, of course, does not require this. For both types of policies just described, the particular functions g^i are tedious to write out but they are sufficiently smooth to allow derivative calculations.

3.7 *An approximation using tail probabilities*

The IPA method could be computationally prohibitive for large problems, or for problems where the starting solution for the simulation is far way from its eventual optimal. So a quick (and accurate!) approximation is presented by Glasserman & Tayur (1996): Recall that our objective is to find base-stock levels that approximately minimize holding and backorder costs. The key step in our procedure approximates the distribution of echelon inventory by a sum of exponentials; the parameters of the exponentials are chosen to match asymptotically exact expressions. The computational requirements of the method are minimal. In a test bed of 72 problems, each with five production stages, the average relative error for our approximate optimization procedure is 1.9%.

Much of the technical development of this approximation is based on the work by Glasserman (1994), who uses techniques from Asmussen (1987).

4. Brief summary of related topics

(1) *Re-entrant flow shops.* How do we handle multi-product *re-entrant* flow lines, a very typical topology in semi-conductor fabrication facilities? (Re-entrant flow lines have attracted significant interest from the research community in recent years because of their direct applicability to semi-conductor fabrication.) This question was motivated by a fab visit at AMD and is studied by Bispo & Tayur (1996). Along the way we are able to study serial multi-product capacitated systems. The framework is similar to that described in § 2 above, but requires a far more detailed analysis.

(2) *Lead time quotation.* To compete effectively, suppliers are realizing that three aspects of lead times are important to customers – short, accurate and consistent – and that this 'service' has to be provided at minimum cost. Several improvements along the just-in-time (JIT) philosophy (and quality paradigms) have made production processes fairly reliable at many supplier plants. While negotiations with customers have reduced the variability of order streams, these variabilities continue to remain significant and further reductions seem unlikely given that the customers themselves face an intensely competitive marketplace with increasingly fickle and demanding (end) customers.

We concentrate on (in Kapuscinski & Tayur 1996b) the uncertainty of demand from multiple customer classes ($1 \leq i \leq M$) and consider production times to be deterministic. It is costlier to quote a longer lead time to ·a customer from a higher class than from a lower one; in particular, the cost is linear with rate m_i for class i customer. We impose the constraint that we will ship the order to the customer with a 100% reliability within the quoted lead time; we may ship earlier at no penalty, but we gain no benefit from the customer for this early shipment. Our original motivation to study this problem came from a company that produces 'rolls' that are used in steel mills. The goal was to find an easily implementable lead-time quoting rule, preferably graphical, that the sales force could use.

Briefly, we consider a finite-horizon, discrete-time production-inventory model with a single stage, single product that faces a stochastic demand from many customer classes in any period. Processing time is deterministic. In each period, after the demands are realized:

(a) we quote lead times to these demands,

(b) we make production decisions for this period, and

(c) we ship some material (sometimes earlier than its due date).

Items (2) and (3) above determine the inventory level at the end of the period, on which we pay a holding cost of h per unit (per period). Note that quoting a large lead time for a low margin customer (and so having a lot of capacity in anticipation of future higher margin customers) causes a penalty from this low margin customer, while a short quoted lead time would force a higher margin customer arriving in the near future to wait a little longer. This is the basic trade-off. We find a policy that is simultaneously optimal for 100% reliable quotation, production, inventory management and shipping. We show that for some cases the policy is easily implementable.

A simple model relating lead times, inventories and batching is studied by Karmarkar (1987); however, no quotation of due dates is considered there.

(3) *Co-operation in supply chains.* Industrial supplier–customer relations have undergone radical changes in recent years as the philosophy behind managing manufacturing systems is influenced by several Japanese manufacturing practices. As more organizations realize that successful in-house implementation of just-in-time alone will have limited effect, they are seeking other members of their supply chain to change their operations. This has resulted in a certain level of co-operation, mainly in the areas of *supply contracts* and *information sharing*, that was lacking before. This is especially true when dealing with customized products, and is most commonly seen between suppliers and their larger customers.

We incorporate (in Gavirneni *et al* 1996) information flow between a supplier (or producer) and a customer in a capacitated setting of a simple supply chain. The customer faces i.i.d. end-product demand, and the supplier has a finite capacity in each period. We consider three situations: (1) a traditional model where there is no information to the supplier prior to a demand to him except from past data; (2) the supplier has the information of the (s, S) policy used by the customer as well as the end-product demand distribution; and (3) full information about the state of the customer. Each of these leads to different

non-stationary demand processes as seen by the supplier. Study of these three models enables us to understand the relationships between capacity, inventory and information at the supplier level. We show that order up-to policies continue to be optimal for models with information flow for the finite horizon, the infinite horizon discounted and the infinite horizon average cost cases. We develop a quick recursive solution procedure to compute the optimal parameters when capacity is infinite. For the finite capacity case, we develop and validate Infinitesimal Perturbation Analysis (IPA), as well as show how the solution for the uncapacitated system can be easily modified to obtain approximate values. Using these solution procedures we estimate the savings at the supplier due to information flow and study when information is most beneficial by varying capacity, holding costs, demand distributions and $S - s$ values.

5. Summary

Several significant advances have occurred in the study of capacitated systems since 1991. This paper informally provides an introduction to topics, and some techniques. The reference list below is not exhaustive at all, but should provide a good starting point. At least two very interesting developments (within a single product setting) are not considered in this informal review: (1) supply contracts and (2) international supply chains. See Scheller-Wolf & Tayur (1997), for example.

References

Asmussen S 1987 *Applied probability and queues* (New York: Wiley)

Bertsekas D P 1987 *Dynamic programming: Deterministic and stochastic models* (Englewood Cliffs, NJ: Prentice-Hall)

Bispo C, Tayur S 1996 Re-entrant flow lines. GSIA Working Paper

Buzacott J, Shanthikumar J G 1993 *Stochastic models of manufacturing systems* (New York: Prentice-Hall)

Clark A, Scarf H 1960 Optimal policies for a multi-echelon inventory problem. *Manage. Sci.* 6: 474–490

Ciarallo F, Akella R, Morton T E 1994 A periodic review, production-planning model with uncertain capacity. *Manage. Sci.* 40: 320–332

Cohen M A, Lee H L 1990 Scale economics, manufacturing complexity, and transportation costs on supply chain facility networks. *J. Manuf. Oper. Manage.* 3: 269–292

Cohen M A, Moon S 1990 Impact of production: Scale economics, manufacturing complexity, and transportation costs on supply chain facility networks. *J. Manuf. Oper. Manage.* 3: 269–292

Debodt M, Graves S C 1985 Continuous review policies for a multi-echelon inventory problem with stochastic demand. *Manage. Sci.* 31: 1286–1299

Federgruen A, Zipkin P 1984a Approximations of dynamic multilocation production and inventory problems. *Manage. Sci.* 30: 69–84

Federgruen A, Zipkin P 1984b Computational issues in an infinite horizon multi-echelon inventory model. *Oper. Res.* 32: 218–236

Federgruen A, Zipkin P 1986a An inventory model with limited production capacity and uncertain demands I: The average-cost criterion. *Math. Oper. Res.* 11: 193–207

Federgruen A, Zipkin P 1986b An inventory model with limited production capacity and uncertain demands II: The discounted-cost criterion. *Math. Oper. Res.* 11: 208–215

Federgruen A, Schweitzer P, Tijms H 1983 Denumerable undiscounted decision processes with unbounded rewards. *Math. Oper. Res.* 8: 298–314

Gavirneni S, Kapuscinski R, Tayur S 1996 Value of information in capacitated supply chains. GSIA Working Paper, CMU, Pittsburgh, PA 15213

Glasserman P 1991 *Gradient estimation via perturbation analysis* (Norwell, MA: Kluwer)

Glasserman P 1994 Bounds and asymptotics for planning critical safety stocks. Columbia University Working Paper (to appear in *Oper. Res.*)

Glasserman P, Tayur S 1994 The stability of a capacitated, multi-echelon production-inventory system under a base-stock policy. *Oper. Res.* 42: 913–925

Glasserman P, Tayur S 1995 Sensitivity analysis for base stock levels in multi-echelon production-inventory system. *Manage. Sci.* 41: 263–281

Glasserman P, Tayur S 1996 A simple approximation for a multistage capacitated production-inventory system. *Naval Res. Log.* 43: 41–58

Graves S C, Rinnooy Kan A, Zipkin P (eds) 1992 *Logistics of production and inventory: Handbooks in operations research and management science* (Amsterdam: North Holland, Elsevier) vol. 4

Kamesam P V, Tayur S 1993 Algorithms for multistage, capacitated assembly systems with stochastic demand. IBM Report

Karlin S 1960a Dynamic inventory policy with varying stochastic demands. *Manage. Sci.* 6: 231–258

Karlin S 1960b Optimal policy for dynamic inventory process with stochastic demands subject to seasonal variations. *J. Soc. Indust. Appl. Math.* 8: 611–629

Kapuscinski R, Tayur S 1996a A capacitated production-inventory model with periodic demand. GSIA Working Paper (to appear in *Oper. Res.*)

Kapuscinski R, Tayur S 1996b 100% Reliable quoted lead times. GSIA Working Paper

Karmarkar U 1987 Lot sizes, lead times and in-process inventories, *Manage. Sci.* 33: 409–418

Langenhoff L J G, Zijm W H M 1989 An analytical theory of multi-echelon production/distribution systems. Working Paper, Eindhoven University of Technology, Department of Mathematics and Computer Sciences (to appear in *Stat. Neerlandica*)

Lee H L, Billington C, Carter B 1991 Gaining control of inventory and service through design for localization. Working Paper

Morton T 1978 The non-stationary infinite horizon inventory problem. *Manage. Sci.* 24: 1474–1482

Morton T, Pentico D 1995 The finite horizon non-stationary stochastic inventory problem. *Manage. Sci.* 41: 334–343

Muckstadt J, Lambrecht M, Luyten R 1984 Protective stocks in multistage production systems. *Int. J. Prod. Res.* 6: 1001–1025

Prabhu U 1965 *Queues and inventories* (New York: Wiley)

Prabhu U 1980 *Stochastic storage processes* (New York: Springer-Verlag)

Rosling K 1989 Optimal inventory policies for assembly systems under random demands. *Oper. Res.* 37: 565–579

Scheller-Wolf A, Tayur S 1997 Reducing international risk through quantity contracts. GSIA Working Paper, CMU, Pittsburgh

Schmidt C, Nahmias S 1985 Optimal policy for a two-stage assembly system under random demand. *Oper. Res.* 33: 1130–1145

Speck C J, Van der Waal J 1991 The capacitated multi-echelon inventory system with serial structure: The average cost criterion. COSOR 91–39, Dept. of Math. and Comp. Sci., Eindhoven Institute of Technology, Eindhoven, The Netherlands

Tayur S 1993 Computing the optimal policy in capacitated inventory models. *Stochast. Models* 9:

Van Houtum G J, Zijm W H M 1990 Computational procedures for stochastic multi-echelon production systems, Centre for Quantitative Methods, Nederlandse Philips Bedrijven, B.V.

Van Houtum G J, Zijm W H M 1994 On multi-stage production/inventory systems under stochastic demand. *Int. J. Prod. Econ.* 35: 391–400

Van Houtum G J, Inderfuth K, Zijm W H M 1995 Materials co-ordination in stochastic multi-echelon systems, University of Twente, Enschede

Zipkin P 1989 Critical number policies for inventory models with periodic data. *Manage. Sci.* 35: 71–80