International Journal of Production Economics, 26 (1992) 257–264
Elsevier

Hierarchical production planning and multi-echelon inventory management

W.H.M. Zijm

University of Twente, Department of Mechanical Engineering, Laboratory of Production Management and Operations Research, P.O. Box 217, 7500 AE Enschede, The Netherlands

Abstract

In this paper we present a framework for the planning and control of the materials flow in a multi-item production system. Our prime objective is to meet a prespecified customer service level at minimum overall costs. In order to motivate our study we first outline the basic architecture of a logistic control system developed at Philips Electronics. Guided by this exposure, we next describe the basic algorithmic framework which is needed to turn the conceptual ideas into operational procedures. The theory is extended with hierarchical planning procedures recognizing the need to first plan on a product family level before disaggregating into plans for end-items.

1. Introduction

The purpose of this paper is to provide a framework for the planning and control of the materials flow in a multi-item integrated production system. The system covers functions such as components and raw materials purchasing, components manufacturing, assembly and finally sales. The goal of the planning and control system is to balance a desired customer service level (the definition of which depends on the particular production situation) against reasonable total costs, with an emphasis on inventory holding costs (here inventory includes all stocks of raw materials, components, subassemblies and final products, as well as all work-in-process inventory). As an example, consider the logistic chain for television manufacturing pictured in Fig. 1.

The need for more integrated logistic control systems became apparent in the early sixties already by the work of Forrester [1] on the cyclical variation of stocks in long production–distribution chains. The growing diversity and, associated with that, the shorter commercial life cycles of consumer products in particular, further emphasized the need to rethink manufacturing and logistics strategies. In response, a number of new control concepts have emerged such as MRP (cf. Orlicky [2], Wight [3], JIT [4] and, for shop-floor control, OPT [5]. Analytic approaches for inventory control in multi-echelon systems were initiated by Clark and Scarf [7], and further developed by many authors (compare e.g., Schwarz [8] and the work of Federgruen and Zipkin [9,10]. Another important analytic contribution to the control of complex production systems started with the work of Hax and Meal [11] (see also Bitran, Haas and Hax [12,13]). Recognizing the presence of product family structures as well as the relatively homogeneous character of capacity resources with respect to different items within such a product family, these authors developed a hierarchical control system. An alternative, approximate, method for stochastic systems was suggested by De Kok [14].

We also note the important progress made in product design, in particular by developing more "manufacturing-friendly" products (Design For Assembly, Design For Manufacture). More specific, by achieving a larger modularity of products, it is possible to postpone the so-called "diversity explosion" to the end of the production line (or at least further downstream), thus sav-

0925-5273/92/$05.00

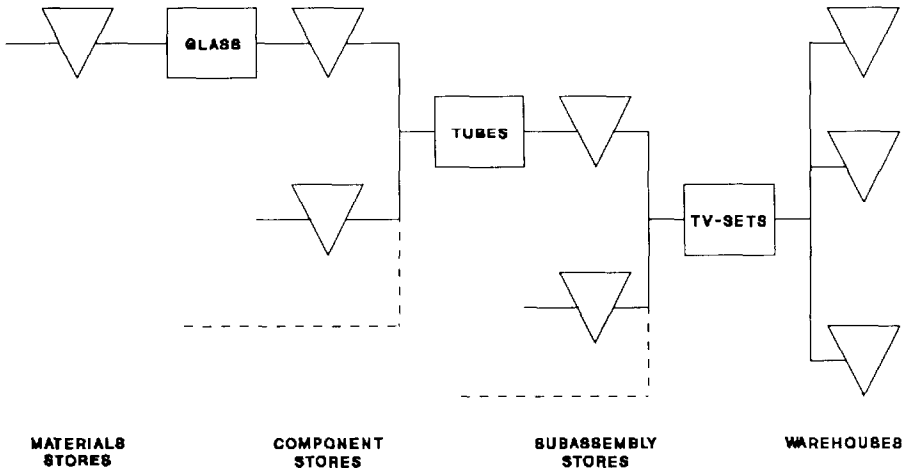

Fig. 1. A logistic chain for television manufacturing.

ing significant costs on work-in-process inventories.

We will not attempt to review all these approaches in detail, this would require an entire paper in itself. Instead, we will briefly describe some main concepts and characteristics of a logistic control system developed for the Consumer Products Division of Philips Electronics in the Netherlands. This system incorporates several new concepts mentioned above, in particular multi-echelon structures and hierarchical planning procedures, based on a product family structure. The description in Section 2 will serve as a reference for the analytic framework discussed in subsequent sections.

The material presented in this paper reviews and extends results of a research program carried out by the author and others (cf. Langenhoff and Zijm [12], Van Houtum and Zijm [16,17]). In section 3 we discuss the basic analytic results of Langenhoff and Zijm [15], i.e., an average cost analysis of general multi-echelon systems under stationary stochastic demand, albeit in a slightly different setting as in the original paper. In fact, this work can be seen as the average cost analogue of the work of Clark and Scarf on the control of inventories in serial systems, but extended with disaggregation procedures for product families. Section 4 highlights the relationship for multi-echelon systems between pure cost models and models in which we wish to determine control policies with minimum inventory holding costs, satisfying a predetermined target service level constraint. In Section 5, we conclude the paper and discuss extensions to both capacitated systems and models with a nonstationary demand structure.

1. A basis architecture for Consumer goods Planning and Logistics (CPL)

The CPL-architecture has been designed to control the entire logistics chain for consumer products. These products typically are made and shipped to stock, i.e., the Customer Order Decoupling Point or shortly Decoupling Point (DP) is placed at the downstream end of the chain. In general, different product–market combinations require different logistic structures (e.g., make and assemble to stock, assemble to order, make to order), leading to different locations of the DP (Fig. 2). The first logistic structure (DP1) covers the situation for consumer products (all planning procedures are forecast driven).

The philosophy underlying the CPL system is based on a number of principles which can be summarized as follows:

- Its *prime logistic objective* is to meet a target *customer survive level* (CSL) at *minimum overall costs.*
- Recognizing the multi-echelon character of the chain, it calls for an integrated control of subsequent production and distribution phases,

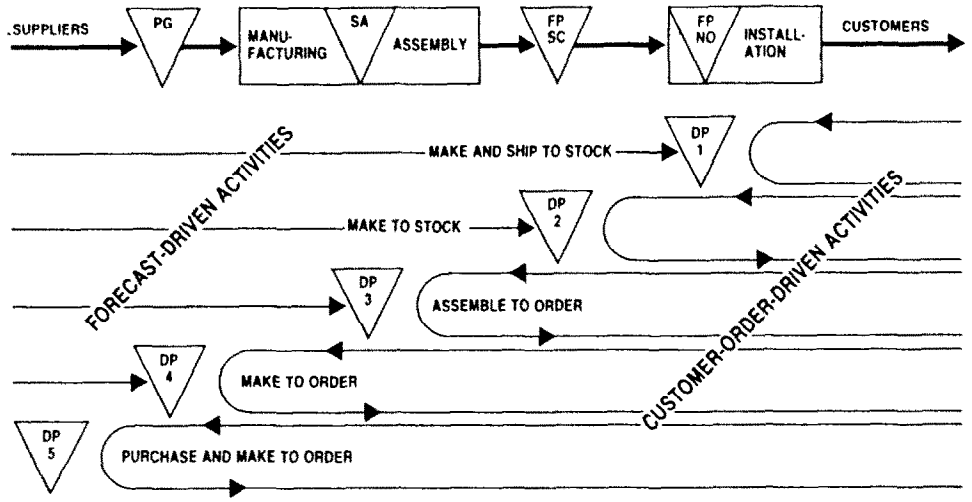

Fig. 2. Five decoupling point positions representing five logistic structures.

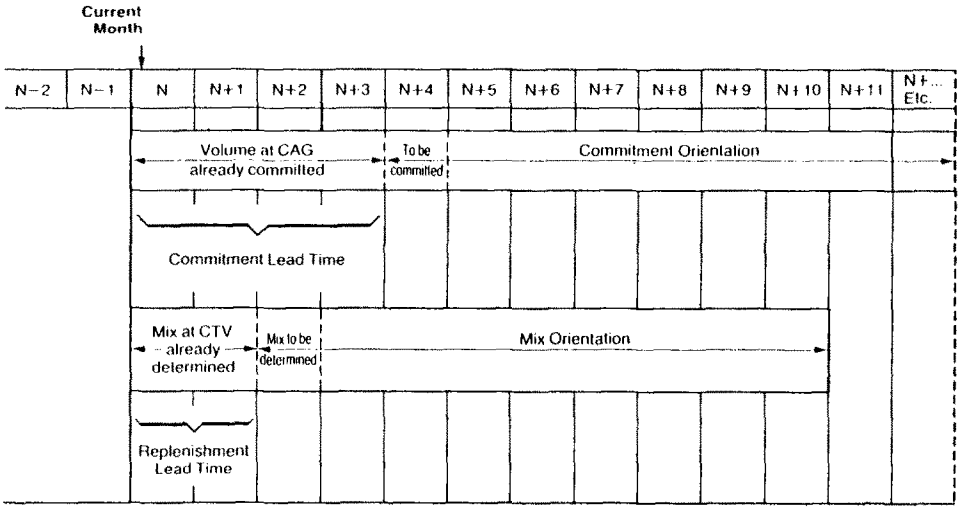

Fig. 3. Leadtimes and horizons for volume commitments and mix replenishments.

hence for a coordinated inventory control system for subsequent stockpoints.

- It is based on a *top–down planning approach*, reflecting the need for volume planning with a longer horizon than needed for the detailed mix planning.
- It is a *periodic review system*, based on *rolling sales forecasts and plans.*
- At each decision moment (at the beginning of each period), *volume commitments* are made on a high aggregation level called Commitment Article Group (CAG). This determines production volumes for each product family for a horizon of typically three to four months. The purpose of this procedure is twofold: to allocate capacity and to initiate purchasing and/or fabrication of long leadtime components which are not type-specific.
- At each decision moment, the earlier committed total volume is disaggregated to determine the actual production mix (item level), typically for a short horizon of one or two periods (Fig. 3 further illustrates the rolling horizon

character of the volume commitment and mix planning procedures).

Summarizing, CPL has been designed as a periodic, multi-echelon control system, exploiting hierarchical procedures, based on naturally arising part families (as a result of a highly modular product structure). An analytic framework for such a system will be outlined in the next section.

3. Average cost analysis of multi-echelon systems: Theoretical results

For ease of exposition we first consider a simple serial system with two successive stockpoints (or installations), consisting e.g., of a manufacturing and an assembly phase (Fig. 4). The system is subject to stationary stochastic demand (this assumption is relaxed in Section 5) which originates at the downstream installation. Products present at installation 2 or in transfer between installation 2 and installation 1 are charged at a rate h_2 per unit per period, products present at installation 1 are charged at a rate h_1. A penalty cost p per unit per period is incurred if installation 1 is unable to meet market demand. Both penalty and inventory costs are charged at the end of a period. In the next section we discuss how to translate this model into a model based on inventory holding costs and service degrees.

The system is (re)planned periodically (with a fixed review period). We assume that no fixed production and distribution costs are present. Although this assumption has been criticized sometimes [7] it is common practice in most industrial companies, to decide on replanning frequencies on a higher, tactical level. Therefore, these fixed costs are already accounted for on that higher level and thus do not play any further role in the daily operation.

Finally, we assume that excess demand is backlogged. Since the variable production and distribution costs are again linear, it follows from this last assumption that they may be ignored (all requested products are produced and distributed eventually), i.e., these costs do not play a role in determining optimal (with respect to *average* costs) control policies. hence, we may focus on inventory holding costs.

Let F_l denote the distribution of the l-period cumulative demand u_l, for all l. If $l=1$, we suppress the index. Furthermore, let l_1 denote the delivery leadtime for goods ordered by installation 1 from installation 2 *if these goods are available* and let l_2 denote the delivery leadtime for goods ordered by installation 2 from the (infinite capacity) initial supplier.

Following Clark and Scarf [6], we define the *echelon stock* of a given installation as all stock at that installation plus in transit to or on hand at any installation downstream minus the backlogs at the most downstream installations (which do not have a successor). The chain under consideration is called the *echelon*. An echelon stock may be negative, indicating that the backlogs are larger than the total inventory in that echelon. The *echelon inventory position* of an echelon denotes the echelon stock plus the materials already ordered but not yet available in its most upstream installation.

The following theorem is fundamental. Its proof (albeit with slightly different definitions) can be found in Langenhoff and Zijm [15].

Theorem 3.1. Consider a policy which, at the beginning of every period, increases the echelon inventory position of echelon n to y_n $(n=1,2)$. Let $D^{(2)}(y_1,y_2)$ be the associated average costs (de-

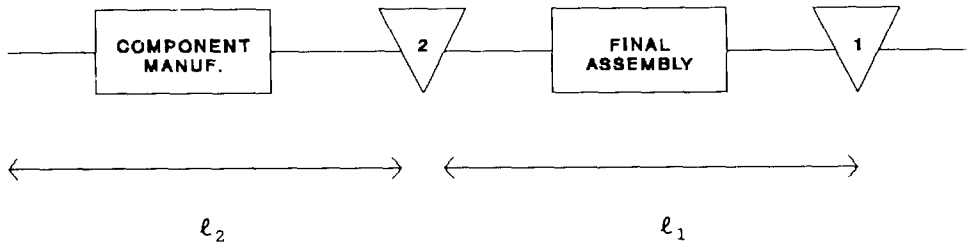

Fig. 4. A serial multi-echelon system.

fined on $\{(y_1,y_2) \mid y_1 \leqslant y_2\}$ only). The function $E^{(2)}(y_1,y_2)$ can be determined recursively by

$$D^{(1)}(y_1) = \int_0^\infty L_1(y_1 - u_{l_1})\, \mathrm{d}F_{l_1}(u_{l_1})$$

$$D^{(2)}(y_1,y_2) = D^{(1)}(y_1) + \int_0^\infty L_2(y_2 - u_{l_2})\, \mathrm{d}F_{l_2}(u_{l_2}) + \int_{y_2-y_2}^\infty [D^{(1)}(y_2 - u_{l_2}) - D^{(1)}(y_1)]\, \mathrm{d}F_{l_2}(u_{l_2})$$

Here, the functions L_n are defined by

$$L_1(x_1) = h_1 \int_0^\infty (x_1 - u)\, \mathrm{d}F(u) + (p + h_1 + h_2) \int_{x_1}^\infty (u - x_1)\, \mathrm{d}F(u) \quad \text{for all } x_1$$

$$L_2(x_2) = h_2 \int_0^\infty (x_2 - u)\, \mathrm{d}F(u) \quad \text{for all } x_2$$

The last two terms in the formula for $D^{(2)}$ are the additional holding costs in installation 2 and the penalty incurred by installation 2 due to its unability to completely satisfy the demand of installation 1. The restriction of its domain to $\{(y_1,y_2) \mid y_1 \leqslant y_2\}$ reflects that the echelon inventory position of installation 2 can never be smaller than the echelon inventory position of installation 1.

The importance of the functions $D^{(1)}(y_1)$ and $D^{(2)}(y_1,y_2)$ is established by the following results (Langenhoff and Zijm [15]):

Theorem 3.2. The overall average cost optimal policy is contained in the class of policies defined in Theorem 3.1. The order-up-to levels of this policy are determined by (S_1,S_2), in which $D^{(2)}(y_1,y_2)$ takes its absolute minimum. This absolute minimum in turn can be found by subsequently minimizing the functions $D^{(1)}(y_1)$ and $D^{(2)}(S_2,y_2)$, where S_1 denotes the minimizing value of $D^{(1)}(y_1)$.

The above two theorems together describe the basic analysis of a two-stage serial multi-echelon system. The extension of the results to an N-stage serial system is straightforward (compare Langenhoff and Zijm [15], Van Houtum and Zijm [17] and will not be discussed here. The key element is the decomposition result (established initially by Clark and Scarf [6] in a discounted cost framework), stating that an N-echelon serial system can be optimized by first optimizing a single stage system, then a 2-echelon system, a 3-echelon system, etc, by using recursive structures as in Theorem 3.1.

Note that we essentially determined an optimal control policy for only one product type. Now, let us consider the situation where a total volume for a product family has been committed (compare Section 2) which l_2 periods later has to be disaggregated into production quantities for end items (which then are completed l_1 periods later). Recall from Section 2 that a volume commitment enables us to order (or make) already certain components which are *not* type-specific. Without loss of generality we may assume that each particular item contains just one *non*specific component. As in the serial case we may define an echelon stock for this nonspecific component as the total amount of components in stock which still have to be allocated to an end-item plus all components already allocated to any end-item (we even count nonspecific components in final products and compensate for eventual backlogs). In the same way, we may define an echelon inventory position for the nonspecific component as its echelon stock plus all components already ordered but not yet available at the moment of disaggregation.

Let the holding charge of product n ($n = 1,\ldots,N$) be denoted by $h + h_n$ (where h is charged for the nonspecific components available when the mix replenishment decision is taken). Let furthermore p_n denote the penalty costs incurred in the case of a possible (temporary) shortage of product n. The l-period demand distribution for product n will be denoted by $F_l^{(n)}$.

We will first explain the disaggregation procedure (i.e., how to split up a total volume into end-

item quantities). let the inventory position of product n, just prior to the mix replenishment decision, be denoted by b_n. Let furthermore b denote the echelon stock of nonspecific components. Clearly, since any final product requires exactly one nonspecific component, we have $\sum_{n=1}^{N} b_n \leqslant b$. Optimal order-up-to levels are determined by solving the following nonlinear program

$$(P[b]) \quad \min \sum_{n=1}^{N} D^{(n)}(y_n)$$

$$\text{under} \sum_{n=1}^{N} y_n \leqslant b \qquad (*)$$

$$y_n \geqslant b_n \quad \text{for } n=1,...,N$$

The functions $D^{(n)}$ are defined by

$$D^{(n)}(y_n) = \int_0^{\infty} L_n(y_n - u_{l_1})\, \mathrm{d}F_{l_1}^{(n)}(u_{l_1})$$

$$\text{for } n=1.,,,.N,$$

where

$$L_n(x_n) = h_n \int_0^{\infty} (x_n - u)\, \mathrm{d}F^{(n)}(u)$$

$$+ (p_n + h_n + h) \int_{x_n}^{\infty} (u - x_n)\, \mathrm{d}F^{(n)}(u)$$

if $x_n \geqslant 0$. Langenhoff and Zijm [15] give an efficient procedure to solve program $P[b]$, based on a subsequent application of Langrangean techniques on certain subproblems. If the set of constraints $(*)$ is not really restrictive to the optimum (which is generally the case, see Eppen and Schrage [18], Van Donselaar and Wijngaard [19]) we say that the inventory positions of the final products are balanced; in this case the solution procedure resembles the one given by Eppen and Schrage [18] (the equivalence of program $P[b]$ with the Eppen and Schrage approach is discussed in Langenhoff and Zijm [15]. For the sequel we assume that indeed the set of constraints $(*)$ are not restrictive, i.e., that the inventory positions of final products are always balanced.

This latter assumption appears to be the key condition to establish a decomposition result similar to the one previously discussed for th serial system [10,15,17]. First we solve single stage systems (with leadtime l_1) for each individual end-item n, to determine order-up-to levels $S^{(n)}$ after which the solutions $(y_1,...,y_N) = (z_1[b],...z_N[b])$ of program $P[b]$ (parameterized by b and, since we assume balance, without the constraints $(*)$) are used to determine an overall average cost function. The solutions $(z_1[b],...z_N[b])$ can be analytically determined as functions of b if the demand per period is normally distributed (e.g., Eppen and Schrage [18]) while good approximations for these functions can be obtained if demand can be characterized by a mixture of Erlang distributions [17]. These results permit us to calculate an overall order-up-to level S for the complete system. As a result, the echelon inventory position of the nonspecific components is returned to S, thereby determining an aggregate order for the complete product family, while l_2 periods later the mix-replenishment decisions are determined as a solution of the program $P[S-u_{l_2}]$, *where* u_{l_2} denotes the cumulative demand for all products (and hence all nonspecific components) during l_2 periods.

Summarizing, the above theory provides a basic framework to analyze, under an average cost criterion, multi-echelon systems where in the final phase a gross volume is disaggregated into end-item quantities for different final products. In the next section we briefly indicate how to proceed from a pure cost analysis to models in which we focus on customer service levels and inventory holding costs solely.

4. Numerical evaluations and relations to service degrees

In the preceding section we have seen how the minimum average cost policies in multi-echelon systems under stationary demand can be determined by subsequently minimizing a series of one-dimensional functions. Although this latter property substantially facilitates the minimization of $D(y_1,...,y_N)$, the minimization of these one-dimensional functions is far from trivial since they are defined only implicitly (by recursion). Van Houtum and Zijm [16] have developed procedures, based on moment approxima-

tions of appropriate distribution functions, to carry out these minimizations. Exact procedures to determine the optimum costs have been developed recently for the case where the demand distribution can be characterized as a mixture of Erlang distributions (Van Houtum and Zijm [17]).

A detailed treatment of these procedures is beyond the scope of this paper, here it suffices to remark that they enable us to calculate optimal order-up-to levels for fairly large systems in a few seconds up to a few minutes on an IBM-AT compatible PC. For serial systems, these procedures moreover reveal as a side result the expected average shortage (EAS) per period. If we define a customer service level δ as the fraction of demand satisfied directly from stock (at installation 1) then clearly

$$\delta = 1 - \mathrm{EAS}/\mu$$

where μ denotes the expected one-period demand.

Now consider the serial system where the penalty cost p is a control parameter and let $\delta(p)$ be the associated service level, given the optimal average cost policy. Then $\delta(p)$ appears to be monotone increasing as a function of p. Since we have developed quick minimization procedures for the analysis of serial systems, it is easy to fit a penalty cost p such that a target customer service level δ is met (for example, by using bisection). Moreover, the resulting policy appears to be optimal within the class of all strategies satisfying this target customer service level δ, with respect to the criterion of minimizing the average inventory holding costs solely. All these results can be found in Van Houtum and Zijm [16,17], and are easily extended to the situation with product families where both volume and mix decisions have to be made.

5. Conclusions and suggestions for further research

All results discussed so far have been derived for uncapacitated systems under stationary demand. Since in particular consumer products suffer from still decreasing product life cycles as well as seasonal fluctuations, and since in most cases capacities are limited simply because of cost considerations, these assumptions have to be relaxed. However, it is easy to verify that, as long as capacities are not involved, the analysis in the preceding sections remains valid for non-stationary demand processes. For instance, when we focus on ultimate demand satisfaction in period t (in an N-serial system), we have to replace S_1 by $S_{1,t-l_1}$, S_2 by $S_{2,t-l_1-l_2}$, etc. Clearly, these parameters will change with t, depending on the demand forecasts for the demand in the interval $[t,t-l_1]$, in the interval $[t,t-l_2-l_1]$ and so on. All derivations however remain essentially unchanged. Note that for the situation where we first have to agree on volume commitments on a product family level, also demand forecasts are needed only on that aggregate level.

More serious problems however arise when we introduce finite capacities in our model and in particular, when we combine nonconstant finite capacities with highly fluctuating demand patterns. Here it may be necessary to shift production to preceding periods because even average demand in some period cannot be covered by normal production capacity. A stochastic version of Land's algorithm (cf. Silver and Peterson [20]) is investigated to cope with this situation. Alternative capacity resources might also be considered. This is the subject of future research.

References

1 Forrester, J.W., 1961. Industrial Dynamics. M.I.T. Press, Cambridge, MA.

2 Orlicky, J.A., 1975. Materials Requirements Planning. McGraw-Hill, New York.

3 Wight, O., 1981. MRP II - Unlocking America's Productivity Potential. Oliver Wight Limited Publications, Essex Junction.

4 Schonberger, R.J., 1982. Japanese Manufacturing Techniques. Free Press (MacMillan), New York.

5 Goldratt, E.M., 1988. Computerized shop floor scheduling. Int. J .Prod. Res., 26: 443–455.

6 Clark, A.J. and Scarf, H., 1960. Optimal policies for a multi-echelon inventory problem. Manage. Sci., 6: 475–490.

7 Clark, A.J. and Scarf, H., 1962. Approximate solutions to a simple multi-echelon inventory problem. In: K.J. Arrow et al., eds., Studies in Applied Probability and Management Science. Stanford University Press, Stanford.

8 Schwarz, L.B., ed., 1981. Multi-level Production/Inventory Control Systems: Theory and Practice, Studies in the Management Sciences, Vol. 16. North-Holland, Amsterdam.

9 Federgruen, A. and Zipkin, P., 1984. Approximations of dynamic, multilocation production and inventory problems. Manage. Sci., 30: 69–84.
10 Federgruen, A. and Zipkin, P., 1984. Computational issues in an infinite horizon, multi-echelon inventory model. Oper. Res., 32: 818–836.
11 Hax, A.C. and Meal, H.C., 1975. Hierarchical integration of production planning and scheduling. In: M.A. Geisler, ed., Logistics, Studies in the Management Sciences, Vol. 1. Elsevier, North-Holland.
12 Bitran, G.R., Haas, E.A. and Hax, A.C., 1981. Hierarchical production planning: A single stage system. Oper. Res., 29: 717–743.
13 Bitran, G.R., Haas, E.A. and Hax, A.C., 1982. Hierarchical production planning: A two stage system. Oper. Res., 30: 232–251.
14 De Kok, A.G., 1988. Hierarchical production planning for consumer goods, CQM-Note nr. 71, Centre for Quantitative Methods, Philips, Eindhoven.
15 Langenhoff, L.J.G. and Zijm, W.H.M., 1990. An analytical theory of multi-echelon production/distribution systems. Statist. Neerlandica, 55(3): 149–174.
16 Van Houtum, G.J. and Zijm, W.H.M., 1990. Computational approaches for stochastic multi-echelon production systems CQM-note 088, Centre for Quantitative Methods, Philips, Eindhoven; (appears in Eng. Costs Prod. Econ).
17 Van Houtum, G.J. and Zijm, W.H.M., 1991. Analysis of multi-echelon production–distribution systems. In preparation.
18 Eppen, G. and Schrage, L., 1981. Centralized ordering policies in a multi-warehouse system with lead times and random demand. In: L. Schwarz ed., Multi-level Production/Inventory Control Systems: Theory and Practice, Studies in the management Sciences, Vol. 16. North-Holland, Amsterdam.
19 Van Donselaar, K. and Wijngaard, J., 1987. Commonality and safety stocks. Eng. Costs Prod. Econ., 12: 197–204.
20 Silver, E.A. and Peterson, R., 1985. Decision Systems for Inventory management and Production Planning, 2nd edition, Wiley, New York.