OR Spektrum (2000) 22: 313–345

OR Spektrum

Towards intelligent manufacturing planning and control systems

Perspektiven intelligenter Produktionsplanungs- und Produktionssteuerungssysteme

W.H.M. Zijm

Faculty of Technology Management, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

Received: August 3, 1999 / Accepted: October 30, 1999

Abstract. In this paper, we review some well-known manufacturing planning and control (MPC) systems and models, and highlight both their advantages and major drawbacks. The analysis indicates that various important planning and control problems, as they arise in industry, are not properly addressed by current MPC systems. A well-known production system typology, illustrated by industrial examples, is briefly discussed to further highlight these planning and control problems. Next, we define a basic framework architecture for planning and control in both make-to-stock and make-to-order systems. The emphasis in this framework is on an integration of technological and logistics planning, and on an integration of capacity planning and materials coordination issues. In addition to this architecture, we further define an algorithmic framework that explicitly aims at the latter integration. To complete the architecture, we suggest a variety of procedures and algorithms to implement in the various modules.

Zusammenfassung. In dieser Arbeit betrachten wir einige bekannte Systeme und Modelle der Produktionsplanung und -steuerung (PPS) und stellen ihre jeweiligen Vor- und Nachteile heraus. Die Analyse zeigt, daß verschiedene in der Industrie anzutreffende Planungs- und Steuerungsprobleme von den gegenwärtigen PPS-Systemen nur unzureichend unterstützt werden. Um diese Planungsprobleme eingehender zu erläutern, wird auf eine bekannte Typologie von Produktionssystemen sowie auf Beispiele aus der Industrie zurückgegriffen. Daran anschließend wird eine grundlegende Architektur von Planungs- und Steuerungssystemen sowohl für lager- als auch für auftragsorientierte Produktionssysteme vorgestellt. Der Schwerpunkt liegt hierbei auf der Integration technologischer und logistischer Planungsaspekte sowie auf der Integration von kapazitäts und materialflußorientierten Gesichtspunk-

ten. Darüber hinaus wird das Grundkonzept eines Algorithmus entwickelt, der insbesondere den letztgenannten Integrationsaspekt berücksichtigt. Die vorgeschlagene Systemarchitektur wird durch eine Reihe von Verfahren und Algorithmen zur praktischen Umsetzung einzelner Systemmodule ergänzt.

Key words: Manufacturing Planning and Control – Architecture of PPC systems – Algorithms

Schlüsselwörter: Produktionsplanung und -steuerung – Architekturen von PPS-Systemen – Algorithmen

1 Introduction

Within the last three decades, manufacturing and logistics managers have faced major challenges as a result of market globalization and in particular a fierce competition from industries in emerging Asian countries. An immediate positive effect was the renewed awareness of manufacturing as a major competitive economic weapon (Skinner, 1985). At the same time, industrial managers began to realize the potential applications of the rapid advances in information and communication technology. As a result, the attention for new concepts and solution methodologies increased dramatically, not only in business management, but also in the scientific community.

Unfortunately, a strong economic pressure on developing and implementing new solutions may have some negative aspects as well. The tendency to uncritically embrace a solution concept, developed for a rather specific manufacturing environment, as the panacea for a variety of other problems in totally different environments has led to many disappointments. Undoubtedly, sales managers of IT companies have done a better marketing job than scientists have; as a result, companies have spent billions of dollars to implementing costly information systems, without realizing or even coming close to the promised benefits (see e.g. Hopp and Spearman, 1996).

All this does not mean that the many three letter acronym systems that have passed during the last couple of decades, did not have any merit or did not bring any advantage. Almost each particular solution concept may yield significant improvements, *at least if the (often hidden) conditions of the underlying models are fulfilled*. Unfortunately, these conditions were often not explicitly stated but nevertheless turned out to be rather severe. They include a highly reliable manufacturing system and a perfectly predictable demand, a high data accuracy, unlimited production capacities, a limited product variety and a completely known product structure upon order acceptance, and manufacturing lead times that are independent of both workload and order mix. Obviously, these assumptions seldom reflect reality, but managers aware of this were often advised to change manufacturing practice in order to fit the needs of the system. Although seemingly unnatural, such an approach may occasionally help if it leads to a reduction of manufacturing system complexity and to a growing awareness of the importance of high product and process quality.

More general, the positive message is that many MPC systems at least helped to articulate the need of sound inventory accounting records and a well-defined product structure. In addition, the extremely important notion of dependent versus independent demand was underlined. But despite these changes, most MPC systems require an almost perfect manufacturing environment in order to fulfill their promises. The results were both negative and positive: many implementation failures, but also a growing number of companies now having a better idea of what they really need, unfortunately at a high price.

Advances in manufacturing practice

The impact of automation on manufacturing and logistics can hardly be overestimated. Before focusing on planning and control, let us briefly delineate the full picture of changes in manufacturing practice that we have witnessed in the last three decades. In principle, we may distinguish three different fields:

- hardware automation (Computer Numerically Controlled Machines, Flexible Manufacturing and Assembly Systems, Automatic Transport Systems, Automated Storage and Retrieval Systems),
- Design and Process Planning (Computer Aided Design, Computer Aided Process Planning, Rapid Prototyping techniques),
- Manufacturing Planning and Control Systems (Materials Requirements Planning, Manufacturing Resources Planning, Optimized Production Technology, Workload Control Systems).

Besides, as a fourth impact factor with a primarily organizational background (although sometimes enabled by information technology), we mention:

- System Complexity reduction (Just in Time production, Production Flow Analysis, Cellular and Team-based Manufacturing, Business Process Re-engineering).

Business Process Re-engineering is included here although it addresses a much wider area: not only manufacturing but also many administrative processes in the service sector (Hammer and Champy, 1993). Some researchers debate the question whether Just in Time production is not merely another control procedure. That may be true but, as we will point out below, JIT control only leads to desired performance improvements in relatively simple manufacturing system structures. For that reason, we chose to list it under the fourth impact factor. More general, all approaches above have in common that they attempt to increase manufacturing capabilities, to support manufacturing planning and control, or sometimes to diminish the planning and control complexity. The ultimate goal is to improve system performance, not only in terms of efficiency and costs, but in particular on aspects of product and process quality, flexibility and speed (cf. Deming, 1982; Blackburn, 1991).

Limited impact of Operations Research on manufacturing planning and control

In the last 50 years, Operations Research has developed into a mature science, with successful applications in a variety of fields such as manufacturing and logistics, telecommunication and computer sciences, military studies and many domains in the service sector. Despite these successes, the impact of OR models in the developments outlined above was limited. Even more, in the first book on Materials Requirements Planning, Orlicky (1975) explicitly stated that there was no need for OR models but, instead, that information systems formed the key to better planning and control procedures in manufacturing. It should be said that at that time most OR studies indeed concentrated on developing exact solutions for relatively simple problems, as opposed to addressing more realistic problems in their full scope. And although there is nothing wrong with working on well-structured simple problems in order to gain insight into more complex problems, it does not immediately help to gain credibility from managers who are in need for systems to master their factory planning problems.

In contrast to Operations Research models and methods, concepts like Manufacturing Resources Planning, Just in Time production, Lean Manufacturing and their followers were embraced almost immediately after their introduction, for several reasons. One reason is their apparent procedural simplicity. The logic behind MRP is easily explained to any manager and the same holds for JIT. A second reason is indeed due to the computational power of computers, allowing these systems to address industrial planning problems in a scale that was never achieved before. *What is less obvious, and indeed a major cause of so many implementation failures, is the fact that a great number of conditions have to be fulfilled before the apparently simple logic can work successfully*. In addition, many managers easily overlook the fact that most systems primarily are developed for advanced bookkeeping. Indeed, despite the name, *MRP does not really plan, nor does JIT*. But most managers in the past also did not *expect* advanced planning support, or even worse, could not believe that an automated system might be able to support intelligent decision making. And exactly that is the "raison d'être" of Operations Research models.

However, the times have changed. In the last two decades, the scope of researchers in quantitative modelling has broadened considerably towards more realistic models for which sound approximate models have been defined. New techniques, such as taboo search, simulated annealing and approximate stochastic network algorithms, have been developed to analyze large size models while research in aggregation and decomposition techniques has been multiplied. In addition, many managers begin to recognize the major drawbacks of most current planning systems and are indeed demanding more intelligent solutions instead of administrative routine procedures. Under the name Advanced Planning Systems, various software developers are responding already by integrating OR models and techniques in their planning and control procedures. The group of models used is still limited, in particular stochastic models are not widespread so far in Advanced Planning Systems, but there are many indications that their use will grow as well.

The goal of this paper is to contribute to the development of more intelligent manufacturing planning and control systems by explicitly outlining a number of useful models and methods. In discussing MPCS's, we will occasionally touch on topics of engineering and process planning, but only incidentally mention system complexity reduction issues. We start with a more in depth evaluation of some dominant approaches in manufacturing planning and control, pointing out both their merits and their drawbacks. In addition, we briefly review a well-known typology of manufacturing systems, highlighted by examples from industry. Next, we emphasize the importance of a hierarchical planning and control framework, in view of the many uncertainties that arise in various phases of the manufacturing cycle. An integral manufacturing planning and control architecture is discussed, focussing on the integration of technological and logistic planning and control, and on the integration of capacity planning and material coordination issues. Using this typology as a reference model, we propose a variety of models and algorithmic approaches that fit different groups of systems within that typology. A basic algorithmic framework that explicitly addresses the integration of resource constraints, workload control and material coordination, is discussed in some more detail. We end the paper with conclusions and directions for future research.

2 An evaluation of current manufacturing planning and control systems

In this section, we discuss some of the major advantages and disadvantages of the most relevant manufacturing planning and control approaches as they are found in practice today. In order to limit the scope of this paper (and to not repeat many textbooks), we assume familiarity of the reader with the basic procedures of the systems that we discuss. We start with a brief account on MRP systems.

Push systems: material requirements planning and manufacturing resources planning

The development of *Material Requirements Planning* (MRP) in the late sixties and early seventies meant a revolution to manufacturing planning and control in the Western world. Although the ideas were not new, their implementation was made possible for the first time because of the computational power of (mainframe) computers. The key of MRP is the recognition of the role of a product structure in generating demand profiles. Instead of the Statistical Re-Order Point procedures used so far, that trigger production once the inventories of parts fall below a specified level, MRP derives demand on subassembly and parts levels directly from a production schedule for final products. This is called *dependent demand*, as opposed to the independent demand arising from external customers. A key role in MRP is played by the *Master Production Schedule* (MPS), specifying the production schedule for end-items, the *Bill of Material* (BOM), describing the product structure, and the *fixed off-set lead times*, being the time windows to be *reserved* for manufacturing a particular part or subassembly or final product. Off-set lead times are usually set

such that they incorporate the effects of lot sizes, waiting times and the like. MRP systems normally plan on a periodic basis (using discrete time periods called time buckets), although there exist examples of continuous time systems. For a detailed account on MRP, the reader is referred to Orlicky (1975), Vollmann et al. (1997) or Hopp and Spearman (1996).

MRP systems are often called *push control systems*, since the MPS drives the scheduling activities (pushes the material through the system), without regarding actual work loads. Initially, MRP did not consider any capacity constraint at all, nor did it account for process uncertainty. As a consequence, safety stocks were not allowed anywhere in the system, except perhaps at the Master Production Scheduling level. Basically, all uncertainties and possible capacity problems should be absorbed by defining the off-set lead times appropriately.

In particular the lack of capacity considerations was early recognized as an important shortcoming, leading to enhancements such as closed loop MRP and in particular *Manufacturing Resources Planning* (MRP II), see Wight (1981) and Vollmann et al. (1997). The latter is a more elaborate functional framework for planning and control of manufacturing systems, with Material Requirements Planning still as the engine driving the lower level production schedules. In addition, at a more aggregate level we find functions like demand management and in particular Rough Cut Capacity Planning (RCCP), specifying at a global level what capacity is needed, whereas next to material requirements planning a Capacity Requirements Planning (CRP) function is defined for more detailed capacity calculations at the operational level. However, these names are quite misleading. *Both functions have nothing to do with finite capacity loading*, i.e. with automatically matching required and available capacities. At best they check, in less or more detail, whether available capacity is sufficient to make a proposed MPS/MRP scheme feasible. If not, a planner may decide to generate an alternative Master Production Schedule, or possibly to adjust capacities. At a low level finally, MRP II allows for the inclusion of Shop Floor Control and vendor control systems. Instead of MRP II, we currently often find *Enterprise Resource Planning* (ERP) Systems but they merely extend the system architecture, by integrating the MRP II functionality with other business functions such as financial accounting and manpower planning, without adding any intelligent planning.

There is no doubt that MRP systems have dramatically changed the view on how to control large-scale manufacturing systems. Their success is in the first place due to the increasing power of modern computer systems, enabling detailed requirement calculations based upon a dependent demand structure. Modern MRP systems request many parameters to be set by the user, regarding lot sizes, safety stocks, off-set lead times and others, but do not provide any help in setting these parameters. Above all, even MRP II does not really integrate material and capacity planning. It does not plan against finite capacity, and moreover, it does not generate alternative production schedules in case some materials or parts do not become available as planned (wrong quantities, inferior quality). Moreover, it requires a rather detailed knowledge on what resources, materials and parts are needed when accepting customer orders, a condition often not fulfilled in Make-to-Order, let

alone in Engineer-to-Order companies. Basically, MRP treats the world as being deterministic, where any possible uncertainties should be covered by sufficiently long off-set lead times. As a result, these lead times tend to grow longer and longer.

Many authors have criticized the obvious deficiencies of MRP. Very early already, Whybark and Williams (1976) advocated the use of safety stocks and safety lead times, in case of demand uncertainty or process uncertainty, respectively. On an aggregate level (basically to support Rough Cut Capacity Planning), linear programming models have been suggested to smooth demand over a longer period. Karmarkar (1987) exploited a queueing analysis to determine off-set lead times as a function of available capacity, balancing Work in Process inventory holding costs against set-up costs and time. Several authors have advocated the use of workload control as a means to watch over internal lead times (see e.g. Bechte, 1987; Bertrand et al., 1990; Wiendahl, 1993; Spearman et al., 1989). They use different procedures to determine acceptable workload norms, given a measure of available capacity. We return to workload control when discussing Pull Systems. At a low level, Shop Floor scheduling systems can be used to cope with capacity issues, since they indeed consider machine capacities and job routings simultaneously. However, once a shop is not properly loaded, even a mathematically optimal solution to a scheduling problem may represent an unacceptable schedule. Günther (1986) proposes a hierarchical model for production planning and scheduling; however, a complete algorithmic framework that simultaneously considers both capacities and materials at all levels in a hierarchical planning system, is still lacking.

Pull systems: just in time, Kanban and workload control

The flood of cheap and reliable products that reached Europe and the United States after the two oil crises in the mid and late seventies, forced manufacturing leaders to study the causes of the undeniable successes of Japanese manufacturing. What they discovered was a set of procedures that soon became known as the *Just in Time* (JIT) system. What JIT clearly distinguishes from MRP systems is that is does not in the first place rely on computerized planning procedures, but, on the contrary, on organizational changes at the shop floor and a basic principle that can be summarized as: *deliver parts or materials to a work station only when they are needed.*

Unfortunately, as with MRP systems, many managers were initially misled by the apparent simplicity of the underlying principle of JIT. This principle can best be explained by studying one particular implementation: a Kanban System (Kanban is the Japanese word for card). Consider products that are manufactured by means of a set of workstations that are visited in a predetermined sequence. In front of each workstation, only a few parts or components of each particular product running over the line are available. On each part or component a card (Kanban) is attached. Once a workstation starts a process step for a particular product, it picks the required components or parts from its input stock. The cards attached to these parts are removed and sent to the preceding workstation, triggering production of similar parts at that workstation to replenish the input stock of its successor. The preceding

workstation in turn picks the relevant parts from its own input stock, removes their Kanbans which are sent backwards, etc. Hence, in this way, a final demand generates a sequence of *replenishment orders* all the way backwards through a sequence of workstations, and finally to material procurement. If the in-process-inventories are low or almost zero, one may indeed speak of delivery on demand or, alternatively, of Just in Time production.

Although the basic procedure outlined above is extremely simple, the system only works provided a number of rather severe conditions are satisfied. Clearly, the number of products running on a single production line has to be limited, in order to prevent stocks of many different parts between work stations. In addition, large set-up times between product changes or significant breakdowns may ruin the system's performance, since any machine stop quickly propagates through the system in case of minimal buffer stocks. Although the argument can be turned around as well (and has been turned around: reducing buffer stocks helps to reveal and next to solve production problems such as breakdowns and rejects), many companies have to work a long way to satisfy the conditions. But indeed, one of the key success factors of the Toyota production system has been the reduction of set-up times and the smoothing and subsequently freezing of monthly production schedules (Monden, 1998). In summary, a Kanban control system requires an extremely high technical flexibility of the production system, and then still performs well only in a relatively stable repetitive manufacturing environment, certainly not in a small batch or one-of-a-kind Make-to-Order system. However, once these conditions are fulfilled, they result in relatively short and highly stable internal manufacturing lead times, hence allowing for very reliable Master Production Schedules (Schonberger, 1982).

The stability of internal manufacturing lead times is typically due to the fact that parts are only released when needed at subsequent stations. For that reason, we speak of *pull systems*, as opposed to the push schedules in MRP systems. It is this internal lead time stability that has led many authors to promote *workload control* as a guiding principle in Manufacturing Planning and Control (Bertrand et al., 1990; Wiendahl, 1993; Spearman et al., 1989). Basically, they suggest to release production orders to a work cell or job shop, consisting of multiple work stations, only when the work load already present in the cell, drops below a certain level. In this sense, Kanban control is a highly specific implementation of workload control, on the level of individual work stations. Variants of workload control distinguish between the load of various products, or between different workstations within a cell. An easy way to determine the relationship between the load and the resulting internal manufacturing lead times within a shop is by modelling and analyzing this shop as a multiclass Closed Queueing Network (e.g. Buzacott and Shanthikumar, 1993). Unfortunately, most authors advocating workload control emphasize the stability of the *internal lead times*, without paying attention *to the time that production orders spend waiting before being released.* Some authors avoid these external waiting time problems by simply denying their existence (an order becomes an order only when it is released), but nevertheless they require materials or parts to be available in stock (and storage time is part of the total lead time as well,

and certainly contributes to internal stocks). As a result, stable internal lead times at departmental levels do not necessarily induce stable lead times at the factory level. This holds in particular in a Make-to-Order company, in which customer order delivery dates are often set without taking into account actual or future work loads, if known at all. Hence, although stabilizing internal work loads clearly has its merits, it does not remove the discrepancy between meeting customer order due dates in a highly fluctuating market on the one hand, and having stable internal loads on the other hand. It seems there is a need for *integrating workload control and resource availability planning on a higher level, or even supporting order acceptance by sophisticated load-based procedures.* We will come back to this issue in the next sections.

Hierarchical production planning and multi-echelon inventory systems

Both MRP II and JIT, and to a less extent, workload control, are established systems in industry that have been implemented at numerous places (albeit often with limited success at best, due to reasons discussed above). That does not hold for Hierarchical Production Planning which differs from the earlier mentioned approaches in some important aspects. Hierarchical Production Planning (HPP) is based on mathematical programming models in the first place, including Linear Programming and combinatorial optimization models (e.g. knapsack problems). Second, it is strongly capacity-oriented, as opposed to the material orientation of both MRP and JIT. But most important, its philosophy is based on the hierarchical nature in which production decisions in a firm are often made, specifying global production quantities at an aggregate level first, and decomposing these quantities later to detailed item production lots. That may look similar to the MRP approach but it is not. Where MRP has to specify in an early stage which products will be made in a certain period, HPP only determines for which product *types* capacity has to be reserved, at some later point in time disaggregates these capacity reservations to time slots reserved for particular product *families* within each type, and finally determines how much time in each slot should be spent to the production of particular *items* within each family. An early description of HPP has been given by Hax and Meal (1975), a detailed description of the different procedures can be found in Hax and Candea (1984), while an extension to a two-stage system (for instance, fabrication followed by assembly) is discussed by Bitran et al. (1982).

The strong capacity orientation or HPP makes it particularly suitable for the batch processing or semi-process industries where the material complexity is often lower than in discrete manufacturing, whereas resources are often expensive and therefore have to be highly utilized. It is able to deal with set-up times when considering capacities at the family planning level but on the other hand it cannot handle uncertainty properly. Basically, the planner is asked to set safety stock levels, in particular when production decisions at the item level are based upon runout times. The initial versions of HPP did not consider lead times at all, but these were necessarily included in the extension to a two-stage system (Bitran et al., 1982). This extension however immediately showed the complexities that arise in multi-stage

systems. An attempt to integrate HPP and MRP for the production of computers is presented by Meal et al. (1987). Still however, the basic drawbacks of HPP systems are the complexities arising in multi-stage systems and the fact that uncertainty at the various levels is not incorporated systematically.

Multi-echelon systems are seldom mentioned when discussing various production planning systems. These models were initially developed to handle material co-ordination problems in a multi-stage system. Under rather strict assumptions, Clark and Scarf (1960) proved the optimality of base stock systems for serial systems in a stochastic setting. Later, inverse arborescent systems (distribution systems with one central depot and multiple local warehouses) were studied by many authors, e.g. Eppen and Schrage (1981), Federgruen and Zipkin (1984), Van Houtum et al. (1996), and Diks et al. (1996). Rosling (1989) and Langenhoff and Zijm (1990) showed the correspondence between arborescent (assembly) systems and serial systems. In the context of hierarchical production planning the following observation is crucial. In a centralized two-echelon distribution system, one first decides upon a replenishment quantity at the central depot stock, and later determines how to distribute this stock to the local warehouses, using the most recent demand information. Exactly the same procedure can be used for a product family in which the various items have a common, sometimes expensive, component (e.g. a common cathode ray tube in a family of television sets). Again, one may at an early stage decide upon the number of tubes to be produced, and only later determine what quantities are allocated to the different television sets. This idea has been worked out in detail by De Kok (1990), and is described in more general terms by Zijm (1992). A similar remark holds in principle for assembly structures and more general for mixed convergent and divergent structures. The central idea is that decisions on production quantities of common components can be made, *while postponing the decision on how to allocate these quantities to specific products as long as possible.* Similarly, in assembly systems one may relate the decisions on various components to each other. If the production of a particular component of an end-item is delayed, one may just as well reconsider the decisions on other components needed for the same end-item, or perhaps re-allocate these components to other end-items. A full exploration of these ideas is given by De Kok (1999). Essential in the analysis is the *timing of the various decisions on production quantities of parts or subassemblies in a particular product structure*. The decision on a particular quantity of a component still does allow quite some freedom on how and where to eventually use that component. This is entirely opposite to the planning logic of MRP where all production decisions are driven by end-item demand.

The theory of multi-echelon systems essentially covers demand uncertainty at different levels and allows for random lead times (although in that case the theory is less fully developed). On the other hand, it is primarily materials oriented; attempts to include capacity constraints have had limited success so far. We will return to capacitated multi-echelon systems in Section 5.

Supply chain management and advanced planning systems

To complete the account on major developments in the last three decades, we mention the strong focus of both companies and researchers on the management and control of complex *supply chains*, comprising of various suppliers that may operate in a network. In the case of *centralized* control, *Supply Chain Management* may be seen as an extension of integrated logistic control, although SCM encompasses not just production and distribution planning and control but also aspects of product design (co-makership), marketing and cost accounting. In a network that consists of several independent companies (*decentralized* decision making) a number of interesting additional questions arise, e.g. on the amount of information to be shared or, more general, on possible contracts between various partners in order to generate mutual benefits. Also, the duration of a possible partnership and the fact that companies often operate in more than one supply chain creates additional challenging research questions. SCM systems that are available today can be characterized in a way similar to ERP systems: they link a wide variety of business functions (purchasing, logistics, marketing, finance), but focus almost exclusively on centralized controlled environments, thereby concentrating on information management and hence without more intelligent, quantitatively based decision support functions. At the same time, the number of models and algorithms addressing a variety of problems in supply chain management (information management, supply contracts, design for postponement in quick response chains, international aspects) is rapidly growing, for a recent overview we refer to Tayur et al. (1999). Models of multi-echelon systems often provide a useful starting point for the analysis of more complex supply chains. Until now, the far majority of models concentrates exclusively on stock-based production.

Both ERP and SCM systems provide an information backbone that is minimally needed as a basis for sound planning procedures (although even the architecture may be criticized). But it is not enough, and indeed the lack of intelligent planning and decision support functions has been noticed today by a wide variety of users, as well as researchers. Recently, various software vendors have responded by developing so-called *Advanced Planning Systems*. The current state of the art shows the integration of hierarchical planning architectures with Linear Programming tools for aggregate production and capacity planning, and sometimes advanced shop floor scheduling systems at a low level in Make to Order production environments. Stochastic models that explicitly address demand or process uncertainties (e.g. stochastic multi-echelon models or models based on queueing networks) are absent in APS's to date.

In our view, *a sound hierarchical planning and control system should explicitly recognize and model the many uncertainties that arise during early planning phases* (e.g. during order acceptance and rough cut capacity planning). In an assemble-to-stock environment this uncertainty traces back to demand volume and mix variability. In a make-to-order system there still may be a great deal of uncertainty on the exact product characteristics and product structure (and hence on capacity and material requirements) in the early stages. As time progresses, more information

on the basic planning and scheduling ingredients (quantities, routings, processing times) becomes available, e.g. due to better demand forecasts or the completion of the process plan, allowing for the use of deterministic planning and scheduling tools on the short term. And even at a short term many random events occur (breakdowns, part rejects, rush orders), that may be treated in two alternative ways: either by modeling the stochastic behavior explicitly or by generating *robust* schedules based upon deterministic methods. The realization of such a hierarchical planning and control system, using models and algorithms that at each level recognize both the information and knowledge structure generated so far, as well as the still existing uncertainties, is far from realized. This paper is an attempt to fill in some parts of such an MPCS. An architecture will be developed in Section 4. A highly important question however concerns the selection of models and algorithms needed to fill in this architecture. The answer to this question in turn depends on the characteristics of the underlying manufacturing system. For that reason we first briefly discuss a manufacturing system typology in the next section.

3 A manufacturing system typology

In this section, we briefly review a well-known manufacturing system typology that serves as a reference when discussing specific algorithms. Manufacturing systems are usually classified along two dimensions. On the one hand we distinguish between various possible logistic product/market relations, while furthermore the internal organization often is used as a second classification criterion. Some authors further discern capacity- and materials-oriented manufacturing organizations. Let us first discuss logistic product/market structures.

Make and assemble to stock (MATS). This is the typical production philosophy for the majority of consumer products such as electronic equipment, food and drugs. The relationship between the manufacturer and the market is only indirect; typically the wholesaler and the retail sector perform the ultimate service to the market.

Make to stock, assemble to order (MTS/ATO). When a large variety of different products is built up from a limited number of components, it makes sense to produce components to stock but to perform the final assembly based on customer orders (catalogue products). In this way one avoids high final product inventories, while still being able to react relatively fast. The manufacturing of cars and trucks is a good example of an MTS/ATO system.

Make to order (MTO). Companies facing a high diversity of end-items in small quantities (small batch manufacturing) where the diversity originates already at the component level, typically operate in a Make-to-order mode. Most metal working (machine) factories belong to this category (e.g. large hydraulic pumps). In principle, materials are universal and often procured on the basis of forecasts.

Engineer to order (ETO). An engineer-to-order company typically designs and engineers products based upon a functional specification of the customer, and in

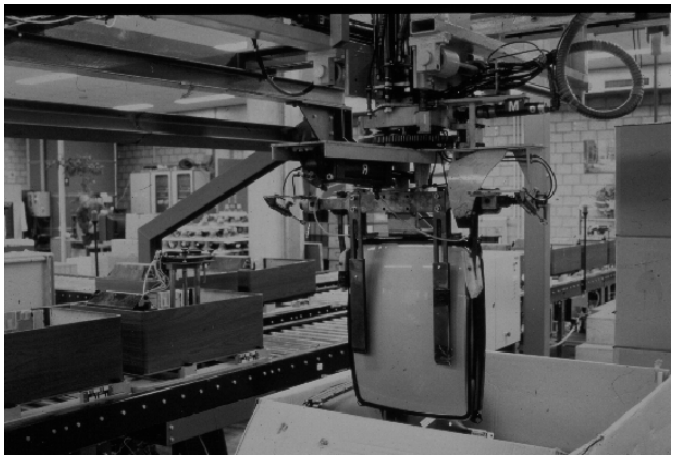

Fig. 1. Television set assembly: an example of a MATS manufacturing system (courtesy of Philips Electronics, Eindhoven, the Netherlands)

Fig. 2. Truck Manufacturing: an example of MTS/ATO (courtesy of Scania Corporation, Zwolle, the Netherlands)

close co-operation with the latter. Only when agreement on the design is reached, the company starts to purchase materials, next manufactures parts and components and finally assembles, tests and installs the product. Highly specialized equipment is typically produced in an ETO mode.

The selection of a particular logistic product/market structure is often based on a trade-off between delivery times on the one hand, and minimum WIP and finished goods inventories on the other hand. Aspects such as the product life cycle, the diversity of the product spectrum and the degree of customization of end-items, as

Fig. 3. Precision Machining of Actuators: an example of MTO production (courtesy of Morskate Machine Factories, Hengelo, the Netherlands)

well as the times needed for procurement, manufacturing and assembly, are key parameters in such a trade-off.

A second classification dimension concerns the internal structure of the manufacturing and assembly system. This structure may differ per department. The three basic structures are (see also Fogarty et al., 1991):

Dedicated (mixed model) flow lines. A classical example is the assembly line designed for a family of television or personal computer sets, or for car assembly. But also manufacturing processes in which products follow a more or less common route along a number of workstations can be set up as flow lines. The primary criteria to set up a dedicated line are a sufficiently large volume and a limited product variety. Also in the (batch) process industry we often find (continuous) flow lines. Machines and equipment in flow lines are often designed for the specific processes.

Job shops. Systems that are typically designed for manufacturing a broad spectrum of products, usually in small quantities, are called job shops. A wide variety of processes can be performed in a job shop; typically for each process one or more universal machines are installed. Often job shops are characterized by a functional layout and are largely process-oriented (as opposed to the product orientation of flow lines).

On site manufacturing. Examples of on site manufacturing include the realization of complex infrastructural works (bridges, tunnels) or the completion of a major industrial facility. These processes are characterized by the fact that all equipment needed to realize the product is transferred to the product's site, instead of the other way around (as in a machine shop). Typically, on site manufacturing relates to the completion of larger projects, usually requiring more engineering and process planning work than in job shop or flow manufacturing.

Fig. 4. Military ship production, an example of ETO (courtesy of Royal Netherlands Navy, Den Helder, the Netherlands)

These three basic structures represent the extremes of an almost continuous spectrum of hybrid structures. For instance, manufacturing cells and self-directed teams are an attempt to combine the diversity of the job shop with the efficiency and short lead times of dedicated flow lines. The trade-off underlying the selection of a manufacturing structure typically depends on efficiency criteria on the one hand and speed (and short feedback loops) as they occur in more dedicated systems on the other hand.

Some authors make a distinction between *materials-oriented* and *capacity-oriented* manufacturing systems. Companies concerned with mass-assembly of items, based on purchased parts and components, usually add limited value to the product and therefore are typically materials oriented. Semi-process and batch process companies (e.g. pharmaceutical companies) on the other hand often use a limited variety of basic materials, from which a large diversity of final products are fabricated, often requiring highly capital-intensive processes. The value added in these manufacturing processes can be substantial.

When combining the various criteria, and including more hybrid structures, we end up with a significant variety of manufacturing system classes. Clearly, some combinations are however more dominant than others are. Dedicated flow lines are often found in large-scale production and therefore typically arise in a Make and

Assemble to Stock environment. Machine shops often operate as job shop in either a Make to Stock/Assemble to Order or a Make to Order environment. Engineer-to-order companies more often work on a project basis but also a job shop structure may occur. As to materials versus capacity orientation, the situation is less clear. A flow line can be both primarily materials oriented (as is the case with an assembly line) or capacity oriented (as is often the case in process industries).

4 A manufacturing planning and control reference architecture

Various authors have presented frameworks of a manufacturing planning and control architecture, almost exclusively hierarchical in nature (e.g. Bertrand et al., 1990; Vollmann et al., 1997; Hopp and Spearman, 1996). The reason for this hierarchy, representing aggregate decisions in an early stage while later disaggregating (similar to the decision structure in Hierarchical Production Planning, for instance), is very natural; it reflects the increasing information that comes to or is gathered by the manufacturing organization as time progresses. Similarly, decomposition can often be naturally applied since many manufacturing organizations show a departmental or modular structure (often called production units, see e.g. Bertrand et al., 1990). However, when it comes to filling in the various modules of the architecture with algorithms or dedicated procedures, the literature is far less complete, and seems to concentrate primarily on mass production in make-to-stock systems. For instance, the role of process planning that is so dominant in make-to-order companies, is almost entirely neglected in the literature. Local decision making is often based on decomposition but how to decompose a decision problem is far from clear, and depends heavily on the underlying production typology. And, as mentioned earlier, uncertainty manifests itself in many different ways, again depending on the production environment. In this section, we sketch a manufacturing planning and control architecture that serves as a reference framework for the algorithmic developments presented in Section 5.

Figure 5 depicts a general architecture for manufacturing planning and control. We have deliberately chosen not to use the MRP terminology, to avoid any possible confusion. This architecture, together with the classification presented in Section 4, will serve as the basis for algorithmic enhancements to be discussed in the next two sections. An explanation of the various terms and modules is given below; in particular we motivate the deviations of this architecture with those appearing elsewhere (e.g. Vollmann et al., 1997).

Product and process design. This is the most important function in particular for many OEM's (Original Equipment Manufacturers). Often a new product range and the required processes are designed simultaneously, in order to prevent highly attractive products that are extremely hard or costly to manufacture. This has motivated the development of Design for Manufacture and Design for Assembly techniques (Boothroyd et al., 1994), as well as a more general approach known as Concurrent Engineering. Alternatively, companies that produce licensed product designs at best pay limited attention to process design.

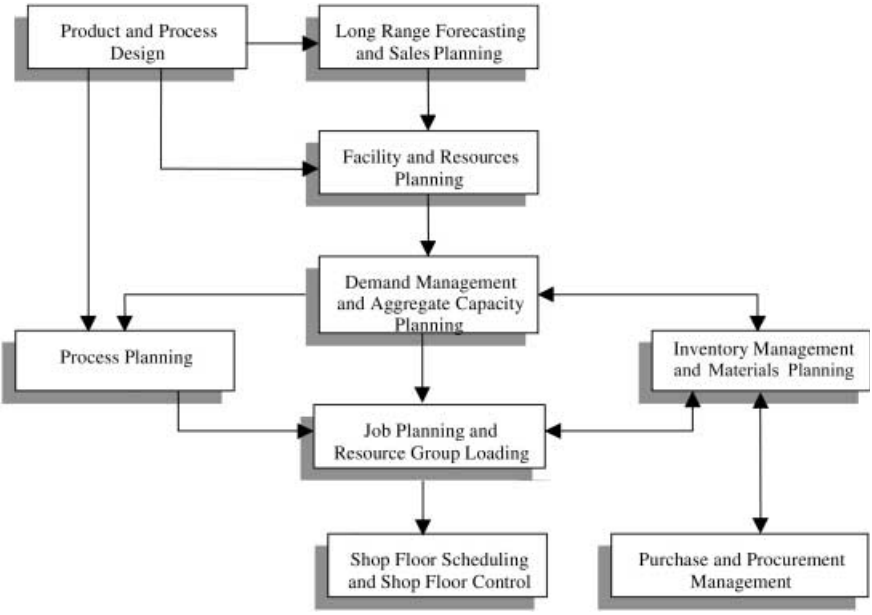

Fig. 5. A manufacturing planning and control architecture

Long range forecasting and sales planning. This function typically deals with long term estimation of a particular market share and, based on that, the planning of sales volumes for particular product ranges on a highly aggregate basis.

Facility and resources planning. Both the technological designs and the commercial planning serve as input to the planning and possible acquisition of the facilities and resources that are needed. Hence, in this phase required resources, including manpower, machines and auxiliary equipment, are specified to enable the planned sales volumes to be realized.

Demand management and aggregate capacity planning. The implementation of this function clearly depends on the logistic product/market structure of the company. In medium to large volume manufacturing of assemble-to-stock products, it encompasses the short term demand forecasting, its translation into prospective orders and finally order acceptance. Aggregate capacity planning in MATS systems involves the synchronization of production requirements with available resource capacities. Also the planning of additional shifts during certain periods and the decision to temporarily subcontract the production of certain components, may be a part of aggregate capacity planning. Next, orders are accepted on a routine basis. In an MTO or ETO environment, the order acceptance function is generally much more complicated, including the specification of functional and technical requirements, quality definitions, a delivery time and a price. Now, the role of aggregate capacity planning as a vital part of order acceptance is in particular to add in quoting realistic customer order due dates. In ETO companies, aggregate capacity planning does not only relate to the manufacturing divisions, but also to the design and engineering

department. In all cases, a clear insight into the relations between the available resource capacities, a possible workload and the resulting manufacturing lead times is essential in order to determine sound inventory policies and to generate realistic customer order delivery times.

Process planning. Surprisingly few authors pay explicit attention to the function of process planning, or more general to technological planning, when designing a MPC architecture. Usually, we distinguish between *macro* and *micro process planning.* Macro process planning concerns the selection of product routings (selection of equipment and resource groups, as well as the sequencing of process steps) and the global estimation of processing times. Micro process planning operates at the more detailed machine level and for instance selects cutting tools, cutting patterns and speed and feed rates, on the basis of which numerical control (NC) programs are generated. Typically, in large volume mass production, process planning is closely linked to process design (also in time), but in small batch manufacturing it is an extremely useful instrument to add flexibility to the shop floor. Using modern CAPP (Computer Aided Process Planning) systems, process specification typically occurs only a couple of days before actual manufacturing in a Make-to-Order company, thereby allowing for the use of alternative process plans, for instance to better balance the load on various machines at the shop floor (Zijm, 1995).

Job planning and resource group loading. Once customer orders or replenishment orders have been accepted and macro process plans have been determined, jobs can be constructed at the resource group level. Basically, a job can be seen as the restriction of an order to a specific department, work cell or resource group (also called production units, see Bertrand et al., 1990). However, several customer order related jobs may be combined into a composite job (batching) or one large job may be split into several smaller jobs, e.g. to balance the load among several work cells or to speed up work (lot splitting). In addition, the availability of parts or components in stock may also alter the lot size of a job (this is called *netting* in the MRP terminology). It is important to notice that *jobs are the operational entities to be controlled at the shop floor*, starting with their release and ending with their formal completion. The *simultaneous* loading of the various resource groups aims at matching the required and available capacity *within each group, by considering effective resource group capacities as well as routing constraints of jobs between the groups, but without specifying in detail routing and precedence constraints of a job within a group.* Planning is based on either customer order delivery dates, or inventory runout-times, and in turn defines internal release and due dates for each separate production job.

Inventory management and materials planning. Inventory management plays an essential role at both an aggregate and detailed level. When smoothing aggregate production plans, inventories naturally arise as temporary (e.g. seasonal) capacity stocks. A second source of inventories is the production in batches, as discussed above. Finally, safety stocks again represent additional inventory, while also safety lead times may lead to additional storage needs. However, inventory management exclusively deals with stocks that are stored in warehouses or storage systems, i.e.

work-in-process stocks at the shop floor are not considered here. It is important to notice that, within this framework, inventory management performs the materials supply function to each department or resource group, and hence represents an essential input to Job Planning and Resource Group Loading, as well as to Purchase and Procurement Management.

Purchase and procurement management. This function takes care of the procurement of all components and materials that are purchased from external suppliers. It receives instructions from inventory management while the allocation of production jobs to time windows naturally depends on the availability of these externally procured materials.

Shop floor scheduling and shop floor control. This is the level where the detailed scheduling of jobs on all workstations in a resource group takes place. The goal is to meet the internal due dates set at the higher production order planning level. Hence, at this level we typically deal with the sequencing of job-operations on individual workstations, but *not* with lot sizing aspects (these have been covered at the Job Planning level already). Shop Floor Control deals with the monitoring and diagnostics of all operations, reporting on quality aspects, and signalling major disruptions that may require a rescheduling or replanning phase.

This concludes the description of a general architecture. Note that, as opposed to most authors, more attention is given to technological planning, next to logistic planning and scheduling. In addition, a hierarchical structure similar to HPP can be recognized in the modules: demand management (order acceptance) and aggregate capacity planning, job planning and resource group loading, and shop floor scheduling. Also note that, in contrast to most MRP-based architectures, *we do not separate material requirements planning and resource group loading*. We believe such a separation is not only artificial, but in fact the source of many problems. To illustrate an alternative, we discuss in the next section a basic algorithmic framework proposed initially by Buzacott (1989) that explicitly integrates these aspects, and hence may serve as a reference for parameter setting.

5 An algorithmic framework for parameter setting in general manufacturing systems

As we have seen in Section 2, several approaches have been developed to relate the internal manufacturing lead times in a resource group to the available resource capacities. In a push framework, these lead times are often based on experience; in a pull framework they can be determined by using a workload control procedure of which the parameters are based upon a Closed Queueing Network analysis, for instance. However, when products have to visit several resource groups sequentially, or when a product is built up from parts that are manufactured at the same time in several parallel resource groups, the question arises how group-based workload norms influence the overall manufacturing lead time, including the time that parts wait to be released to a workcell (due to either the workload control rule or to assembly matching problems). The availability of parts or subassemblies in intermediate

inventory banks may further complicate the question of how quickly a particular customer order can be delivered. In other words: we are seeking for algorithms that predict overall manufacturing lead times (including the just mentioned intra-cell waiting times) as a function of workload norms for the various resource groups. By taking into account intermediate inventory banks, these algorithms should also be able to generate reliable customer order response times.

Below we briefly outline a general algorithmic framework that at least partially responds to the above requirements. The underlying idea for this framework is due to Buzacott (1989), who calls it a Generalized Kanban system, and is discussed to some extent in Chapter 10 of Buzacott and Shanthikumar (1993). Indeed, this framework in principle allows to analyze models that combine the presence of inventory banks between work cells, limited resource capacities and a workload control rule. Approximate queueing network models are applied to determine internal lead times in the work cells, overall manufacturing lead times as well as response times to customer orders by taking into account possible inventories and intra-cell waiting times. Note that the overall manufacturing lead times as well as the response times may vary significantly over time, as a result of both waiting times and varying stocks. However, opposite to MRP practice, these variances manifest themselves primarily *between work cells, much less within the work cells*, and hence can be managed more easily, e.g. by adequate priority setting. The small variances within the work cells in turn can be handled at the shop floor scheduling level. The reader may note that this framework also extends the pull framework to multiple cell systems, where cells may be arranged in a network structure.

Now, we present a simple example and discuss methods to determine lead times and service levels as a function of workload norms and inventory levels (cf. Figure 6).

Consider a Make and Assemble to Stock manufacturing system that works according to a Base-Stock Control policy. To explain its operation as well as the parameter setting in more detail, let's follow the flow of orders and materials through the system. Assume that external demand (each demand requires one product) ar-

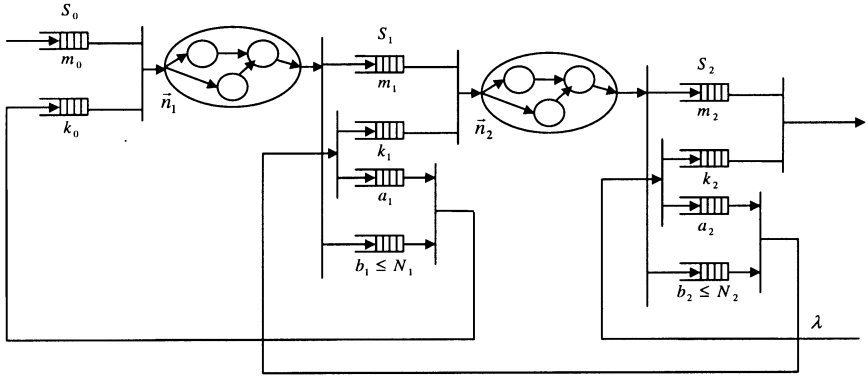

Fig. 6. A two cell serial system with workload control at each cell

rives according to a Poisson process with rate λ. Furthermore assume that both the manufacturing department and the assembly department follow a Base Stock policy, i.e. each time a product or part is required, the system immediately attempts to replenish the stock by assembling a product from its constituting parts, or by fabricating a new part from basic materials or components. The base stock levels of the assembly and fabrication department are equal to S_1 and S_2, respectively. However, the amount of work-in-process in each department is limited to N_1 products and N_2 parts, with $N_i \geq S_i$, *for* $i = 1, 2$. One may for instance assume that the actual production in each department is regulated by Kanbans, however, as soon as a part or product is completed, the attached Kanban is removed and stored in a separate storage, while the product itself is stocked (or delivered if there is a backlog). When an external demand arrives, it is immediately split into two requirements, one for the actual product and one that triggers a stored Kanban to be transferred to the next upstream stockpoint to pick a part and subsequently start the assembly of one product. Note that, when $N_2 = S_2$, we have a classical Kanban system, in which the demand for one product triggers its replenishment as long as the final product stock is not depleted. However, if customers have to wait then in a classical Kanban system also the replenishment production is delayed, since first a product has to be completed before a Kanban becomes available and can be transferred to the preceding stockpoint. However, when $N_2 > S_2$, splitting the external demand into two actions, as described above, allows us to already start the replenishment production even if the final stockpoint is depleted, as long as there are still Kanbans available. The extreme case in which $N_2 = \infty$ represents the classical open base stock system without any workload control, in which any external demand is immediately transferred to all relevant stages. Hence, we conclude that a Generalized Kanban System includes both the classical Kanban system $(N_2 = S_2)$ and the classical Base Stock system with one for one replenishment $(N_2 = \infty)$ as special cases.

So far we have only described the working of a single cell-stockpoint combination, i.e. the assembly cell. It will be clear that the preceding manufacturing cell works in the same way, where the release of a new production order for a part to this cell depends on the requirement of a part from stockpoint 1, the availability of Kanbans at the same stockpoint, and the availability of raw materials at stockpoint 0. Again, it is important to note that the actual availability of a part at stockpoint 1 and the availability of Kanbans at that stockpoint are decoupled if $N_1 > S_1$.

The two-stage system described above is essentially an open system, with workload constraints on each department, cell or resource group. The question is whether it can be analyzed. The answer is yes: at least in an approximate sense. The key to the solution is the observation that each cell in principle operates as a semi-open queueing network, i.e. a network with a limited internal population where jobs not admitted are waiting externally until they are allowed to be released to the cell. Such semi-open queueing networks can be analyzed by using closed queueing networks in combination with a flow-equivalent server approach, see e.g. Dallery (1990), Buzacott and Shanthikumar (1993) and Buitenhek (1998). Extensions to multi-class systems, i.e. systems with different parts or products that are produced

in the same cell, are discussed by Baynat and Dallery (1996) and Buitenhek et al. (1997, 1999); the latter distinguish between a general workload norm simultaneously for all product classes, and class-dependent workload norms. The multi-cell arrangement has been analyzed by DiMascolo et al. (1996) for some special cases, by means of an iteration over the cells. Although the scope of this paper does not allow a detailed analysis, we sketch a basic solution procedure which draws on a combination of ideas from the above references.

Consider the system depicted in Figure 6. If we initially assume that stockpoint 1 (containing intermediate parts) is never depleted, then the assembly cell can be analyzed as a semi-open queueing network, i.e. a network with a population constraint, preceded by a (possibly empty) queue of part/Kanban combinations waiting to be released, and followed by a stock of final products and a Kanban inventory bank. By reversing the role of jobs and Kanbans, we can remodel this system as a *Closed Queueing Network with one synchronization queue*, which can be analyzed by a combination of Mean Value Analysis (MVA) and a Synchronization Queue approach, due to Dallery (1990). As a result, we obtain among other things the probability distribution for the stock and the backlog at stockpoint 2, as well as a (state-dependent) arrival rate of Kanbans at stockpoint 1. The latter is used as input to analyze the preceding manufacturing cell, yielding among other things a probability distribution for the stock and backlog at stockpoint 1. Next, we return to the assembly cell which is now analyzed with two synchronization queues, one for the external demand matching the Kanbans at stockpoint 2, and one for the Kanbans arriving at stockpoint 1 that have to match the parts available at this stockpoint. Again, by reversing the role of pallets and jobs, this system can be analyzed approximately by MVA, leading to state-dependent arrival patterns of Kanbans at stockpoint 1. Next, we return to the manufacturing cell, etc. We iterate until convergence occurs.

The key point of the analysis is that it provides us with both internal lead times and pre-release waiting times at both the manufacturing and assembly cell. Even more, due to the way Mean Value Analysis works, we obtain these lead times and intra-cell waiting times for each possible load, not only the maximum workloads. In addition, we obtain the distribution of the number of parts and final products stored in the intermediate and final storage rooms. Now, it is relatively easy to find parameter combinations (base stock levels, number of Kanbans or workload norms) such that, for given arrival rates, the desired lead times and market service levels arise. Alternatively, one may verify under what arrival rates predetermined response times can be met, thus determining the *effective system capacity* as a function of the desired market response, nominal resource capacities and workload limits. And although the system as a whole clearly operates as an open network, it is not hard to determine for each workcell what throughput can be obtained as a function of the machine and operator capacities and the number of Kanbans, again by analyzing it as a closed queueing network. This throughput in turn determines the *effective capacity of the workcell*.

The above analysis can be extended in many ways. First, as mentioned already, we can handle multiple product and part classes simultaneously in each cell (each

part or product having its own routing), under both general and class-dependent workload restrictions. Note that this includes the case where parallel cells are working independently on parts that next are assembled into the same final product. Another extension concerns the case where divergent structures occur, i.e. when parts are common in multiple final products. Again, in principle these structures can be analyzed, although some care should be taken in cases, where for one product all constituting parts are available while assembly of another product has to wait because of missing parts. With respect to the parts allocation, priority should in such a case be given to the assembly of the first product. The inclusion of set-ups and lotsizing aspects, as well as machine breakdowns can all be handled by including these effects in the processing times, see Hopp and Spearman (1996) for a more detailed treatment. Finally, the models discussed here can also be used in a periodic review planning and control framework, by translating the *effective throughputs* of the workcells, discussed above, to an effective capacity per period. In this way, a periodic resource loading procedure automatically accounts for the many dynamic interactions that take place at the shop floor, since these are explicitly considered when modelling workcells as a queueing network.

The reader may note that the above framework is also general in the sense that it includes MATS, ATO/MTS and MTO systems. The MATS system has been treated, the ATO/MTO system follows by defining the final product base stock levels equal to zero (hence $S_2 = 0$ in Figure 2), while the MTO follows from setting both the final and the intermediate stock levels to zero ($S_1 = S_2 = 0$ in Figure 2). The ETO case is harder to deal with, essentially since capacity in an engineering department cannot just be modeled as a job shop or assembly line, while also the outcome of the work by engineering determines the parameters (routings and processing times for instance) in the subsequent departments. This is a topic for further research (see also the last section).

The algorithmic framework discussed in this section will play an essential role when discussing how to fill in the planning and control architecture for each of the production typologies discussed earlier. This will be the topic of the next section.

6 Algorithms for manufacturing planning and control

In this section, we propose methods to fill in the general framework discussed in Section 4. Depending on the underlying production system typology, different procedures or algorithms may appear to be useful. Clearly, any eventual selection of a method will depend on more detailed structural properties of the underlying manufacturing system. Hence, what follows are proposals for methods that are generally believed to be useful but by no means provide the ultimate answer. Below, we discuss methods for all modules defined in the general architecture, and relate them to the different system typologies discussed in Section 3.

Product and process design

In an MATS system, the impact of a sound product and process design is lasting longer than in almost any other system. Hence, for this reason *Design for Assembly* and *Design for Manufacture* methods deserve special attention (*cf.* Boothroyd et al., 1994). The logistic consequences of a particular product design are still often neglected, although it is obvious that a high degree of modularity (common parts in different products) and standardization may result in important inventory savings and increased flexibility (design for postponement). A similar observation holds for MTS/ATO systems. In fact, one might even turn the argument around. It only makes sense to produce to stock if two basic conditions are fulfilled: a rather predictable demand, *and a small or at most moderate product variety*. In any case, a sound product design should reflect the logistic product/market structure of a company.

It is well-known that the far majority of costs made during manufacturing are determined by decisions made during the product and process design phase (Ulrich and Eppinger, 1995). A general guideline underlying for example *Concurrent Engineering* is to postpone decisions as long as possible, in order to increase flexibility during the latter process planning and capacity allocation processes. Important decisions underlying the process selection are related to scale. If a product family is expected to run for a sufficiently long time in large volumes, it makes sense to install dedicated equipment, e.g. a specialized assembly line. When volumes are only moderate, the product mix increases or product life cycles are relatively short, it is sensible to invest in more universal equipment.

An Engineer to Order company represents one end of this continuous spectrum. Resources are almost always universal. The time needed for engineering may represent a significant part of the total lead time. Companies that are able to rapidly introduce new products and hence to reduce the time to market have a significant competitive advantage (Blackburn, 1991; Suri, 1998). In order to limit this time, stochastic project planning and scheduling methods are often exploited (e.g. PERT). Other, more technically oriented methods to quickly evaluate a proposed design, include rapid prototyping techniques (see e.g. Kalpakjian, 1992). A very important development concerns feature-based design; here features represent basic physical elements, i.e. a combination of material, physical shapes and tolerance measures. Various CAD systems exploit elementary and compound features as their main building blocks. For an overview, see Kalpakjian (1992) or Kusiak (1990). Within ETO systems, design is closely related to technological process planning which closes the gap between design and actual manufacturing. We return to process planning later.

Long range forecasting and sales planning

Long range forecasting aims at the prediction of a market as a whole and is partly a method of expert judgement. A qualitative method to estimate future market volumes is e.g. the *Delphi Method*. With respect to quantitative forecasting methods,

we distinguish between *causal models* and *time series models*. For the prediction of market volumes in the long run, causal models are usually exploited. In particular, *(multiple) regression models* are often used, based upon earlier observed relations between a number of identified causal factors and the realized sales volume *for similar products* (see Makridakis et al., 1998, for an overview). Sales planning relates to the estimated market share that a company expects to achieve (or can handle in terms of resource capacity) and is based on the market analysis already mentioned, as well as an assessment of the power of competitors. Also, price setting constitutes an important instrument in gaining a particular market share; here again, a well-thought and quickly manufacturable product design may result in a significant competitive advantage.

Medium term sales planning becomes more easy in Make to Order and Engineer to Order companies, simply because contracts with customers often cover longer time periods. Also, many MTO and ETO companies gain significant additional incomes from after-sales service contracts. In the long run, sales in an MTO or ETO environment heavily depends on specific customer relations for which general forecasting methods have less value. Primary contract winners often are those companies that are also frontrunners in technological process design.

Facility and resources planning

With respect to the planning of facilities and resources, the main criteria are often throughput, lead time, quality and costs. With respect to costs, not only the cost of equipment and personnel, but also the capital tied up in inventories should be accounted for. At this level, closed queueing networks are often proposed as a method to evaluate the impact of alternative equipment and an alternative layout on the overall performance (see e.g. Suri et al., 1993). For an assembly department, one may for instance choose between several dedicated assembly lines or one mixed model assembly line that requires additional set-up's. Also, Assembly Line Balancing techniques may often help to design such lines in order to maximize its productivity (throughput). With respect to parts manufacturing department(s), Closed Queueing Network analysis again appears to be highly useful, both in a MTS and MTO environment. In the latter, CQN models have to be applied at a slightly more aggregate level, to represent the uncertainty of the actual product mix. In these systems, often a job shop and sometimes a cellular manufacturing structure applies (depending on volume and/or mix).

When a functional departmental structure has been selected, a variety of models is available to support the subsequent layout planning. These models primarily focus on the minimization of the costs of the material flow between various departments (e.g. Francis et al., 1992; Tompkins et al., 1996). However, in a functional layout, transportation times and batching may often be significant and hence induce high work-in-process inventories and overall manufacturing lead times. For that reason, many authors are advocating a more product-focussed layout, of which a cellular manufacturing system is a profound example (Wemmerlov and Hyer,

1989). Burbidge (1975) has been a promotor of the application of production flow analysis as a means to arrive at a more group-technological process layout.

Demand management and aggregate capacity planning

Demand management for MATS systems in the first place relates to time-based forecasting of which the various versions of exponential smoothing are by far the most widespread methods. Trends and seasonal fluctuations are easily included in such methods. For a detailed account on time-based forecasting methods the reader is referred to Box and Jenkins (1970), see also Makridakis et al. (1998). The next step is to translate these forecasts into an aggregate capacity plan, taking into account manufacturing lead times. To determine these lead times, as well as the effective capacities of resource groups, we advocate the use of models such as discussed in Section 5. We stress the importance of the use of *effective capacities*, instead of nominal available time, to capture the effects of dynamic interactions of jobs within resource groups without considering them in detail (after all, we are still planning at an aggregate level). Using these effective capacities, Linear Programming methods are now an excellent tool to smooth a production plan as far as needed, and to further investigate the temporary use of additional capacity (overtime work, hiring temporary personnel, subcontracting). The goal usually is to minimize the sum of the total costs of non-regular capacity and the inventory holding costs needed to match available capacity and perceived demand. Hence, a close link exists with aggregate inventory management. Hopp and Spearman (1996) provide a brief overview of LP-models for aggregate capacity planning (although they restrict themselves to the use of nominal capacities). Finally, the actual generation of replenishment orders is often based on inventory management policies such as base stock or fixed order quantity policies (using either fixed or random lead times). For an overview, the reader is referred to Silver et al. (1998). These models however do not consider capacity limitations; in order to include the latter, the use of models such as discussed in the preceding section is advocated.

With respect to order-based production, demand management is primarily related to customer order management. Major concerns are due date and cost management, where due date refers to the order delivery time. In an ATO/MTS environment, this due date is determined by the assembly lead time and the available parts inventories. In an MTO system, and even more in an ETO system, a significant time has to be reserved for technical order specification, engineering and process planning activities. Since usually time does not permit a detailed engineering and process planning phase before order acceptance, management has to rely on a rough estimation of the impact of such an order on resource utilization, and eventually has to adjust capacity. Again, linear programming models can be used here. The use of effective capacities instead of nominal capacities however is questionable in MTO systems, in particular since the steady state analysis of queueing models is often not applicable. When already at an aggregate level complex precedence relations between different job groups within an order can be distinguished, it makes sense to incorporate these relations in the aggregate capacity planning procedure. To this

end, integer programming models based on column generation and Lagrange relaxation, have been developed. For details, the reader is referred to Hans et al. (1999).

Process planning

Process Planning is often said to represent the slash between CAD and CAM and indeed it specifies all the technical information needed before a production order, a job or part of a job can be executed. Usually we distinguish between *macro and micro process planning*, where the first concerns all decisions at a shop level, while the latter deals with the detailed machine and tool instructions. The way in which a production order is split into a number of potential production jobs, to be loaded on the various resource groups, is a macro-process planning decision, and the same holds for the specification of resource requirements (operators, machines, auxiliary equipment) and product routings. The determination of cutting patterns, and of tool speed and feed rates at machines are micro process planning decisions, leading for instance to the generation of Numerical Control programs.

Within an MATS environment, process planning is already performed during the process design phase and hence hardly plays a separate role. The same holds for an MTS/ATO system but for MTO and ETO environments the situation becomes quite different. In order to speed up the process planning activities, databases are needed of possible processes, machining methods and tool combinations, from which a selection can be made after which a CAPP system automatically generates the NC programs. These CAPP systems in turn are often based on the use of *process planning features* (not to be confused with the design features discussed earlier) that specify basic material processing patterns (e.g. bending, material removal, welding patterns). The combination of many process planning features yields a complete machine instruction (see e.g. Kusiak, 1990).

It is important to realize that in principle much freedom exists in the selection of machining methods and hence of routings and process plans. Currently, having more advanced CAPP systems available, process planning in many metal working factories is performed only a couple of days (and sometimes less) before actually processing a job. Consequently, it makes sense to take into account the actual work load on the shop floor when developing process plans for a new order, for instance with the aim to balance the load among various workstations. One way to significantly increase the loading flexibility on the shop floor is by developing several alternative job routings (Zijm, 1995).

Job planning and resource group loading

As mentioned earlier, a job is in principle the restriction of either a replenishment order or a customer order to a specific resource group. However, its size may be adapted by either combining jobs of the same product (*batching*), in order to save

set-up times, or by dividing a large batch into smaller lots (*splitting*), to accelerate the flow through the system. Another important aspect concerns the *netting* procedure, where gross requirements are translated into net requirements by taking into account the available inventories at different levels. With respect to batching and set-up's, some further remarks are in place. Final assembly lines are often dedicated to particular product families and require at best minor set-up times. Within parts manufacturing however, a trade-off between the loss of time due to set-up's, and the amount of parts inventories should be made. We emphasize the need to consider set-up *times*, instead of costs, because it is time, related to the actual utilization of resources (workers and equipment), that counts. For that reason, set-up costs are generally less relevant in discrete manufacturing (in process industries the situation may be different because there a set-up may induce a significant loss of material or require a significant start-up time). Hence, EOQ models and their time-dependent variants such as the Wagner and Whitin algorithm or the Silver and Meal heuristic (Silver et al., 1998) are considered less appropriate in discrete manufacturing. An alternative is provided by Karmarkar (1987) who studies the trade-off between time lost to set-up's, resource utilization, and the resulting internal lead times.

In a MTS/ATO environment, jobs usually concern the manufacturing of parts, i.e. the stock-based production and hence the above remarks also apply here. With respect to MTO and ETO systems, jobs are often directly derived from customer orders. The availability of raw materials and purchased parts naturally has to be ensured again by inventory management.

Once the various production jobs have been determined, these jobs have to be loaded to the corresponding resource groups. Recall that *jobs are the operational entities to be controlled at the shop floor.* The usual way these groups are loaded in a MATS environment is by using the MRP-based time-phasing procedure, using fixed off-set lead times (see Section 2). As long as the load can be kept relatively stable, this is a reasonable procedure but the question arises what are realistic lead times. Here, the models developed by Buzacott (1989) and Buzacott and Shanthikumar (1993) discussed in the preceding section can play a key role, since these models and their extensions (see Section 5) *essentially combine capacity loading and material control in one integral system.* In particular, when observing the inventory position (possibly a backlog) in front of each resource group, and knowing its effective capacity (throughput) and internal lead time, it is easy to establish state-dependent release and due dates for jobs at these resource groups. By adjusting the base stock parameters, the same models can be used for job planning and resource group loading in a MTS/ATO and in a MTO system.

Inventory management and materials planning

Inventories basically arise from the impossibility to synchronize all phases in a productive system, due to reasons of economies of scale. Although often considered as a secondary management function, inventory management is therefore in many companies closely tied to most tactical and operational planning and control functions (see also Figure 6), and hence at the heart of a logistic control system.

This holds even more for a Make and Assemble to Stock environment where inventory management not only controls internal stocks but also serves to match internal production capacities with external demand. Various aspects of inventory planning and control have already been mentioned when discussing other planning and control modules. Production smoothing requires capacity inventories; lot sizing in both production and in purchasing (see below) results in batch-related inventories. But also in MTS/ATO and in MTO systems, inventory management plays a key role in the control of parts inventories and raw materials stocks. The key contribution of MRP has undoubtedly been the recognition of dependent demand for which classical statistical inventory control procedures are not in place. On the other hand, buffer stocks are still advisable as long as process uncertainties (product or process quality problems) occur (see also Whybark and Williams, 1976). Periodic review planning and control policies constitute another source of batch-related inventories (see Silver et al., 1998). The tight interplay between resource capacities, inventories and lead times is clearly demonstrated at an aggregate level by the Linear Programming models discussed earlier, and at the job planning level through the generalized Kanban model of Buzacott (1989) and its extensions, discussed in Section 5.

Shop floor scheduling and shop floor control

Much research has been devoted to the concept of Shop Floor Scheduling and Shop Floor Control. In our framework, this function is executed within the resource groups. In assembly departments in a MATS environment, shop floor scheduling usually boils down to a simple input-output control system and hence primarily deals with sequencing the entire jobs on a single resource (e.g. an assembly line). In MTO environments, where production is performed with more universal equipment, we often are confronted with job shop scheduling problems. In a job shop, with many different jobs each having their own routing, jobs have to be sequenced on all individual workstations, thereby dramatically increasing the number of possible schedules. A variety of different policies have been developed to deal with complex job shop problems (see e.g. Pinedo and Chao, 1999; Morton and Pentico, 1993). In particular in MTO manufacturing systems, a primary goal is often to meet due dates set on the higher production job planning and resource group loading level. Adams et al. (1988) developed the Shifting Bottleneck Procedure that has inspired a lot of research into more practical oriented shop floor scheduling methods (see e.g. Schutten (1998). In some cases, this research has led to the development of commercial shop floor scheduling systems (Meester et al., 1999).

Purchase and procurement management

Purchase and procurement management deals with aspects of vendor selection on a strategic level, the definition of purchase contracts on a mid-term tactical level, and the determination of batch sizes on the operational level. Here the trade-off between fixed costs related to actual procurement (administration, transport, receiving) and

the inventory holding costs come into play, hence at this place Economic Order Quantities and their time-dependent equivalents can be used (Silver et al., 1998). A way to diminish these fixed costs is to establish framework contracts with suppliers over a longer period, specifying minimum and maximum purchase quantities, as well as reliable delivery times. Just-in-time delivery agreements are almost always based on such framework contracts.

This concludes the discussion on algorithms and methods that can be used to complete the framework architecture discussed earlier. In the next section, we draw conclusions and suggest directions for future research.

7 Conclusions and directions for future research

In this paper, we have presented a critical review on existing planning and control systems, in particular MRP-based systems, Just in Time and Workload Control procedures and Hierarchical Production Planning. We have argued the need for a system that integrally considers material and capacity constraints and combined a general architecture with more intelligent procedures and algorithms. Also, a better integration of engineering and technological planning issues has been discussed. In addition, we have defined such a planning and control architecture. Subsequently, the algorithmic framework developed by Buzacott has been discussed in some detail, because this framework explicitly considers the above required integration of capacity and material constraints. Finally, possible methods and algorithms for the various modules in the architecture have been suggested.

The continuous time planning models, discussed in Section 5, can be used at a strategic level to study the relations between resource capacities and lead times, at a tactical level to define customer order lead times, and at an operational level to release jobs to the shop floor (using the pull control mechanism). The extension of these models to general networks of resource groups however requires further investigation. In particular the combination of divergent multi-echelon models and capacitated semi-open queueing networks seems to be promising.

Various functions in manufacturing planning and control systems still operate on a periodic (e.g. a weekly) basis. For instance, capacity adjustment procedures based upon linear programming models normally use time buckets as a basis for planning. We advocate the use of effective resource group capacities (based on a throughput analysis of workload controlled systems) within LP models, to define realistic weekly capacity profiles that capture dynamic effects as they occur on the shop floor. This is a subject for further research.

Modelling the work flow of engineering departments and the relations between engineering and manufacturing also requires substantial further research. In addition, the integration of aggregate planning, process planning and job planning deserves further attention. In particular the gradually increasing amount of information from process planning has not been modeled explicitly so far. Still, this is at the heart of hierarchical planning. Typically, aggregate capacity planning considers product families that share the same resources while only later the disaggregation

to individual items should be made. Again, we believe that the stochastic models discussed in Section 5 apply, since these models can be used at different levels of aggregation, by adjusting the stochastic parameters.

References

Adams J, Balas E, Zawack D (1988) The Shifting Bottleneck procedure for job shop scheduling. Management Science 34: 391–401

Baynat B, Dallery Y (1996) Approximate techniques for general closed queueing networks with subnetworks having population constraints. European Journal of Operational Research 69: 250–264

Bechte W (1987) Theory and practice of load-oriented manufacturing control. International Journal of Production Control 26: 375–396

Bertrand JWM, Wortmann JC, Wijngaard J (1990) Production control: a structural and design oriented approach. Elsevier, Amsterdam

Bitran GR, Haas EA, Hax AC (1982) Hierarchical production planning: a two stage system. Operations Research 30: 232–251

Blackburn JD (1991) Time-based competition: the next battleground in American Manufacturing. Irwin, Homewood, IL

Boothroyd G, Dewhurst W, Knight W (1994) Product design for manufacture and assembly. Marcel Dekker, New York

Box GEP, Jenkins GM (1970) Time series analysis, forecasting and control. Holden-Day, San Francisco

Buitenhek R (1998) Performance evaluation of dual resource manufacturing systems. Ph D Thesis, University of Twente, The Netherlands

Buitenhek R, van Houtum GJ, Zijm WHM (1999) An open queueing model for flexible manufacturing systems with multiple part types and general purpose pallets. In: Ashayeri J, Sullivan WG, Ahmad MM (eds) Flexible automation and intelligent manufacturing. pp 713–734. Begell House, New York

Buitenhek R, van Houtum GJ, Zijm WHM (1997) Approximate MVA algorithms for open queueing networks with population constraints. Annals of Operations Research (in press)

Burbidge JL (1990) Production control: a universal conceptual framework. Production Planning and Control 1 (1): 3–16

Buzacott JA (1989) Generalized Kanban/MRP systems. Technical report. Department of Management Sciences, University of Waterloo

Buzacott JA, Shanthikumar JG (1993) Stochastic models of manufacturing systems. Prentice-Hall, Englewood Cliffs, NJ

Clark AJ, Scarf H (1960) Optimal policies for a multi-echelon inventory problem. Management Science 6: 475–490

Dallery Y (1990) Approximate analysis of general open queueuing networks with restricted capacity. Performance Evaluation 11: 209–222

Deming WE (1982) Quality, productivity and competitive position. MIT Center for Advanced Engineering Study, Cambridge, MA

Diks EB, de Kok AG, Lagodimos AG (1996) Multi-echelon systems: A service measure perspective. European Journal of Operational Research 95: 241–263

Di Mascolo M, Frein Y, Dallery Y (1996) An analytical method for performance evaluation of Kanban-controlled production systems. Operations Research 44 (1): 50–64

Eppen G, Schrage L (1981) Centralized ordering policies in a multi-warehouse system with lead times and random demand. In: Schwarz L (ed) Multi-level production/inventory

control systems: theory and practice (Studies in the Management Sciences, Vol 16), pp. 51–67. North-Holland, Amsterdam

Federgruen A, Zipkin P (1984) Approximations of dynamic, multi-location production and inventory problems. Management Science 30: 69–84

Fogarty DW, Blackstone JH, Hoffmann TR (1991) Production and inventory management, 2nd edn. South-Western Publishing Company, Cincinnati

Francis RL, McGinnis (Jr) LF, White JA (1992) Facility layout and location: an analytical approach, 2nd edn. Prentice Hall, Englewood Cliffs, NJ

Günther H-O (1986) The design of a hierarchical model for production planning and scheduling. In: Axsäter S, Schneeweiss C, Silver EA (eds) Multi-stage production planning and inventory control, pp 227–260, Springer, Berlin Heidelberg New York

Hammer M, Champy J (1993) Reengineering the Corporation. HarperCollins, New York

Hans EW, Gademann AJRM, van de Velde SL, Zijm WHM (1999) Capacity loading by column generation techniques. BETA working paper, Eindhoven University of Technology, Faculty of Technology Management

Hax AC, Candea D (1984) Production and inventory management. Prentice-Hall, Englewood Cliffs, NJ

Hax AC, Meal HC (1975) Hierarchical integration of production planning and scheduling. In: Geisler MA (ed) Logistics (Studies in the Management Sciences, Vol 1). Elsevier, North-Holland

Hopp WJ, Spearman ML (1996) Factory physics: foundations of manufacturing management. Irwin, Homewood, IL

van Houtum GJ, Inderfurth K, Zijm WHM (1996) Materials coordination in stochastic multi-echelon systems. European Journal of Operational Research 95: 1–23

Kalpakjian S (1992) Manufacturing engineering and technology, 2nd edn. Addison-Wesley, Reading, MA

Karmarkar US (1987) Lot sizes, lead times and in-process inventories. Management Science 33 (3): 409–423

Kusiak A (1990) Intelligent manufacturing systems. Prentice-Hall, Englewood Cliffs, NJ

de Kok AG (1990, Hierarchical production planning for consumer goods. European Journal of Operational Research 45: 55–69

de Kok AG (1999) Demand availability management. Forthcoming

Langenhoff LJG, Zijm WHM (1990) An analytical theory of multi-echelon production/distribution systems. Statistica Neerlandica 44 (3): 149–174

Makridakis S, Wheelwright SC, Hyndman RJ (1998) Forecasting: methods and applications. 3rd edn. Wiley, New York

Meal HC, Wachter MH, Whybark DC (1987) Material requirements planning in hierarchical production planning systems. International Journal of Production Research 25 (7): 947–956

Meester GJ, Schutten JMJ, van de Velde SL, Zijm WHM (1999) Shop floor scheduling in discrete parts manufacturing. In: Villa A, Brandimarte P (eds) Modeling manufacturing systems: from aggregate planning to real-time control. Springer, Berlin Heidelberg New York

Monden Y (1998) Toyota production system: an integrated approach to just-in-time, 3rd edn. Engineering Management Press, Norcross, GA

Morton TE, Pentico DW (1993) Heuristic scheduling systems: with applications to production systems and project management. Wiley, New York

Orlicky J (1975) Material requirements planning: the new way of life in production and inventory management. McGraw-Hill, New York

Pinedo M, Chao X (1999) Operations scheduling with applications to manufacturing and services. Irwin/McGraw-Hill, Boston

Rosling K (1989) Optimal inventory policies for assembly systems under random demands. Operations Research 37 (4): 565–579

Schonberger RJ (1982) Japanese manufacturing techniques: nine hidden lessons in simplicity. The Free Press, New York

Schutten JMJ (1998) Practical job shop scheduling. Annals of Operations Research 83: 161–177

Silver EA, Pyke DF, Peterson R (1998) Inventory management and production planning and scheduling. Wiley, New York

Skinner W (1985) Manufacturing: the formidable competitive weapon. Wiley, New York

Spearman ML, Woodruff DL, Hopp WJ (1989) CONWIP: a pull alternative to Kanban. International Journal of Production Research 28 (5): 879–894

Suri R (1998) Quick response manufacturing: a companywide approach to reducing lead times. Productivity Press, Portland

Suri R, Sanders JL, Kamanth M (1993) Performance evaluation of production networks. In: Graves SC, Rinnooy Kan AHG, Zipkin PH (eds) Logistics of production and inventory (Handbooks in operations research and management science, Vol 4). North- Holland, New York

Tayur S, Ganeshan R, Magazine M (eds) (1999) Quantitative models for supply chain management. Kluwer, Boston

Tompkins JA, White JA, Bozer YA, Frazelle EH, Tanchoco JMA, Trevino J (1996) Facilities planning, 2nd edn. Wiley, New York

Ulrich KT, Eppinger SD (1995) Product design and development. McGraw-Hill, New York

Vollmann TE, Berry WL, Whybark DC (1997) Manufacturing planning and control systems, 4th edn. Irwin/McGraw-Hill, New York

Wemmerlov U, Hyer NL (1989) Cellular manufacturing in the US industry: a survey of users. International Journal of Production Research 27: 1511–1530

Whybark DC, Williams JG (1976) Material requirements planning under uncertainty. Decision sciences 7 (4): 595–606

Wiendahl H-P (1993) Load-oriented manufacturing control. Springer, Berlin Heidelberg New York

Wight O (1981) MRP II: unlocking america's productivity potential. CBI Publishing, Boston

Zijm WHM (1992) Hierarchical production planning and multi-echelon inventory management. International Journal of Production Economics 26: 257–264

Zijm WHM (1995) The integration of process planning and shop floor scheduling in small batch part manufacturing. Annals of the CIRP 44/1: 429–432