Planning of Combined Make-to-Order and Make-to-Stock Production

Beemsterboer, Bartholomeus Jacobus

Publication date:
2016

Chapter 1

Introduction

Production companies face the daily challenge of producing at minimal cost, while they are expected to achieve a high degree of delivery reliability. At the same time, companies have to deal with uncertain demand. This is why the planning of production is in the constant attention of firms and has been a major research area for the past decades as well.

1.1 Make-to-stock production

One approach to deal with stochastic demand is to produce in advance, so that customers can be served instantly. Once a customer places an order, the product is supplied from inventory. This way of organizing production is known as *make-to-stock* (MTS) production.

A common way of organizing MTS production is to replenish stock when the inventory level of a product is below a certain level. Once this occurs, a replenishment order is sent to the production department, which then initiates production of additional units of the product so that the firm does not run out of inventory. Logically, the production takes some time so the units are not available immediately. Hence, replenishment orders should ideally be placed at a time when there is sufficient inventory available to satisfy the demand during the time the replenishment order is still under production. But as future demand is not known exactly, how should the firm determine when to reorder an item at the production department? If we reorder a product when there is still much inventory available, it is stored in inventory long before it is sold, and all the time in between the firm has to cover the investments with its own capital. This is reflected by the *holding costs*, which are the costs of holding inventory of products. On the other hand, if we reorder a product when we expect to have just enough products in inventory to satisfy the demand until the reordered items become available, we might very well run out of stock in case more units are demanded than we expect.

Moreover, most machines cannot start working on a different product right away. Instead, machine settings have to be changed; the machines have to be *set up*. Hence, switching production from one product to another takes time. This reduces the production capacity and should therefore not happen too often. This is why production companies generally produce in batches. However, while large batches might save production capacity, they also take longer to finish so that demand has to be supplied from the remaining inventory for a larger amount of time. Larger batches also occupy the machines for longer periods, reducing the ability of the firm to react quickly to eventual unforeseen high demand for other products.

Companies face many more challenges than just these when they apply MTS production. These and other challenges concerning MTS production are studied within the field of Inventory Control. This field was established more than a century ago by Harris (1913), who showed companies how to determine how many parts to make at once. His result is known as the *Economic Order Quantity*. It forms the basis of what is known as the (r, Q) policy, which states that a fixed amount of Q items is ordered when the inventory level reaches the reorder point given by r. An alternative but similar policy is given by the (s, S) policy. This model also works with a reorder point, here referred to as s, but instead of ordering a fixed amount, it orders the difference between the actual inventory level and a maximum level given by S.

If we take into account that replenishments are not immediately available, the stochasticity of the demand comes into play. If we have the possibility that the demand during the lead time is larger than the amount of products we left in inventory, which might run out of stock, in which case new customers cannot be served. If they are willing to wait for the product, it is *backordered*, but this reduces the satisfaction of the customers and therefore damages the competitiveness of the firm. If customers are not willing to wait and prefer to buy the product at another firm, we incur *lost sales*. As both options are unwanted, companies generally hold more products in inventory than the exact amount of demand that is expected to arrive during the lead time, and the difference is called the *safety stock*. We can distinguish two main approaches for obtaining these safety stocks. One approach considers *service levels* and determines optimal order quantities under the condition that the probability of a stock-out is not larger than a certain level. The other approach takes the actual costs of a stock-out (backorder and/or lost sales costs) into account. The essence of both approaches is however the same, as a higher stock-out cost results in a higher optimal service level.

The above approaches support firms in making decisions on the production quantities. They also provide guidance in deciding when to release replenishment orders of products at their production facilities. However, almost every firm offers more than one product to customers, and firms often produce multiple products on the same production facility. Hence, replenishment orders of products compete for the capacity. If the firm ignores this, when planning its production, it may very well occur that multiple products are reordered simultaneously so that at least some of them are completed later than anticipated. By contrast, if no replenishments are ordered for a while, the production capacity stands unutilized during that period. Hence, in order to obtain a model suitable for planning production of make-to-stock production companies, it is necessary to take the presence of multiple products into account. This is taken care of by the *Economic Lot Scheduling Problem*. It includes both decisions on production quantities and timing decisions, the latter in terms of a sequence. If in addition the stochasticity of demand is taken into account, the problem is extended to what is known as the *Stochastic Economic Lot Scheduling Problem* (SELSP). The solution procedures developed for this problem offer production companies guidance in organizing their production in a sufficiently realistic setting. SELSP solution procedures help companies select which product is produced next after the capacity becomes available and how many units of that product should be made. With this, the procedures account for the (uncertain) demand until the replenishment is delivered and for the demand for other products as well.

This overview is just a glimpse of the extensive research in the field of inventory control. We refer the reader to Axsäter (2007) for an excellent overview of many models, methods and theory in this field, including those discussed above.

1.2 Make-to-order production

Another approach to deal with stochastic demand that avoids the need for inventories is to agree on a delivery time window with customers, i.e. to deliver before a *due date*. This allows the company to manufacture products after the moment of ordering, so that it is not required to invest in production long before the revenue arrives. It also avoids the risk of obsolescence. In addition, this approach allows the firm to utilize its capacity efficiently. Machines do not have to be left idle in order to be able to respond immediately to any possible future demand. Instead, incoming orders may wait until sufficient capacity is freed at the production department. However, customers are generally not prepared to wait for a long period of time, so offering long lead times would damage the competitiveness of the firm. Many firms are therefore bound to deliver before tight due dates, imposing challenges to the production planning. This approach of organizing production is known as *make-to-order* (MTO) production.

An important advantage of applying MTO production is that a firm may allow customers to *customize* their products. For instance, a firm could offer customers to select the color, size or product modules. Such customizability is seen most often in high variety production systems, which are characterized by multiple workstations that perform operations in order to produce a high variety of products. These products generally do not undergo the same operations in the same sequence. Instead, the routings of the product may vary substantially. Such production systems are known as *job shops*.

When firms apply MTO production, the production planning comes with multiple challenges. First of all, several approaches for the production planning organization exist. One possibility is to make a detailed production schedule, by determining starting (and finishing) times for each of the operations of the jobs at all workstations mathematically. The research field that deals with this approach is called *scheduling*. Another approach is to determine production decisions locally, based on priority rules. We will discuss each of these approaches in more detail.

The field of scheduling formulates planning problems, such as those of production companies, as combinatorial optimization problems. These require definition of fixed sets of jobs and machines, and fixed sequences of operations and due dates for each job. The aim of the optimization problem is to obtain a feasible schedule with respect to a preselected objective, for instance minimizing the amount of tardiness. Although the formulation of such problems appears rather simple, solving them to optimality has been shown to be rather hard in many cases. The field of scheduling derives solution methods for a rich variety of these combinatorial optimization problems. Many studies in this field conclude that realistically modeled and sized scheduling problems cannot be solved to optimality within an acceptable amount of time. For more information on scheduling theory, we refer the reader to Pinedo (2012).

The approach with local decisions works differently. Each job is sent to the production department (for job shops also referred to as the *shop floor*) and goes to the first workstation in its routing. If this workstation is idle, it starts working on the job. If not, the job waits for processing in the queue of the workstation. If the workstation finishes a job, it proceeds to the next operation in its routing and a new job is selected from the queue of the workstation. This approach requires selecting the rule according to which it is determined which of the jobs is selected, called the *dispatching rule*. A simple rule is to process the jobs in the queue of a workstation in the order of arrival, i.e. on a First Come, First Served (FCFS) basis. Another simple rule is to select the job with the Shortest Processing Time (SPT). Selecting the job with the Earliest Due Date (EDD) is a method that is in many cases more closely related to the production objective. A more advanced approach is to assign fictituous due dates to each of the operations in the routing of a job, and, when a workstation becomes available, select the job with the earliest due date for the concerning operation. This method, known as Operation Due Dates (ODD), hence allows accounting for the remaining processing (and waiting) times that jobs incur after the concerning workstation. Besides these four, many more dispatching rules and methods have been developed. An extensive evaluation of over 100 of such dispatching rules is provided by Panwalkar & Iskander (1977).

Workload Control is a more sophisticated planning concept. Besides working with machine queues and dispatching, the concept uses a 'queue' for the entire production system known as the *order pool*. Incoming orders are held in the pool until workload limits on the shop floor permit them to be *released*. Although letting orders wait may seem counterintuitive when the objective is to complete as many orders on time as possible, the concept is actually able to reduce the number of jobs that finish late by taking advantage of central load balancing. See Land (2004) for details.

1.3 Which strategy to apply?

Having two production strategies at its disposal, how does a production company choose which of them to apply? Selecting the right strategy is not always an easy task, as MTO and MTS production each have their own advantages and requirements.

Producing to stock avoids letting customers wait until the product undergoes the production process of the company, which is especially beneficial when this process is long and/or hard to control. Producing to stock then makes these products more attractive for customers as these generally prefer short and certain lead times. If by contrast a production company can produce items fast and reliably, the company would be not so much hindered in producing in an MTO fashion. It also matters in which market a firm operates. If competitors are able to respond to demands instantly by applying MTS, a firm would lose many customers if it produces in an MTO fashion.

Producing to stock also involves a holding cost, as pointed out above, as the company has to pre-invest in the production of the product. Logically, the value of the investment and the length of the period between costs and revenues determine how much holding costs a company should account for. These depend on the value of the products and on the demand, which therefore influence the decision whether to make to stock or make to order as well. For products that are relatively cheap to produce and that are regularly demanded, producing beforehand can be affordable. It may become too expensive when products have more value and have a smaller and/or less predictable demand.

Moreover, if a company produces its products before they are sold, can it be certain that the products are ever sold at all? When it concerns food products, the products have a limited shelf life after which they cannot be sold anymore, for instance due to government regulations. Products that are subject to rapid technological developments may become obsolete as well. Hence, the risk of obsolescence should be taken into account when deciding which production strategy to apply. With this, the value of the products and especially the uncertainty of demand again play an important role.

The decision whether to make to stock or to order becomes a lot easier when customers are given the opportunity to select certain properties of the product. Obviously, we cannot produce as long as we do not know exactly what or how. Hence, a company that produces customizable products only has MTO as a real option, regardless of the other factors.

Besides pure make-to-order and make-to-stock production, other options exist. A production company could decide to perform some stages of the production process before products are demanded. The intermediate products are then stored in a buffer until they are demanded, after which they undergo the final stages of the production process. This strategy is known as *assemble-to-order*. We could also make the distinction between make-to-order and variants that not only produce after demand occurs, but also buy the raw materials or even develop the entire product after it is demanded. These latter two strategies are known as *purchase-to-order* and *engineer-to-order*.

The place in the process where stages that take place before demand occurs are separated from stages that take place after demand occurs is known as the *Customer Order Decoupling Point* or *Order Penetration Point*. We refer the reader to Van Donk (2001) for a qualitative framework of relevant factors on selecting this point.

1.4 Hybrid production

A production company may very well be confronted with products that differ in value, demand, demand uncertainty, required production time and/or production time uncertainty. Hence, if a company determines whether its products should be made to stock or to order, it may end up in applying the MTS strategy to some products and applying the MTO strategy to others. Producing on an MTS basis and on an MTO basis on the same facility is known as *hybrid production*. Besides the possibility that product and market characteristics advocate different production strategies, hybrid production systems could also arise in other ways.

Some companies produce a high variety of products and therefore have to deal with variable job routings, order quantities and processing times within their production process. Such a company may decide to produce the products that are demanded the most regularly on an MTS basis, so that they can be produced in periods in which there are relatively few orders in the production system. The company could then deliver these products from inventory in busy periods, leaving sufficient production capacity for the remaining products. Hence, this method affects the delivery time and reliability of *all* products that the company offers. If the stock of the concerning products is empty, newly arriving demands can be made to order as usual.

High variety production companies also have to deal with large amounts of idle time of workstations. If these workstations could be used to produce a different type of product, a so-called 'filler', the company could add this product to the product mix without incurring a large investment (i.e. this differs from the previous situation, in which the production strategy of an existing product is switched from MTO to MTS). These fillers can again be produced in periods of low demand, avoiding wasting of production capacity. If such filler products are standardized products which customers can easily buy at competitors, the company would lose demand for a short period of time when no stock is available, but this is not a major problem if the main reason for offering this product is that it avoids idle production capacity.

Make-to-stock production companies also can opt for adding a few make-to-order products and produce in a hybrid fashion. For instance, if a company has spare production capacity available, producing a number of additional items may contribute to the company's profits, or just make the company less dependent on their current markets. If the new items require an MTO approach, for instance because they are customizable, the production system becomes hybrid.

1.5 Differences between MTO and MTS production planning

Although applying a hybrid production strategy may allow a company to better utilize its production capacity or gain an additional margin on the products, it leaves the production planning with a complex 'puzzle' of combining production strategies that require very different ways of planning and control. Indeed, how to best organize the planning of a hybrid production system?

Although we can argue that both production types aim at finishing products on time, the interpretation of 'finishing on time' is different. For MTO, it is clearly defined when jobs should be finished because each job has a due date. For MTS, 'on time' has the interpretation of completing items before a stock-out occurs, so that future demands are not lost or do not have to be backordered. Using demand forecasting techniques, we can estimate when we expect a stock-out and hence when product replenishments should be completed. However, these estimates come with uncertainty. If we plan to complete a replenishment just in time based on the demand forecast, we will encounter a stock-out already if the demand is just a little larger than forecasted. The probability of this occurrence may be unacceptably large. On the other hand, completing a replenishment a little later than the demand forecast prescribes does not automatically imply that a stock-out will occur; it has a larger probability of occurring.

Moreover, finishing products later than intended may have a different consequence as well. For MTS, we have that the longer a stock-out lasts, the longer the period during which demands are lost or backordered. Hence, delayed deliveries have a proportional effect on the performance of the system. For MTO, it may also hold that the period of time by which a due date is violated matters, but this need not be proportional. It may even be the case that the amount of lateness is irrelevant, i.e. late is just late. Also delivering earlier than agreed may have various consequences. If early deliveries are not allowed, the completed products have to be stored until the shipment, which may come at a cost. In many cases, however, early deliveries are not so harmful. Hence, when it comes to the consequences of completing products late or early, MTO and MTS production differ substantially.

Perhaps the most striking difference between the two production types relates to the degree of freedom in production timing and amounts. For MTO production, the products that have to be produced, their amounts, and the time before which they should be completed, are all given by the customer order. Decisions on the produced amount are therefore not applicable to MTO. For MTS, however, a planner for instance has the freedom to decide how many products are stored. This allows controlling the delivery performance for these products. If the planner holds a larger amount of inventory, stock-outs occur less often. Hence, if we only consider delivery performance, it is optimal to hold large amounts of inventory. Logically, this is not realistic. Holding inventory comes at a cost and stocking decisions should balance these costs with the corresponding delivery performance. This forms a sharp contrast with MTO, for which stocking decisions do not apply. Although cost-based planning approaches exist for MTO, the main approach is to consider delivery performance in isolation, whereas this cannot be applied to MTS production.

Besides the amount of inventory held, controlling production amounts also creates flexibility for MTS production. For MTO, these amounts are given by the customer orders, whereas for MTS these can be decided upon. This, in turn, allows including efficiency considerations in the production planning. If machines require a setup between different products, it is beneficial to produce them in larger lots so that these setup times are incurred less often. However, producing in large lots also has disadvantages as machines are then occupied for a larger amount of time. This reduces the possibility to quickly respond to demands for other products, increasing the probability of stock-outs or late deliveries for these. This trade-off does not apply to MTO.

If we now apply hybrid production by combining MTO and MTS on the same facility, the production planning of the two must be combined as well. However, because of the differences between MTO and MTS in the way production planning and control is organized, it is not at all obvious how production planning of a hybrid system should be organized. We should simultaneously make decisions upon timing and/or sequencing of MTO jobs and determine when, or, alternatively, at which stock level, an MTS replenishment is ordered for production. Moreover, it is obvious that the performances of the two product types compete with each other. Hence, organizing production in a hybrid system brings a number of substantial challenges.

1.6 Literature on planning and control of hybrid production systems

We do not provide a complete literature review in this chapter, but refer the reader to Sections 2.2, 3.1 and 4.2. These sections discuss the literature on hybrid production planning more extensively and each from the perspective of the corresponding chapter. For this chapter, we confine ourselves to the most notable contributions in the field of hybrid production planning.

Carr & Duenyas (2000) and Iravani et al. (2012) both provide basic, two-product models of hybrid production processes. The basic setting allows them to model the hybrid systems as mathematical optimization problems known as *Markov Decision Processes*. They use these models to derive structural results and the effects of multiple model parameters. Although the models themselves may be considered too basic for practical use, they provide useful insights into hybrid production planning and this work can be considered a building block for future studies.

The work by Soman et al. (2004, 2006, 2007) provides a major contribution to the field by conducting a number of studies on hybrid production focused on the food processing industry. Soman et al. (2004) review the existing literature in the field of hybrid production. This review is not restricted to planning and includes studies on different research questions, such as the question whether products should be made to stock or to order. The authors further provide a hierarchical planning framework for hybrid production systems in the remainder of this contribution. Soman et al. (2006) compare the performance of four MTS planning procedures applied to a hybrid production environment. The authors continued their work by conducting a case study (Soman et al., 2007) in which they propose an extensive planning approach.

Chang et al. (2003) and Wu et al. (2008) develop sophisticated planning procedures for the control of a semiconductor foundry that applies hybrid production. Chang et al. propose a planning method with three submodules; a module for identifying the bottleneck workstation, and two modules for controlling order release and dispatching at workstations. The release and dispatching modules aim at maintaining work at the bottleneck station, so as to maximize the throughput. Wu et al. adapt a control method by Glassey & Resende (1988) called Starvation Avoidance. They propose a method that releases MTO just-in-time, aiming to avoid releasing earlier than required to meet the due date, so that the corresponding production capacity is reserved for MTS. This allows the method to achieve a high delivery performance for both product types simultaneously.

We finally discuss the work of Schönsleben (2011), who proposes an advanced method for organizing hybrid production known as Capacity Oriented Materials Management (Corma). Corma consists of multiple parts that cover order release (part 1), shop floor control (part 2), and a mechanism that couples floor scheduling with materials management (part 3). The third part smartly deals with some important differences between MTO and MTS production planning. It defines fictitious due dates for MTS items that are used for shop floor control, but allows these fictitious due dates to be changed based on later events. For instance, if the demand for the MTS item is larger than foreseen, the MTS due date is redefined more tightly. In the same way, if the stock level drops slower than foreseen, the MTS replenishment order may be finished later and the due date is adapted accordingly. Moreover, the method also reacts to changes in the MTO workload. Its level of sophistication for hybrid production makes Corma a promising production planning concept for practitioners. For a detailed description of Corma, we refer to Schönsleben (2011, p. 731-739).

Although these and other authors have provided useful contributions in the field of hybrid production planning, the amount of research in this field is still rather limited, especially when it comes to planning approaches that do right to the differences between MTO and MTS. In this thesis, we provide a number of studies in order to bridge this gap.

1.7 Contribution and outline

In the preceding sections, we discussed the production strategies make-to-order and make-to-stock and the way they are planned. As the two production strategies differ substantially, especially in the way they are planned, it is not obvious how the planning and control of a hybrid MTO-MTS production system should be organized. The existing literature offers some background on the planning of hybrid production systems, but especially when it comes to fully considering the crucial characteristics of each of the two production systems, the amount of research is rather limited. In this thesis, we therefore provide three studies in this field in which we try to do right to the characteristics of MTO and MTS and the differences between them. This section discusses the contribution of each of the studies in more detail.

In Chapter 2, we explore the benefits of a hybrid planning approach using Markov Decision Process modeling. We confine ourselves to the most basic form of a hybrid production system by considering a minimal amount of details, but, contrary to existing models, we include a lead time for MTO items as we consider that a crucial characteristic of the production strategy that hybrid production planning should take care of. We show how production decisions should be based on both inventory level and backlog state of MTO products, and that especially discriminating between states with and without backlog of MTO orders is important.

In Chapter 3, we explore lot sizing decisions for MTS. As pointed out above, the MTS production strategy allows for an efficient capacity usage in case machines require a changeover time when switching from one product to another. In this way, we could use the production capacity efficiently by producing MTS in batches. However, in a hybrid production system, the size of these batches does not only affect the state of MTS, but also the responsiveness to incoming MTO orders, which in our view should be taken into account in hybrid production systems. We therefore extend the model presented in Chapter 2 with a setup time for the MTS items, so that lot sizing decisions come into play. We propose and analyze a flexible lot sizing policy, where the lot size is driven by make-to-order backlogs as well as stock levels.

In Chapter 4, we consider a more specific hybrid production situation, namely that of an MTO job shop to which an MTS item is added as a 'filler'. We assume that the job shop is controlled using Operation Due Dates; a method that is specifically designed for controlling MTO production. We propose and compare four ways of integrating MTS into the control of this job shop, all based on fictitious due dates for MTS. We further embed the insights that we obtained from Chapters 2 and 3, by taking care of the key differences between MTO and MTS.

Chapter 5 is, in some respects, the odd one out. Here, we still address the sharing of resources between different products, but rather than continuing the research on sharing production capacity between MTO and MTS we consider the division of storage capacity. In line with Chapters 2 and 3, we study a stylized setting with two products that share a limited storage capacity. We aim at finding an ordering policy that minimizes the cost per time unit. We distinguish from regular ordering policies by allowing the order quantity of one of the products to vary, so that in the same way as in Chapter 3 the flexibility of lots comes into play. As the subject of this chapter deviates from the rest of this thesis, we avoid discussing it in more detail and refer the reader to Section 5.1 for a more extensive introduction of this study.

1.8 Journal publications

The chapters in this thesis are based on the following journal publications and papers under review.

Chapter 2
Beemsterboer, B., Land, M., & Teunter, R. (2016). Hybrid MTO-MTS production planning: An explorative study. *European Journal of Operational Research*, 248, 453-461.

Chapter 3
Beemsterboer, B., Land, M., & Teunter, R. (2016). Flexible lot sizing in hybrid make-to-order/make-to-stock production planning. *Paper under review*.

Chapter 4
Beemsterboer, B., Land, M., & Teunter, R. (2016). Integrating make-to-order and make-to-stock in job shop control. *Paper under review*.

Chapter 5
Beemsterboer, B., Teunter, R., & Riezebos, J. (2016). Two-product Storage-capacitated Inventory Systems: A Technical Note. *International Journal of Production Economics*, forthcoming.