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

Beemsterboer, Bartholomeus Jacobus

Chapter 3

Flexible lot sizing in hybrid make-to-order/make-to-stock production planning

Abstract. *Hybrid make-to-order/make-to-stock production systems are difficult to control. Batch production of make-to-stock products allows for efficient capacity usage, but fixed batch sizes can make the system less responsive to make-to-order customers. 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. We evaluate the performance of this policy for a two-product hybrid system using Markov Decision Process modeling. We find that it leads to savings of up to 23% compared to policies that use either completely or partly fixed lot sizes. We also find that flexible lot sizing is especially beneficial for systems with a low load and where make-to-stock is important in terms of production capacity usage and cost.*

3.1 Introduction

Make-to-order (MTO) and make-to-stock (MTS) production require different planning approaches. MTO production is triggered by demand, allowing for customization of products, and usually focuses on a timely delivery of the products. Standardized MTS products are preferably delivered of the shelf to guarantee high service levels and can be produced in lots to limit setup time and/or costs. Because of these different objectives, hybrid MTO-MTS systems are difficult to control.

Hybrid production systems are common practice as many firms do not apply a single production strategy for their entire product mix. Fast moving standard-

ized products are delivered of the shelf, whereas specialized products are made-to-order. When it comes to production planning of hybrid systems, many firms may be tempted to restrict to simple rules that are known to perform well in pure MTO or MTS systems, for instance by using an (r, Q) reordering policy for MTS items that triggers a replenishment batch of size Q when an item's inventory position drops to a level r. However, using fixed MTS lot sizes may be problematic for hybrid production systems. Batch production is generally applied for efficiency reasons, allowing a machine setup to be used for a series of products. In hybrid systems, however, working on large MTS batches may lead to conflicts as MTO orders have to be delivered within a restricted lead time. Put differently, a focus on a timely delivery of MTO orders may impose restrictions on the MTS lot size. Hence, intuitively it would be preferred to use a more flexible lot sizing approach that takes both the stock level and the MTO order book into account, and even during the production of the batch.

Lot sizing was first considered in the literature for pure MTS production systems. In early models, demand and production rates are often assumed non-random, so it is logical to use fixed lot sizes as in the traditional (r, Q) model. See, for instance, Elmaghraby (1978) for a review on the multi-item case, known as the Economic Lot Scheduling Problem (ELSP). Later models consider random demand, so that it becomes beneficial to let lot sizes depend on the state of the production system, giving rise to the Stochastic ELSP (SELSP). See Sox et al. (1999) and Winands et al. (2011) for review studies.

One of the first to consider hybrid production systems was Williams (1984), who addresses several questions, such as how to prioritize MTO and MTS items, assuming the traditional (r, Q) structure with a fixed lot size. Federgruen & Katalan (1999) analyze the impact of adding an MTO item to an MTS production system. The production runs for the MTS items are assumed to last until a predefined order-up-to level is attained. Hence, the resulting batch size depends on the state of the system, but only through the inventory level of the product that is currently manufactured. Chang et al. (2003) and Wu et al. (2008) propose special-purpose planning methods for hybrid production systems, based on a strict priority of MTO orders and on the rationale that MTS production is used to fill the remaining capacity. Hence, the lot sizes depend on the amount of workload in the system.

Soman et al. (2006) compare four existing SELSP procedures in the context of a hybrid production environment: the method known as the Economic Manufac-

turing Quantity and those proposed by Vergin & Lee (1978), Leachman & Gascon (1988), and Fransoo et al. (1995). In a follow-up study, Soman et al. (2007) execute a case study in a food processing company, in which they propose an extensive planning approach. In both papers, Soman et al. allow the MTS lot size to depend on the state of all products through the run-out times of MTS units and remaining lead times of MTO orders. Several other articles consider hybrid production planning without addressing MTS lot sizing (e.g. Sox et al. (1997), Hadj Youssef et al. (2004)).

Contrary to the previous approaches, the following models implicitly use a more flexible way of determining lot sizes. Carr & Duenyas (2000) and Iravani et al. (2012) consider contract manufacturers, who produce on an MTS basis for contracted customers and on an MTO basis for others. Beemsterboer et al. (2016) determine the potential benefits of a planning approach for hybrid production systems that fully considers the state of both MTS and MTO. All use simple, two-product Markov Decision Process models to analyze MTO order acceptance and/or planning decisions in hybrid production systems. Carr & Duenyas and Iravani et al. determine decisions based on the MTS inventory level and the amount of MTO products, whereas Beemsterboer et al. additionally take the remaining MTO lead times into account. In these models, the MTS lot size does not have to be determined at the start of the batch, but for every decision epoch, the production decision is reviewed so that the eventual lot size is the result of a series of production decisions that can be based on demands during the production of the batch. However, Carr & Duenyas, Iravani et al. and Beemsterboer et al. do not include machine setup times in their models. As a consequence, the capacity advantage, which is the most important reason to produce standardized products in batches, is not considered in these models, and the benefits of flexible lot sizing cannot be determined in this respect.

Our paper aims at demonstrating the benefits of a flexible lot sizing approach for MTS products in hybrid production systems. It is flexible in that lot sizes can be state dependent but also in that the size does not have to be determined at the start of the batch. We analyze a Markov Decision Process model of a hybrid production system in order to determine optimal production decisions for each period, based on the most up-to-date state of the system. From a modeling perspective, we follow the approach of Carr & Duenyas (2000), Iravani et al. (2012) and Beemsterboer et al. (2016). However, we include a setup time in our model so that efficiency considerations of batch production come into play. Moreover, to account for the timing of MTO orders we include an MTO lead time allowance,

Table 3.1: Notations

Demand parameters	d_o	Average MTO demand per period
	d_s	Average MTS demand per period
	d_o^{max}	Maximum MTO demand per period
	d_s^{max}	Maximum MTS demand per period
Cost parameters	h	MTS holding costs (per unit, per period)
	q	MTO lateness costs (per unit, per period)
	b_s	MTS lost sales costs (per unit)
	b_o	MTO lost sales costs (per unit)
System parameters	K	Maximum # MTO orders in the system
	L	MTO lead time allowance
Demand probabilities	$p_o(j)$	Probability of j MTO demands in a period
	$p_s(j)$	Probability of j MTS demands in a period
State space	i	MTS inventory level
	k_l	# MTO orders in the system for l periods ($l = 0, \ldots, L-1$)
	k_L	# late MTO orders
	m	Machine setup status
Other	k	Total number of MTO orders

whereas Carr & Duenyas and Iravani et al. instead use a proxy cost so that all MTO orders are modeled as being 'late' from the moment they arrive.

The remainder of this paper is organized as follows. We present our model and Markov Decision Process formulation in Section 3.2. In Section 3.3, we discuss the optimal policy for an illustrative example and examine the resulting MTS lot sizes. We conduct a numerical investigation in Section 3.4, in which we compare our flexible model with two reference models that use fixed batch sizes. Conclusions and directions for further research are given in Section 3.5.

3.2 Model formulation

In this section, we model a hybrid production system. In Subsection 3.2.1, we provide the framework of the model, which we then formulate as a Markov Decision Process in Subsection 3.2.2. A list of notations is given in Table 3.1.

3.2.1 Production system

We consider a system that manufactures two products, one on an MTO basis and one on an MTS basis. The system can only manufacture one product at a time. For

both products, the system requires a setup. We assume that MTO products are customized and that the system therefore requires a setup for each product. By contrast, MTS products are standardized products and only one setup is needed to produce a batch.

In order to formulate our problem as a Markov Decision Process in the next subsection, we make the following simplifying assumptions. We use a discrete time framework and we assume that the setup and unit processing time of both products are all equal to each other and given as one period. In order to ensure finiteness of the state space, we will use bounded demand distributions and we define a maximum MTS inventory level I and a maximum amount of MTO orders in the system K. We assume that excess MTO demand is lost. The maximum inventory level I is not treated as a system control parameter, but set large enough so as not to affect the numerical results. With these simplifications, we retain the core of the problem so that we can obtain the insights we aim for.

Demands for MTO and MTS are random and follow discrete and bounded probability distributions, and we assume that these are independent. The probability of j demands in a period is denoted as $p_o(j)$ for MTO and $p_s(j)$ for MTS. We denote the average demands per period as d_o for MTO and d_s for MTS and maximum demands as d_o^{max} (MTO) and d_s^{max} (MTS), so we have $d_o = \sum_{j=0}^{d_o^{max}} jp_o(j)$ and $d_s = \sum_{j=0}^{d_s^{max}} jp_s(j)$. MTS demands have to be satisfied from inventory on hand or are otherwise lost. MTO orders have a promised lead time of L periods (not including the period in which the order arrives). Late orders should still be delivered, but incur a penalty cost per time unit. In our system, MTO orders differ only in their remaining time until the due date, and therefore it is optimal to produce them in a first come, first served fashion to realize the lowest lateness costs.

In each period, the production system may either perform a setup (MTO, MTS) or manufacture a product (MTO, MTS). Logically, the system may only manufacture a product when it is set up for the concerning product type. Producing a unit of MTO brings the system back to a state where no setup is completed, while producing a unit of MTS maintains the setup status so that another MTS unit can be produced in the succeeding period. An MTO setup, which concerns customized products, can only be performed when there is at least one MTO order in the system. By contrast, an MTS setup (concerning the standardized product) can always be performed. We also allow the system to maintain the MTS setup status without producing a unit, which makes idling redundant.

This is modeled by doing another setup. Finally, we assume that performing or maintaining a setup does not constrain the machine to produce the concerning product type in the succeeding period, i.e. production may be 'aborted' after a setup.

The order of events in each period is as follows. First, the system decides on an action. Second, if this action is to produce a unit of MTO or MTS, it becomes available. Third, demands occur and are either accepted/satisfied or lost.

The objective is to minimize the average costs per period. These costs are composed of per period holding costs for MTS products, lateness costs for MTO products, and lost sales costs for MTO and MTS products. The MTS holding costs are given as h per unit per period, late MTO orders are penalized by a cost of q per period and lost sales costs are b_o per missed MTO demand and b_s per missed MTS demand.

As the production decision is determined period by period, and the MTS lot size is not taken in advance, our model follows the flexible approach described in Section 3.1.

3.2.2 Markov Decision Process

We model the problem described above as a discrete-time Markov Decision Process, which requires the definition of a state space $\mathcal{S}$, an action space $\mathcal{A}$, costs, and transition probabilities.

State space

The state of the system is described by the MTS inventory level i, the numbers of accepted MTO orders k_l, $l = 0, \ldots, L-1$, that have been in the system for l periods (not including the period of arrival) and so have $L-l$ periods remaining until the due date, the number of outstanding late orders k_L, and the machine setup status m, indicating whether the machine is not set up ($m = 1$) which happens after MTO production has finished, set up for MTO ($m = 2$), or set up for MTS ($m = 3$). Denoting the state space by $\mathcal{S}$, a state $s \in \mathcal{S}$ is described by the tuple $s = (i, k_0, \ldots, k_L, m)$. For convenience, we refer to the MTO part of the state space $(k_0, \ldots, k_L)$ as the *order state*. We further define k as the total number of MTO orders in the system, i.e. $k = \sum_{j=0}^{L} k_j$.

Action space

The action space is denoted by $\mathcal{A} = \{1, 2, 3, 4\}$, where $a = 1, \ldots, 4$ corresponds to, respectively, an MTO setup, MTO production, an MTS setup, and MTS production. Action $a = 1$ (MTO setup) is admissible when the system is not set up for MTO yet, and only when MTO orders are at hand, i.e. in states $s = (i, k_0, \ldots, k_L, m)$ with $m \neq 2$ and $k \geq 1$. Action $a = 2$ (MTO production) is admissible when the system is set up for MTO and again when MTO orders are at hand, i.e. in states s with $m = 2$ and $k \geq 1$. Action $a = 3$ (MTS setup) is admissible in any state and action $a = 4$ (MTS production) is admissible when the system is set up for MTS and when the inventory level is below its maximum, i.e. in states s with $m = 3$ and $i = 0, \ldots, I - 1$.

Costs

We now define the expected costs in state s if action a is taken, denoted as $c^a(s)$. Writing $[x]^+$ for $\max\{x, 0\}$, these are given as

$$c^1(s) = hi + qk_L + b_s \sum_{j=0}^{d_s^{max}} \left(p_s(j)[j-i]^+\right) + b_o \sum_{j=0}^{d_o^{max}} \left(p_o(j)[k+j-K]^+\right),$$

$$c^2(s) = hi + qk_L + b_s \sum_{j=0}^{d_s^{max}} \left(p_s(j)[j-i]^+\right) + b_o \sum_{j=0}^{d_o^{max}} \left(p_o(j)[k+j-K-1]^+\right),$$

$$c^3(s) = hi + qk_L + b_s \sum_{j=0}^{d_s^{max}} \left(p_s(j)[j-i]^+\right) + b_o \sum_{j=0}^{d_o^{max}} \left(p_o(j)[k+j-K]^+\right),$$

$$c^4(s) = hi + qk_L + b_s \sum_{j=0}^{d_s^{max}} \left(p_s(j)[j-i-1]^+\right) + b_o \sum_{j=0}^{d_o^{max}} \left(p_o(j)[k+j-K]^+\right).$$

The first and second terms represent the holding and lateness costs, respectively. The third and fourth terms are, respectively, the expected MTS and MTO lost sales costs. For $a = 2$ (MTO production) and $a = 4$ (MTS production), the product that is manufactured becomes available before the demand occurs so that the expected excess demand of the concerning product type decreases by one.

Transition probabilities

We next discuss the transition probabilities $\pi^a(s, s')$, denoting the probability that the state at the end of a period is $s' = (i', k'_0, \ldots, k'_L, m')$ given that the state at the beginning is s and action a is taken. For simplicity, we restrict the discussion to transitions for which lost sales can be ignored, i.e. transitions to states for which $i' > 0$ and $\sum_{j=0}^{L} k'_j < K$. For other transition probabilities, which we provide in Appendix B, the possibility of lost sales leads to slightly different expressions, but the dynamics are essentially the same. We first consider the setup status transitions for each action, followed by the order state transitions, and then combine those into complete system state transitions with corresponding probabilities.

Following the definitions above, $a = 1$ (MTO setup) can be applied if $m \neq 2$ and sets the setup status to $m' = 2$. Action $a = 2$ (MTO production) can be applied if $m = 2$, resetting the setup status to $m' = 1$ afterward. Action $a = 3$ (MTS setup) can be applied regardless of the setup status but sets it to $m' = 3$ afterward, and, finally, $a = 4$ (MTS production) applies if $m = 3$ and maintains that status. Accordingly, we define setup status transition sets for each action:

$$\mathcal{M}_1 = \{(s, s') | m \neq 2, m' = 2\}, \tag{3.1}$$

$$\mathcal{M}_2 = \{(s, s') | m = 2, m' = 1\}, \tag{3.2}$$

$$\mathcal{M}_3 = \{(s, s') | m' = 3\}, \tag{3.3}$$

$$\mathcal{M}_4 = \{(s, s') | m = 3, m' = 3\}. \tag{3.4}$$

The order state transitions only depend on whether MTO production takes place ($a = 2$) or not ($a = 1, 3, 4$). If it does not, then the order state transitions are as described by the following set.

$$\mathcal{O}_I = \{(s, s') | k'_l = k_{l-1} \text{ for } l = 1, \ldots, L-1 \text{ and } k'_L = k_L + k_{L-1}\} \tag{3.5}$$

If production does take place, then an order with the smallest remaining lead time will be satisfied (recalling from Subsection 3.2.1 that first come, first served is optimal). Denoting 'the longest time that an order has been in the system' by $l^* = \max\{l | k_l \geq 1\}$, and defining $k''_l = k_l$ for $l \in \{0, \ldots, L\} \backslash l^*$ and $k''_{l^*} = k_{l^*} - 1$, we obtain the second order transition set.

$$\mathcal{O}_{II} = \left\{(s, s') | k'_l = k''_{l-1} \text{ for } l = 1, \ldots, L-1 \text{ and } k'_L = k''_L + k''_{L-1}\right\} \tag{3.6}$$

Combining the transition sets of setup status and order state, and using independence of MTO and MTS demands, we obtain the following complete state transition probabilities.

$$\pi^1(s,s') = p_o(k_0')p_s(i-i') \qquad \text{for } (s,s') \in \mathcal{M}_1 \cap \mathcal{O}_I, \tag{3.7}$$

$$\pi^2(s,s') = p_o(k_0')p_s(i-i') \qquad \text{for } (s,s') \in \mathcal{M}_2 \cap \mathcal{O}_{II}, \tag{3.8}$$

$$\pi^3(s,s') = p_o(k_0')p_s(i-i') \qquad \text{for } (s,s') \in \mathcal{M}_3 \cap \mathcal{O}_I, \tag{3.9}$$

$$\pi^4(s,s') = p_o(k_0')p_s(i-i'+1) \qquad \text{for } (s,s') \in \mathcal{M}_4 \cap \mathcal{O}_I, \tag{3.10}$$

$$\pi^a(s,s') = 0 \qquad \text{elsewhere.}$$

3.3 Policy structure

Before performing an extensive numerical exploration using various parameterizations, we start by discussing the typical structure of the optimal policy for an illustrative example that has a relatively small state space, allowing all states to be represented in a table. We select the smallest possible (but positive) values for the maximum demands per period d_o^{max} and d_s^{max}, i.e. $d_o^{max} = d_s^{max} = 1$. This means that we define the demand per period of both product types as Bernoulli distributions. For the Bernoulli distribution, the mean equals the probability of a success, and so we omit introducing additional notation and use d_o and d_s as model parameters. For this example their values are selected as 0.25 for both MTO and MTS, i.e. $d_o = d_s = 0.25$. Notice that producing an MTO unit (setup and production) takes a total of two periods, so an average MTO demand of $d_o = 0.25$ leads to a 'machine load' contribution of 50%, i.e. 50% of the machine capacity is needed to satisfy this demand if no demand would be lost. For MTS production, the required number of periods per unit lies between one, in the extreme case of producing an infinitely long batch, and two, in case we would perform a setup for each unit. So $d_s = 0.25$ implies a machine load contribution of MTS between 25% and 50% and hence the total machine load is 75 - 100%. As a total machine load of 100% would be too large, some degree of MTS batching is needed.

The remaining parameters are as follows. We set the MTO lead time allowance to $L = 3$, which allows making a distinction between more urgent and less urgent MTO orders. The maximum amount of MTO orders in the system is set to $K = 5$. Finally, we normalize the MTS holding costs to $h = 1$ and we set $q = 8$ and $b_s = b_o = 250$, i.e. the unit MTO lateness costs is 8 times larger and the lost

sales costs of both product types are 250 times larger than the MTS holding costs. These cost ratios guarantee that the resulting optimal policy avoids MTO lateness before lost sales come into view, and also sufficiently avoids MTS lost sales. This parameterization yields 36 different order states and we have a resulting optimal policy that requires a maximum inventory level of 5. As we also have three setup status values, there are 648 states altogether.

Table 3.2 shows the optimal policy for this example. The optimal actions are split into three columns, each corresponding to a machine setup status (not set up, set up for MTO, and set up for MTS). Each column contains six smaller columns corresponding to all (attainable) inventory levels. Each row corresponds to an order state, containing, from left to right, the numbers of MTO orders just received, received one period earlier, received two periods earlier, and received longer ago so that they are late. Note that, because performing an MTO setup is not allowed when there are no orders in the system, the order states $(0, 0, 0, 0)$ and $(1, 0, 0, 0)$ in the MTO setup status cannot be attained by definition.

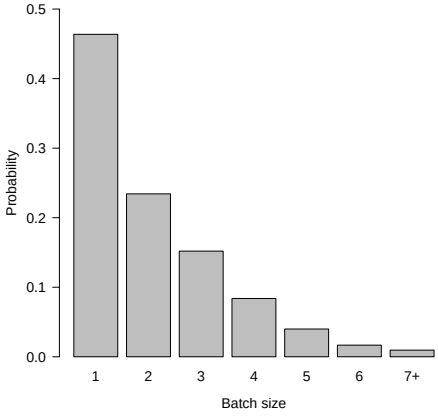
The optimal policy shown in Table 3.2 reveals several insights. First of all, when the machine is not set up yet, the system performs an MTS setup either when there is no MTS inventory or when there are no MTO orders left in the system, and performs an MTO setup in all other cases. Second, the system always continues producing MTO when an MTO setup is completed. This is logical, as a switch to MTS would imply that the MTO setup that had just been performed is wasted, spoiling valuable production capacity. When set up for MTS, MTS production continues until a particular inventory level is reached, which depends on the order state. For the order state $(0, 0, 0, 0)$, the system stops producing MTS from an inventory level of five, after which the system maintains its MTS setup status. In the order state $(1, 0, 0, 0)$, MTS production continues until an inventory level of four is reached, after which the system switches to MTO. In all other order states, the system switches to MTS at an inventory level of three.

Although the differences in these 'switching' levels are small (they can be much larger for other examples), these reveal important benefits of a state-dependent switching level. In particular the difference in switching level between 'no MTO orders' (first row) and 'MTO orders present' (all other rows) is beneficial; if there are no MTO orders, the production capacity is used to restock up to high inventory levels, allowing the production of MTS to be less urgent when MTO orders arrive. In addition, we observe a different switching level for the second row and all lower rows. This is because in the order state $(1, 0, 0, 0)$, the single

Table 3.2: Optimal policy for illustrative example
Actions: o = MTO setup, p = MTO production, s = MTS setup, q = MTS production
Parameter settings: $d_o = d_s = 0.25$, $L = 3$, $K = 5$, $q = 8$, $b_s = b_o = 250$

Setup status →	Not set up ($m = 1$)						For MTO ($m = 2$)						For MTS ($m = 3$)					
Inventory →	0	1	2	3	4	5	0	1	2	3	4	5	0	1	2	3	4	5
Order state ↓																		
(0, 0, 0, 0)	s	s	s	s	s	s							q	q	q	q	q	s
(1, 0, 0, 0)	s	o	o	o	o	o							q	q	q	q	o	o
(0, 1, 0, 0)	s	o	o	o	o	o	p	p	p	p	p	p	q	q	q	o	o	o
(1, 1, 0, 0)	s	o	o	o	o	o	p	p	p	p	p	p	q	q	q	o	o	o
(0, 0, 1, 0)	s	o	o	o	o	o	p	p	p	p	p	p	q	q	q	o	o	o
(1, 0, 1, 0)	s	o	o	o	o	o	p	p	p	p	p	p	q	q	q	o	o	o
(0, 1, 1, 0)	s	o	o	o	o	o	p	p	p	p	p	p	q	q	q	o	o	o
(1, 1, 1, 0)	s	o	o	o	o	o	p	p	p	p	p	p	q	q	q	o	o	o
(0, 0, 0, 1)	s	o	o	o	o	o	p	p	p	p	p	p	q	q	q	o	o	o
(1, 0, 0, 1)	s	o	o	o	o	o	p	p	p	p	p	p	q	q	q	o	o	o
(0, 1, 0, 1)	s	o	o	o	o	o	p	p	p	p	p	p	q	q	q	o	o	o
(1, 1, 0, 1)	s	o	o	o	o	o	p	p	p	p	p	p	q	q	q	o	o	o
(0, 0, 1, 1)	s	o	o	o	o	o	p	p	p	p	p	p	q	q	q	o	o	o
(1, 0, 1, 1)	s	o	o	o	o	o	p	p	p	p	p	p	q	q	q	o	o	o
(0, 1, 1, 1)	s	o	o	o	o	o	p	p	p	p	p	p	q	q	q	o	o	o
(1, 1, 1, 1)	s	o	o	o	o	o	p	p	p	p	p	p	q	q	q	o	o	o
(0, 0, 0, 2)	s	o	o	o	o	o	p	p	p	p	p	p	q	q	q	o	o	o
(1, 0, 0, 2)	s	o	o	o	o	o	p	p	p	p	p	p	q	q	q	o	o	o
(0, 1, 0, 2)	s	o	o	o	o	o	p	p	p	p	p	p	q	q	q	o	o	o
(1, 1, 0, 2)	s	o	o	o	o	o	p	p	p	p	p	p	q	q	q	o	o	o
(0, 0, 1, 2)	s	o	o	o	o	o	p	p	p	p	p	p	q	q	q	o	o	o
(1, 0, 1, 2)	s	o	o	o	o	o	p	p	p	p	p	p	q	q	q	o	o	o
(0, 1, 1, 2)	s	o	o	o	o	o	p	p	p	p	p	p	q	q	q	o	o	o
(1, 1, 1, 2)	s	o	o	o	o	o	p	p	p	p	p	p	q	q	q	o	o	o
(0, 0, 0, 3)	s	o	o	o	o	o	p	p	p	p	p	p	q	q	q	o	o	o
(1, 0, 0, 3)	s	o	o	o	o	o	p	p	p	p	p	p	q	q	q	o	o	o
(0, 1, 0, 3)	s	o	o	o	o	o	p	p	p	p	p	p	q	q	q	o	o	o
(1, 1, 0, 3)	s	o	o	o	o	o	p	p	p	p	p	p	q	q	q	o	o	o
(0, 0, 1, 3)	s	o	o	o	o	o	p	p	p	p	p	p	q	q	q	o	o	o
(1, 0, 1, 3)	s	o	o	o	o	o	p	p	p	p	p	p	q	q	q	o	o	o
(0, 1, 1, 3)	s	o	o	o	o	o	p	p	p	p	p	p	q	q	q	o	o	o
(0, 0, 0, 4)	s	o	o	o	o	o	p	p	p	p	p	p	q	q	q	o	o	o
(1, 0, 0, 4)	s	o	o	o	o	o	p	p	p	p	p	p	q	q	q	o	o	o
(0, 1, 0, 4)	s	o	o	o	o	o	p	p	p	p	p	p	q	q	q	o	o	o
(0, 0, 1, 4)	s	o	o	o	o	o	p	p	p	p	p	p	q	q	q	o	o	o
(0, 0, 0, 5)	s	o	o	o	o	o	p	p	p	p	p	p	q	q	q	o	o	o

Figure 3.1: Distribution of the flexible MTS batch sizes in the optimal policy of the example



MTO order is not urgent in the sense that it can still be completed on time if the production is postponed by one period, while in all other order states a lateness penalty will be incurred if the system does not switch to MTO immediately. Other considered examples confirmed these insights. Hence, we observe that an optimal policy may indeed benefit from the possibility to determine a production decision based on the state of both product types.

Figure 3.1 shows a histogram of the distribution of the realized batch sizes for this example. Apparently, a high variation of the MTS batch size is needed to obtain minimum costs. The optimal fixed batch size in this example would have been 3, but the optimal flexible batch size is often much smaller with an average of 2.09 and a standard deviation of 1.35. It equals one in 46% and two in another 23% of the cases, and 15% of the batches become larger than size 3. This suggests that fixing the batch size may be far from optimal. A more extensive numerical study in the next section confirms this.

3.4 Numerical investigation

This section provides a numerical study of the benefits of flexible MTS lot sizing compared to fixed batch sizes, by comparing the performance of the model defined in Section 3.2 with two reference models that use fixed MTS batch sizes. We introduce these reference models in Subsection 3.4.1. We then discuss the design

of a first set of experiments in Subsection 3.4.2 and we present the results in Subsection 3.4.3. As these do not reveal the effect of the demand mix, we conducted additional experiments which we discuss in Subsection 3.4.4.

3.4.1 Reference models

Recall from Section 3.1 that besides the flexible approach, two other lot sizing approaches exist in the literature. These fix the MTS lot sizes either from the start of the batch or throughout the entire policy. We adopt these approaches into the following reference models. In the interest of consistency, the model developed in Section 3.2 will be referred to as *Fully Flexible*.

Partly Flexible. When starting an MTS batch, its size is selected in advance, and switching during production is not possible. Switching after an MTO setup is also not possible for consistency. Hence, for each state, an optimal batch size is selected.

Not Flexible. The MTS batch size is fixed throughout the entire policy, i.e. one optimal batch size is obtained across all states.

The optimization approach for the reference models is very similar to that for Fully Flexible. For Partly Flexible, we extend the Markov Decision Process defined in Subsection 3.2.2 so that each possible MTS batch size is associated with an action. For Not Flexible, we also use an adapted Markov Decision Process, but as the fixed MTS batch size has to be provided beforehand, we repeatedly solve instances of this model using a one-dimensional grid search, each with a different predefined MTS batch size, in order to obtain the overall optimal fixed batch size.

3.4.2 Experimental design

We first select and study a base case, followed by a sensitivity study in which we vary more parameters. The demand and cost parameters of the base case are selected equal to those of the illustrative example in Section 3.3: $d_o = d_s = 0.25$, $h = 1$, $q = 8$ and $b_s = b_o = 250$. Recall that $d_o = d_s = 0.25$ implies a machine load of $75 - 100\%$. We use Bernoulli demand distributions again, implying $d_o^{max} = d_s^{max} = 1$, which keeps the state space and calculation times acceptable, also for the larger values of K and L that we will consider. The maximum number of MTO orders K is mainly included in the model to bound the size of the

Table 3.3: Parameters and results of the experiments

#	Parameters								Machine load	Costs			Savings relative	
	d_o	d_s	L	K	h	q	b_s	b_o		NF	PF	FF	to PF	to NF
1	0.25	0.25	7	8	1	8	250	250	75-100%	5.0	4.8	4.5	6.0%	9.5%
2	**0.20**	0.25	7	8	1	8	250	250	65-90%	3.6	3.4	3.0	10.9%	14.6%
3	**0.30**	0.25	7	8	1	8	250	250	85-110%	7.8	7.6	7.4	2.7%	5.5%
4	0.25	**0.20**	7	8	1	8	250	250	70-90%	3.8	3.6	3.4	6.2%	10.2%
5	0.25	**0.30**	7	8	1	8	250	250	80-110%	6.5	6.3	6.0	4.5%	7.9%
6	0.25	0.25	**6**	8	1	8	250	250	75-100%	5.3	5.1	4.9	5.4%	8.5%
7	0.25	0.25	**8**	8	1	8	250	250	75-100%	4.6	4.5	4.2	6.6%	10.3%
8	0.25	0.25	7	**6**	1	8	250	250	75-100%	5.0	4.8	4.5	6.0%	9.6%
9	0.25	0.25	7	**10**	1	8	250	250	75-100%	5.0	4.8	4.5	6.0%	9.4%
10	0.25	0.25	7	8	**0.5**	8	250	250	75-100%	3.2	3.1	2.9	5.1%	8.6%
11	0.25	0.25	7	8	**2**	8	250	250	75-100%	7.6	7.4	6.9	6.8%	10.1%
12	0.25	0.25	7	8	1	**4**	250	250	75-100%	4.3	4.1	3.8	8.2%	11.7%
13	0.25	0.25	7	8	1	**16**	250	250	75-100%	5.4	5.3	5.0	4.7%	7.2%
14	0.25	0.25	7	8	1	8	**125**	250	75-100%	4.1	4.0	3.8	6.3%	8.1%
15	0.25	0.25	7	8	1	8	**500**	250	75-100%	5.4	5.2	4.9	6.4%	10.1%
16	0.25	0.25	7	8	1	8	250	**125**	75-100%	5.0	4.8	4.5	6.1%	9.7%
17	0.25	0.25	7	8	1	8	250	**500**	75-100%	5.0	4.8	4.5	5.9%	9.4%

state space and we prefer to have a value that is large enough so that it has little influence on the behavior of the optimal policies. Pretesting showed that a value of $K = 8$ is large enough for our purposes. To allow for completion of larger MTS batches we set our lead time allowance for MTO to $L = 7$.

'In order to examine the dependence of the benefits of flexible lot sizing on the parameters, we conduct a sensitivity analysis in which we select a smaller and a larger value for each of the parameters d_o, d_s, L, K, h, q, b_s and b_o, as specified in Table 3.3. The parameters are varied one by one, i.e. all other parameters are fixed at their base value. The demand distribution type is not varied.

3.4.3 Results

The right part of Table 3.3 overviews the results of the full experimental design. The experimental setup allows us to distinguish two types of savings resulting from: (1) the flexibility to select various lot sizes, and (2) the flexibility to determine lot sizes once production has started. The latter type, represented by the savings of Fully Flexible relative to Partly Flexible, yields around 6% and varies considerably, ranging from 3 to 11% for these experiments. The flexibility to select various lot sizes adds another 3 to 4%, bringing the total savings of the fully flexible versus the not flexible model to around 10% for most of our experiments, ranging from 6 to 15%.

Before discussing the key sensitivity findings, we observe that varying the maximum amount of MTO orders (experiments 8 and 9) has a very limited effect on the costs and savings. This is what we aimed for as we introduced this parameter for technical reasons (see Section 3.2).

Examining the results of the individual experiments, we distinguish the following two main effects. Considering experiments 2 to 5, in which we vary the demand rates, we observe that lower demand rates lead to higher savings. The effect is stronger for a varying MTO demand rate, but essentially applies to both. Production systems with a larger total demand rate have a higher machine load and, as a result, incur more costs. The savings in absolute terms, however, are not notably lower than systems with a lower machine load. However, as these systems incur overall more costs, the *relative* savings become lower.

Another effect is revealed by considering the experiments in which we vary the cost parameters. Experiments 11, 12, 15 and 16 lead to larger savings than the base case, while their counterparts (10, 13, 14 and 17) lead to smaller savings. Either when an MTO cost parameter is decreased, or when and MTS cost parameter is increased, the savings become larger. From these observations, we learn that whenever MTS becomes more costly, more is to be gained by varying lot sizes. This is intuitive: the models differ in the way they handle MTS production and hence when this type becomes more important from a cost perspective, more is to be gained. Considering experiments 6 and 7, where the MTO lead time allowance is varied, the same explanation applies. When MTO orders have more time to be produced, MTO lateness costs are incurred less often, and MTS becomes more important from a cost perspective.

The experiments so far do not allow us to draw clear conclusions with respect to the demand mix. Experiments 2 and 3 show that an increasing MTO demand rate, obviously leading to larger share of MTO, leads to smaller savings. By contrast, experiments 4 and 5 show that a decreasing MTS demand rate, also leading to a larger share of MTO, leads to increased savings. These seemingly contradictory results are related to the fact that the experiments could not isolate the demand mix (variations), because changing the demand rates not only affects the demand mix but also the machine load, for which we have observed effects. In order to find the effect of the demand mix, we conduct a second set of experiments.

3.4.4 Effect of demand mix

It is impossible to keep both the total demand rates and the machine load bounds constant when varying the product mix, because only MTS products can be batch produced. Therefore, in these additional experiments we apply a two-dimensional setting in which we vary demand rates as shown in Table 3.4. We increase the share of MTS in the demand mix in two ways: with a fixed total demand rate and a decreasing minimum machine load (from left to right), and with a fixed minimum machine load and an increasing total demand rate (from top to bottom). The two ways have opposite effects on the actual machine load. However, both ways show increasing savings. Hence, we may conclude that, in general, an increasing share of MTS leads to higher savings.

Table 3.4: Effect of varying demand mix

	Demand parameters (d_o, d_s)			Savings relative to NF		
Total demand rate	Min. machine load 80%	75%	70%	Min. machine load 80%	75%	70%
0.4	$(0.40, 0.0)$	$(0.35, 0.05)$	$(0.30, 0.1)$	0.0%	1.0%	7.8%
0.45	$(0.35, 0.1)$	$(0.30, 0.15)$	$(0.25, 0.2)$	4.4%	7.1%	10.2%
0.5	$(0.30, 0.2)$	$(0.25, 0.25)$	$(0.20, 0.3)$	5.7%	9.5%	13.5%
0.55	$(0.25, 0.3)$	$(0.20, 0.35)$	$(0.15, 0.4)$	7.9%	11.0%	16.8%
0.6	$(0.20, 0.4)$	$(0.15, 0.45)$	$(0.10, 0.5)$	9.7%	13.5%	22.7%

The savings range from 0% (with $d_o = 0.4$ and no MTS demand) to 23% (with $d_o = 0.1$ and $d_s = 0.5$). Logically, when there is no MTS demand, no batching can be applied at all so there are no savings. By contrast, when MTS constitutes a large share of the total demand, the savings are substantial. Again, it is explained through the main difference of the models, which differ especially in the way they plan MTS production. Fully Flexible allows for better control of the inventory level than the other models, leading to both lower inventory holding costs and lower MTS lost sales costs. Thus, overall, we have shown that when MTS becomes more important either from a cost or from a demand perspective, the benefits of flexible lot sizing increase.

3.5 Conclusion

We studied the benefits of varying MTS lot sizes in hybrid production systems. We proposed a 'fully flexible' approach, where lot sizes are not fixed at any time but rather evolve in response to demands during the production of the batch.

Analysis of this approach with a Markov Decision Process has shown that the optimized fully flexible approach leads to considerable variation in lot sizes in order to reduce costs. For example, in a situation where an optimal fixed batch size of 3 would be applied, optimization under flexibility was shown to bring the average lot size down to 2.09 with a standard deviation of 1.35.

We compared our approach with two approaches that consider lot sizes that are fixed either from the start of the batch (partly flexible) or throughout the entire policy (not flexible) for a two-product hybrid system. The numerical experiments showed that this fully flexible approach may save up to 23% of the costs. Our design allowed us to attribute this percentage to two types of flexibility. The majority of the savings, up to 19%, is due to flexibility created by the possibility to review the production decision in each period. A smaller part of the benefits was attributed to the flexibility to differentiate lot sizes based on the state of the production system.

The savings are especially large in three situations. First, relatively high savings are obtained for low load situations. Second, a large share of MTS demand corresponds with higher savings, and third, when the MTS costs are relatively large, we obtain high savings. This confirms the intuition that savings will relate strongly to MTS characteristics, as lot sizing also relates to the MTS products.

We conducted this study in a setting with a number of simplifying assumptions that allowed us to use an analytical optimization procedure and study the lot sizing decision in isolation. This helped us clarify the mechanisms that determine the advantage of flexible lot sizes. Future research could be directed at the performance of a fully flexible approach in more complex settings. Promising extensions include: multiple products; variations in the production times, setup times or MTO lead times; or by considering a more detailed representation of a production facility, e.g. a job shop. Secondly, to fully exploit the savings potential we found, this isolated lot sizing method should be embedded in planning approaches for hybrid production systems, providing another promising research opportunity.