mathematics

MDPI

Article

Performance Analysis of Hybrid MTS/MTO Systems with Stochastic Demand and Production

Dieter Fiems [1,*], Eline De Cuypere [1], Koen De Turck [1,2] and Dieter Claeys [3]

[1] Department of Telecommunications and Information Processing, Ghent University, 9000 Ghent, Belgium; eline.decuypere@gmail.com (E.D.C.); Koen.DeTurck@UGent.be (K.D.T.)
[2] Laboratoire des Signaux et Systèmes, CNRS, CentraleSupélec, 91190 Gif-sur-Yvette, France
[3] Department of Industrial Systems Engineering (ISyE), Flanders Make, Ghent University, 9000 Ghent, Belgium; Dieter.Claeys@UGent.be
* Correspondence: Dieter.Fiems@UGent.be

Received: 14 September 2020; Accepted: 27 October 2020; Published: 2 November 2020

Abstract: We present a comprehensive numerical approach with reasonably light complexity in terms of implementation and computation for assessing the performance of hybrid make-to-stock (MTS)/make-to-order (MTO) systems. In such hybrid systems, semi-finished products are produced up front and stored in a decoupling inventory. When an order arrives, the products are completed and possibly customised. We study this system in a stochastic setting: demand and production are modelled by random processes. In particular, our model includes two coupled Markovian queues: one queue represents the decoupling inventory and the other the order backlog. These queues are coupled as order processing can only occur when both queues are non-empty. We rely on matrix analytic techniques to study the performance of the MTO/MTS system under non-restrictive stochastic assumptions. In particular, we allow for arrival correlation and non-exponential setup and MTS and MTO processing times, while the hybrid MTS/MTO system is managed by an (s, S)-type threshold policy that governs switching from MTO to MTS and back. By some numerical examples, we assess the impact of inventory control, irregular order arrivals, setup and order processing times on inventory levels and lead times.

Keywords: decoupling inventory; performance evaluation; stochastic modelling; Markov process

1. Introduction

The boundary between forecast-driven and demand-driven activities is a key strategic decision in supply chain management. This boundary is referred to as the order decoupling point, and its position considerable influences the market and operational performance of the supply chain [1,2]. Make-to-stock (MTS) and make-to-order (MTO) are the best known inventory management strategies. Under make-to-stock management, all production is purely forecast driven. The products are manufactured up front without allowing for customer customisation. As these products are stored till ordered, high holding costs are unavoidable. Moreover, one can expect frequent stock-outs when demand fluctuates considerably [2]. Some main performance metrics of MTS systems include the fill rate and the average work-in-process, as well as the demand forecasting accuracy [3]. While MTS is purely forecast driven, MTO is purely order driven. In MTO, manufacturing of products only starts when a customer order is placed. This of course leads to long response times during periods where there is considerable demand for the product. Performance metrics for MTO systems therefore include the mean response time, order delay and manufacturing lead time; see [3] and the references therein. Incorporating the benefits from both systems, hybrid MTS/MTO strategies have been proposed in the literature [4,5]. In such hybrid strategies, production is split up into two states. The production of semi-finished items in the first production stage is forecast driven, and the semi-finished products are stored in a decoupling inventory, much like for the MTS strategy. In contrast and in line with MTO, the second production stage only starts upon reception of an order.

Various authors have studied such MTS/MTO systems. Ghrayeb et al. [6] proposed a deterministic optimisation model of a hybrid MTS/MTO system, in the context of an assemble-to-order manufacturing environment. These authors showed that the hybrid case inherits the strengths and conceals the weaknesses of both pure MTS and pure MTO systems. Köber and Heinecke [7] considered hybrid production strategies when demand is volatile and seasonal. The costs and benefits of delayed product differentiation were studied by Gupta and Benjaafar [8] when the order lead times are load dependent. It was observed that later differentiation is preferred when the load is higher. Liu et al. [9] studied a similar problem and found that a push-pull strategy is optimal when the system load is high. Hierarchical production planning was investigated by Soman et al. [3] in the context of production management decisions for MTS/MTO operations in food processing. Their framework includes strategic, tactical and operational levels, which correspond to the MTS/MTO decision, the capacity coordination and the scheduling, respectively. These authors [10] also investigated an economic lot scheduling problem, again in the context of a food production system. Finally, Rafiei and Rabbani [11] focused on the tactical level of the hierarchical production planning framework for hybrid MTS/MTO production systems with pure MTS, pure MTO and hybrid MTS/MTO products and applied their capacity coordination model to a realistic industrial case.

In this work, we propose a stochastic inventory model for assessing the performance of hybrid MTS/MTO systems. Stochastic models explicitly account for uncertainty in demand, inventory replenishment and order processing. Such stochastic inventory models most closely relate to the present study. We mention [12], where a variety of combined make-to-order (MTO) and make-to-stock (MTS) supply chains was investigated in a stochastic framework. These authors found that hybrid MTS/MTO systems can dramatically cut costs, although the information exchange between suppliers and manufacturers is key for effective lead time quotations. Further, Adan and Van der Wal [13] showed that the combination of pure MTS and pure MTO strategies in a production system can significantly reduce lead times, albeit under the restrictive stochastic assumption of exponentially distributed production times. Ohta et al. [14] and Arreola-risa and Decroix [15] proposed conditions that make MTO and MTS policies optimal using a base-stock inventory policy. In these works, a single server queueing model represents the production system. Ohta et al. [14] also relied on results from queueing theory in their study of a multi-product inventory system. Demand arrives in accordance with a Poisson process, and the production times follow an Erlang distribution, as the authors noted that queueing theory does not provide explicit expressions for the queue size probabilities for general production time distributions. Their key result was an optimality condition that classifies products into MTS and MTO. Finally, Beemsterboer et al. [16] studied the optimal lot sizing when make-to-stock products are batch produced, where the system can only manufacture one product at a time (either MTS or MTO) and when a setup is required to switch between MTS and MTO.

In continuous make-to-stock operations, MTS production only halts when the inventory is full. When this happens, there can be a non-negligible setup time to restart production. If this is the case, starting production whenever there is space in the product inventory is inefficient. Efficiency can be attained by continuous review (s, S)-policies. For such a policy, replenishment starts when the inventory level drops to the threshold value s and stops when level S is attained. Such policies need constant monitoring of the inventory level, hence the term 'continuous review'. Most work conducted on (s, S)-policies assumes that any amount of inventory can be replenished all at once; see, e.g., [17–21] and the references therein. However, for real manufacturing systems, the production time to produce a single item is non-negligible, and it takes considerable time to replenish the inventory one item at a time. Motivated by this observation, different authors have also investigated hybrid MTS/MTO systems with replenishments on an item-by-item basis [22–26].

The present study differs from previous work on stochastic hybrid MTS/MTO systems by its inclusion of an order backlog. This modification considerably complicates performance assessments, as the corresponding stochastic model now consists of two coupled 'queues': a decoupling inventory and an order backlog. See, e.g., [27–30] for other applications of coupled queueing systems and [31–34] for Markovian systems that include both queueing and inventory management. The performance of this hybrid MTS/MTO system is assessed when the same production capacity is used for MTS and MTO. Such an assumption is for example natural in a job shop [4]. Other applications where the production system is shared by MTO and MTS operations include the aluminium profiles manufacturing industry [35] and the garments industry [36]. In our study, production alternates between MTS and MTO by a modified (s, S)-policy: production switches from MTS to MTO at inventory level S and switches from MTO to MTS when the inventory is empty or when the inventory is below level s and there are no outstanding orders. While the inclusion of the additional queue complicates the Markovian description of the system, we can exploit structural properties of the Markov model and solve the system with reasonable computational complexity by matrix-analytic methods. Furthermore, the stochastic assumptions are non-restrictive, including Markovian production and order arrivals. Specifically, the Markovian production process at hand allows for introducing phase-type distributed setup times and MTO and MTS production times. Specifically, the Markovian production process at hand allows for introducing phase-type distributed setup times and MTO and MTS production times (cf. Section 3).

The remainder of this paper is organised as follows. In Section 2, the Markovian hybrid MTS/MTO model is introduced and analysed. The balance equations are studied in Section 2.1 and solved in Section 2.2. Expressions for various performance measures of interest are found in Section 2.3, while Section 2.4 studies the MTO/MTS system in overload. To illustrate our approach, Section 3 considers some numerical examples. In particular, we study inventory levels, backlog sizes and lead times, as well as the optimal (s, S)-policy when holding cost and lead times are combined in a single cost function. Finally, conclusions are drawn in Section 4.

2. Queueing/Inventory Model

We consider a stochastic inventory model, which supports MTS and MTO operations. Figure 1 depicts the main components of the system. There is a product inventory, an order backlog, as well as MTO and MTS production units. The product inventory can store up to S semi-finished products, whereas the infinite-capacity order backlog keeps track of the orders that have not yet been delivered. Arriving orders are processed first-come-first-served. Each order takes a semi-finished product from the decoupling inventory and sends it to the order processing unit to complete the product in accordance with order specifications. Note that the two 'queues'—the product inventory and the order backlog—in the model at hand are tightly coupled. Departures from the inventory are only possible when there are backlogged orders. Similarly, departures from the order backlog are only possible if there are semi-finished products in the product inventory. There is also coupling between the MTS and MTO production units. We assume there is only a single production facility for MTS and MTO operations. Therefore, the production facility either works on finishing semi-finished products (MTO) or on producing semi-finished products (MTS). Production switches from MTO to MTS when the inventory is empty or when there are no backlogged orders and the inventory is at most at level s. Oppositely, production switches from MTS to MTO when the inventory level hits S.

We study the inventory model at hand in a continuous-time Markovian framework, which combines modelling versatility with computationally efficient analysis techniques. Orders arrive at the system in accordance with a Markov arrival process with state space $\mathcal{A} = \{1, \ldots, A\}$. Let α^1_{ij} and α^0_{ij} denote the transition rates from state i to state j with and without an order arrival, respectively, for $i, j \in \mathcal{A}$ and with $\alpha^0_{ii} = 0$ for $i \in \mathcal{A}$ as usual.

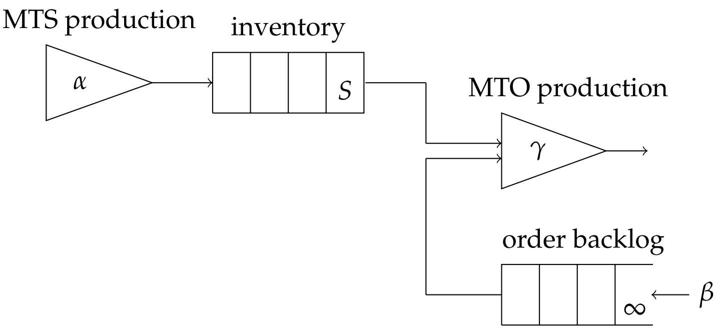

Figure 1. Generic inventory model. MTS, make-to-stock; MTO, make-to-order.

The MTS production process is modelled by a Markov process with finite state space $\mathcal{B} = \{1, \ldots, B\}$. Let β^1_{ij} and β^0_{ij} denote the transition rates from state i to state j with and without the completion of a semi-finished product, respectively, for $i, j \in \mathcal{B}$ and with $\beta^0_{ii} = 0$ for $i \in \mathcal{B}$. When the production system switches to MTS production, the MTS process starts in a random state, independent of the system state. Let p_i denote the probability that MTS production starts in state $i \in \mathcal{B}$. Note that the distribution of the initial state of the MTS production process can be chosen freely and typically differs from the stationary distribution of the Markov process (with transition rates $\beta^0_{ij} + \beta^1_{ij}$); see Section 3.

Furthermore, the MTO production process is modelled by a Markov process. As the production process is either making to order or making to stock, let $C = \{B + 1, \ldots, B + C\}$ denote the finite state space of the MTO production process. Further, let γ^1_{ij} and γ^0_{ij} denote the transition rates from state i to state j with and without the completion of an order for $i, j \in \mathcal{C}$. As for the MTS process, when the production system switches to MTO production, the MTO process starts in a random state. Let q_i denote the probability that MTO production starts in state $i \in \mathcal{C}$. The distribution of the initial state of the MTO production process can be again chosen freely and typically differs from the stationary distribution of the MTO Markov process. Finally, it is possible that MTO production temporarily stops, without switching to MTS. This happens when there are more than s semi-finished products in the inventory and there are no outstanding orders. In this case, the MTO production process is temporarily stopped: the state of the process does not change till there is a new order arrival if the production state is in any of the states in $\mathcal{C}_0 \subset \mathcal{C}$. Transitions from states in $\mathcal{C}_1 = \mathcal{C} \setminus \mathcal{C}_0$ are possible, but only transitions without order completions are allowed: $\gamma^1_{ij} = 0$ for $i \in \mathcal{C}_1$. This partition of the state space $\mathcal{C}$ offers the versatility to separate production from setup times and maintenance times in the MTO production process. A concrete example with setup times will be introduced in Section 3.

2.1. Balance Equations

In view of the assumptions above, the system at hand can be described by a four-dimensional continuous-time Markov chain. To be precise, the state of the inventory system at each point in time is described by a tuple (n, m, i, j) where $n \in \mathbb{N}$ is the number of backlogged orders, $m \in \mathcal{S} = \{0, \ldots, S\}$ denotes the inventory level, $i \in \mathcal{A}$ denotes the state of the order arrival process and $j \in \mathcal{B} \cup \mathcal{C}$ denotes the state of the production process. Noting that (i) production is MTO when the inventory is full, (ii) that production is MTS when the inventory is empty and (iii) that production is MTS when the inventory is at most s and there are no outstanding orders, the state space of the Markov process is,

$$\mathcal{X} = (\{0\} \times \{s+1, \ldots, S\} \times \mathcal{A} \times \mathcal{C}) \cup (\mathbb{N} \times (\mathcal{S} \setminus \{S\}) \times \mathcal{A} \times \mathcal{B}) \cup (\mathbb{N}^+ \times (\mathcal{S} \setminus \{0\}) \ \times \mathcal{A} \times \mathcal{C})\,,$$

with $\mathbb{N} = \{0, 1, 2, \cdots\}$ and $\mathbb{N}^+ = \{1, 2, \cdots\}$.

We now summarise the description of the inventory system at hand, by listing the possible transition rates from a fixed state $(n, m, i, j) \in \mathcal{X}$:

- Transitions related to the order arrival process: A new order arrives with rate $\alpha^1_{ii'}$, which invokes a transition to state $(n+1, m, i', j)$. The state of the order arrival process changes without arrivals with rate $\alpha^0_{ii'}$ invoking a transition to state (n, m, i', j).
- Transitions related to the MTS production process ($j \in \mathcal{B}$): Note that $j \in \mathcal{B}$ implies $m < S$ as there is no MTS production when the inventory is full. For $m < S$, there is a possible transition to state (n, m, i, j') with rate $\beta^0_{jj'}$ for $j' \in \mathcal{B}$. For $m < S - 1$, there is a possible transition to state $(n, m+1, i, j')$ with rate $\beta^1_{jj'}$ for $j' \in \mathcal{B}$. For $m = S - 1$, the production of a new semi-finished product induces a switch from MTS to MTO. Hence, we have a transition to state (n, S, i, k) for $k \in \mathcal{C}$ with rate:
$$\sum_{j' \in \mathcal{B}} \beta^1_{jj'} q_k \,.$$
- Transitions related to the MTO production process ($j \in \mathcal{C}$): Note that $j \in \mathcal{C}$ implies $m > 0$, as there is no MTO production when the inventory is empty. When $n = 0$, there is a possible transition to state (n, m, i, j') with rate $\gamma^0_{jj'}$ for $j \in \mathcal{C}_1$ and $j \in \mathcal{C}$. When $n > 0$, there is a possible transition to state (n, m, i, j') with rate $\gamma^0_{jj'}$ for $j' \in \mathcal{C}$. For $n > 1$ and $m > 1$ and for $n = 1$ and $m > s$, there is a possible transition to state $(n-1, m-1, i, j')$ with rate $\gamma^1_{jj'}$ for $j' \in \mathcal{C}$. Finally, for $n = 1$ and $m \leq s$ (no backlogged orders and inventory level below s) and for $n > 0$ and $m = 1$ (empty inventory), MTO switches to MTS upon completion of a product: there is a transition to state $(n-1, m-1, i, k)$ for $k \in \mathcal{B}$ with rate:
$$\sum_{j' \in \mathcal{C}} \beta^1_{jj'} p_k \,.$$

Let $\{\pi(m, n, i, j), (m, n, i, j) \in \mathcal{X}\}$ denote the invariant probability distribution of the Markov chain at hand. Accounting for the transmission rates defined above and assuming that the Markov process is ergodic (see below), we find that the invariant distribution satisfies the following set of balance equations:

$$\begin{aligned}
\pi(n, m, i, j)\chi(n, m, i, j) = & \sum_{k \in \mathcal{A}} \pi(n-1, m, k, j)\alpha^1_{ki} + \sum_{k \in \mathcal{A}} \pi(n, m, k, j)\alpha^0_{ki} \\
& + \mathbb{1}_{\{j \in \mathcal{B}\}} \sum_{k \in \mathcal{B}} \pi(n, m, i, k)\beta^0_{kj} + \mathbb{1}_{\{j \in \mathcal{B}\}} \sum_{k \in \mathcal{B}} \pi(n, m-1, i, k)\beta^1_{kj} \\
& + \mathbb{1}_{\{m = S\}} \sum_{k \in \mathcal{B}} \sum_{\ell \in \mathcal{B}} \pi(n, S-1, i, k)\beta^1_{k\ell} q_j + \mathbb{1}_{\{j \in \mathcal{C}, n=0\}} \sum_{k \in \mathcal{C}_1} \pi(n, m, i, k)\gamma^0_{kj} \\
& + \mathbb{1}_{\{j \in \mathcal{C}, n>0\}} \sum_{k \in \mathcal{C}} \pi(n, m, i, k)\gamma^0_{kj} + \mathbb{1}_{\{j \in \mathcal{C}\}} \sum_{k \in \mathcal{C}} \pi(n+1, m+1, i, k)\gamma^1_{kj} \\
& + \mathbb{1}_{\{j \in \mathcal{B}, m=0\}} \sum_{k \in \mathcal{C}} \sum_{\ell \in \mathcal{C}} \pi(n+1, 1, i, k)\gamma^1_{k\ell} p_j + \mathbb{1}_{\{j \in \mathcal{B}, s \geq m > 0, n=0\}} \sum_{k \in \mathcal{C}} \sum_{\ell \in \mathcal{C}} \pi(1, m+1, i, k)\gamma^1_{k\ell} p_j \,, \qquad (1)
\end{aligned}$$

with,

$$\chi(n, m, i, j) = \sum_{k \in \mathcal{A}} (\alpha^0_{ik} + \alpha^1_{ik}) + \mathbb{1}_{\{j \in \mathcal{B}\}} \sum_{k \in \mathcal{B}} (\beta^0_{jk} + \beta^1_{jk}) + \mathbb{1}_{\{j \in \mathcal{C}, n>0\}} \sum_{k \in \mathcal{C}} (\gamma^0_{jk} + \gamma^1_{jk}) + \mathbb{1}_{\{j \in \mathcal{C}_1, n=0\}} \sum_{k \in \mathcal{C}} \gamma^0_{jk} \,,$$

for $(n, m, i, j) \in \mathcal{X}$ and where we set $\pi(n, m, i, j) = 0$ for $(n, m, i, j) \notin \mathcal{X}$ to simplify notation. Here, $\mathbb{1}_{\{.\}}$ denotes the indicator function, which evaluates to one if its argument is true and to zero if this is not the case.

Let $\mathcal{X}_0$ and $\mathcal{X}_1$ denote the set of states of the inventory, the arrival and the production process when there are no orders and when there are outstanding orders, respectively:

$$\begin{aligned}
\mathcal{X}_0 &= (\{s+1, \ldots, S\} \times \mathcal{A} \times \mathcal{C}) \cup ((\mathcal{S} \setminus \{S\}) \times \mathcal{A} \times \mathcal{B}) \,, \\
\mathcal{X}_1 &= ((\mathcal{S} \setminus \{S\}) \times \mathcal{A} \times \mathcal{B}) \cup ((\mathcal{S} \setminus \{0\}) \ \times \mathcal{A} \times \mathcal{C}) \,.
\end{aligned}$$

These sets are finite and countable. In the remainder, we assume a fixed order when we iterate over the elements of these sets to construct vectors and matrices. We collect the invariant probabilities of states with the same order backlog sizes in row vectors $\boldsymbol{\pi}_0 = [\pi(0,m,i,j)]_{(m,i,j)\in\mathcal{X}_0}$ and $\boldsymbol{\pi}_n = [\pi(n,m,i,j)]_{(m,i,j)\in\mathcal{X}_1}$ $(n = 1,2,\ldots)$. The system of balance equations can then be rewritten as follows,

$$\boldsymbol{\pi}_0\mathbf{F}_0 + \boldsymbol{\pi}_1\mathbf{F}_1 = 0\,, \quad \boldsymbol{\pi}_0\widehat{\mathbf{G}}_0 + \boldsymbol{\pi}_1\mathbf{G}_1 + \boldsymbol{\pi}_2\mathbf{G}_2 = 0\,, \quad \boldsymbol{\pi}_{n-1}\mathbf{G}_0 + \boldsymbol{\pi}_n\mathbf{G}_1 + \boldsymbol{\pi}_{n+1}\mathbf{G}_2 = 0\,, \tag{2}$$

where the matrices:

$$\mathbf{F}_0 = [f_0(m,i,j|m',i',j')]_{(m',i',j')\in\mathcal{X}_0,(m,i,j)\in\mathcal{X}_0}\,, \quad \mathbf{F}_1 = [f_1(m,i,j|m',i',j')]_{(m',i',j')\in\mathcal{X}_1,(m,i,j)\in\mathcal{X}_0}\,,$$

collect the transition rates,

$$\begin{aligned}
f_0(m,i,j|m',i',j') &= \mathbb{1}_{\{m=m'\}}\mathbb{1}_{\{j=j'\}}\alpha^0_{i'i} + \mathbb{1}_{\{m=m'\}}\mathbb{1}_{\{i=i'\}}\mathbb{1}_{\{j\in\mathcal{B}\}}\mathbb{1}_{\{j'\in\mathcal{B}\}}\beta^0_{j'j} \\
&+ \mathbb{1}_{\{m'=m-1\}}\mathbb{1}_{\{i=i'\}}\mathbb{1}_{\{j\in\mathcal{B}\}}\mathbb{1}_{\{j'\in\mathcal{B}\}}\beta^1_{j'j} \\
&+ \mathbb{1}_{\{m=S\}}\mathbb{1}_{\{m'=S-1\}}\mathbb{1}_{\{i=i'\}}\mathbb{1}_{\{j\in\mathcal{C}\}}\mathbb{1}_{\{j'\in\mathcal{B}\}}\sum_{\ell\in\mathcal{B}}\beta^1_{j'\ell}q_j \\
&+ \mathbb{1}_{\{m=m'\}}\mathbb{1}_{\{i=i'\}}\mathbb{1}_{\{j'\in\mathcal{C}_1\}}\mathbb{1}_{\{j\in\mathcal{C}\}}\gamma^0_{j'j} - \mathbb{1}_{\{m=m'\}}\mathbb{1}_{\{i=i'\}}\mathbb{1}_{\{j=j'\}}\sum_{k\in\mathcal{A}}(\alpha^0_{ik} + \alpha^1_{ik}) \\
&- \mathbb{1}_{\{m=m'\}}\mathbb{1}_{\{i=i'\}}\mathbb{1}_{\{j=j'\}}\mathbb{1}_{\{j\in\mathcal{B}\}}\sum_{k\in\mathcal{B}}(\beta^0_{jk} + \beta^1_{jk}) - \mathbb{1}_{\{m=m'\}}\mathbb{1}_{\{i=i'\}}\mathbb{1}_{\{j=j'\}}\mathbb{1}_{\{j\in\mathcal{C}_1\}}\sum_{k\in\mathcal{C}}\gamma^0_{jk}\,, \\
f_1(m,i,j|m',i',j') &= \mathbb{1}_{\{m'=m+1\}}\mathbb{1}_{\{i=i'\}}\mathbb{1}_{\{j\in\mathcal{C}\}}\mathbb{1}_{\{j'\in\mathcal{C}\}}\gamma^1_{j'j} + \mathbb{1}_{\{m=0\}}\mathbb{1}_{\{m'=1\}}\mathbb{1}_{\{i=i'\}}\mathbb{1}_{\{j\in\mathcal{B}\}}\mathbb{1}_{\{j'\in\mathcal{C}\}}\sum_{\ell\in\mathcal{C}}\gamma^1_{j'\ell}p_j \\
&+ \mathbb{1}_{\{s\geq m>0\}}\mathbb{1}_{\{m'=m+1\}}\mathbb{1}_{\{i=i'\}}\mathbb{1}_{\{j\in\mathcal{B}\}}\mathbb{1}_{\{j'\in\mathcal{C}\}}\sum_{\ell\in\mathcal{C}}\gamma^1_{j'\ell}p_j\,,
\end{aligned}$$

and where the matrices:

$$\widehat{\mathbf{G}}_0 = [g_0(m,i,j|m',i',j')]_{(m',i',j')\in\mathcal{X}_0,(m,i,j)\in\mathcal{X}_1}\,, \quad \mathbf{G}_i = [g_i(m,i,j|m',i',j')]_{(m',i',j')\in\mathcal{X}_1,(m,i,j)\in\mathcal{X}_1}\,,$$

for $i \in \{0,1,2\}$ have elements,

$$\begin{aligned}
g_0(m,i,j|m',i',j') &= \mathbb{1}_{\{m=m'\}}\mathbb{1}_{\{j=j'\}}\alpha^1_{i'i}\,, \\
g_1(m,i,j|m',i',j') &= \mathbb{1}_{\{m=m'\}}\mathbb{1}_{\{j=j'\}}\alpha^0_{i'i} + \mathbb{1}_{\{m'=m\}}\mathbb{1}_{\{i'=i\}}\mathbb{1}_{\{j\in\mathcal{B}\}}\mathbb{1}_{\{j'\in\mathcal{B}\}}\beta^0_{j'j} \\
&+ \mathbb{1}_{\{m'=m-1\}}\mathbb{1}_{\{i'=i\}}\mathbb{1}_{\{j\in\mathcal{B}\}}\mathbb{1}_{\{j'\in\mathcal{B}\}}\beta^1_{j'j} \\
&+ \mathbb{1}_{\{m=S\}}\mathbb{1}_{\{m'=S-1\}}\mathbb{1}_{\{i'=i\}}\mathbb{1}_{\{j\in\mathcal{C}\}}\mathbb{1}_{\{j'\in\mathcal{B}\}}\sum_{\ell\in\mathcal{B}}\beta^1_{j'\ell}q_j + \mathbb{1}_{\{m'=m\}}\mathbb{1}_{\{i=i'\}}\mathbb{1}_{\{j\in\mathcal{C}\}}\mathbb{1}_{\{j'\in\mathcal{C}\}}\gamma^0_{j'j} \\
&- \mathbb{1}_{\{m=m'\}}\mathbb{1}_{\{i=i'\}}\mathbb{1}_{\{j=j'\}}\sum_{k\in\mathcal{A}}(\alpha^0_{ik} + \alpha^1_{ik}) \\
&- \mathbb{1}_{\{m=m'\}}\mathbb{1}_{\{i=i'\}}\mathbb{1}_{\{j=j'\}}\left(\mathbb{1}_{\{j\in\mathcal{B}\}}\sum_{k\in\mathcal{B}}(\beta^0_{jk} + \beta^1_{jk}) + \mathbb{1}_{\{j\in\mathcal{C},m>0\}}\sum_{k\in\mathcal{C}}(\gamma^0_{jk} + \gamma^1_{jk})\right)\,, \\
g_2(m,i,j|m',i',j') &= \mathbb{1}_{\{m'=m+1\}}\mathbb{1}_{\{i'=i\}}\mathbb{1}_{\{j\in\mathcal{C}\}}\mathbb{1}_{\{j'\in\mathcal{C}\}}\gamma^1_{j'j} \\
&+ \mathbb{1}_{\{m=0\}}\mathbb{1}_{\{m'=1\}}\mathbb{1}_{\{i'=i\}}\mathbb{1}_{\{j\in\mathcal{B}\}}\mathbb{1}_{\{j'\in\mathcal{C}\}}\sum_{\ell\in\mathcal{C}}\gamma^1_{j'\ell}p_j\,.
\end{aligned}$$

2.2. *Quasi-Birth-Death Process*

From Equation (2), it is easily seen that the Markov process under consideration is a homogeneous quasi-birth-and-death process (QBD); see [37]. A QBD is a Markov process, where the state is described by a level and a phase. The phase can only take a finite number of values, and transitions between any two phases are allowed. In contrast, the level is a non-negative integer number, and only transitions between adjacent levels are allowed. In the present setting, we identify the level with the number of backlogged orders while the phase indicates both the content of the decoupling inventory, the states of the arrival process and the state production processes. The balance equations now immediately show that transitions are indeed restricted to states in the same level (from state $(n, *, *, *)$ to state $(n, *, *, *)$) or in two adjacent levels (from state $(n, *, *, *)$ to state $(n + 1, *, *)$ or state $(n - 1, *, *, *)$). Intuitively, this is also clear: orders arrive and are processed one-by-one such that the order backlog increases and decreases in unit steps.

The QBD is ergodic and therefore admits an invariant distribution provided that the drift towards the level down exceeds the drift towards the level up. This drift condition can be expressed as follows,

$$\boldsymbol{\omega}\mathbf{G}_2\mathbf{e}_1' > \boldsymbol{\omega}\mathbf{G}_0\mathbf{e}_1' , \tag{3}$$

where $\mathbf{e}_1 = [1]_{(m,i,j)\in\mathcal{X}_1}$ is a row vector of ones, where $\mathbf{e}'$ is the transpose of $\mathbf{e}$ and where $\boldsymbol{\omega}$ is the normalised solution of:

$$\boldsymbol{\omega}(\mathbf{G}_0 + \mathbf{G}_1 + \mathbf{G}_2) = 0 .$$

In general, the ergodicity of the system depends on the inventory policy, e.g., the modelling assumptions allow for introducing setup times (see below) when production switches from MTS to MTO or vice versa. In this case, the stability rule from queueing theory that states that the load should not exceed the service capacity depends on the switching rates between MTS and MTO, which in turn depend on the inventory policy as some production capacity is lost on setup times. In some instances, however, a simple condition for the ergodicity of the process can be found. For example, when the MTO and MTS production times constitute sequences of independent random variables and switching between MTS and MTO is immediate, the ergodicity condition can be expressed as follows,

$$\rho(m_p^{\mathrm{mts}} + m_p^{\mathrm{mto}}) < 1 ,$$

where ρ denotes the average number of order arrivals per time unit and where m_p^{mts} and m_p^{mto} denote the mean MTS and MTO production times.

Assuming that the stability condition (3) holds, a well-known method for finding the stationary distribution of QBD processes is the matrix-geometric method. The method proposes the following solution of the set of balance Equation (2),

$$\boldsymbol{\pi}_k = \boldsymbol{\pi}_1\mathbf{R}^{k-1} ,$$

where the so-called rate matrix $\mathbf{R}$ is the minimal non-negative solution of the non-linear matrix equation:

$$\mathbf{R}^2\mathbf{G}_2 + \mathbf{R}\mathbf{G}_1 + \mathbf{G}_0 = \mathbf{0} . \tag{4}$$

The rate matrix $\mathbf{R}$ records the rates of sojourn in states at a level ℓ per unit of the local time of the preceding level $\ell - 1$. This can be re-expressed as $\mathbf{R} = \mathbf{G}_0\mathbf{N}$, where the matrix $\mathbf{N}$ records the expected sojourn times at a level ℓ before the first visit to the lower level $\ell - 1$ if one starts at level ℓ; see, e.g., ([37], section 6.4). Several iterative procedures exist for solving Equation (4) above. For example, Gun [38] uses the following simple recursion:

$$\mathbf{R} \leftarrow -(\mathbf{G}_0 + \mathbf{R}^2\mathbf{G}_2)\mathbf{G}_1^{-1} .$$

In the numerical section, we compute the rate matrix by implementing the improved iterative algorithm of [37] (Chapter 8, pp. 179–187). Once the rate matrix has been determined, the vectors $\boldsymbol{\pi}_0$ and $\boldsymbol{\pi}_1$ can be found by solving,

$$\boldsymbol{\pi}_0\mathbf{F}_0 + \boldsymbol{\pi}_1\mathbf{F}_1 = 0 , \quad \boldsymbol{\pi}_0\widehat{\mathbf{G}}_0 + \boldsymbol{\pi}_1\mathbf{G}_1 + \boldsymbol{\pi}_1\mathbf{R}\mathbf{G}_2 = 0 , \quad \boldsymbol{\pi}_0\mathbf{e}_0' + \boldsymbol{\pi}_1(\mathbf{I} - \mathbf{R})^{-1}\mathbf{e}_1' = 1 ,$$

where $\mathbf{e}_0 = [1]_{(m,i,j)\in\mathcal{X}_0}$ is again a row vector of ones. After some simple matrix manipulations, we find,

$$\begin{aligned}
\boldsymbol{\pi}_0 &= -\boldsymbol{\pi}_1\mathbf{F}_1\mathbf{F}_0^{-1}\,,\\
\boldsymbol{\pi}_1 &= \mathbf{e}_1(-\mathbf{F}_1\mathbf{F}_0^{-1}\widehat{\mathbf{G}}_0 + \mathbf{G}_1 + \mathbf{R}\mathbf{G}_2 - \mathbf{F}_1\mathbf{F}_0^{-1}\mathbf{e}_0'\mathbf{e}_1 + (\mathbf{I}-\mathbf{R})^{-1}\mathbf{e}_1'\mathbf{e}_1)^{-1}\,.
\end{aligned}$$

Here, the irreducibility of the Markov process implies that $\mathbf{F}_0$ is invertible.

2.3. Performance Measures

We now express a number of interesting performance measures for the decoupling inventory system in terms of $\boldsymbol{\pi}_0$, $\boldsymbol{\pi}_1$ and $\mathbf{R}$. The marginal probability mass function of the content of the number of orders in the order backlog equals,

$$\pi^{(o)}(0) = \boldsymbol{\pi}_0\mathbf{e}_0'\,, \quad \pi^{(o)}(n) = \boldsymbol{\pi}_n\mathbf{e}_1' = \boldsymbol{\pi}_1\mathbf{R}^{n-1}\mathbf{e}_1'\,,$$

for $n = 1, 2, \ldots$, while the marginal probability mass function of the inventory equals,

$$\pi^{(p)}(m) = \boldsymbol{\pi}_0\mathbf{h}_{0,m}' + \boldsymbol{\pi}_1(\mathbf{I}-\mathbf{R})^{-1}\mathbf{h}_{1,m}'\,,$$

for $m = 0, \ldots, S$, with $\mathbf{h}_{0,m} = [\mathbb{1}_{\{m'=m\}}]_{(m',i',j')\in\mathcal{X}_0}$ and $\mathbf{h}_{1,m} = [\mathbb{1}_{\{m'=m\}}]_{(m',i',j')\in\mathcal{X}_1}$ and where $\mathbf{I}$ is the identity matrix. We can further express the mean and the variance of the number of orders in the order backlog as follows,

$$\begin{aligned}
\mathsf{E}[Q_o] &= \sum_{n=1}^{\infty} n\boldsymbol{\pi}_1\mathbf{R}^{n-1}\mathbf{e}_1' = \boldsymbol{\pi}_1(\mathbf{I}-\mathbf{R})^{-2}\mathbf{e}_1'\,,\\
\mathsf{Var}[Q_o] &= \sum_{n=1}^{\infty} \pi^{(o)}(n)n^2 - \mathsf{E}[Q_o]^2 = \boldsymbol{\pi}_1(\mathbf{I}-\mathbf{R})^{-3}\mathbf{e}_1' + \mathsf{E}[Q_0] - \mathsf{E}[Q_o]^2\,.
\end{aligned}$$

Moreover, as the inventory is finite, the mean and variance of the inventory content can be simply calculated as follows,

$$\begin{aligned}
\mathsf{E}[Q_p] &= \sum_{m=1}^{S} \pi^{(p)}(m)m\,,\\
\mathsf{Var}[Q_p] &= \sum_{m=1}^{S} \pi^{(p)}(m)m^2 - \mathsf{E}[Q_p]^2\,.
\end{aligned}$$

We can further calculate the mean lead time by Little's result. The mean lead time LT is the average amount of time between the placement of an order and the completion of a finished product,

$$\mathsf{LT} = \frac{\mathsf{E}[Q_o]}{\eta}\,,$$

with η the order arrival rate,

$$\eta = \boldsymbol{\theta}\mathbf{A}_1\mathbf{e}_a'$$

Here, $\mathbf{e}_a$ is a row vector with A ones and $\boldsymbol{\theta}$ is the unique normalised solution of $\boldsymbol{\theta}(\mathbf{A}_0 + \mathbf{A}_1) = 0$.

Finally, we also calculated the switching rate ξ between MTO and MTS or, equivalently, the switching rate between MTS and MTO. Both are of course identical as each switch to MTO is followed by a switch to MTS. As a switch from MTS to MTO occurs when the inventory level reaches S, we find:

$$\xi = \sum_{n=0}^{\infty}\sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{B}}\sum_{k\in\mathcal{B}} \pi(n, S-1, i, j)\beta_{jk}^{1} = \boldsymbol{\pi}_0\mathbf{v}_0' + \boldsymbol{\pi}_1(\mathbf{I}-\mathbf{R})^{-1}\mathbf{v}_1'\,,$$

with $\mathbf{v}_0 = [\mathbb{1}_{\{m=S-1\}}\mathbb{1}_{\{j\in\mathcal{B}\}}\sum_{k\in\mathcal{B}}\beta^1_{jk}]_{(m,i,j)\in\mathcal{X}_0}$ and $\mathbf{v}_1 = [\mathbb{1}_{\{m=S-1\}}\mathbb{1}_{\{j\in\mathcal{B}\}}\sum_{k\in\mathcal{B}}\beta^1_{jk}]_{(m,i,j)\in\mathcal{X}_1}$.

2.4. Overload Analysis

When the arrival load of new orders exceeds the production capacity of the MTO/MTS system, the order queue grows unbounded. This particularly implies that there always will be outstanding orders for the MTO/MTS system. In this case, the state of the production process and the inventory size can be studied in isolation, without accounting for the state of the arrival process or the state of the order backlog. The balance Equation (1) then considerably simplifies. With a slight abuse of notation, let $\pi(m, j)$ denote the invariant probability to have inventory level m and production state j. These probabilities can then be found by solving the following set of balance equations,

$$\begin{aligned}\pi(m,j)&\left(\mathbb{1}_{\{j\in\mathcal{B}\}}\sum_{k\in\mathcal{B}}(\beta^0_{jk}+\beta^1_{jk})+\mathbb{1}_{\{j\in\mathcal{C}\}}\sum_{k\in\mathcal{C}}(\gamma^0_{jk}+\gamma^1_{jk})\right)\\ &=\mathbb{1}_{\{j\in\mathcal{B}\}}\sum_{k\in\mathcal{B}}\pi(m,k)\beta^0_{kj}+\mathbb{1}_{\{j\in\mathcal{B}\}}\sum_{k\in\mathcal{B}}\pi(m-1,k)\beta^1_{kj}\\ &+\mathbb{1}_{\{j\in\mathcal{C}\}}\sum_{k\in\mathcal{C}}\pi(m,k)\gamma^0_{kj}+\mathbb{1}_{\{j\in\mathcal{C}\}}\sum_{k\in\mathcal{C}}\pi(m+1,k)\gamma^1_{kj}\\ &+\mathbb{1}_{\{m=S\}}\sum_{k\in\mathcal{B}}\sum_{\ell\in\mathcal{B}}\pi(S-1,k)\beta^1_{k\ell}q_j+\mathbb{1}_{\{j\in\mathcal{B},m=0\}}\sum_{k\in\mathcal{C}}\sum_{\ell\in\mathcal{C}}\pi(n+1,1,i,k)\gamma^1_{k\ell}p_j\,, \qquad (5)\end{aligned}$$

for $(m, j) \in \mathcal{X}_2 = (\mathcal{S} \setminus \{S\}) \times \mathcal{B} \cup (\mathcal{S} \setminus \{0\}) \times \mathcal{C}$. The size of this system of equations is considerably smaller and is easily solved numerically. As usual, the system of equations determines the probabilities $\pi(m, j)$ up to a normalising constant, which follows from the normalisation condition:

$$\sum_{(m,j)\in\mathcal{X}_2}\pi(m,j)=1\,.$$

The mean and variance of the inventory size and the MTO/MTS switching rate can then be expressed in terms of these probabilities as follows,

$$\mathsf{E}[Q_p]=\sum_{(m,j)\in\mathcal{X}_2}\pi(m,j)m\,,\quad \mathsf{Var}[Q_p]=\sum_{(m,j)\in\mathcal{X}_2}\pi(m,j)m^2-\mathsf{E}[Q_p]^2\,,\quad \xi=\sum_{j\in\mathcal{B}}\sum_{k\in\mathcal{B}}\pi(S-1,j)\beta^1_{jk}\,.$$

Finally, note that the parameter s does not appear in the system of Equation (5) above. Hence, for increasing load, one can expect that the influence of the parameter s on the system's performance measures disappears. This is not unexpected. The threshold s determines switching to MTS when the order backlog is empty, which never occurs in overload.

3. Numerical Results

We now illustrate our approach by means of some numerical examples. In particular, we focus on a system where the MTO and MTS production times constitute sequences of independent and identically phase-type distributed random variables and where a PH-distributed setup time is required for switching from MTO to MTS and back. Moreover, new orders arrive in accordance with an interrupted Poisson process.

To limit the number of parameters, we further assume that all PH-type distributions (for the production and setup times) are generalised Erlang(2) distributions. The generalised Erlang(2) distribution is the distribution of the sum of two exponential random variables and therefore completely characterised by the rates of these exponential distributions. We write GE(α_1, α_2) for the generalised

Erlang distribution with rates α_1 and α_2. For values $m > 0$ and $s^2 > 0$, such that $0.5 \le s^2/m^2 < 1$, let x_1 and x_2 be the (positive) solutions of the quadratic equation:

$$x^2 + (m - x)^2 - s^2 = 0\,.$$

Then, $\mathrm{GE}(1/x_1, 1/x_2)$ is a generalised Erlang distribution with mean m and variance s^2. See also Telek and Heindl [39] who discussed the moment characterisation of the more general ACPH(2)distribution, which allows for higher variances without introducing additional states.

Let the setup time of MTO be $\mathrm{GE}(\phi_1^o, \phi_2^o)$, and let the MTO production times be $\mathrm{GE}(\psi_1^o, \psi_2^o)$. Similarly, let the setup time of MTS be $\mathrm{GE}(\phi_1^s, \phi_2^s)$, and let the MTS production times be $\mathrm{GE}(\psi_1^s, \psi_2^s)$. The corresponding MTO and MTS arrival processes then each have four states (two setup states and two production states), such that the production process has eight states in total. The transition probabilities of these production processes can then be expressed in terms of the parameters of the different GE distributions as follows. Assume that States 1 and 2 (States 3 and 4) correspond to the setup time (production time) of MTO, while States 5 and 6 (States 7 and 8) correspond to the setup time (production time) of MTS. We then have that all transition rates of the production process are zero apart from the following non-zero transition rates,

$$\begin{array}{llll} \beta_{12}^0 = \phi_1^o\,, & \beta_{23}^0 = \phi_1^o\,, & \beta_{34}^0 = \psi_1^o\,, & \beta_{43}^1 = \psi_1^o \\ \gamma_{56}^0 = \phi_1^s\,, & \gamma_{67}^0 = \phi_1^s\,, & \gamma_{78}^0 = \psi_1^s\,, & \gamma_{87}^1 = \psi_1^s\,. \end{array}$$

Moreover, the MTO and MTS production processes start in production States 1 ($p_1 = 1$) and 5 ($q_5 = 1$), respectively.

As mentioned above, we model order arrivals by an interrupted Poisson process, which allows for assessing the impact of correlation in the order arrival process. The interrupted Poisson process is a two-state Markov arrival process. There is an active state (say State 1) during which new orders arrive in accordance with a Poisson process with rate λ. In contrast, there are no arrivals when it is in its inactive state (say State 2). Let a and b denote the transition rates from the active to the inactive state and from the inactive to the active state, respectively. Hence, we have that all transition rates of the arrival process are zero apart from the following non-zero transition rates,

$$\alpha_{12}^0 = a\,, \quad \alpha_{21}^0 = b\,, \quad \alpha_{11}^1 = \lambda\,.$$

For convenience, we replace the parametrisation (a, b, λ) of the interrupted Poisson process by the more intuitive parametrisation (σ, κ, ρ), with:

$$\sigma = \frac{b}{a+b}\,, \quad \kappa = \frac{1}{a} + \frac{1}{b}\,, \quad \rho = \lambda\sigma\,.$$

Here, σ denotes the fraction of time the interrupted Poisson process is active, while the absolute time parameter κ is the average time between the start of two consecutive active periods and ρ is the arrival load of the orders.

We are now ready to explore the impact of the parameters on the various performance measures. To this end, we consider the base case of Table 1 with m_s^{mts} and v_s^{mts} the mean and variance of the MTS setup times. Similar notation is used for the mean and variance of the production times (with subscript p) and for the corresponding quantities for MTO (with superscript mto). In the numerical examples below, we only mention the parameters that deviate from this base case. It is implicit that parameter values that are not mentioned are taken from the base case. The parameters of the base case correspond to a system load of 63%.

Table 1. Parameters for the base case used in the numerical examples.

Inventory	Orders	MTO	MTS
$s = 5$	$\sigma = 0.2$	$m_s^{\text{mto}} = 2$	$m_s^{\text{mts}} = 2$
$S = 20$	$\kappa = 10$	$v_s^{\text{mto}} = 2$	$v_s^{\text{mts}} = 2$
	$\rho = 0.15$	$m_p^{\text{mto}} = 2$	$m_p^{\text{mts}} = 2$
		$v_p^{\text{mto}} = 3$	$v_p^{\text{mts}} = 3$

The left pane of Figure 2 studies the impact of the load on the system's performance. In particular, the mean inventory level $\mathsf{E}[Q_p]$, the mean backlog size $\mathsf{E}[Q_o[$ and the mean lead time LT are depicted vs. the order arrival load ρ for different values of the threshold s. In line with expectations, the mean backlog size and lead time increase when ρ increases. When there are more orders, it is more likely that there are many uncompleted orders when a new order is placed. In contrast, the mean product inventory mostly decreases when the load increases. However, for $s = 1$, the mean product inventory first decreases and then again increases with increasing load. More load means that more semi-finished products are produced, but also that more semi-finished are made to order. It is hence hard to predict its influence on the mean inventory level. The effect of changing the threshold s is easier to explain. For higher s, production more easily switches from MTO to MTS. Therefore, it is not unexpected that the mean inventory level increases for higher s. Larger values of s further lead to lower values of the mean order backlog and the mean lead time. This can be explained by noting that having more semi-finished products ready lowers the backlog and lead times. Finally, note that for higher load, the influence of the parameter s disappears. This is in line with our findings in Section 2.4 where it is noted that performance measures do not depend on s in overload situations.

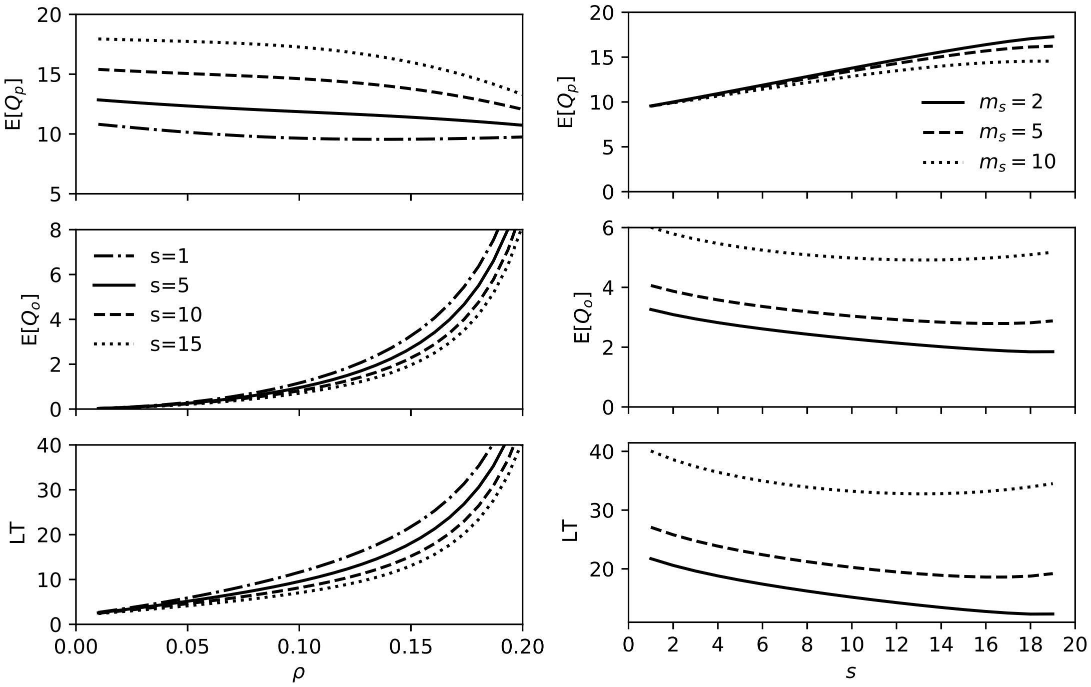

Figure 2. Mean inventory level, mean backlog and mean lead times vs. the order arrival load for different values of the threshold s (**left**) and vs. the threshold s for different values of the mean MTO/MTS switching time m_s (**right**).

The influence of the threshold s on the performance is further explored in the right pane of Figure 2. The mean inventory level $\mathsf{E}[Q_p]$, the mean backlog size $\mathsf{E}[Q_o]$ and the mean lead time LT are depicted vs. the threshold s for different values of the mean MTO/MTS setup time m_s. The mean setup times from MTO to MTS and MTS to MTO are assumed to be Erlang(2) distributed with the same mean value $m_s = m_s^{\mathrm{mto}} = m_s^{\mathrm{mts}}$ (and with variance $v_s = m_s^2/2$). In line with the observations above, the mean inventory level increases when s increases. Moreover, longer setup times lead to lower inventory levels. This is not unexpected; it is less likely that there are no outstanding orders as the system's load is higher as a larger part of the production capacity is lost on setup. The threshold s affects the mean order backlog and the mean lead time as well. For increasing s, these quantities first decrease and then increase again. The shape of these curves can be explained by two opposing effects. For higher s, the system more easily switches when there are no orders, such that it is more likely that there are semi-finished products ready when orders arrive. However, switching incurs also a cost as some production capacity is lost on setups.

To study the effect of order arrival correlation, the mean $\mathsf{E}[Q_p]$ and variance $\mathsf{Var}[Q_p]$ of the inventory level, the mean $\mathsf{E}[Q_o]$ and variance $\mathsf{Var}[Q_o]$ of the order backlog size, the mean lead time LT and the MTO/MTS switching rate ξ are depicted vs. the arrival parameter σ for different values of the threshold s in Figure 3. Recall that the parameter σ denotes the fraction of time with order arrivals. Hence, for a fixed load, low σ means that the order arrivals are more concentrated in time compared to higher values of σ. In line with findings from queueing theory, we find that the mean and variance of the order backlog and the mean lead time are negatively affected by arrival correlation. Indeed, these values increase when σ decreases. The variance of the inventory level also decreases for increasing σ, while, depending on the threshold s, the mean inventory level either increases or decreases when there is more correlation. More correlation means that there are longer periods without orders during which the inventory level is high, but this also means there is less overhead from switching between MTS and MTO. The latter is clearly confirmed by the plot of the switching rate, which shows that there is less switching when there is more correlation.

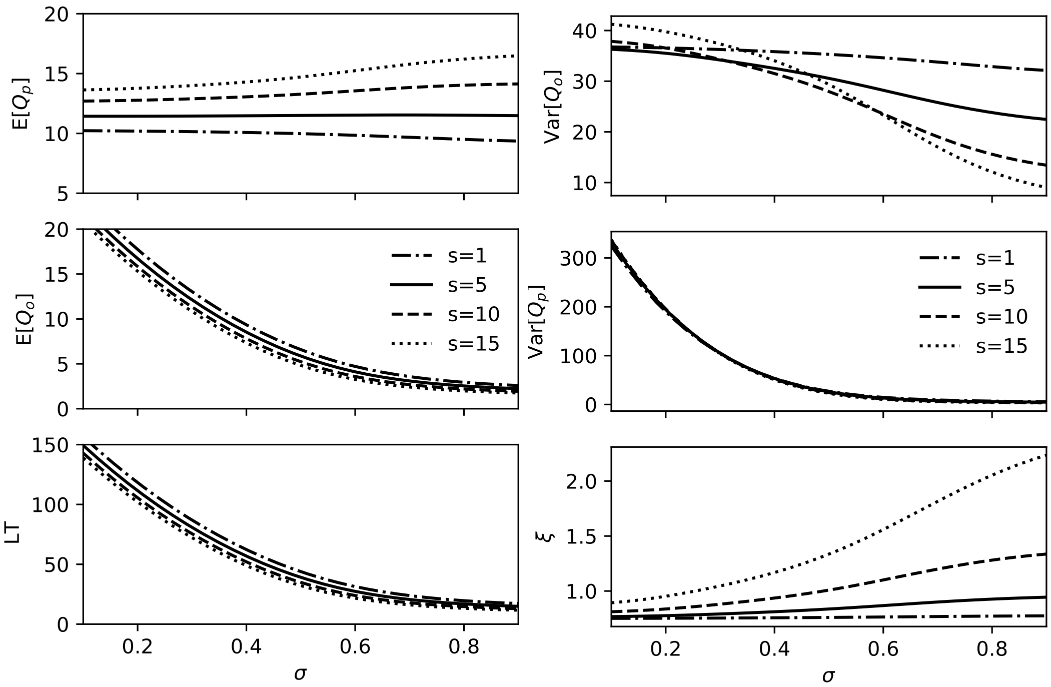

Figure 3. Mean of the order backlog, the inventory level and the lead time vs. the arrival parameter σ for different values of the threshold s (**left**) and variance of the inventory level and order backlog and setup rate vs. the arrival parameter σ for different values of the threshold s (**right**).

The same performance measures are also considered in Figure 4. These measures are plotted for different values of the threshold s vs. the inventory capacity S, for $S \geq s$. Both the mean and variance of the inventory level increase for increasing S. This is not unexpected, S being the inventory level when MTS switches back to MTO. However, larger S are not necessarily beneficial for the mean and variance of the order backlog and for the mean lead time. If S is small, increasing S will lead to less switching (and setup) such that the mean and variance of the order backlog and the mean lead time decrease. However, if one increases S further, the cost of switching is already low, while the MTS operation interrupts the MTO operation for longer times. As the backlog typically increases during the MTO operation, one finds larger backlogs and longer lead times.

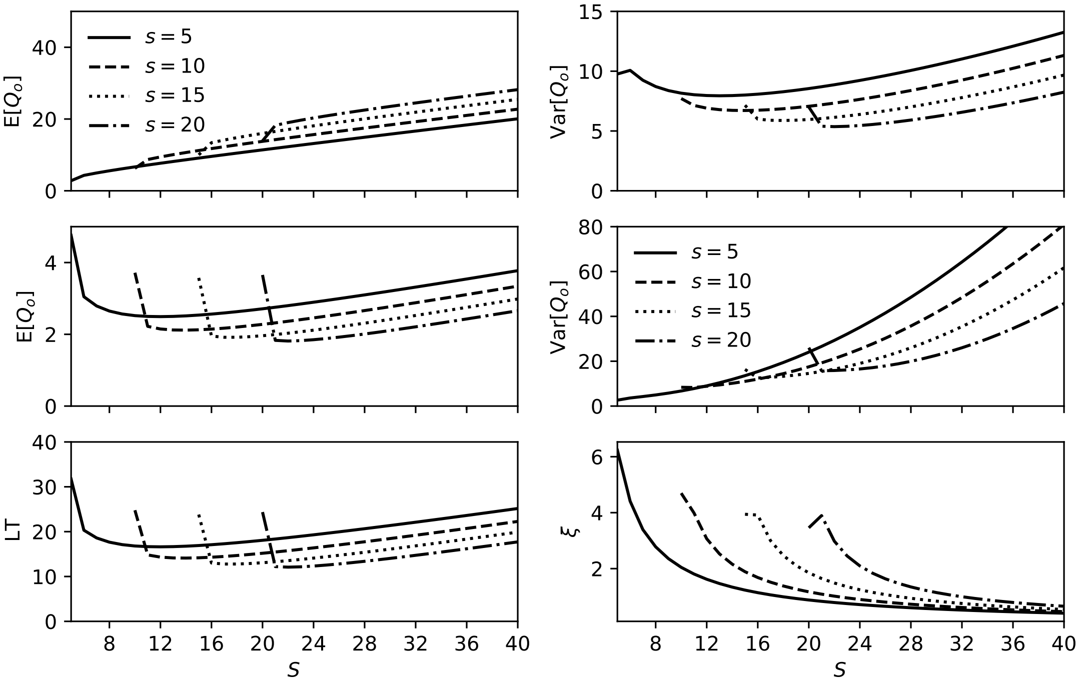

Figure 4. Mean of the order backlog, the inventory level and the lead time vs. the inventory size S for different values of the threshold s (**left**) and variance of the inventory level and order backlog and setup rate vs. the inventory size S for different values of the threshold s (**right**).

Finally, we study the optimal (s, S)-policy, accounting for both holding costs and lead times. To be more precise, we seek the optimal strategy that minimises the cost C_ν,

$$C_\nu = \mathsf{E}[Q_p] + \nu \mathsf{LT}\,.$$

Here, the parameter ν determines the relative weight of the holding cost and the lead time. Figure 5 shows the optimal s and S vs. the cost parameter ν. The left pane considers the optimal policy for different values of the setup time m_s as indicated (we again assume Erlang(2) setup times as in Figure 2). In contrast, the right pane depicts the optimal policy for different values of the load ρ. Increasing ν corresponds to shifting weight from the holding cost to the lead times. As holding costs are lower for smaller s and S, it is expected that s and S increase with ν. For large ν, the holding cost only marginally affects the cost C_ν, so that the policy converges to the policy that only optimises the lead time. Moreover, if the mean setup time increases, it is optimal to keep more inventory. Long setup times make switching between MTO and MTS expensive, while keeping more inventory reduces the need for switching. Finally, in line with expectations, it is also beneficial to keep more inventory if there are more orders.

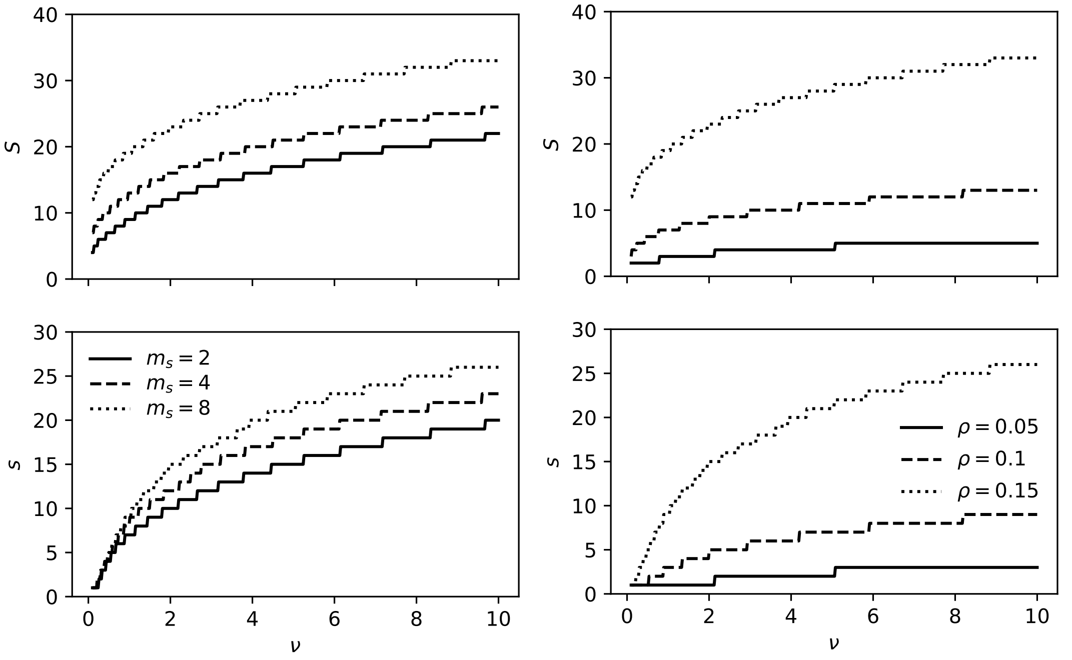

Figure 5. Optimal (s, S)-policy vs. the cost weight parameter ν for different mean setup times m_s (**left**) and for different load ρ (**right**) as indicated.

4. Conclusions

We introduced an analytically tractable stochastic model for studying the performance of a hybrid make-to-stock/make-to-order system. Our model includes two "queues": the semi-finished product inventory and the order backlog. We rely on matrix-analytic solution techniques to evaluate the performance of this system. Our approach allows accounting for uncertainty in demand and MTO and MTS setup and production times under non-restrictive stochastic assumptions. Switching between MTO and MTS operations is governed by an (s, S)-type policy. By some numerical examples, we show how one can find the optimal policy that balances holding costs and lead times.

The model at hand can be used to support MTO/MTS decisions in production environments. For an accurate performance assessment, it is however essential that the stochastic characteristics of the production processes and of the demand are accurately estimated. The statistical inference of the model parameters can either be based on prior measurements or can be learned while the production system is running using an exploration-exploitation approach [40,41], and this will be considered in future work.

We finally mention some alternative hybrid MTO/MTS systems that can be studied by a similar methodological approach: (i) systems with dedicated MTS and MTO production, (ii) systems with multiple production units, either dedicated or units that can switch between MTS and MTO, and (iii) systems where production capacity can be arbitrarily divided between MTS and MTO.

Author Contributions: Conceptualization and methodology, E.D.C., K.D.T., D.F.; formal analysis, all authors; writing—original draft preparation, all authors; writing—review and editing, D.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by UGent-BOF under grant number B/11010-BOF10/STA/043.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hoekstra, S.; Romme. *Integral Logistic Structures: Developing Customer-Oriented Goods Flow*; McGraw-Hill: New York, NY, USA, 1992.

2. Rafiei, H.; Rabbani, M. An MADM Framework toward Hierarchical Production Planning in Hybrid MTS/MTO Environments. *World Acad. Sci. Eng. Technol.* **2009**, *58*, 462–466.
3. Soman, C.A.; Van Donk, D.P.; Gaalman, G. Combined make-to-order and make-to-stock in a food production system. *Int. J. Prod. Econ.* **2004**, *90*, 223–235. [CrossRef]
4. Beemsterboer, B.; Land, M.; Teunter, R.; Bokhorst, J. Integrating make-to-order and make-to-stock in job shop control. *Int. J. Prod. Econ.* **2017**, *185*, 1–10. [CrossRef]
5. Peeters, K.; van Ooijen, H. Hybrid make-to-stock and make-to-order systems: A taxonomic review. *Int. J. Prod. Res.* **2020**, *58*, 4659–4688. [CrossRef]
6. Ghrayeb, O.; Phojanamongkolkij, N.; Tan, B.A. A hybrid push/pull system in assemble-to-order manufacturing environment. *J. Intell. Manuf.* **2008**, *20*, 379–387. [CrossRef]
7. Köber, J.; Heinecke, G. Hybrid Production Strategy between Make-to-order and Make-to-stock—A Case Study at a Manufacturer of Agricultural Machinery with Volatile and Seasonal Demand. In Proceedings of the 45th CIRP Conference on Manufacturing Systems, Athens, Greece, 16–18 May 2012; Volume 3, pp. 453–458.
8. Gupta, D.; Benjaafar, S. Make-to-order, make-to-stock, or delay product differentiation? A common framework for modeling and analysis. *IIE Trans.* **2004**, *36*, 529–546. [CrossRef]
9. Liu, L.; Xu, H.; Zhu, S.X. Push verse pull: Inventory-leadtime tradeoff for managing system variability. *Eur. J. Oper. Res.* **2020**, *287*, 119–132. [CrossRef]
10. Soman, C.A.; van Donk, D.P.; Gaalman, G. Comparison of dynamic scheduling policies for hybrid make-to-order and make-to-stock production systems with stochastic demand. *Int. J. Prod. Econ.* **2006**, *104*, 441–453. [CrossRef]
11. Rafiei, H.; Rabbani, M. Capacity coordination in hybrid make-to-stock/make-to-order production environments. *Int. J. Prod. Res.* **2012**, *50*, 773–789. [CrossRef]
12. Kaminsky, P.; Kaya, O. Combined make-to-order/make-to-stock supply chains. *IIE Trans.* **2009**, *41*, 103–119. [CrossRef]
13. Adan, I.; van der Wal, J. Combining make to order and make to stock. *OR Spectr.* **1998**, *20*, 73–81. [CrossRef]
14. Ohta, H.; Hirota, T.; Rahim, A. Optimal production-inventory policy for make-to-order versus make-to-stock based on the $M/E_r/1$ queuing model. *Int. J. Adv. Manuf. Technol.* **2007**, *33*, 36–41. [CrossRef]
15. Arreola-risa, A.; DeCroix, G.A. Make-to-order versus make-to-stock in a production inventory system with general production times. *IIE Trans.* **1998**, *30*, 705–716. [CrossRef]
16. Beemsterboer, B.; Land, M.; Teunter, R. Flexible lot sizing in hybrid make-to-order/make-to-stock production planning. *Eur. J. Oper. Res.* **2017**, *260*, 1014–1023. [CrossRef]
17. Parlar, M. Continuous-review inventory problem with random supply interruptions. *Eur. J. Oper. Res.* **1997**, *99*, 366–385. [CrossRef]
18. Nielsen, C.; Larsen, C. An analytical study of the Q(s, S) policy applied to the joint replenishment problem. *Eur. J. Oper. Res.* **2005**, *163*, 721–732. [CrossRef]
19. Bensoussan, A.; Liu, R.H.; Sethi, S.P. Optimality of an (s, S) policy with compound Poisson and diffusion demands: A quasi-variational inequalities approach. *SIAM J. Control Optim.* **2009**, *44*, 1650–1676. [CrossRef]
20. Gürler, U.; Ozkaya, B.Y. Analysis of the (s, S) policy for perishables with a random shelf life. *IIE Trans.* **2008**, *40*, 759–781. [CrossRef]
21. Gao, Y.; Wen, M.L.; Ding, S.B. (s, S) policy for uncertain single period inventory problem. *Int. J. Uncertain. Fuzziness Knowl. Based Syst.* **2013**, *21*, 945–953. [CrossRef]
22. Haghighi, A.M.; Mishev, D.P. *Queueing Models in Industry and Business*; Nova Science Publishers, Inc.: New York, NY, USA, 2008; pp. 129–148.
23. Srinivasan, M.M.; Lee, H.S. Random review production/inventory systems with compound Poisson demands and arbitrary processing times. *Manag. Sci.* **1991**, *37*, 813–833. [CrossRef]
24. Lee, H.S.; Srinivasan, M.M. *The Continuous (s, S) Review Policy for Production/Inventory Systems with Compound Poisson Remands*; Technical Report; Industrial & OE University of Michigan: Ann Arbor, MI, USA, 1988.
25. Lee, H.S.; Srinivasan, M.M. *The Continuous (s, S) Review Policy for Production/Inventory Systems with Poisson Demands and Arbitrary Processing Times*; Technical Report; Industrial & OE University of Michigan: Ann Arbor, MI, USA, 1987.

26. Rafiei, H.; Rabbani, M.; Vafa-Arani, H.; Bodaghi, G. Production-inventory analysis of single-station parallel machine make-to-stock/make-to-order system with random demands and lead times. *Int. J. Manag. Sci. Eng. Manag.* **2017**, *12*, 33–44. [CrossRef]
27. De Cuypere, E.; De Turck, K.; Fiems, D. A Maclaurin-series expansion approach to multiple paired queues. *Oper. Res. Lett.* **2014**, *42*, 203–207. [CrossRef]
28. De Cuypere, E.; De Turck, K.; Wittevrongel, S.; Fiems, D. A Maclaurin-series expansion approach to coupled queues with phase-type distributed service times. In Proceedings of the 10th EAI International Conference on Performance Evaluation Methodologies and Tools, Taormina, Italy, 26–28 October 2016; p. 8.
29. Evdokimova, E.; De Turck, K.; Fiems, D. Coupled queues with customer impatience. *Perform. Eval.* **2018**, *118*, 33–47. [CrossRef]
30. Evdokimova, E.; Wittevrongel, S.; Fiems, D. A Taylor series approach for service-coupled queueing systems with intermediate load. *Math. Probl. Eng.* **2017**, *2017*, 3298605. [CrossRef]
31. Ozkar, D.; Uzunoglu Kocer, U. Two-commodity queueing-inventory system with two classes of customers. *Opsearch* **2020**. [CrossRef]
32. Sun, B.; Dudin, A.; Dudin, S. Queueing system with impatient customers, visible queue and replenishable inventory. *Appl. Comput. Math.* **2018**, *17*, 161–174.
33. Melikov, A.Z.; Ponomarenko, L.A.; Bagirova, S.A. Markov Models of Queueing-Inventory Systems with Variable Order Size. *Cybern. Syst. Anal.* **2017**, *53*, 373–386. [CrossRef]
34. Ko, S.; Kang, J.; Kwon, E. An (s, S) inventory model with level-dependent G/M/1-type structure. *J. Ind. Manag. Optim.* **2016**, *12*, 609–624.
35. Fernandes, N.O.; Silva, C.; Carmo-Silva, S. Order release in the hybrid MTO–FTO production. *Int. J. Prod. Econ.* **2015**, *170*, 513–520. [CrossRef]
36. Elmehanny, A.M.; Abdelmaguid, T.F.; Eltawil, A.B. Optimizing Production and Inventory Decisions for Mixed Make-to-order/Make-to-stock Ready-made Garment Industry. In Proceedings of the IEEE International Conference on Industrial Engineering and Engineering Management, Bangkok, Thailand, 16–19 December 2018.
37. Latouche, G.; Ramaswami, V. *Introduction to Matrix Analytic Methods in Stochastic Modeling*; SIAM: Philadelphia, PA, USA, 1999.
38. Gun, L. Experimental results on matrix-analytical solutions techniques—Extensions and comparisons. *Stoch. Model.* **1989**, *5*, 669–682. [CrossRef]
39. Telek, M.; Heindl, A. Matching Moments For Acyclic Discrete And Continuous Phase-Type Distributions of Second Order. *Int. J. Simul. Syst. Sci. Technol.* **2003**, *3*, 47–57.
40. Cardinaels, E.; Borst, S.C.; van Leeuwaarden, J.S.H. Job assignment in large-scale service systems with affinity relations. *Queueing Syst.* **2019**, *93*, 227–268. [CrossRef]
41. Yekkehkhany, A.; Nagi, R. Blind gb-pandas: A blind throughput-optimal load balancing algorithm for affinity scheduling. *IEEE/ACM Trans. Netw.* **2020**. [CrossRef]

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.