View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Ghent University Academic Bibliography

A queueing theoretic approach to decoupling inventory

De Cuypere E., De Turck K., and Fiems D.

Department of Telecommunications and Information Processing, Ghent University,
St-Pietersnieuwstraat 41, B-9000, Belgium
(eline.decuypere, koen.deturck, dieter.fiems)@telin.ugent.be

Abstract. This paper investigates the performance of different hybrid push-pull systems with a decoupling inventory at the semi-finished products and reordering thresholds. Raw materials are 'pushed' into the semi-finished product inventory and customers 'pull' products by placing orders. Furthermore, production of semi-finished products starts when the inventory goes below a certain level, referred to as the threshold value and stops when the inventory attains stock capacity. As performance of the decoupling stock is critical to the overall cost and performance of manufacturing systems, this paper introduces a Markovian model for hybrid push-pull systems. In particular, we focus on a queueing model with two buffers, thereby accounting for both the decoupling stock as well as for possible backlog of orders. By means of numerical examples, we assess the impact of different reordering policies, irregular order arrivals, the set-up time distribution and the order processing time distribution on the performance of hybrid push-pull systems.

1 Introduction

In a make-to-stock system (push type), products are stocked in advance, while in a make-to-order system (pull type), a product only starts to be manufactured when a customer order is placed, see a.o [24,16,12,25,7]. Nowadays, as a means to respond quickly to growing variety, shorter product life cycles while keeping inventory costs as low as possible, hybrid push-pull systems are introduced [23]. An important issue in the overall performance of such hybrid systems is the position of the decoupling point [23,20]. Hoekstra et. al [10] defined the customer order decoupling point (CODP) concept. These authors considered market, product and production related factors as well as the desired service level and associated inventory costs to locate the optimal decoupling point. Under different hybrid push/pull control policies, Pandey and Khokhajaikiat [19] conducted a case study concerning the design and performance evaluation of a multistage production system. Results indicated that the choice of the optimal decoupling positions changes with the extent of raw material constraint operating at the stages and the demand lead time variabilities. To account for a degree of customisation and short delivery times, Blecker and Abdelkafi [2] considered a decoupling point at the inventory of semi-finished products. Here, after an order is received, only the final completion step still needs to be done. A case study at Phoenix showed that, by a hybrid approach, the company would save 20 to 25 percent of the total late costs and inventory costs compared to a pure push approach, which was at that time being used [5]. Research on the performance of the decoupling inventory in a hybrid push-pull system is therefore of main importance. This is the subject of the present paper.

In the present setting, we use a queuing theoretic approach to study the hybrid push-pull system. Queuing theory has already been successfully applied to assess decoupling points. Kaminsky and Kaya [13] considered a variety of combined make-to-order (MTO) and make-to-stock (MTS) supply chains with a single manufacturer and a single supplier in order to minimise a function of the total inventory, lead times and tardiness. The arrival process at the manufacturer is treated as a single facility with multiple classes of Poisson arrivals scheduled FCFS. As in previous research, they concluded that costs can be cut dramatically by using a combined system instead of pure MTO or MTS systems. Ohta and al. [18] analysed a multi-product inventory system where demand for each item arrive according to a Poisson process and the production time has an Erlang distribution. An optimality condition that specifies whether each product should be produced MTS or MTO is proposed. Bell [1] investigated a decoupling inventory between two successive production stages, the demand at stage 2 being independent from production at stage 1. The stages are decoupled by storing intermediate products. Limits on the available storage capacity and the rates of flow production into and out of the decoupling inventory are set, which enables the firm to determine the optimum capacities for the storage facility and to determine the value of an additional supply of intermediate product. Chang and Lu [4] studied a one-station production system consistent with MTO and MTS productions and dealing with two types of random demands: ordinary demand and specific demand. In this system, both types of demand arrive according to a Poisson process and production times of the workstation are exponentially distributed. Specific demand has a higher priority with respect to ordinary demand and the performance of this system is studied by means of matrix-geometric methods.

The present study of the decoupling stock closely relates to literature on two-part assembly systems, sometimes termed paired queues or kitting processes. For such systems, there are two queues, each storing a specific part, and production only starts when both part buffers are non-empty. In the current setting, one part-buffer corresponds to the decoupling stock, while the other corresponds to the list of backlogged orders. Also, production only starts when both buffers are non-empty. Indeed, each delivery of a finished product requires both the order specifications and a semi-finished product and can only be satisfied if both are present. If both part-buffers have unlimited capacity, Harrison [9] was the first to prove that, assuming no arrival control strategy, this queueing system is inherently unstable. In particular, he studied the multiple-input extension of the GI/G/1 queue in which arrivals in each stream are described by an independent renewal process and service times are independent and identically distributed.

He showed that part waiting times converge to non-defective limiting distributions only if the buffer capacities are bounded. This was also demonstrated by Latouche [14] who termed the two-part assembly system as waiting lines with paired customers. He considered a system of infinite capacity queues with Poisson arrivals for both parts and exponential services. The steady state is attained, i.e. the system is stable, if the arrival rates depend on the difference between queue lengths. [3] extended Latouche's research by considering two exponential distributions, one for the part processing distribution, i.e. the synchronisation phase, and the other for the assembly operation distribution. Approximations for the throughput rate and average queue length were given. Lipper and Sengupta [17] is another extension of the work of Latouche. In this paper, multiple Poisson input streams arrive in buffers with finite capacity. A more general structure in which parts are withdrawn from infinite pools and processed prior to assembly has been studied by Hopp et. al [11] and Som et. al [22]. Som and Wilhem [21] studied a two-queue system in which each part is processed according to an exponential distribution and the assembly operation times are generally distributed. They follow a matrix-geometric approach to numerically determine the marginal distributions of both kit and end-product inventory positions. Finally, assuming finite part-buffers, a two-part assembly system in a Markovian environment is studied in [6] by numerically solving the corresponding Markov chains by the generalized minimal residual method.

Furthermore, this article analyses hybrid push-pull systems with a threshold inventory: once the stock of semi-finished products drops below some level, this is either communicated to the production department if the parts are produced in-house or an order is placed with a third-party company if this is not the case. In both cases, it may take some time, the reordering time, before the inventory is replenished. Then, production stops when the semi-finished product inventory level attains stock capacity. The studied inventory control system closely relates to the well-known economic order quantity (EOQ) model [8]. This is a deterministic fluid-model for a single inventory and determines optimal reordering policies which balance the purchase, order and storage costs. While the single-part inventory problem is well understood, both in a deterministic and a stochastic setting, many issues of optimal inventory management in the multi-queue inventory case remain unresolved, most prominently in the stochastic setting.

In contrast to previous research, this paper investigates a two-queue system with one finite and one infinite buffer. Indeed, to limit involved costs, the decoupling stock needs to be sufficiently small. Hence, finite capacity is assumed. However, no such assumption is imposed for the other queue: the order backlog queue has an infinite capacity. Assuming a finite capacity product queue also assures the existence of a steady-state solution, provided that the arrival rate of orders is limited. In particular, this article analyses hybrid push-pull systems under different threshold policies, assuming that production stops when the inventory level reaches maximum capacity. Comparing versatility and numerical tractability, we study the decoupling stock in a Markovian environment as in [6]. This approach allows for assessing the effect of variability in the production process of semi-finished products, the ordering process and the delivery process on the performance of the decoupling stock.

The remainder of this paper is organised as follows. Section 2 describes the decoupling stock model at hand. In section 3, the decoupling inventory system is analysed as a quasi-birth-and-death-process (QBD) and a number of specific application scenarios for the decoupling inventory system are introduced. Also, the numerical solution methodology is discussed and relevant performance measures are determined. To illustrate our approach, section 4 considers some numerical examples. Finally, conclusions are drawn in section 5.

2 Model description

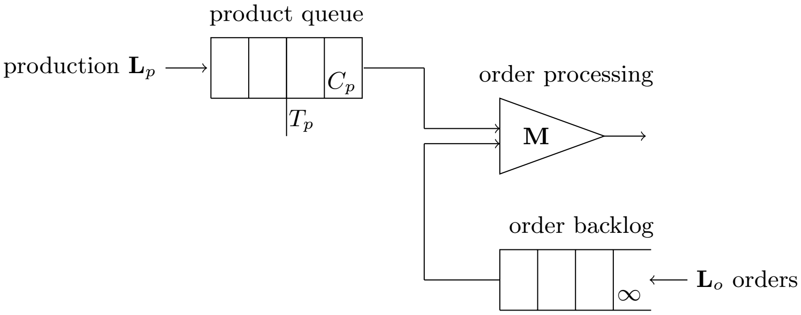


Fig. 1. Decoupling inventory of semi-finished products in a hybrid push-pull system.

The decoupling stock is modelled as a queueing model with two queues, as depicted in Figure 1. The product queue has finite capacity C_p and stores the semi-finished products prior to being processed to finished products. Moreover, production of semi-finished starts when the inventory goes below the threshold value T_p and stops when the inventory level reaches capacity C_p. The order queue keeps track of the orders that have not yet been delivered and has infinite capacity. Arriving orders are served in accordance with a first-come-first-served queueing discipline. Each order takes a semi-finished product from the product queue and completes the product in accordance with order specifications. Note that the two queues in the model at hand are tightly coupled. Departures from the product queue are only possible when there are orders. Similarly, departures from the order queue are only possible if there are semi-finished products in the product queue.

Arrivals at both queues are modelled according to possibly dependent arrival processes and order completion is not instantaneous. For ease of modelling, it is assumed that there is a modulating Markov chain, arrival and service rates depending on the state of this modulating chain. To be more precise, the decoupling inventory system is a three-dimensional continuous-time Markov Chain with infinite state space $\mathbb{N} \times \{0, 1, 2, \ldots, C_p\} \times \mathcal{K}$, $\mathcal{K} = \{0, 1, \ldots, K\}$ being the state space of the modulating chain. At any time, the state of the decoupling inventory system is described by the triplet $[n, m, i]$, n being the number of backlogged orders, m being the number of semi-finished products and i being the state of the modulating chain. We now describe the state transitions.

- The state of the modulating chain can change when there are neither arrivals nor departures. Let α_{ij} denote the transition rate from state i to state j ($i, j \in \mathcal{K}$, $i \neq j$). Further, for ease of notation, let

  $$\alpha_{ii} = -\sum_{j \neq i} \alpha_{ij}.$$

  and let $\mathbf{A} = [\alpha_{ij}]_{i,j \in \mathcal{K}}$ denote the corresponding generator matrix. Further, it is assumed that when either of the queues is empty, different transition rates (when there are neither arrivals nor departures) can be specified: let $\hat{\alpha}_{ij}$ and $\hat{\mathbf{A}}$ denote the transition rate from state i to state j and the corresponding generator matrix, respectively.
- The state of the modulating chain may remain the same or may change when there is an arrival. Let $\lambda_{ij}^{(p)}$ and $\lambda_{ij}^{(o)}$ denote the (marked) transition rate from state i to state j when there is an arrival at the product queue and the order queue, respectively. Moreover, let $\boldsymbol{\Lambda}_p = [\lambda_{ij}^{(p)}]_{i,j \in \mathcal{K}}$ and $\boldsymbol{\Lambda}_o = [\lambda_{ij}^{(o)}]_{i,j \in \mathcal{K}}$ denote the corresponding generator matrices. Note that marked self transitions from state i to state i are allowed.
- Analogously, the state of the modulating chain may remain the same or may change when there is a departure (in each buffer). Let μ_{ij} and $\underline{\mathbf{M}}$ denote the corresponding transition rate and generator matrix respectively.

Remark 1. The transition rates are dependent on the product queue size, the state of the modulating chain and whether the order queue is empty, e.g. there are no product arrivals when the queue is full, production starts only when the semi-finished product inventory level goes below the threshold value and there are only departures if both queues are non-empty.

3 Analysis

3.1 Quasi-birth-death process

The studied Markov process is a homogeneous quasi-birth-and-death process (QBD), see [15]. In the present setting, the so-called level or block-row number, indicates the number of backlogged orders while the phase, i.e. the index within a block element, indicates both the content of the decoupling stock and the state of the Markovian environment. The one-step transitions are 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 $(n - 1, *, *)$).

We then find that the generator matrix of the Markov chain has the following block matrix representation,

$$\mathbf{Q} = \begin{bmatrix} \mathbf{L}'_p & \mathbf{L}_o & \mathbf{0} & 0 & \cdots \\ \mathbf{W} & \mathbf{L}_p & \mathbf{L}_o & 0 & \cdots \\ 0 & \mathbf{W} & \mathbf{L}_p & \mathbf{L}_o & \cdots \\ 0 & 0 & \mathbf{W} & \mathbf{L}_p & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} . \tag{1}$$

The blocks are given by,

$$\mathbf{L}_o = \begin{bmatrix} \boldsymbol{\Lambda}_o^{(0)} & 0 & 0 & \cdots & 0 \\ 0 & \boldsymbol{\Lambda}_o^{(1)} & 0 & \cdots & 0 \\ 0 & 0 & \boldsymbol{\Lambda}_o^{(2)} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \boldsymbol{\Lambda}_o^{(C_p)} \end{bmatrix}, \quad \mathbf{L}_p = \begin{bmatrix} \underline{\mathbf{D}}^{(0)} & \boldsymbol{\Lambda}_p^{(0)} & 0 & \cdots & 0 \\ 0 & \mathbf{D}^{(1)} & \boldsymbol{\Lambda}_p^{(1)} & \cdots & 0 \\ 0 & 0 & \mathbf{D}^{(2)} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \mathbf{D}^{(C_p)} \end{bmatrix}, \tag{2}$$

$$\mathbf{L}'_p = \begin{bmatrix} \underline{\mathbf{D}}^{(0)} & \boldsymbol{\Lambda}_p^{(0)} & 0 & \cdots & 0 \\ 0 & \underline{\mathbf{D}}^{(1)} & \boldsymbol{\Lambda}_p^{(1)} & \cdots & 0 \\ 0 & 0 & \underline{\mathbf{D}}^{(2)} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \underline{\mathbf{D}}^{(C_p)} \end{bmatrix}, \quad \mathbf{W} = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ \mathbf{M}^{(1)} & 0 & \cdots & 0 & 0 \\ 0 & \mathbf{M}^{(2)} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \mathbf{M}^{(C_p)} & 0 \end{bmatrix}. \tag{3}$$

with $\mathbf{D}^{(m)} = \mathbf{A}^{(m)} - \partial \boldsymbol{\Lambda}_o^{(m)} - \partial \boldsymbol{\Lambda}_p^{(m)} - \partial \mathbf{M}^{(m)}$ and $\underline{\mathbf{D}}^{(m)} = \hat{\mathbf{A}}^{(m)} - \partial \boldsymbol{\Lambda}_o^{(m)} - \partial \boldsymbol{\Lambda}_p^{(m)}$ with $m = (0, 1, 2, \ldots, C_p)$ being the number of semi-finished products in the buffer. Note that $\partial \mathbf{X}$ represents a diagonal matrix with diagonal elements equal to the row sums of $\mathbf{X}$. Intensities in the generator matrices Λ_o, Λ_p, $\underline{\mathbf{D}}$, $\mathbf{D}$ and $\mathbf{M}$ are dependent of the product buffer content m. Therefore, we introduce the superscript $^{(m)}$ to make this dependence explicit. Note that if no superscript is indicated, the intensities in the generator matrix are equal for all numbers of semi-finished products in the queue.

To simplify notation, states representing an inactive production and a product queue content equal or less than the threshold value, are taken into account in the generator matrix structure. However, as production is always active when the semi-finished product inventory level is below the threshold value, the next transition changes the given inactive background state to an active one. The matrix structure also considers states where the number of semi-finished products equals capacity and the background state is active. Again, the next transition changes the background state into an inactive state.

In the general case, arrivals and departures at both queues are modelled according to possibly dependent Markovian arrival processes (MAP) and phase-type distributed order processing times, respectively. The Markovian arrival processes are described by the generator matrices $\mathbf{B}_1^{(m)}$ and $\mathbf{B}_3^{(m)}$ with arrivals of semi-finished products and orders respectively and the generator matrices $\mathbf{B}_0^{(m)}$ and $\mathbf{B}_2^{(m)}$ without arrivals at the decoupling stock and the queue of backlogged orders respectively. The phase-type distribution is completely characterised by an initial probability vector $\boldsymbol{\tau}$ and the matrix $\mathbf{T}$ which corresponds to non-absorbing transitions [15]. Let $\mathbf{t}' = -\mathbf{T}\mathbf{e}'$ be the column vector with the rates to the absorbing state with $\mathbf{e}$ a row vector of ones. We have,

$$\boldsymbol{\Lambda}_p = \mathbf{B}_1^{(m)}\,, \quad \boldsymbol{\Lambda}_o = \mathbf{B}_3^{(m)}\,, \quad \mathbf{A} = \mathbf{B}_0^{(m)} + \mathbf{B}_2^{(m)} + \mathbf{T}\,,$$
$$\hat{\mathbf{A}} = \mathbf{B}_0^{(m)} + \mathbf{B}_2^{(m)}, \quad \mathbf{M} = \mathbf{t}'\boldsymbol{\tau}\,.$$

Before proceeding, we introduce a number of specific application scenarios of the decoupling inventory system at hand.

Example 1. In the most basic setting, when the semi-finished product inventory level goes below a given threshold value, semi-finished products arrive in the queues in accordance with an independent Poisson process with parameter λ_p and production stops when the stock capacity is reached. Orders arrive according to an independent Poisson process with parameter λ_o and order processing times are exponentially distributed with parameter μ. In this case, the modulating state indicates whether the production of semi-finished products is active or not. We have,

$$\boldsymbol{\Lambda}_p = \lambda_p\mathbf{I}\,, \quad \boldsymbol{\Lambda}_o = \lambda_o\mathbf{I}\,, \quad \mathbf{M} = \mu\mathbf{I}\,, \quad \mathbf{A} = \hat{\mathbf{A}} = \mathbf{0}\,.$$

Here $\mathbf{I}$ denotes the identity matrix.

Example 2. To account for variability in the production times of semi-finished products, we consider a Markovian arrival process with the generator matrices $B_0^{(m)}$ and $B_1^{(m)}$ as described above. Assuming Poisson arrivals of orders with parameter λ_o and order processing times exponentially distributed with parameter μ, we have,

$$\boldsymbol{\Lambda}_p = \mathbf{B}_1^{(m)}\,, \quad \boldsymbol{\Lambda}_o = \lambda_o\mathbf{I}\,, \quad \mathbf{A} = \hat{\mathbf{A}} = \mathbf{B}_0^{(m)}\,, \quad \mathbf{M} = \mu\mathbf{I}\,.$$

Example 3. Unreliability in the ordering process can also be modelled by a Markovian arrival process. Here, the MAP is described by the generator matrix $\mathbf{B}_3^{(m)}$ of transitions with order arrivals and the generator matrix $\mathbf{B}_2^{(m)}$ without arrivals. Retaining exponentially distributed order processing times and assuming Poisson arrivals of semi-finished products, we have,

$$\boldsymbol{\Lambda}_p = \lambda_p\mathbf{I}\,, \quad \boldsymbol{\Lambda}_o = \mathbf{B}_3^{(m)}\,, \quad \mathbf{A} = \hat{\mathbf{A}} = \mathbf{B}_2^{(m)}\,, \quad \mathbf{M} = \mu\mathbf{I}\,.$$

Example 4. As for the arrival processes, the model at hand is sufficiently flexible to include phase-type distributed order processing times. The phase-type distribution is completely characterised by an initial probability vector $\boldsymbol{\tau}$ and the matrix $\mathbf{T}$ which corresponds to non-absorbing transitions [15]. Let $\mathbf{t}' = -\mathbf{T}\mathbf{e}'$

be the column vector with the rates to the absorbing state with $\mathbf{e}$ a row vector of ones. Assuming Poisson arrivals in both buffers (with rate λ_p and λ_o, respectively), we get the following matrices,

$$\boldsymbol{\Lambda}_p = \lambda_p \mathbf{I}\,, \quad \boldsymbol{\Lambda}_o = \lambda_o \mathbf{I}\,, \quad \mathbf{A} = \mathbf{T}\,, \quad \hat{\mathbf{A}} = \mathbf{0} \quad \mathbf{M} = \mathbf{t}'\boldsymbol{\tau}\,.$$

3.2 Methodology: the matrix-geometric technique

Consider the above defined Markov process on the three-dimensional state space $\{(n, m, i) \mid n \geq 0, 0 \leq m \leq C_p, i = 0, 1, \ldots, K\}$ where i denotes the state of the modulating chain, as the phase set i is defined in the finite state space $\mathcal{K}$ (see section 2). A well-known method for finding the stationary distribution of QBD processes is the matrix-geometric method. With $\pi(n, m, i)$ the stationary probability of the process being in state (n, m, i), and using the vector notation $\boldsymbol{\pi}_k = (\pi(k, 0, 0), \pi(k, 0, 1), \ldots, \pi(k, C_p, K))$, the probability vectors can be expressed as,

$$\boldsymbol{\pi}_k = \boldsymbol{\pi}_0 \mathbf{R}^k. \tag{4}$$

where the so-called rate matrix $\mathbf{R}$ is the minimal non-negative solution of the non-linear matrix equation $\mathbf{R}^2\mathbf{W} + \mathbf{R}\mathbf{L}_p + \mathbf{L}_o = \mathbf{0}$. Here, we compute the rate matrix by implementing the iterative algorithm of [15, chapter 8].

3.3 Performance measures

Once the steady state probabilities have been determined numerically, we can calculate a number of interesting performance measures for the decoupling inventory system. For ease of notation, we introduce the marginal probability mass functions of the queue content of the product queue and the order queue: $\pi^{(p)}(m) = \sum_{i\in\mathcal{K}} \sum_{n=0}^{\infty} \pi(n, m, i)$ and $\pi^{(o)}(m) = \sum_{i\in\mathcal{K}} \sum_{m=0}^{C_p} \pi(n, m, i)$.

Note that as the queue of the backlogged orders is infinite, the throughput of the decoupling inventory system η equals the order arrival rate λ_o. In addition, we have the following performance measures.

- The mean semi-finished product queue and the order backlog content: $\mathrm{E}\, Q_p$ and $\mathrm{E}\, Q_o$ respectively,

$$\mathrm{E}\, Q_p = \sum_{m}^{C_p} \pi^{(p)}(m) m\,, \quad \mathrm{E}\, Q_o = \sum_{n}^{\infty} \pi^{(o)}(n) n\,.$$

- The variance of the semi-finished product queue and the order backlog content: $\mathrm{Var}\, Q_p$ and $\mathrm{Var}\, Q_o$ respectively,

$$\mathrm{Var}\, Q_p = \sum_{m}^{C_p} \pi^{(p)}(m) m^2 - (\mathrm{E}\, Q_p)^2\,,$$

$$\mathrm{Var}\, Q_o = \sum_{n}^{\infty} \pi^{(o)}(n) n^2 - (\mathrm{E}\, Q_o)^2\,.$$

– The mean lead time LT (calculated based on Little's theorem) is the average amount of time between the placement of an order and the completion to a finished product:

$$\mathrm{LT} = \frac{\mathrm{E}\, Q_o}{\lambda_o}$$

– As the product queue has finite capacity, production prior to the decoupling stock may be blocked. This happens when there is a product arrival and the queue is full. Hence, blocking corresponds to the loss probability in the product buffer. The loss probability is most easily expressed in terms of the throughput η. We have,

$$b_p = \frac{\lambda_p - \eta}{\lambda_p} = \frac{\lambda_p - \lambda_o}{\lambda_p} \, .$$

We now illustrate our approach by means of some numerical examples.

4 Numerical examples

4.1 Poisson arrivals and exponential order processing times

As a first example, the difference between the mean semi-finished product queue and the mean order backlog content versus the threshold value of the semi-finished product inventory is depicted in figure 2(a). We assume that semi-finished products and orders arrive according to a Poisson process with parameter $\lambda_p = 1$ and $\lambda_o = 0.85$, respectively. The inventory capacity C_p equals 20 and order processing times are exponentially distributed with service rate μ equal to 1 for all curves. The described model is a decoupling inventory system with Poisson arrivals and exponential order processing times as described in example 1 of section 3. As the figure shows, the threshold value of 5 results on average in no backlogged orders and no semi-finished products in stock. Under and above this level, products and orders are on average backlogged, respectively. Obviously, there is on average more stock of semi-finished products and less backlog of orders as the threshold value increases.

Figure 2(b) represents the trade-off between the maximum probability to have the lead time higher or equal to 30 (left side) and the average stock of the semi-finished products (right side). Note that we calculated the lead time distribution by using the one-sided Chebyshev's inequality. Under the same parameter assumptions of figure 2(a), the maximum probability to have the lead time higher or equal to 30 decreases and the average stock increases as the inventory capacity increases for each threshold value. Indeed, if more buffer capacity is available, it will be used – the mean semi-finished product queue increases such that there is on average less time required to deliver an order. Finally, in this numerical example, we observe that the highest threshold value give the average best results: the intersection between the two performance measures and the necessary stock capacity have the lowest value.

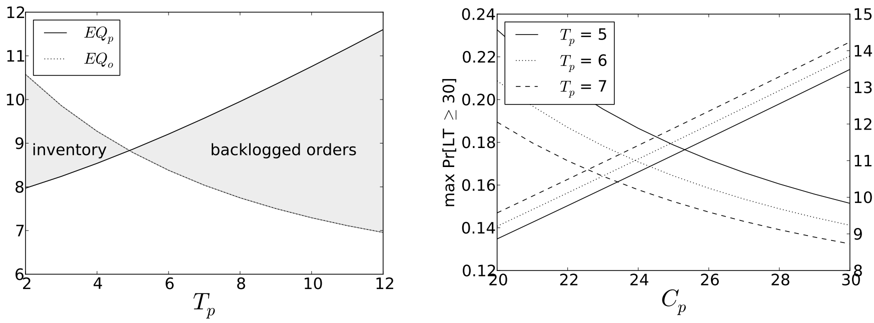
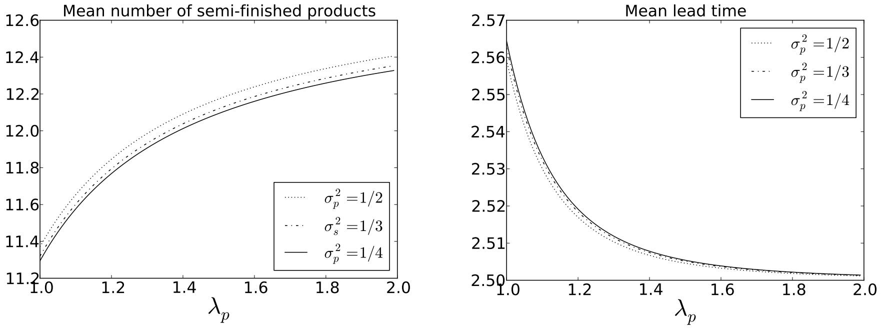


Fig. 2. There is a trade-off between the average stock of the semi-finished products and the average number of backlogged orders and between the lead time.

4.2 Erlang distributed set-up times

The second numerical example quantifies the impact of variability in the production process of semi-finished products on the decoupling inventory system. In particular, we here study Erlang-distributed set-up times – the set-up time starts when the semi-finished product inventory goes below a certain level and stops after some Erlang distributed time. Then, the semi-finished products arrive according to a Poisson process with arrival rate λ_p until the stock capacity is reached. The described model is a decoupling inventory system with Markovian arrivals as described in example 2 of section 3.


Mean number of semi-finished products

Mean lead time

Fig. 3. The shape of the set-up time distribution has a small effect on the mean number of semi-finished products and on the mean lead time.

Figure 3(a) 3(b) show the mean number of semi-finished products in the buffer and the mean lead time of the system with a buffer capacity equal to 20 and a threshold value equal to 5. In both figures, the arrival rate is varied and different values of the variance of the set-up time process are assumed as indicated. The order arrival rate λ_o equals 0.6, order processing times are assumed to be exponentially distributed with service rate μ equal to 1 and the mean set-up time equals 1. As expected, the mean number of semi-finished products increases and the mean lead time decreases as the arrival rate of the semi-finished products λ_p increases. Furthermore, the shape of the set-up time distribution has a very small effect on both performance measures. In particular, the mean number of semi-finished products and the mean lead time show respectively a slight decrease and increase as the variance of the set-up time distribution σ_p^2 increases. This is due to the fact that the regularly the set-up time, the less semi-finished products are on average in stock and the more orders are on average backlogged.

4.3 Markovian arrival process for orders

We also quantify the impact of irregular order arrivals. To this end, we compare both buffers with Poisson arrivals to corresponding decoupling inventory systems with interrupted Poisson arrivals for the orders and Poisson arrivals for the semi-finished products. The arrival interruptions account for inefficiency in the ordering process. Order processing times are assumed to be exponentially distributed with service rate μ equal to 1, this value being independent of the number of products and orders in the different buffers. This numerical example fits example 3 of section 3.

The interrupted Poisson process considered here is a two-state Markovian process. In the active state, new orders arrive in accordance with a Poisson process with rate λ_o whereas no new orders arrive in the inactive state. Let α and β denote the rate from the active to the inactive state and vice versa, respectively. We then use the following parameters to characterise the interrupted Poisson process (IPP),

$$\sigma = \frac{\beta}{\alpha+\beta}\,, \quad \kappa = \frac{1}{\alpha} + \frac{1}{\beta}\,, \quad \rho_o = \lambda_o \sigma\,.$$

Note that σ is the fraction of time that the interrupted Poisson process is active, the absolute time parameter κ is the average duration of an active and an inactive period, and ρ_o is the arrival load of the orders.

Figure 4 shows the mean number of backlogged orders versus the arrival rate of semi-finished products with buffer capacity C_p equal to 20 and the threshold value T_p equal to 5 and 7 for Poisson arrivals as well as for interrupted Poisson arrivals of orders. Order processing times are exponentially distributed with service rate μ equal to one for all curves. In addition, we set $\sigma = 0.8$ and $\kappa = 10$ for the interrupted Poisson processes (λ_o equals 0.6 for Poisson arrivals and 0.75 for interrupted Poisson arrivals). As expected, the mean number of backlogged orders decreases as the arrival rate of semi-finished products increases. Furthermore, the impact of the threshold value on the average number of backlogged orders decreases as the arrival rate of semi-finished products increases – both Markovian models converge to some value for T_p equal to 5 and 7. Finally, comparing interrupted Poisson and Poisson processes, burstiness in the ordering

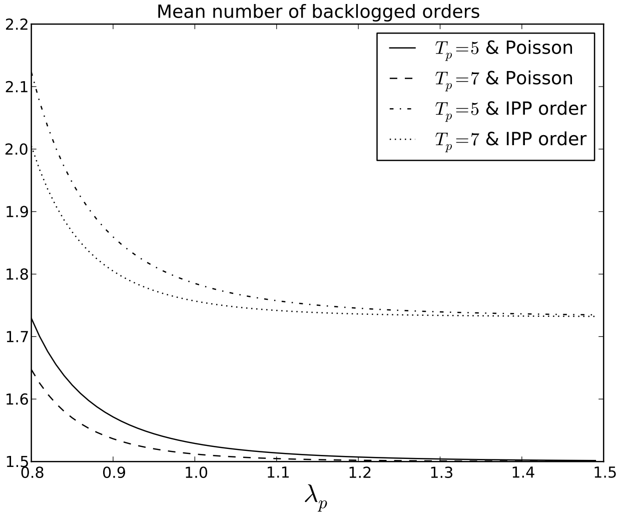


Fig. 4. Irregular order arrivals result in a higher average number of backlogged orders.

process has a negative impact on performance – there is on average more time required to deliver an order.

4.4 Phase-type distributed order processing times

The last numerical example quantifies the impact of the distribution of the order processing times on the decoupling inventory performance. In particular, we here study Erlang-distributed order processing times. This numerical example fits example 4 of section 3.

Figure 5(a) and 5(b) depict the mean number of semi-finished products in the buffer and the mean lead time of the decoupling inventory system. In both figures, the arrival rate of semi-finished products is varied and different values of the order processing time distribution are assumed as indicated. The service rate μ equals 1 for all curves, the order arrival rate λ_o equals 0.6, the inventory capacity C_p equals 20 and the threshold value T_p is equal to 5. Clearly, figure 5(a) and 5(b) show respectively that the buffer content of semi-finished products converges to capacity and the lead time decreases until a certain value as the arrival rate of semi-finished products increases. Concerning the mean number of semi-finished products, we can conclude that the order processing time distribution has no significant effect on this performance measure. Indeed, we observe that the difference is very small and that it decreases as the arrival rate of semi-finished products increases. However, the difference between a variance σ_s^2 equal to 1, 1/2 and 1/4 for the mean lead time remains constant and is significant,

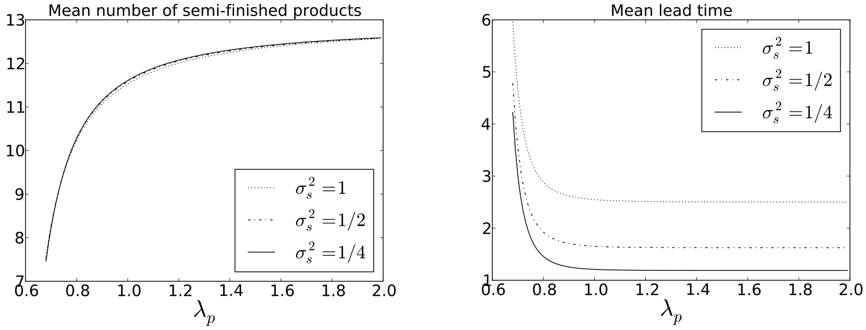


Fig. 5. The shape of the order processing time distribution is not significant for the mean number of semi-finished products and is significant for the mean lead time.

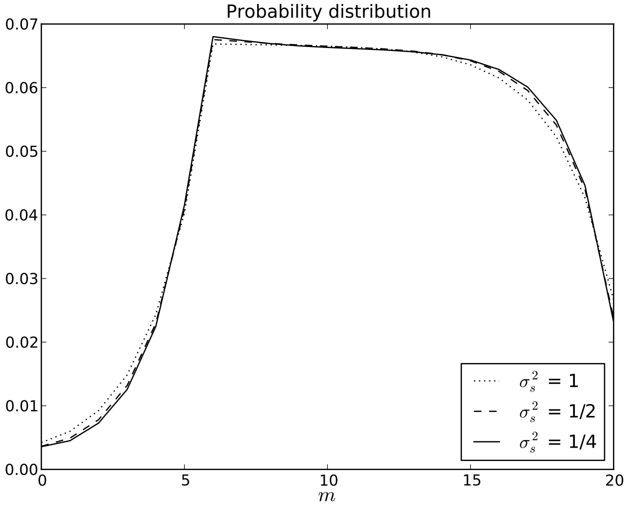


Fig. 6. The zero probability increases when the variance of the order processing time distribution decreases.

especially when the arrival rate λ_p is smaller than 0.7. Furthermore, in this numerical example, the mean number of semi-finished products decreases and the mean lead time increases as the variance increases. Indeed, the results of figure 6 show that the zero probability increases slightly as the variance of the order processing time distribution increases. As for Erlang distributed set-up times in section 4.2, we have a coupling effect between both performance measures – the mean number of semi-finished products increases such that the mean number of backlogged orders (and thus the mean lead time) decreases.

5 Conclusion

In this paper, we evaluate the performance of different hybrid push-pull systems with a decoupling inventory at the semi-finished products and reordering thresholds. In particular, we investigate the impact of different reordering policies, irregular order arrivals as well as the set-up time distribution and the order processing time distribution on the performance of hybrid push-pull systems. In the studied hybrid push-pull systems, production of semi-finished products starts when the inventory goes below the so-called threshold value and stops when the inventory attains stock capacity. Decoupling means that the completion of a semi-finished product is only possible when there is an order. These orders are backlogged and can be satisfied only when the semi-finished products are available. Therefore, the studied push-pull system is modelled as a homogeneous quasi-birth-and-death process (QBD) and solved with matrix-analytic methods.

As our numerical examples show, there is trade-off to be made between the inventory cost and the service level, as expected – e.g. a higher threshold value causes on average a higher inventory cost and a smaller lead time. Furthermore, irregular order arrivals have a negative impact on the performance of the hybrid push-pull system. However, system performance is relatively insensitive to variation in the set-up time distribution and partially insensitive to variation in the order processing time distribution. Future work will focus on determining the total cost of the studied hybrid push-pull systems.

References

1. Bell, P.: A decoupling inventory problem with storage capacity constraints. Operations Research 28, 476–488 (1980)
2. Blecker, T., Abdelkafi, N.: Complexity and variety in mass customization systems: analysis and recommendations. Management Decision 44(7), 908–929 (2006)
3. Bonomi, F.: An approximate analysis for a class of assembly-like queues. Queueing Systems Theory and Applications pp. 289–309 (1987)
4. Chang, K., Lu, Y.: Queueing analysis on a single-station make-to-stock/make-to-order inventory-production system. Applied Mathematical Modelling 34, 978–991 (2010)

5. Cochran, J., Kim, S.: Optimizing a serially combined push and pull manufacturing system by simulated annealing. In: Second International Conference on Engineering Design and Automation, (1998)
6. De Cuypere, E., Fiems, D.: Performance evaluation of a kitting process. In: Proceedings of the 17th International Conference on analytical and stochastic modelling techniques and applications, Lecture Notes in Computer Science. vol. 6751, pp. 175–188. Venice, Italy (2011)
7. Ghrayeb, O., Phojanamongkolkij, N., Tan, B.: A hybrid push/pull system in assemble-to-order manufacturing environment. Journal of Intelligent Manufacturing 20, 379–387 (2009)
8. Harris, F.: How many parts to make at once. Factory, the magazine of Management 10(2), 135 – 136 (1913)
9. Harrison, J.: Assembly-like queues. Journal Of Applied Probability 10(2), 354–367 (1973)
10. Hoekstra, S., Romme, J., Argelo, S.: Integral logistic structures : developing customer-oriented goods flow. McGraw-Hill (1992)
11. Hopp, W.J., Simon, J.T.: Bounds and heuristics for assembly-like queues. Queueing Systems 4, 137 – 156 (1989)
12. Hopp, W., Spearman, M.: Factory physics: Foundations of manufacturing management. The McGraw-Hill Companies, Inc. (2000)
13. Kaminsky, P., Kaya, O.: Combined make-to-order/make-to-stock supply chains. IIE Transactions 41, 103 – 119 (2009)
14. Latouche, G.: Queues with paired customers. Journal of Applied Probability 18, 684–696 (1981)
15. Latouche, G., Ramaswami, V.: Introduction to Matrix Analytic Methods in Stochastic Modeling. SIAM (1999)
16. Lee, C.: A recent development of the integrated manufacturing system: A hybrid of mrp and jit. International Journal of Operations and Production Management 13(4), 3–17 (1993)
17. Lipper, E., Sengupta, B.: Assembly-like queues with finite capacity: bounds, asymptotics and aproximations. Queueing Systems: Theory and Applications 18, 684 (1986)
18. Ohta, H., Hirota, T., Rahim, A.: Optimal production-inventory policy for make-to-order versus make-to-stock based on the m/er/1 queuing model. International Journal of Advanced Manufacturing Technologies 33, 36 – 41 (2007)
19. Pandey, P., Khokhajaikiat, P.: Performance modeling of multistage production systems operating under hybrid push-pull control. International Journal Production Economics 43, 115–126 (1995)
20. Ramachandran, K., Whitman, L., Ramachandran, A.: Criteria for determining the push-pull boundary. In: Industrial Engineering Research Conference. Orlando, FL, USA (2002)
21. Som, P., Wilhelm, W.: Analysis of stochastic assembly with GI-distributed assembly time. INFORMS Journal on Computing 11, 104 – 116 (1999)
22. Som, P., Wilhelm, W., Disney, R.: Kitting process in a stochastic assembly system. Queueing Systems 17, 471 – 490 (1994)
23. Soman, C., van Donk, D., Gaalman, G.: Combined make-to-order and make-to-stock in a food production system. International Journal of Production Economics 90, 223 – 235 (2004)
24. Spearman, M., Zazamis, M.: Push and pull production systems: Issues and comparisons. Operations Research 3, 521–532 (1992)

25. Takahashi, K., Nakamura, N.: Push pull, or hybrid control in supply chain management. International Journal of Computer Integrated Manufacturing 17(2), 126–140 (2004)