Analysis of a Production/Inventory System with Multiple Retailers

Ann M. Noblesse [1], Robert N. Boute [1,2], Marc R. Lambrecht [1], Benny Van Houdt [3]

[1] Research Center for Operations Management, University KU Leuven, Naamsestraat 69,
B-3000 Leuven, Belgium, ann.noblesse@kuleuven.be, robert.boute@kuleuven.be, marc.lambrecht@kuleuven.be
[2] Technology & Operations Management Area, Vlerick Business School, Vlamingenstraat 83,
B-3000 Leuven, Belgium, robert.boute@vlerick.com
[3] Department of Mathematics and Computer Science, University of Antwerp, Middelheimlaan 1,
B-2020 Antwerp, Belgium, benny.vanhoudt@ua.ac.be

Abstract

We study a production/inventory system with one manufacturing plant and multiple retailers. Production lead times at the plant are stochastic and endogenously determined by the orders placed by the different retailers. Assuming stochastic (phase-type distributed) production and setup times, we make use of matrix analytic techniques to develop a queuing model that is capable to compute the distribution of the time orders spend in the production facility, depending on the retailer's lot sizing decisions. The time orders spend in the production facility influences holding and backlogging costs at the retailers. Given the distribution of the time spent in the production facility, the distribution of inventory levels at each retailer can be computed. The goal is to compute total costs given the inventory parameters for every retailer, taking the endogeneity of their order policy on production lead times into account. Thanks to this procedure we will be able to analyse the interactions between the order policies of the retailers.

Keywords: Production/inventory system with multiple retailers, lot sizing, joint replenishment, can order policy.

1. Introduction

Supply chains tend to be composed of several firms, where the outcome of the decisions of one firm depends on the other firm's decisions. Many researchers took up the challenge to incorporate this dependency in their models, for example, by analyzing a multi-retailer system instead of a single-retailer system (Dror et al. 2012; Timmer et al. 2013). Note that studying a multi-retailer system is not only a more realistic, but unfortunately also a more difficult problem.

In this paper, we study a production/inventory system with multiple retailers and one manufacturing plant. We assume that the orders placed by the retailers are sent to the manufacturing plant, which produces on order. After the endogenously determined production lead times, the finished order is sent to the retailer and its inventory is replenished.

In the inventory management literature, the lot size of the replenishment orders is usually determined by the retailer to balance the fixed cost per order against the holding costs, whereby lead times are treated exogenously with respect to the inventory policy. It is justified to treat lead times as exogenous variables when transportation lead times are significantly longer than production times, when the manufacturer guarantees fixed delivery data, or when the manufacturer produces on stock (Benjaafar et al. 2005). However, in many settings, these conditions do not hold. We consider an integrated production/inventory supply chain, where the lot size of the replenishment policy determines the lead times of the production facility. When, for instance, a setup time per lot size exists, placing very small orders increases the number of setup times, the utilization rate at the production facility, and long waiting times and queues will follow. At the same time, when there is a production time per unit, placing very large orders also causes an increase in production lead times (Karmarkar 1987). In this paper, we take the impact of the order sizes on production lead times into account. Higher lead times impact holding and backlogging costs, and will therefore also determine the cost minimizing parameters of the inventory policy. In a setting with multiple retailers, the order process of each retailer also influences the utilization rate of the manufacturing plant, and therefore it also influences the production lead time of the orders placed by other retailers. We expect that if one retailer ignores the impact of the ordering policies of the other retailers on the utilization rate (and the resulting lead times) at the production facility, this retailer might have significantly higher costs compared to the situation where he opts for the best ordering policy while taking the impact of the other retailers into account.

We assume that a fixed order cost and a fixed setup time at the production facility exist. Therefore, we will assume that each individual retailer uses a continuous review (s, S) policy to manage his inventory. The (s, S) inventory policy, in which an order point s and an order-up-to level S are established, was first introduced by Arrow et al. (1951): a replenishment order is made as soon as the inventory position reaches the order point s, at that moment an order is placed to restore the inventory position to the level S.

In a setting with multiple retailers, the ordering cost and the lead times can be reduced by placing joint orders. This means that if two retailers place an order simultaneously, the fixed order cost and the major setup time are charged only once, whereas, if both retailers place an order at a different moment, the fixed order cost and the major setup time are charged twice (i.e., once for every retailer). To avoid this, a "can order" policy can be adopted by the retailers. This order policy was first introduced by Balintfy (1964). A (s, c, S) can order inventory policy has three parameters: the order point s, the can order level c, and the order-up-to level S. If the inventory of one of the retailers reaches its order point, the inventory positions of all other retailers are evaluated. Every retailer whose inventory position is at or below the can order level places an order. Order quantities are such that the inventory position of every retailer who placed an order is raised to the order-up-to level. A disadvantage of the can order policy is the size of the optimisation problem: for a setting with only two retailers, one needs to optimize six parameters (Özkaya, Gürler, and Berk 2006).

2. Model assumptions and notations

In this paper we study a continuous review production/inventory system. We assume that two retailers hold inventory of a single item. At each retailer j, a compound Poisson demand arrives (with arrival rate $\lambda^{(j)}$), which is independent of the demand arrival rate of the other retailer. Demand sizes per arrival are independent and identically distributed and follow a general discrete, finite distribution with maximum demand size m_j. Let $d_i^{(j)}$ denote the probability of a demand of size i at retailer j. Retailers place orders at a manufacturing plant, which produces on order. The manufacturing plant is a finite capacity production system, where orders are produced on a first-come-first-served basis on one processor which produces the units sequentially. Every order undergoes a phase-type distributed major setup time, all units of the order undergo one by one a phase-type distributed production time. If an order was placed by two retailers, an additional phase-type distributed minor setup/change-over time is needed (see Figure 1).

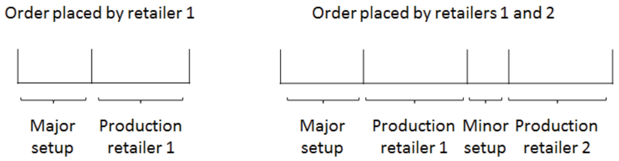

Figure 1: Sequence of events at the production facility.

The major setup time is assumed to have an order $n_s^{(1)}$ phase-type representation (δ_1, V_1), where V_1 is the $n_s^{(1)} \times n_s^{(1)}$ subgenerator matrix, and δ_1 is the $1 \times n_s^{(1)}$ vector with its entries equal to the initial probabilities to start in any of the 1 to $n_s^{(1)}$ states. If two retailers place an order at the same time, the additional minor setup/change-over time at the manufacturing plant before production starts has a phase-type distribution (δ_2, V_2) of order $n_s^{(2)}$. The production time of one unit has an order n_p phase-type representation with parameters (γ, U).

Only when the last unit of the order is produced, the order is replenished in the retailers' inventory. If two retailers place a joint order, the order is only delivered at the retailers as soon as the production of the joint order is completed. As clustering orders of several retailers together reduces fixed order costs and setup times, a coordination strategy such as a (s_j, c_j, S_j) can order policy can be considered by the retailers. (Note that if $s_j = c_j$, the can order policy reduces to an (s_j, S_j) policy.) We define $o_{max}^{(j)}$ as the maximum order quantity placed by retailer j: $S_j - s_j + m_j - 1$. If inventory is not sufficient to fulfil demand, unmet demand is backlogged. We assume that a fixed cost per order is charged (independent of whether the order was placed by one or multiple retailers), called the major setup cost K, and a fixed cost per retailer is charged, called the minor setup cost k_j, as this is commonly the case in joint replenishment problems (e.g., if only retailer j places an order, the resulting fixed ordering cost equals $K + k_j$, whereas, if both retailers place a joint order, the resulting fixed order cost equals $K + k_1 + k_2$). Furthermore, a holding h_j (resp. backlogging p_j) cost per unit that retailer j has in inventory (resp. backlog) per unit of time is charged. We denote the probability of having $S_1 - i$ units on hand as ns_i.

In this paper, we compute the expected total cost per time unit of retailer 1 denoted as $C(s_1, c_1, S_1)$, as the sum of major and minor fixed order costs, holding costs, and backlogging costs:

$$C(s_1, c_1, S_1) = (K + k_1)(rate_1 + rate_{CAN1}) + k_1 rate_{CAN2} + h_1 [NS]_1^+ + p_1 [NS]_1^-, \qquad (1)$$

where $[NS]_1^+$ refers to the expected number of units in inventory at retailer 1, and $[NS]_1^-$ denotes the expected number of units backlogged at retailer 1 at a random point in time. We define $rate_1$ as the expected number of orders per time unit which were only placed by retailer 1 only and $rate_{CANj}$ as the expected number of joint orders per time unit which were triggered by retailer j (with $j \in \{1,2\}$).

The major difficulty in the computation of the expected total cost per time unit is to find $[NS]_1^+$ and $[NS]_1^-$. We show the derivation of the inventory level distribution at retailer 1 in Section 5.

Remark that the can order policy (s, c, S) causes the order arrival process at the manufacturing plant to have stochastic order quantities and a stochastic time between orders. We want to point out that the order quantities and the time between orders can be correlated (depending on the demand size distribution). Therefore, no standard queuing formulas can be applied. In the next section, we define a Markov process which characterizes the queuing model. Thanks to this Markov process, we are able to derive the joint probability that the inventory position of retailer 1 was equal to $S_1 - k$ at the moment when the order (which is currently in production) was placed and this order has spent a time x in the system (Section 3.2), we can compute the utilization rate of the production facility (Section 4), and in the end, we are able to compute the inventory level distribution and the expected total costs per time unit (Section 5).

3. Characterization of the queuing model as a Markov process

We set up a model to determine the expected total cost for each retailer independently (in the paper we focus on retailer 1). If we would set up a model to compute the expected total cost for both retailers simultaneously, the number of states would be much larger, and so does the computation time. Note that, although we only focus on the performance measures of one retailer, obviously we do include the impact of the orders of both retailers in our model.

3.1 Defining the Markov process

We start by setting up a continuous-time Markov process $(X_t, L_t)_{t\geq 0}$ that observes the system whenever the server at the manufacturer is busy, where X_t is the time the order has spent in service at time t at the production facility ($X_t \geq 0$). We define L_t as the state of the order which is in service at time t, which tracks the following information:

- Did retailer 1 initiate the order?
    - If so, what was the inventory position of retailer 1 when the order (that is currently in service) was placed? The inventory position before placing an order may range from s_1 to $s_1 + 1 - m_1$, so that m_1 possibilities exist. Note that this information is relevant when we want to compute the expected total cost of retailer 1. To be able to do so, we need to derive the joint probability $q_{k,n}$ that the inventory position of retailer 1 was equal to $S_1 - k$ at the moment when the order was placed, and that during the time the current order has spent in the production facility, n customers have arrived at retailer 1 (we will derive this in Section 5).
    - Is the order a joint order?
        - What is the current phase of the unit in service? We assume that the major setup time per order has a phase-type distribution with rang $n_s^{(1)}$, the minor change-over time has a phase-type distribution with rang $n_s^{(2)}$, and the production time per unit has a phase-type distribution with rang n_p. How many units still need to start/complete production? Notice that the number of units which needs to be produced may range from zero to $o_{max}^{(1)} + S_2 - s_2 - 1$ (which is equivalent to the maximum size of a joint order). The number of units which still need to start/complete production can attain the value zero, because the Markov process is defined such that it performs the setup phase (and the change-over phase) after producing the units of an order (whether we perform the setup and the change-over phase first or last has no impact on the performance measures of interest).
    - Is the order not a joint order?
        - What is the current phase (which can be a phase of the major setup time or of the unit production time) of the unit in service? No states refer to a minor change-over time, as the order is not a joint order. How many units still need to start/complete production? Notice that the maximum number of units now equals $o_{max}^{(1)}$.
        - What was the inventory position of retailer 2 at the moment when retailer 1 placed the order? We need to keep track of this information to determine the future orders of retailer 2. In our Markov process, if future demand arrivals occur at retailer 2, his inventory position is depleted from this position onwards (Note that if retailer 2 also placed an order (i.e., the order is a joint order), we do not need to track this information as we know that the inventory position then increases to S_2).
- Did retailer 2 trigger the order?
    - Is the order a joint order?

    - What was the inventory position of retailer 1 when the order was placed? As the joint order was initiated by retailer 2, $c_1 - s_1$ possible values for the inventory position of retailer 1 exist: $\{s_1 + 1, \cdots, c_1\}$. Although retailer 2 placed the order, we focus on the performance measures of retailer 1. Therefore, we need to know how much the inventory position of retailer 1 has depleted since the production facility has been busy producing the current order (placed by retailer 2). Thanks to this information, we are able to compute the inventory level distribution of retailer 1 in Section 5.
    - What is the current phase (which can be a phase of the major setup time, of the minor change-over time or of the unit production time) of the unit in service? How many units still need to be produced? Notice that the maximum number of units which need to be produced equals $o_{max}^{(2)} + S_1 - s_1 - 1$.
- Is the order not a joint order?
    - What was the inventory position of retailer 1 when the order was placed? As the order is not a joint order, the inventory position of retailer 1 at the moment when retailer 2 placed the order was between $c_1 + 1$and S_1, which implies that one should account for $S_1 - c_1$ possible values. Like before, thanks to this information, we are able to compute the inventory level distribution of retailer 1 in Section 5.
    - What is the current phase (which can be a phase of the major setup time or of the unit production time) of the unit in service? How many units still need to be produced? Notice that the maximum number of units equals $o_{max}^{(2)}$.

Based on this information, we define the state space of the Markov process $(X_t, L_t)_{t \geq 0}$ as $\mathbb{R}^+ \times (\{P_1\} \cup \{P_2\})$, where P_1 and P_2 are the state space when the order is triggered by retailer 1 and retailer 2, respectively. We define P_1 as:

$$P_1 = m_1 \times \left(\left\{C_0^{(1)}\right\} \cup \left\{C_1^{(1)}\right\}\right), \tag{2}$$

$$\text{with } C_0^{(1)} = (S_2 - c_2) \times \left(n_s^{(1)} + o_{max}^{(1)} n_p\right), \tag{3}$$

$$\text{and } C_1^{(1)} = n_s^{(1)} + n_s^{(2)} + \left(o_{max}^{(1)} + S_2 - s_2 - 1\right) n_p. \tag{4}$$

Remark that Eq. (3) is related to orders which are only placed by retailer 1, whereas Eq. (4) is related to joint orders triggered by retailer 1.(2)

Analogously, if the order was triggered by retailer 2 we define P_2 as:

$$P_2 = \left\{C_0^{(2)}\right\} \cup \left\{C_1^{(2)}\right\}, \tag{5}$$

$$\text{with } C_0^{(2)} = (S_1 - c_1) \times \left(n_s^{(1)} + o_{max}^{(2)} n_p\right), \tag{6}$$

$$\text{and } C_1^{(2)} = (c_1 - s_1) \times \left(n_s^{(1)} + n_s^{(2)} + \left(o_{max}^{(2)} + S_1 - s_1 - 1\right) n_p\right). \tag{7}$$

Eq. (6) refers to individual orders placed by retailer 2. Eq. (7), on the other hand, refers to joint orders which were triggered by retailer 2.

Consider the bivariate Markov process $(X_t, L_t)_{t \geq 0}$, with $X_t \geq 0$ and $L_t \in \{1, \cdots, l\}$ (with $l = P_1 + P_2$). The process evolves as follows: the time X_t an order spent in the system at time t increases linearly unless a downward jump in X_t occurs when production of the order is completed. Assume that one starts in (x, i), which means that the order currently in production has spent a time x in the production facility and its state equals i (from this state we can obtain all the necessary information, which was described above). Three types of jumps can occur from (x,i):

1. A transition to (x,j) with rate $(A_0)_{i,j}$ (for $i \neq j$) when a unit of the order completes production or when the production or setup phase changes (the same order is still in production),
2. A jump in the interval $([x-u,x),j)$, for $0<u<x$, with a rate $A_{i,j}(u)$, where we denote $dA_{i,j}(u)$ as its density function. Production completion occurred. The next order is now in production. This new order in production was already ordered before the production completion occurred, therefore, the new order has spent some time in the queue. If the inter-arrival time between the replenished and the subsequent order is at most u, we know that the subsequent order spent at least $x-u$ time units waiting in queue (which is illustrated in Figure 2).
3. A jump to $(0,j)$ with rate $\int_{u=x}^{\infty} dA_{i,j}(u)$ when the order is replenished and the queue is empty (which occurs if the inter-arrival time with the next order is larger than the lead time of the replenished order).

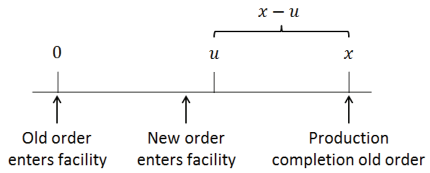
$x-u$
0
u
x
Old order enters facility
New order enters facility
Production completion old order

Figure 2. New order spent at least $x-u$ time units in queue.

Finally, define the (negative) diagonal entries of A_0 such that $\left(A_0 + \int_{u=0}^{\infty} dA(u)\right)e_l = e_l$ and assume that $A = A_0 + \int_{u=0}^{\infty} dA(u)$ is irreducible.

From Sengupta (1989), we know that the Markov process $(X_t, L_t)_{t\geq 0}$ has a matrix exponential distribution. In other words, there exists a $l \times l$ matrix T such that the vector $\pi(x)$, for $x \geq 0$, which contains the steady-state density of the states $(x,1)$ to (x,l) for any time $x \geq 0$ spent in the production facility (if and only if the utilization rate of the production facility $\rho < 1$), can be written as:

$$\pi(x) = \pi(0)\exp(Tx), \tag{8}$$

where T is the smallest non-negative solution to

$$T = A_0 + \int_{x=0}^{\infty} \exp(Tx)\, dA(x), \tag{9}$$

and $\pi(0) = \theta(-T)$, where θ is the unique invariant vector of A, i.e., the non-zero vector θ such that $\theta A = 0$. Remark that we need the vector $\pi(x)$ for any $x \geq 0$ in order to derive the inventory level distribution.

To derive the distribution of the time x an order has spent in the production facility at a random point in time, we need to define matrix T, which is based on matrix A_0 and $dA(x)$ (Eq. (9)).

Matrix A_0 is the rate matrix as long as the order is in service and no production completion occurred. We define this rate matrix as $A_0 = F_{++} = \begin{bmatrix} F_{++}^{(1)} & 0 \\ 0 & F_{++}^{(2)} \end{bmatrix}$, with

$$F_{++}^{(1)} = I_{m_1} \otimes \begin{bmatrix} I_{S_2-c_2} \otimes \begin{bmatrix} V_1 & & & & \\ u\delta_1 & U & & & \\ & u\gamma & \ddots & & \\ & & \ddots & \ddots & \\ & & & u\gamma & U \end{bmatrix} & 0 \\ 0 & \begin{bmatrix} V_2 & & & & \\ v_1\delta_2 & V_1 & & & \\ & u\delta_1 & U & & \\ & & \ddots & \ddots & \\ & & & u\gamma & U \end{bmatrix} \end{bmatrix}. \tag{10}$$

$F_{++}^{(1)}$ is a $P_1 \times P_1$ matrix giving the transition rates in the production system when retailer 1 triggered an order. If only retailer 1 placed an order, one can observe that the upper left part of the matrix keeps track of the inventory position at retailer 2 ($I_{S_2-c_2}$ is the identity matrix of size $S_2 - c_2$). Because we want to compute the performance measures of retailer 1, we need to keep track of the inventory position of retailer 1 at the moment when the order was placed (as is indicated by the identity matrix I_{m_1} of size m_1). Furthermore, we define $u = -Ue$, $v_1 = -V_1 e$, and e the unit vector.

The matrix $F_{++}^{(2)}$ gives the transition rates in the production system when retailer 2 reached his order point. Matrix $F_{++}^{(2)}$ is defined as follows:

$$F_{++}^{(2)} = \begin{bmatrix} I_{S_1-c_1} \otimes \begin{bmatrix} V_1 & & & & \\ u\delta_1 & U & & & \\ & u\gamma & \ddots & & \\ & & \ddots & \ddots & \\ & & & u\gamma & U \end{bmatrix} & 0 \\ 0 & I_{c_1-s_1} \otimes \begin{bmatrix} V_2 & & & & \\ v_1\delta_2 & V_1 & & & \\ & u\delta_1 & U & & \\ & & \ddots & \ddots & \\ & & & u\gamma & U \end{bmatrix} \end{bmatrix}. \tag{11}$$

$F_{++}^{(2)}$ is a $P_2 \times P_2$ matrix with $I_{S_1-c_1}$ the identity matrix of size $S_1 - c_1$, $I_{c_1-s_1}$ the identity matrix of size $c_1 - s_1$, $u = -Ue$, $v_1 = -V_1 e$, and e the unit vector. Note $I_{S_1-c_1}$ and $I_{c_1-s_1}$ because we want to compute the inventory level distribution at retailer 1.

As we explained before, the computation of matrix T is also based on $dA(x)$. $dA(x)$ is the density function of matrix $A(x)$, with $A(x)$ the rate matrix to go from one state (immediately before production completion) to another state which observes the start of production of the subsequent order which was ordered at most u time units after the previous order. $dA(x)$ is defined as:

$$dA(x) = F_{+-}\exp(F_{--}x)F_{-+} \tag{12}$$

Define F_{--} as the following $(S_1 - s_1)(S_2 - s_2) \times (S_1 - s_1)(S_2 - s_2)$ matrix (with its states referring to the inventory positions of retailer 1 and retailer 2: $(S_1, S_2), \cdots, (S_1, s_2 + 1), \cdots, (s_1 + 1, S_2), \cdots, (s_1 + 1, s_2 + 1)$):

$$F_{--} = \begin{bmatrix} -\lambda^{(1)} - \lambda^{(2)} & \cdots & \lambda^{(2)} d^{(2)}_{S_2-s_2-1} & \cdots & \lambda^{(1)} d^{(1)}_{S_1-s_1-1} & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & -\lambda^{(1)} - \lambda^{(2)} & \cdots & 0 & \cdots & \lambda^{(1)} d^{(1)}_{S_1-s_1-1} \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ 0 & \cdots & 0 & \cdots & -\lambda^{(1)} - \lambda^{(2)} & \cdots & \lambda^{(2)} d^{(2)}_{S_2-s_2-1} \\ \vdots & & \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & -\lambda^{(1)} - \lambda^{(2)} \end{bmatrix}. \qquad (13)$$

The matrix F_{--} gives the transition rates of changes in the inventory positions due to arriving demands before the order point of one of both retailers is reached.

Define F_{-+} as the $(S_1 - s_1)(S_2 - s_2) \times l$ matrix which gives the rates of arriving demands which trigger an order: $F_{-+} = \begin{bmatrix} F^{(1)}_{-+} & F^{(2)}_{-+} \end{bmatrix}$ with $F^{(1)}_{-+}$ a $(S_1 - s_1)(S_2 - s_2) \times P_1$ matrix which contains the rates of arriving demands at retailer 1 which reduce the inventory position of retailer 1 to the order point s_1 or below, such that retailer 1 places an order. The matrix is defined as follows:

$$F^{(1)}_{-+} = \sum_{k=0}^{m_1-1} \left(\begin{bmatrix} -\lambda^{(1)} & \lambda^{(1)} d^{(1)}_1 & \lambda^{(1)} d^{(1)}_2 & \lambda^{(1)} d^{(1)}_3 & \cdots & \lambda^{(1)} d^{(1)}_{S_1-s_1+k} \\ 0 & -\lambda^{(1)} & \lambda^{(1)} d^{(1)}_1 & \lambda^{(1)} d^{(1)}_2 & \cdots & \lambda^{(1)} d^{(1)}_{S_1-s_1+k-1} \\ \vdots & \ddots & \ddots & \ddots & \cdots & \vdots \\ 0 & \cdots & 0 & -\lambda^{(1)} & \cdots & \lambda^{(1)} d^{(1)}_{k+1} \end{bmatrix} \otimes I_{S_2-s_2} \right) \Gamma^{(1)}_k, \qquad (14)$$

with $\Gamma^{(1)}_k$ a $(S_2 - s_2) \times P_1$ matrix with all its entries equal to zero, except for the entries ranging from $\left(i, \Delta_1 + (i-1)\left(o^{(1)}_{max} n_p + n^{(1)}_s\right) + 1\right)$ to $\left(i, \Delta_1 + (i-1)\left(o^{(1)}_{max} n_p + n^{(1)}_s\right) + n_p\right)$ for $i \in \{1, \cdots, S_2 - c_2\}$, and from $\left(i, \Delta_1 + C^{(1)}_0 + n^{(2)}_s + (i-1)n_p + 1\right)$ to $\left(i, \Delta_1 + C^{(1)}_0 + n^{(2)}_s + i n_p\right)$ for $i \in \{S_2 - c_2 + 1, \cdots, S_2 - s_2\}$ with $\Delta_1 = k\left(C^{(1)}_0 + C^{(1)}_1\right) + n^{(1)}_s + (S_1 - s_1 + k - 1)n_p$. These entries equal the $1 \times n_p$ initial vector γ which gives the probabilities to start in the n_p different states of the phase-type production time. Matrix $\Gamma^{(1)}_k$ keeps track of the state in which production starts if the size of the order placed by retailer 1 equals $S_1 - s_1 + k$ and if the inventory position of retailer 2 was equal to $S_2 + 1 - i$ at the moment when retailer 1 triggered an order. If only retailer 1 placed an order (or equivalently, if $i \in \{1, \cdots, S_2 - c_2\}$), matrix $\Gamma^{(1)}_k$ also keeps track of the inventory position at retailer 2 at the moment when the order was placed. Note that for $i \in \{1, \cdots, S_2 - c_2\}$, the zero-entries of matrix $\Gamma^{(1)}_k$ correspond to:

- the states of the remaining $S_1 - s_1 + k - 1$ units of the order that need to be produced and the n_p phases,
- the $n^{(1)}_s$ phases of the major setup time,
- the states of the order sizes different from $S_1 - s_1 + k$,
- the states of orders placed by retailer 1 when retailer 2 had an inventory position different from $S_2 + 1 - i$ (for $i \in \{1, \cdots, S_2 - c_2\}$).

For $i \in \{S_2 - c_2 + 1, \cdots, S_2 - s_2\}$, the zero-entries of matrix $\Gamma^{(1)}_k$ correspond to:

- the states of the remaining $S_1 - s_1 + i - 1 + k - 1$ units of the order that need to be produced (and the n_p phases),
- the $n^{(1)}_s$ phases of the major setup time,

- the $n_s^{(2)}$ phases of the change-over time,
- the states of the order sizes different from $S_1 - s_1 + k$,
- the $C_0^{(1)}$ states referring to an individual order of size $S_1 - s_1 + k$.

$F_{-+}^{(2)}$ is a $(S_1 - s_1)(S_2 - s_2) \times P_2$ matrix, which holds the transition rates of demand arriving at retailer 2 which triggers an order of retailer 2.

$$F_{-+}^{(2)} = \left(I_{S_2 - s_2} \otimes \begin{bmatrix} -\lambda^{(2)} & \lambda^{(2)} d_1^{(2)} & \lambda^{(2)} d_2^{(2)} & \lambda^{(2)} d_3^{(2)} & \cdots & \lambda^{(2)} d_{S_2 - s_2 + k}^{(2)} \\ 0 & -\lambda^{(2)} & \lambda^{(2)} d_1^{(2)} & \lambda^{(2)} d_2^{(2)} & \cdots & \lambda^{(2)} d_{S_2 - s_2 + k - 1}^{(2)} \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & -\lambda^{(2)} & \cdots & \lambda^{(2)} d_{k+1}^{(2)} \end{bmatrix} \right) \Gamma_k^{(2)}, \qquad (15)$$

with $\Gamma_k^{(2)}$ a $(S_1 - s_1) \times P_2$ matrix with all its entries equal to zero, except for the entries ranging from $\left(i, \Delta_2 + (i-1)\left(o_{max}^{(2)} n_p + n_s^{(1)}\right) + 1\right)$ to $\left(i, \Delta_2 + (i-1)\left(o_{max}^{(2)} n_p + n_s^{(1)}\right) + n_p\right)$ for $i \in \{1, \cdots, S_1 - c_1\}$, and from $\left(i, \Delta_2 + C_0^{(2)} + (i - S_1 - c_1 - 1)\left(\left(S_1 - s_1 - 1 + o_{max}^{(2)}\right) n_p + n_s^{(1)} + n_s^{(2)}\right) + n_s^{(2)} + (i-1) n_p + 1\right)$ to $\left(i, \Delta_2 + C_0^{(2)} + (i - S_1 - c_1 - 1)\left(\left(S_1 - s_1 - 1 + o_{max}^{(2)}\right) n_p + n_s^{(1)} + n_s^{(2)}\right) + n_s^{(2)} + i n_p\right)$ for $i \in \{S_1 - c_1 + 1, \cdots, S_1 - s_1\}$ with $\Delta_2 = n_s^{(1)} + (S_2 - s_2 - 1 + k) n_p$, which equal the $1 \times n_p$ initial production vector γ. Matrix $\Gamma_k^{(2)}$ describes the transition rates caused by a demand arrival at retailer 2 which triggers retailer 2 to place an order. One needs to keep track of the inventory position at retailer 1 (because we want to compute the inventory level distribution of retailer 1).

Finally, we define the matrix F_{+-} which holds the transition rates (after production completion) to the inventory positions. The retailer who placed an order starts with an inventory position equal to the order-up-to level, the retailer who did not place an order stays at the same inventory position which he had at the moment when the other retailer placed an order (remember that we had to keep track of the inventory position of retailer 2 when only retailer 1 placed an order). Define the $l \times (S_1 - s_1)(S_2 - s_2)$ matrix as follows: $F_{+-} = \begin{bmatrix} F_{+-}^{(1)} \\ F_{+-}^{(2)} \end{bmatrix}$, with $F_{+-}^{(1)}$ a $P_1 \times (S_1 - s_1)(S_2 - s_2)$ matrix and $F_{+-}^{(2)}$ a $P_2 \times (S_1 - s_1)(S_2 - s_2)$ matrix:

$$F_{+-}^{(1)} = e_{m_1} \otimes \begin{bmatrix} \alpha_1 \otimes \left(I_{S_2 - c_2} \otimes \begin{bmatrix} v_1 \\ \vdots \\ v_1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \right) \\ \begin{bmatrix} v_2 \\ \vdots \\ v_2 \\ 0 \\ \vdots \\ 0 \end{bmatrix} (\alpha_1 \otimes \alpha_2) \end{bmatrix} \text{ and } F_{+-}^{(2)} = \begin{bmatrix} I_{S_1 - c_1} \otimes \left(\begin{bmatrix} v_1 \\ \vdots \\ v_1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \alpha_2 \right) \\ \left(e_{c_1 - s_1} \otimes \begin{bmatrix} v_2 \\ \vdots \\ v_2 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \right) (\alpha_1 \otimes \alpha_2) \end{bmatrix}, \qquad (16)$$

with the $1 \times (S_j - s_j)$ matrix $\alpha_j = [1 \quad 0 \quad \cdots \quad 0]$, the $\left(n_s^{(1)} + o_{max}^{(1)} n_p\right) \times 1$ vector $\begin{bmatrix} v_1 \\ \vdots \\ v_1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$ which is a vector filled with zeros, except for the first $n_s^{(1)}$ entries which equal $v_1 = -V_1 e$, and the $C_1^{(1)} \times 1$ vector $\begin{bmatrix} v_2 \\ \vdots \\ v_2 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$ which is filled with zeros, except for the first $n_s^{(2)}$ entries which equal $v_2 = -V_2 e$. One can observe that $F_{+-}^{(1)}$ refers to the order completion of an order which was originally started by retailer 1. Similarly, $F_{+-}^{(2)}$ refers to an order which was placed by retailer 2. In both matrices, the upper part considers the orders for an inventory position of the other retailer larger than the can order level. The lower part of both matrices focuses on the order completion of a joint order. As we explained before, our goal is to determine the inventory level distribution of retailer 1. Therefore, we need to keep track of the inventory position of retailer 1 at the moment when the order was placed, as can be observed in both matrices (i.e. the vector e_{m_1}, the matrix $I_{S_1 - c_1}$, and the vector $e_{c_1 - s_1}$). Another point which was already mentioned earlier, is that one needs to keep the inventory position of retailer 2 in mind if only retailer 1 placed an order (this means that retailer 2 does not raise his inventory position to the order-up-to level, and therefore one needs to remember the correct inventory position), this is taken into account by both matrices by making use of the matrix $I_{S_2 - c_2}$.

3.2. Derivation of the steady state vector of the fluid queue

From A_0 and $dA(x)$ we can compute matrix T iteratively (starting with $T_0 = 0$):

$$T_{n+1} = A_0 + \int_{x=0}^{\infty} \exp(T_n x)\, dA(x). \tag{17}$$

Unfortunately, this method is impractical for high loads (as it results in linear convergence). Therefore, we construct a fluid queue (Latouche 2006) which results in quadratic convergence. We define the fluid as the time the current order in production has spent in the production facility. We know that the time spent in the system increases linearly over time as every unit is being produced and as the order undergoes a setup time. If a production completion occurs, the time spent in the system (which now refers to a new order which is currently in production) starts from zero (if the new order did not spend any time waiting in queue) or starts from a positive value (which is equal to the time the order already spent waiting in queue). The fluid queue is constructed by replacing the immediate downward jumps (after a production completion) by intervals of the appropriate length during which the level decreases linearly. In other words, we obtain a fluid queue with l phases in which the fluid increases (our original l production states) and $(S_1 - s_1)(S_2 - s_2)$ phases in which the fluid decreases (the $(S_1 - s_1)(S_2 - s_2)$ artificial states that are added). Let F_{++}, F_{+-}, F_{-+} and F_{--} hold the rates at which the phase changes while the fluid increases (i.e. production of the same order continues), from an up to a down phase (i.e. production completion occurs), from a down to an up phase (i.e. a new order starts production) and while going down (i.e. demand arrives at the retailers, but both retailers did not reach their order point yet), respectively. Notice, F defined as $F = \begin{bmatrix} F_{++} & F_{+-} \\ F_{-+} & F_{--} \end{bmatrix}$ is the rate matrix of a continuous-time Markov chain. If we take the expression for the steady state of a fluid queue [Latouche 2006] and observe the queue only when the level increases, one finds that its steady state $\pi(x)$ has a matrix exponential form $\pi(x) = \pi(0)\exp(Tx)$, with $T = F_{++} + \Psi F_{-+}$ where Ψ is the minimal nonnegative solution to an algebraic Riccati equation [Latouche 2006]. Thus, to compute T, it suffices to determine Ψ. The computation of Ψ can be done through the algorithm of Guo et al. [2007]. However, as matrix F_{++} is a block diagonal matrix, we apply the algorithm of Meini [2013], which is faster and requires less memory.

$\pi(0)$ equals $\theta(-T)$, where θ is the stochastic vector solving $\theta A = 0$, with $A = A_0 + F_{+-}(-F_{--})^{-1}F_{-+}$. Finally, the vector $\pi(x)$, holding the steady-state density of the states $(x,1)$ to (x,l) for any time $x \geq 0$ spent in the production facility (if and only if the utilization rate of the production facility $\rho < 1$), for $x \geq 0$, can be computed as [Sengupta 1989]: $\pi(x) = \pi(0)\exp(Tx)$.

4. Defining the utilization rate of the production system

Based on Section 3, we are able to define the joint probability $q_{k,n}$. However, if one wants to compute the inventory level distribution and the expected total cost of retailer 1, one also needs to define the utilization rate of the production system: the computation of the inventory level distribution will be different if the production system is busy (with a probability ρ) and if the production system is idle (with probability $1-\rho$). Furthermore, for the computation of the expected fixed ordering cost (which is part of the cost function), one needs the expected number of orders which were only placed by retailer j and the expected number of joint orders which were triggered by retailer j (with $j \in \{1,2\}$). In this Section, we provide a brief explanation on the derivation of the utilization rate and the expected number of orders placed (for further details on the matrices, we refer to our full working paper; Noblesse et al. 2014).

The utilization rate at the production facility is defined as follows:

$$\begin{aligned}\rho = \; & \textstyle\sum_{j=1}^{2}\left[\lambda^{(j)}(\gamma(-U)^{-1}e)\left[\sum_{i=1}^{m_j} id_i^{(j)}\right]\right] + (\delta_1(-V_1)^{-1}e)(rate_1 + rate_2) + \\ & (\delta_1(-V_1)^{-1}e + \delta_2(-V_2)^{-1}e)(rate_{CAN1} + rate_{CAN2}), \end{aligned} \tag{18}$$

with $\lambda^{(j)}$ the expected arrival rate of customers at retailer j, $(\gamma(-U)^{-1}e)$ the expected time to produce one unit, and $\sum_{i=1}^{m_j} id_i^{(j)}$ the expected demand size per customer arriving at retailer j. $(\delta_1(-V_1)^{-1}e)$ equals the expected major setup time, $(\delta_2(-V_2)^{-1}e)$ the expected minor change-over time, $rate_j$ the expected number of orders per time unit which were only placed by retailer j, and $rate_{CANj}$ the expected number of joint orders (per time unit) triggered by retailer j.

We define the expected number of orders per time unit which were only placed by retailer j ($rate_j$ in the equation above) as the inverse of the expected time between two subsequent orders which were only placed by retailer j. The time between two subsequent orders can be described by a phase-type distribution with an initial state vector π_{orderj} and a transition matrix $F_{--orderj}$, such that the expected time between subsequent orders equals $\pi_{orderj}\left(F_{--orderj}\right)^{-1}e$. The expected number of joint orders which were triggered by retailer j ($rate_{CANj}$ in Eq. (18)) can be computed in an analogous way. A detailed definition of the matrices $F_{--order1}$, $F_{--order2}$, F_{--CAN1}, and F_{--CAN2} can be found in the working paper Noblesse et al. [2014].

5. The probability distribution of inventory levels and the expected total cost

We define $q_{k,n}$ as the joint probability that the inventory position of retailer 1 was equal to $S_1 - k$ at the moment when the order (which is currently in production) was placed, and that

n customers have arrived at retailer 1 (each of them having a random demand size) since this order was placed, given that the server at the manufacturer is busy (which is illustrated in Figure 3). The joint probability $q_{k,n}$ is used to compute the probability distribution of inventory levels at retailer 1.

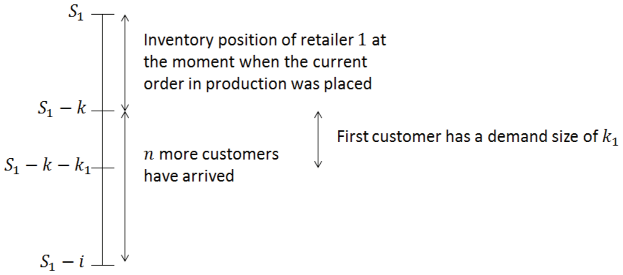
S_1
Inventory position of retailer 1 at the moment when the current order in production was placed
$S_1 - k$
First customer has a demand size of k_1
$S_1 - k - k_1$
n more customers have arrived
$S_1 - i$

Figure 3. Probability of inventory levels if the production facility is busy.

We start by computing the joint probabilities $q_{k,n}$. As the Markov process $(X_t, L_t)_{t\geq 0}$ has the matrix exponential form [Sengupta 1989], we can obtain the joint probability that the inventory position of retailer 1 was $S_1 - k$ when the current order in production was placed and that the time this order has spent in the production facility up till now equals x. As demand arrivals at retailer 1 have a Poisson distribution with arrival rate $\lambda^{(1)}$, the probability of n demand arrivals during a time interval x equals $\frac{\left(\lambda^{(1)}x\right)^n}{n!}\exp\left(-\lambda^{(1)}x\right)$. Define the vector $q_n = \left(q_{0,n}, \cdots, q_{S_1-s_1+m_1-1,n}\right)$. Using the matrix exponential form of the steady state of $(X_t, L_t)_{t\geq 0}$, we find

$$\begin{aligned}
q_{(S_1-s_1+(0:m_1-1)),n} &= \pi(0)\int_{x=0}^{\infty}\exp(Tx)\frac{\left(\lambda^{(1)}x\right)^n}{n!}\exp\left(-\lambda^{(1)}x\right)dx\left(I_{m_1}\otimes e_{C_0^{(1)}+C_1^{(1)}}\right)\\
&= \pi(0)\left(\lambda^{(1)}\right)^n\left(\lambda^{(1)}I_l - T\right)^{-(n+1)}\left(I_{m_1}\otimes e_{C_0^{(1)}+C_1^{(1)}}\right)\\
q_{(S_1-c_1:S_1-s_1-1),n} &= \pi(0)\int_{x=0}^{\infty}\exp(Tx)\frac{\left(\lambda^{(1)}x\right)^n}{n!}\exp\left(-\lambda^{(1)}x\right)dx\left(I_{c_1-s_1}\otimes e_{C_1^{(2)}/(c_1-s_1)}\right)\\
&= \pi(0)\left(\lambda^{(1)}\right)^n\left(\lambda^{(1)}I_l - T\right)^{-(n+1)}\left(I_{c_1-s_1}\otimes e_{C_1^{(2)}/(c_1-s_1)}\right)\\
q_{(0:S_1-c_1-1),n} &= \pi(0)\int_{x=0}^{\infty}\exp(Tx)\frac{\left(\lambda^{(1)}x\right)^n}{n!}\exp\left(-\lambda^{(1)}x\right)dx\left(I_{S_1-c_1}\otimes e_{C_0^{(2)}/(S_1-c_1)}\right)\\
&= \pi(0)\left(\lambda^{(1)}\right)^n\left(\lambda^{(1)}I_l - T\right)^{-(n+1)}\left(I_{S_1-c_1}\otimes e_{C_0^{(2)}/(S_1-c_1)}\right) \qquad (19)
\end{aligned}$$

Next, we compute the probability ns_i that the number of units on hand at retailer 1 equals $S_1 - i$, for $i \geq 0$, at an arbitrary point in time. With probability ρ, the server will be busy and we will compute ns_i from $q_{k,n}$, using the fact that the demand sizes are i.i.d. When the server is idle, we can compute the probabilities of having $S_1 - i$ units on hand, for $i = 0$ to $S_1 - s_1 - 1$ as the steady-state vector θ_{neg} of the fluid queue, given that the amount of fluid is zero. This stochastic vector can be computed as $\theta_{neg}(F_{--} + F_{-+}\Psi) = 0$. Hence, for $i \geq 0$

$$ns_i = (1-\rho)\left(\theta_{neg}\right)_{i+1} + \rho\sum_{k\leq i, n\leq i-k} q_{k,n}\sum_{\substack{k_1,\cdots,k_n>0\\ k_1+\cdots+k_n=i-k}}\left(\prod_{s=1}^{n}p_{k_s}\right) \qquad (20)$$

Where the latter sum corresponds to an n-fold convolution of the demand size distribution.

Based on the inventory level distribution at retailer 1 (Eq. (20)), the expected number of orders per time unit placed by retailer 1, and the expected number of joint orders placed by retailer 2 per time unit, we compute the expected total cost of retailer 1 per period:

$$C(s_1, c_1, S_1) = h_1 \sum_{i=0}^{S_1} n_i (S_1 - i) + p_1 \sum_{i=S_1+1}^{\infty} n_i (i - S_1) + (K + k_1)(rate_1 + rate_{CAN1}) + k_1 rate_{CAN2} \quad (21)$$

7. Conclusion

In this paper, we study a production/inventory model with one production facility and multiple retailers. Applying matrix analytic methods, we are able to compute the distribution of the time an order spent in the production facility, the inventory level distributions and the resulting expected total cost taking endogenous lead times into account. Some first numerical experiments (which we omitted given the page limits) show that taking into account the impact of the replenishment policy of other retailers, who place orders at same production facility, can decrease the expected total cost of a retailer significantly. If a fixed cost and/or setup time per order exists, a can order policy can decrease expected total costs even further (compared to a setting with (s, S) policies).

8. Acknowledgements

We thank Beatrice Meini (Università di Pisa) for her helpful suggestions on the iterative algorithm we used for the computation of the expected costs.

9. References

Arrow, K.J., Harris, T., Marschak, J., 1951. Optimal inventory policy. *Econometrica*, 19 (3), 250-272.

Balintfy, J.L., 1964. On a basic class of multi-item inventory problems. *Management Science*, 10, 287-297.

Benjaafar, S., Cooper, W. L., Kim, J.-S., 2005. On the benefits of pooling in production-inventory systems. *Management Science*, 51 (4), 548-565.

Dror, M., Hartman, B.C., Chang, W., 2012. The cost allocation issue in joint replenishment. *International Journal of Production Economics*, 135, 242-254.

Guo, C.H., Iannazzo, B., Meini, B., 2007. On the doubling algorithm for a (shifted) nonsymmetric algebraic Riccati equation. *SIAM Journal on Matrix Analysis and Applications*, 29, 1083-1100.

Karmarkar, U.S., 1987. Lot sizes, lead times and in-process inventories. *Management Science*, 33 (3), 409-418.

Latouche, G., 2006. Structured Markov chains in applied probability and numerical analysis. *Markov Anniversary Meeting*, 69-78.

Meini, B., 2013. On the numerical solution of a structured nonsymmetric algebraic Riccati equation, *Performance Evaluation*, 70, 682-690.

Noblesse, A.M., Boute, R.N., Lambrecht, M.R., Van Houdt, B., 2014. Coordination of lot sizing decisions in production/inventory systems with multiple retailers, unpublished working paper, KU Leuven, Leuven.

Özkaya, B.Y., Gürler, Ü., Berk, E., 2006. The stochastic joint replenishment problem: A new policy, analysis, and insights. *Naval Research Logistics*, 53, 525-546.

Sengupta, B., 1989. Markov processes whose steady state distribution is matrix-exponential with an application to the GI/PH/1 queue. *Advances in Applied Probability,* 159-180.

Timmer, J., Chessa, M., Boucherie, R.J., 2013. Cooperation and game-theoretic cost allocation in stochastic inventory models with continuous review. *European Journal of Operational Research*, 231, 567-576.