An explicit analysis of the lead time syndrome: Stability condition and performance evaluation

Baris Selcuk, Ivo J.B.F. Adan, Ton G de Kok, Jan C Fransoo

An explicit analysis of the lead time syndrome: Stability condition and performance evaluation

Journal:	*International Journal of Production Research*
Manuscript ID:	TPRS-2006-IJPR-0284.R2
Manuscript Type:	Original Manuscript
Date Submitted by the Author:	11-Apr-2007
Complete List of Authors:	Selcuk, Baris; Technische Universiteit Eindhoven, Technology Management Adan, Ivo; Technische Universiteit Eindhoven, Mathematics and Computer Science De Kok, Ton; Technische Universiteit Eindhoven, Technology Management Fransoo, Jan; Technische Universiteit Eindhoven, Technology Management
Keywords:	MARKOV MODELLING, QUEUEING MODELS, MANUFACTURING MANAGEMENT, DUE-DATE ASSIGNMENT
Keywords (user):	

An explicit analysis of the lead time syndrome: Stability condition and performance evaluation

Barış Selçuk*
Department of Technology Management, Technische Universiteit Eindhoven, The Netherlands, b.selcuk@tm.tue.nl

Ivo J.B.F. Adan
Department of Mathematics and Computer Science, Technische Universiteit Eindhoven, The Netherlands, iadan@win.tue.nl

Ton G. De Kok, Jan C. Fransoo
Department of Technology Management, Technische Universiteit Eindhoven, The Netherlands, a.g.d.kok@tm.tue.nl, j.c.fransoo@tm.tue.nl

Updating planned lead times in response to changing workload levels leads to erratic ordering behavior, resulting in even larger variability in the workload levels and flow times. This phenomenon is called *lead time syndrome*, and describes the cyclic interaction between adaptive planned lead times and order sizes. Although it has been conceptually defined and intuitively accepted, formal analysis with analytical evaluation of the phenomenon has not been conducted. The objective of this paper is to provide a stronger understanding of the lead time syndrome, and to give new insights into the effects of frequently updating planned lead times. We develop a two-dimensional Markov process to model a single-item production process with orders released sensitive to the planned lead time. Using matrix-geometric methods, analytical results on the utilization level and the variability in the system are presented in relation to the frequency of updating the planned lead time. Although the average utilization level is always retained, the lead time syndrome causes an increase in the average workload level and the actual flow times of the completed orders. The variability of the planned lead time increases with the update frequency except at the utilization boundaries, where the relative effect of the update frequency diminishes.

Key words: lead time setting, order release, quasi-birth-and-death processes, matrix-geometric methods

1. Introduction

Planned lead time is one of the key modeling parameters in planning and control of multi-stage production-distribution systems. It refers to the estimated/planned duration of time that elapses between the release time of an order and the time that the order is completed and made available to customers or down-stream production stages. Setting the correct planned lead times that fit the associated production-distribution systems has been of concern both in academia and industry

* Corresponding author.

for decades. This is due to the fact that planned lead time is a fixed parameter that refers to a dynamic and uncertain frame on a continuous time axis.

The research on lead time management can be classified into three related lines: (1) workload control to satisfy fixed planned lead times, (2) due-date assignment associated with the existence of internal due-dates, and (3) setting planned lead times for time-phased order releases in a *Material Requirements Planning* (MRP) context. Workload control is a means to stabilize the flow times based on the intuition derived from Little's law, and it is related to orders to be loaded to the shop or not. Due-date assignment is complementary to workload control, since it is related to the scheduling of the orders that are loaded. MRP can be considered as an order release mechanism that aims to satisfy forecasted demand, and provides inputs for the control of the workload at specific work-centers and assignment of due-dates to specific orders. The planned lead time in this case is used to fine tune the inventory position with demand forecasts. In this paper, we consider modeling and evaluating the situation with dynamic planned lead times.

The due-date assignment literature provides us with numerous techniques for setting the planned lead times dynamically. The most popular ones utilize the order characteristics together with dynamic shop load information. Examples include *Total Work Content* (TWK) - the planned lead time of an order is set in proportion to the total processing time of the order (Conway et al. (1967), and Kanet (1986)) - and *Jobs in Queue* (JIQ) - the flow time of an order is estimated based on a proportion of the total number of orders in queue on its routing (Chang (1994)). The majority of the studies in this line are conducted in a produce-to-order job-shop environment concentrating on service-related performance measures such as the average length of the planned lead time, tardiness, earliness, etc. One of the fundamental assumptions is that the order characteristics are determined externally and independently from the planned lead time, which generally holds true for engineer-to-order situations. However, for example in batch processing industries where the orders are released and processed in batches of production items, the orders are generally released according to some anticipated knowledge on the total demand levels during the planned lead times. In a multi-stage production-inventory system, one would expect that the planned lead times are used for coordination purposes between stages.

In an MRP context, fixed planned lead times have been widely accepted. The concept of fixed planned lead times is further developed by De Kok and Fransoo (2003), and by Spitter et al. (2005) within the context of planning supply chain operations with capacity constraints. Also in various make-to-stock settings fixed planned lead times have been commonly used (e.g., Karaesmen et al. (2002), and Liberopoulos and Koukoumialos (2005)). Most of the studies on setting planned

lead times for MRP systems approach the problem from a static view, and strive to find *fixed* planned lead times that best fit the planned situation (e.g., Yano (1987), Molinder (1997), and Enns (2001)). Hoyt (1978) is the first to criticize the fixed planned lead times, and argues that the planned lead times should be dynamic, in a sense that they reflect the dynamic operational characteristics of a production process, in particular, by looking at the average queue length and the average output realized recently. This discussion is further enhanced by Kanet in a series of papers; he first investigated the various effects that planned lead times have on a multi-stage production-inventory system (Kanet (1982)), then he emphasized the favorable results in terms of order tardiness achieved by the TWK rule (Kanet (1986)). Since then, the research on dynamic planned lead times for supply chain situations has not attracted much attention. The state-of-the-art on this topic models the dynamic planned lead times from the perspective of job-shop scheduling, where the job-arrival patterns are derived from an MRP explosion process (e.g., Zijm and Buitenhek (1996), and Lambrecht et al. (1998)). The available techniques have been proven to make considerable improvement over the fixed planned lead times (see Vandaele et al. (2000) for a case study). Additionally, Homem-de-Mello et al. (1999) presents a method for setting release times for jobs in a stochastic production flow line. However, these studies do not consider the effect of dynamic planned lead times on the order generating process. Recently, Enns and Suwanruji (2004) investigated the use of exponentially smoothed planned lead times, and showed the sensitivity of the system to safety lead time factors and lot-sizing choices. Selçuk et al. (2006) also used exponentially smoothed planned lead times for a capacitated multi-stage make-to-stock system and identified the conditions that generate erratic order release patterns.

The variability introduced by frequent updates of planned lead times has been discussed by various researchers and is generally denoted as the lead time syndrome (cf. Mather and Plossl (1978)). It is argued that closing the gap between planned lead times and actual flow times by updating the planned lead times results in uncontrolled order release patterns. As the planned lead times get larger, orders must be released earlier, queues get longer and flow times get longer which causes again larger planned lead times. It results from the fact that while releasing the orders the relationship between workload and flow times is ignored. The general consensus is that the lead time syndrome causes instability, and must be avoided. This reasoning has become one of the main arguments for controlling manufacturing flow times within the predetermined norms instead of forecasting them (e.g., Tatsiopoulos and Kingsman (1983), Plossl (1988), Kingsman et al. (1989), Zäpfel and Missbauer (1993), and Breithaupt et al. (2002)). By workload control, the mean and the variability of the time that an order spends on the shop floor can be reduced significantly.

However, the total order flow time as seen from an order release perspective may still possess a high level of variability.

Consequently, there are a number of opportunities for a continuing discussion on dynamic planned lead times in different contexts. In this paper, we aim to shed some formal light on this discussion by emphasizing the variability introduced by the lead time syndrome. We argue that an explicit analysis of the phenomenon has been needed to propose stronger, and new insights.

The lead time syndrome is becoming increasingly relevant due to the opportunities of frequent information exchange enabled by recent advances in data processing and storage technologies. Ubiquitous information such as inventory levels, work-in-process, and shop conditions at various stages of the supply chain may continuously be available to update planning parameters. An important question is to what extent all this information should be used to update plans and planning parameters. Insights, based on the lead time syndrome would suggest to make limited use of these updating capabilities due to the increased variability in the order releases in response to the operational changes. Let us describe the lead time syndrome by an example.

Lead Time Syndrome: An Example from MRP

Consider a producer of a single item and a downstream stock point where orders are released based on the MRP logic. Tables 1-3 provide the changes in the planned order releases depending on how the lead time changes. Gross requirements refer to the anticipated future independent/dependent demand for the item during each period. Scheduled receipts are the total quantity of items released previously and scheduled to be replenished by the end of each period. Projected available balance indicates the current and the future anticipated inventory level for the item at the end of each period. Assume that in period 1 the planned lead time is set to two periods. The plan for the order releases in this case is given in Table 1. To keep it clean and simple the gross requirements are expressed in terms of the lot size, which is equal to d in our example. In period 1 and with a planned lead time of two periods, orders of sizes d are planned to be released at the start of periods 1, 2, 3 and 4.

Table 1 MRP record in period 1, planned lead time = 2.

Period		1	2	3	4	5	6
Gross Requirements		d	d	d	d	d	d
Scheduled Receipts		d	0	0	0	0	0
Projected Available Balance	d	d	0	0	0	0	0
Planned Order Release		d	d	d	d		

Assume that due to some temporary deviations in the production the scheduled order to be received within period 1 cannot be finished, which causes the orders in the pipeline to be delayed

for one period. Then, in order to be realistic, the planned lead time is increased to three periods at the start of period 2. As given in Table 2, the previously planned order of size d at the start of period 2 is now changed to $2d$ due to the increase in the planned lead time.

Table 2 MRP record in period 2, planned lead time = 3.

Period		2	3	4	5	6	7
Gross Requirements		d	d	d	d	d	d
Scheduled Receipts		d	d	0	0	0	0
Projected Available Balance	0	0	0	$-d$	0	0	0
Planned Order Release		$2d$	d	d			

On the other hand, it may as well be the case that the production occurs faster than anticipated and the released order of period 1 is finished within period 1. Considering this unexpected decrease in the workload the planned lead time is decreased to one period, and consequently, the previously planned order of size d for period 2 is now canceled.

Table 3 MRP record in period 2, planned lead time = 1.

Period		2	3	4	5	6	7
Gross Requirements		d	d	d	d	d	d
Scheduled Receipts		0	0	0	0	0	0
Projected Available Balance	$2d$	d	0	0	0	0	0
Planned Order Release		0	d	d	d	d	

Our analysis in Tables 1-3 is also in line with the fact that MRP systems are equivalent to somewhat generalized base-stock systems, with the key difference being that MRP systems make decisions at each level using an echelon target stock that includes a forecast of future final demand. A detailed analysis of the equivalence of MRP and base-stock systems can be found in Lambrecht et al. (1984), and Buzacott et al. (1992). In Tables 1-3, one can easily check that the base-stock level is equal to $d\times$(planned lead time $+1$). One should also notice that the effect of updating the planned lead time is only transient (only the order of period 2 is affected) under the assumption that the planned lead time is going to be fixed in the future. The lead time syndrome is then realized when we repeatedly update the planned lead time.

An illustration of the lead time syndrome is provided in Table 4 for a duration of 12 time-periods. The demand is deterministic with a fixed level of d units/period, and the periodic production may vary with an expected quantity of d units/period. In each time-period the following sequence of events occurs: the planned lead time is set according to the total workload divided by the expected production quantity, the order for that period is released, produced items are delivered to the stock

Table 4 Order lead time sheet.

Period	Workload	Inventory	Lead Time	Order	Production
1	$2d$	0	2	d	d
2	$2d$	0	2	d	0
3	$3d$	$-d$	3	$2d$	d
4	$4d$	$-d$	4	$2d$	d
5	$5d$	$-d$	5	$2d$	d
6	$6d$	$-d$	6	$2d$	$2d$
7	$6d$	0	6	d	$2d$
8	$5d$	d	5	0	d
9	$4d$	d	4	0	d
10	$3d$	d	3	0	d
11	$2d$	d	2	0	0
12	$2d$	0	2	d	d

point, and demand is realized. The initial planned lead time is two periods with a total workload of $2d$.

In period 1 a standard order of size d is placed. In period 2, production can not occur due to some temporary breakdowns, and the total workload increases to $3d$ units. Therefore, the planned lead time is increased to three periods. The order in period 3 now includes the static case order plus an additional period's demand, thereby increasing the total workload to $4d$ units. As a result, the planned lead time is increased to four periods, and in period 4 an order of size $2d$ is placed again. The vicious cycle continues, and in period 6 the planned lead time is increased to six periods, and the workload to $6d$ units. In this period $2d$ units are produced due to faster production. Thus, the planned lead time is not changed in period 7, and a static order of size d is placed. The production goes faster again in period 7, decreasing the workload to $5d$ units and the planned lead time to five periods in period 8. The new order now includes the static order minus the excess of one period demand, which results in cancelation of the static order. This causes a further decrease in the workload and therefore in the planned lead time. The cyclic effect continues and in week 11 the planned lead time becomes two periods again.

From Table 4, one should notice that the deviation in the production quantities is amplified and carried onwards the workload levels through erratic order releases caused by updating the planned lead time. This behavior seems quite arbitrary for a rational planner, but the phenomenon is generally considered to be a relevant problem in real life decision support systems (cf. IBM (1972)), and also within industrial dynamics of Forrester (1980).

In this paper, we concentrate our analysis to the production process; specifically the workload level, utilization and the flow time of orders. From the perspective of the *production department* the orders are generated externally by an *inventory planner* which may be an MRP, base-stock or similar system. We model the production process as a single-server queue with continuous

arrival and processing of production orders. A two-dimensional Markov process is developed for this purpose, and using matrix-geometric methods we tackle the following research questions:

- How does the lead time syndrome affect the stability of the production process?
- What is the effect of the update frequency on the performance of the production department?

The outline of this paper is as follows. In Section 2 we describe the essential features of the problem setting, and the lead time setting procedure. In the sequel, we describe the Markov process, its solution and the stability condition. Then, in Section 4 we provide the performance evaluation results in relation to the update frequency. Finally, Section 5 provides the contributions and conclusions of this study with some further research directions.

2. Problem Setting

We model a production department facing a stochastic arrival of production orders. Under static conditions, orders of a single lot-size are placed according to a Poisson process with rate λ. Due to the dynamic planned lead time, orders can be released in integer number of lot-sizes. A change in the planned lead time may yield additional lots to be released for production or some lots to be canceled. During the rest of this paper, each production lot is referred to as a single job in the queueing system. w denotes the number of jobs being processed (WIP) in the shop, and the total workload is $\hat{w} = w + b$, where b is the number of jobs waiting in the production backlog to be loaded to the shop floor. There is a WIP limit, W_0, that indicates the size and the speed of the shop floor with respect to a single job. The shop is assumed to handle at most W_0 jobs at the same time, and when the shop is totally loaded, arriving jobs are put in the backlog queue, and loaded to the shop each time a job is completed and leaves the shop. The backlog queue is modeled as a single-server system with FCFS processing discipline. The WIP in the shop is processed within exponentially distributed time intervals with rate μ jobs/time-unit. An illustration of our queueing system is given in Figure 1. The following examples may further clarify the situation.

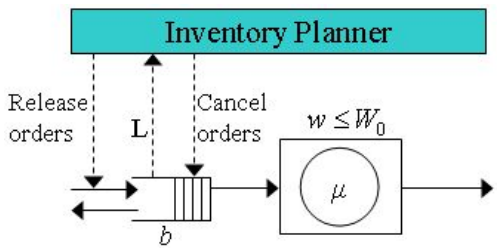


Figure 1 **A single-server queueing system with adaptive planned lead time.**

EXAMPLE 1. The shop has a single machine that operates with exponentially distributed processing times with rate μ. There is a buffer in front of the machine that can store at most $W_0 - 1$ jobs at the same time. □

EXAMPLE 2. The shop has multiple manufacturing centers. Jobs can follow different routes in the shop. The possibility that different jobs interfere (setups, processor sharing, etc.) with each other depends on the WIP level in the shop. Each job is processed fast when the WIP level is low and vice-versa. Jobs are processed in parallel with independent, identically distributed exponential processing times with rate μ/WIP. Since the minimum of the exponentials is also exponential with a rate equal to the sum of the rates, the overall processing rate is μ irrespective of the WIP level in the shop. □

Figure 1 illustrates the relationship between the production department and the inventory planner. The flow of information is represented by dashed lines, and the solid lines refer to the physical flow of jobs. The inventory planner is responsible for determining the number of jobs released or the number of the jobs suspended from the backlog queue. The production department continuously reviews the workload status and quotes a planned lead time to the inventory planner.

The planned lead time is determined based on the expected flow time of the last job currently residing in the backlog. When the backlog queue is empty, $\hat{w} \leq W_0$, the lead time is set to a fixed level, L_0, referring to the minimum estimated flow time. When there are jobs waiting in the backlog, then an estimate of the waiting time in the backlog is added to L_0. From the memoryless property of the exponential processing times it is straightforward that, at any point in time, the expected duration of time to finish b jobs through the production process is b/μ. Here, the planned lead time is determined based on a logic similar to those procedures that have been previously applied in the literature such as TWK, and JIQ (Kanet (1986), and Chang (1994)). The estimated waiting time is set by a management constant multiplied by the term, b/μ. Since μ is constant, without loss of generality, we can write the planned lead time as a function of the total workload level as follows:

$$L = L_0 + \lfloor \alpha \cdot b \rfloor,$$

where α is a management constant, and the planned lead time is an integer. In words, the planned lead time is based on the production department's perceptions on the range of the number of jobs that can be finished in a certain time frame. Implicitly, α refers to update frequency. For greater α, the lead time is updated more frequently, and for smaller α, the lead time is updated less frequently. We define the reciprocal of the update frequency, $r = 1/\alpha$, as the amount of increase or decrease

in the number of jobs in the backlog, b, in order to have one unit of increase or decrease in the planned lead time respectively.

In a realistic setting, although the planning process is a very complex task including human intervention, the underlying relationship as has been described in Section 1 still remains valid. Depending on the situation, the degree of response to the change in the planned lead time may vary, but generally, depends on the traffic intensity through a decision function $h(\lambda)$. When the planned lead time is fixed, orders of standard size are given with rate λ. When the lead time is increased, an amount of $h(\lambda)$ is added to the standard order, and when there is a decrease in the planned lead time, an excess of $h(\lambda)$ is canceled from the production backlog. From the perspective of the production department, jobs are canceled. From the systems perspective jobs do not disappear, but suspended from the production backlog until the time new jobs are released. From the base-stock policy (Base-stock $= L \cdot \lambda$), one may expect that $h(\lambda) = \lambda$. However, due to various factors in the planning process, the response can be different. For example, the planner may anticipate a trendy increase/decrease in the planned lead time and over-react by setting $h(\lambda) > \lambda$. On the other hand, the planner may react to dampen the variability, and smooth the production orders by setting $h(\lambda) < \lambda$. In this study, for modeling purposes we apply $h(\lambda) = 1$.

3. Markov Process

3.1. Description

Define $\hat{w}(t)$ and $b(t)$ as the total number of jobs in the system and the number of jobs in the backlog at time t respectively. We model this queueing system as a two-dimensional Markov process defined by $\{X_r(t), t \geq 0\}$, $X_r(t) = (\hat{w}(t), L_r(t))$, where $L_r(t) = L_0 + \lfloor \frac{b(t)}{r} \rfloor$ is the planned lead time defined by the update parameter r. We use r as a subscript because it determines the characteristics of the process. When $b(t) = r \cdot (L_r(t) - L_0) + r - 1$, an arrival triggers an increase in the planned lead time and an additional job is ordered immediately. When $b(t) = r \cdot (L_r(t) - L_0)$, a departure triggers a decrease in the planned lead time and a job is canceled immediately. As long as $b(t)$ remains in between, the system behaves as an ordinary $M^{\lambda}|M^{\mu}|1$. An illustration of the process $\{X_r(t), t \geq 0\}$ with $r = 4$ is provided in Figure 2.

The process has a Quasi-Birth-and-Death (QBD) structure in the diagonal direction of the $(\hat{w}(t), L_r(t))$ coordinates. For clearness of presentation, we transform the state space of the process $\{X_r(t), t \geq 0\}$ into the standard QBD format. The new process is defined by $\{Y_r(t), t \geq 0\}$, $Y_r(t) = (\hat{L}_r(t), \hat{b}_r(t))$, where $\hat{L}_r(t) = L_r(t) - L_0$, and $\hat{b}_r(t) = \hat{w}(t) - W_0 - r\hat{L}_r(t)$. The state space of the process $\{Y_r(t), t \geq 0\}$ can be partitioned into the set of boundary

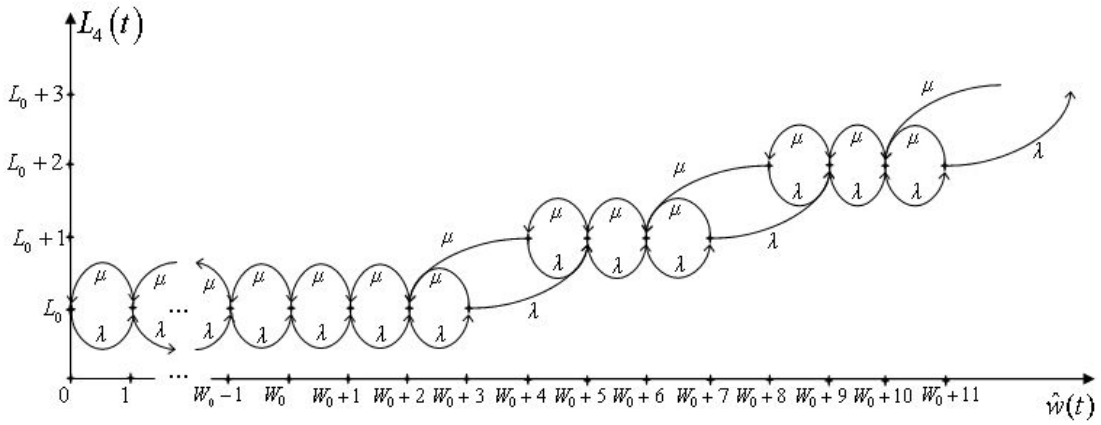


Figure 2 **Transition rate diagram for the process** $\{X_4(t), t \geq 0\}$.

states $\{(0,-W_0),(0,-W_0+1),\ldots,(0,-2),(0,-1)\}$ and levels l, where level l is the set of states $\{(l,0),(l,1),\ldots,(l,r-2),(l,r-1)\}$, $l=0,1,\ldots$; the transition rate diagram for the process $\{Y_4(t), t \geq 0\}$ is shown in Figure 3. Note that $(0,-W_0)$ corresponds to the empty state.

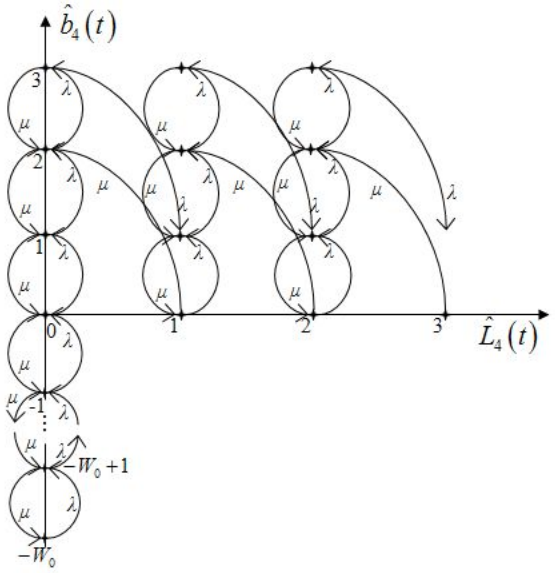


Figure 3 **Transition rate diagram for the process** $\{Y_4(t), t \geq 0\}$.

Throughout the rest of this paper, we will concentrate on the explicit solution of the QBD process $\{Y_r(t), t \geq 0\}$. As can be seen in Figure 3, for every $r=2,3,\ldots,$ the right drift occurs from state $(l,r-1)$ to $(l+1,1)$, and the left drift occurs from state $(l+1,0)$ to $(l,r-2)$, $l=0,1,\ldots$. To conveniently describe the infinitesimal generator $Q^{(r)}$ of the process, $\{Y_r(t), t \geq 0\}$, we will employ the following notation for square matrices with arbitrary dimension $k>0$. We will denote the identity matrix by $I^{(k)}$, the right and left shift matrices as $T_R^{(k)}$ and $T_L^{(k)}$ respectively. So, $\left(T_R^{(k)}\right)_{i,j} = \delta_{i+1,j}$

and $\left(T_L^{(k)}\right)_{i+1,j} = \delta_{i,j}$ for $i,j = 0,1,\ldots,k-1$, where $\delta_{i,j}$ denotes the Kronecker delta. Further, we define the k-dimensional unit column-vector on the j^{th} coordinate by $e_j^{(k)}$, $j = 0,1,\ldots,k-1$. Given that the states are in lexicographic order, the generator $Q^{(r)}$ is:

$$Q^{(r)} = \begin{bmatrix} Z_1^{(W_0)} & Z_0^{(W_0\times r)} & 0 & 0 & \cdots \\ Z_2^{(r\times W_0)} & A_1^{(r)} & A_0^{(r)} & 0 & \cdots \\ 0 & A_2^{(r)} & A_1^{(r)} & A_0^{(r)} & \cdots \\ 0 & 0 & A_2^{(r)} & A_1^{(r)} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \end{bmatrix}.$$

$Q^{(r)}$ is partitioned into matrices that provide the transition rates between and within the levels of the QBD process. $A_0^{(r)}$ is the transition rate matrix from level l to level $l+1$, and $A_1^{(r)}$ provides the transition rates within level l, $l = 0,1,\ldots$. $A_2^{(r)}$ is the transition rate matrix from level l to level $l-1$, $l = 1,2,\ldots$. Further, $Z_0^{(W_0\times r)}$, $Z_1^{(W_0)}$ and $Z_2^{(r\times W_0)}$ are the transition rate matrices from, within and to the set of boundary states. We represent these matrices as follows:

$$\begin{aligned} A_0^{(r)} &= \lambda e_{r-1}^{(r)} \left(e_1^{(r)}\right)^T, \\ A_1^{(r)} &= \lambda T_R^{(r)} + \mu T_L^{(r)} - (\lambda+\mu) I^{(r)}, \\ A_2^{(r)} &= \mu e_0^{(r)} \left(e_{r-2}^{(r)}\right)^T, \\ Z_0^{(W_0\times r)} &= \lambda e_{W_0-1}^{(W_0)} \left(e_0^{(r)}\right)^T, \\ Z_1^{(W_0)} &= \lambda T_R^{(W_0)} + \mu T_L^{(W_0)} - (\lambda+\mu) I^{(W_0)} + \mu e_0^{(W_0)} \left(e_0^{(W_0)}\right)^T, \\ Z_2^{(r\times W_0)} &= \mu e_0^{(r)} \left(e_{W_0-1}^{(W_0)}\right)^T. \end{aligned}$$

3.2. Stability Condition

Theoretically, for $W_0 \to \infty$ the planned lead time is never updated, and the stability condition is $\lambda/\mu < 1$. This is the case where the production department is assumed to have an infinite capacity relative to the size of the orders such that the delivery time is always kept at the level of minimum flow time independent of the workload level. It is a common assumption in classical inventory control theory. The same stability condition also holds for $r \to \infty$, which also yields a fixed planned lead time. Throughout the rest of this paper we define

$$\rho = \lambda/\mu.$$

It is the utilization of the production department in a static system.

For finite r and finite W_0, the stability condition of the QBD process, $\{Y_r(t), t \geq 0\}$, can be derived from *Neuts' mean drift condition* (cf. Neuts (1981)). The Markov process defined by the generator $Q^{(r)}$ is ergodic (stable) if and only if

$$\pi^{(r)} A_0^{(r)} e^{(r)} < \pi^{(r)} A_2^{(r)} e^{(r)}, \tag{1}$$

where $e^{(r)}$ is the r-dimensional column vector of ones, and $\pi^{(r)} = (\pi_0^{(r)},\ \pi_1^{(r)},\ \ldots,\ \pi_{r-1}^{(r)})$ is the steady-state probability vector of the Markov process with generator $A^{(r)} = A_0^{(r)} + A_1^{(r)} + A_2^{(r)}$. So,

$$\pi^{(r)} A^{(r)} = \mathbf{0}^{(r)}, \quad \pi^{(r)} e^{(r)} = 1,$$

where $\mathbf{0}^{(r)}$ is the r-dimensional row vector of zeros.

Condition (1) has an intuitive interpretation. The generator $A^{(r)}$ describes the behavior of the QBD process $\{Y_r(t), t \geq 0\}$ in the (vertical) $\hat{b}_r(t)$-direction. Weighted by the steady state probabilities in the vertical direction, if the mean drift to the left, $\pi^{(r)} A_2^{(r)} e^{(r)}$, is greater than the mean drift to the right, $\pi^{(r)} A_0^{(r)} e^{(r)}$, then the process is stable. Condition (1) reduces to:

$$\frac{\pi_{r-1}^{(r)}}{\pi_0^{(r)}} \cdot \rho < 1, \tag{2}$$

In order to explicitly determine stability condition (2) for every update parameter r, we need to have a detailed look at the Markov process defined by the generator $A^{(r)}$. It has r states, and its transition rate diagram is given in Figure 4. By utilizing the global balance principle, we can derive a relationship between the probabilities $\pi_{r-1}^{(r)}$ and $\pi_0^{(r)}$, which implies that Condition (2) always reduces to $\rho < 1$. This is summarized in Theorem 1.

THEOREM 1. *For every update parameter $r = 2, 3, \ldots$, the Markov process defined by $\{Y_r(t), t \geq 0\}$ is stable if and only if*

$$\rho < 1 \tag{3}$$

Proof See Appendix A.

Our findings in this section lead to the interesting conclusion that the stability condition of the system is independent of whether or not the lead time is updated. If the system is stable under the static policy that employs a fixed planned lead time, we will not see increasing uncontrollable congestion once we start to update the planned lead time in response to changing workload level. The stability is always satisfied for all update frequencies. An intuitive explanation is that the

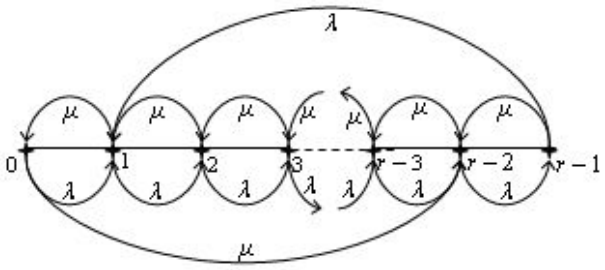


Figure 4 **Transition rate diagram for the Markov process with generator $A^{(r)}$.**

increase in the total workload due to placing additional jobs in response to an increase in the planned lead time is balanced by the decrease in the total workload due to canceling jobs in response to a decrease in the planned lead time. The frequency of updating the planned lead time varies depending on the choice of the update parameter r ($\alpha = 1/r$), but the rates of job addition and job cancelation stay balanced in the long-run, therefore the stability is retained.

3.3. Steady-State Distribution

For the derivations done throughout the rest of the paper we assume the queuing system is stable, that is, $\rho < 1$. Let $p^{(r)}(l, j)$ be the steady state probability that the QBD process $\{Y_r(t), t \geq 0\}$ is in state (l, j). From the equilibrium equations for the boundary states, we readily obtain

$$p^{(r)}(0, j) = p^{(r)}(0, 0)\rho^j, \qquad j = -W_0, -W_0 + 1, \ldots, 0. \tag{4}$$

Let $p_l^{(r)}$ denote the vector of equilibrium probabilities for level l. So,

$$p_l^{(r)} = \left(p^{(r)}(l, 0),\ p^{(r)}(l, 1),\ \ldots,\ p^{(r)}(l, r-2),\ p^{(r)}(l, r-1)\right), \qquad l = 0, 1, \ldots$$

The equilibrium equations for level l are

$$\mu p^{(r)}(0,0)\left(e_0^{(r)}\right)^T + p_0^{(r)}A_1^{(r)} + p_1^{(r)}A_2^{(r)} = \mathbf{0}^{(r)}, \tag{5}$$

$$p_{l-1}^{(r)}A_0^{(r)} + p_l^{(r)}A_1^{(r)} + p_{l+1}^{(r)}A_2^{(r)} = \mathbf{0}^{(r)}, \qquad l = 1, 2, \ldots \tag{6}$$

Note that we have eliminated the boundary probability $p^{(r)}(0, -1)$ in Equation (5) by substituting the expression in Equation (4). The normalization equation is

$$\sum_{j=-W_0}^{-1} p^{(r)}(0, j) + \sum_{l=0}^{\infty} p_l^{(r)} e^{(r)} = 1,$$

which, by substituting Equation (4), reduces to

$$\frac{p^{(r)}(0,0)(1-\rho^{W_0})}{\rho^{W_0}(1-\rho)} + \sum_{l=0}^{\infty} p_l^{(r)} e^{(r)} = 1. \tag{7}$$

Given that the Markov process with generator $Q^{(r)}$ is ergodic, the equilibrium probability vectors are determined by deploying the matrix-geometric form,

$$p_l^{(r)} = p_0^{(r)} \left(R^{(r)}\right)^l, \; l = 0, 1, \ldots, \tag{8}$$

where the r-dimensional rate matrix $R^{(r)}$ is the minimal nonnegative solution of the matrix-quadratic equation,

$$A_0^{(r)} + R^{(r)} A_1^{(r)} + \left(R^{(r)}\right)^2 A_2^{(r)} = \mathbf{0}^{(r)\times(r)}. \tag{9}$$

Matrix geometric methods, initiated by Neuts (1981), serve as a powerful framework to analyze and (approximately) solve large classes of stochastic processes of $G|M|1$ type in a unified manner. In order to solve for the steady state properties of the process, one should determine the rate matrix $R^{(r)}$ that solves Equation (9). The problem of finding an explicit rate matrix is still a developing research area. Structural results have been provided in Ramaswami and Latouche (1986) for the QBD processes with transition matrices of rank 1. Van Leeuwaarden and Winands (2005) describe a class of QBD processes for which an explicit rate matrix can be found. Based on the results of Ramaswami and Latouche (1986), we provide an explicit solution for the rate matrix $R^{(r)}$ of the QBD process $\{Y_r(t), t \geq 0\}$ for every update parameter, r.

THEOREM 2. *Given the Markov process with generator $Q^{(r)}$ is ergodic, the rate matrix $R^{(r)}$ that exactly solves the matrix quadratic equation* (9) *for every $r = 2, 3, \ldots$ is given by*

$$R^{(r)} = \begin{bmatrix} \boldsymbol{0}^{(r)} \\ \boldsymbol{0}^{(r)} \\ \vdots \\ \boldsymbol{0}^{(r)} \\ R_{r-1}^{(r)} \end{bmatrix}, \tag{10}$$

where

$$R_{r-1}^{(r)} = \left(\rho, \; \rho(1+\rho), \; \rho^2(1+\rho), \; \ldots, \; \rho^{r-2}(1+\rho), \; \rho^{r-1}\right). \tag{11}$$

Proof See Appendix B.

In order to determine the steady-state distribution of $\{Y_r(t), t \geq 0\}$ we need to derive $p_0^{(r)}$ from equilibrium equation (5), and normalization equation (7). Using the explicit expression for the rate matrix and the matrix geometric form (8), equilibrium equation (5) is rewritten as:

$$\mu p^{(r)}(0,0)\left(e_0^{(r)}\right)^T + \lambda p_0^{(r)} T_R^{(r)} + \mu p_0^{(r)} T_L^{(r)} - (\lambda+\mu) p_0^{(r)} + \lambda p^{(r)}(0,r-1)\left(e_{r-2}^{(r)}\right)^T = \mathbf{0}^{(r)}.$$

Solving this system of linear equations we get

$$p^{(r)}(0,j) = p^{(r)}(0,0)\rho^j, \qquad j = 1,2,\ldots,r-2, \tag{12}$$

$$p^{(r)}(0,r-1) = p^{(r)}(0,0)\cdot\frac{\rho^{r-1}}{1+\rho}, \tag{13}$$

We will now determine $p^{(r)}(0,0)$ from the normalization equation (7). First, let us rewrite the term $\sum_{l=0}^{\infty} p_l^{(r)} e^{(r)}$ by substituting (8), which is going to simplify the solution.

$$\sum_{l=0}^{\infty} p_l^{(r)} e^{(r)} = p_0^{(r)}\left(I^{(r)} + R^{(r)} + \left(R^{(r)}\right)^2 + \cdots\right) e^{(r)}.$$

Since the rate matrix $R^{(r)}$ has rows of zero, except the last one, the power matrices of $R^{(r)}$ can be expressed as:

$$\left(R^{(r)}\right)^l = \left(R^{(r)}\right)_{r-1,r-1}^{l-1} R^{(r)}$$

The stability of the QBD process $\{Y_r(t), t \geq 0\}$ directly implies $\left(R^{(r)}\right)_{r-1,r-1} = \rho^{r-1} < 1$. This can also be verified by the fact that, if $\{Y_r(t), t \geq 0\}$ is stable, then the largest eigenvalue of the rate matrix $R^{(r)}$ is less than 1 (Theorem 1.7.1 in Neuts (1981)). The lower diagonal structure of $R^{(r)}$ directly implies that the largest eigenvalue is $\left(R^{(r)}\right)_{r-1,r-1}$. Hence,

$$\sum_{l=0}^{\infty} p_l^{(r)} e^{(r)} = p_0^{(r)}\left(I^{(r)} + \frac{R^{(r)}}{1-\rho^{r-1}}\right) e^{(r)}, \tag{14}$$

where the term $\left(I^{(r)} + \frac{R^{(r)}}{1-\rho^{r-1}}\right) e^{(r)}$ is an r-dimensional column vector of ones, except the last row being equal to $1 + \frac{R_{r-1}^{(r)}}{1-\rho^{r-1}} \cdot e^{(r)}$. Employing the explicit expression of $R_{r-1}^{(r)}$ provided in Equation (11),

$$1 + \frac{R_{r-1}^{(r)}}{1-\rho^{r-1}} \cdot e^{(r)} = \frac{1+\rho}{1-\rho}. \tag{15}$$

This result, together with our findings in Equations (12) and (13) to represent $p_0^{(r)}$ in Equation (14), provides us with

$$\sum_{l=0}^{\infty} p_l^{(r)} e^{(r)} = \frac{p^{(r)}(0,0)}{1-\rho}.$$

Then, using this result in the normalization equation and solving the normalization equation for $p^{(r)}(0,0)$ we find

$$p^{(r)}(0,0) = \rho^{W_0}(1-\rho), \tag{16}$$

which directly implies that the probability of an empty system is given by

$$p^{(r)}(0,-W_0) = 1-\rho, \tag{17}$$

As a direct conclusion of Equation (17), the utilization of the system is found to be insensitive to the lead time update and is always equal to ρ. This is in line with our findings on the stability condition of the system. It reveals the fact that updating the planned lead time does not increase or decrease the long-run average traffic intensity in the system. The rate at which the jobs are loaded to the shop is the same as in the static case.

4. Performance Evaluation

In this section we answer our second research question by investigating the effect of the update frequency on the performance of the production department. The key performance indicators are related to the cost performance, the delivery performance, and the nervousness created in the planning system. To avoid intricacy in the notations and the analysis, we should note that the parameter r refers to the reciprocal of the update frequency α.

We have shown in the previous section that the dynamic and static utilization levels are equal. Therefore, keeping a large number of jobs as workload introduces inefficiency, and increased costs are incurred due to, e.g., material handling and inventory holding. Maintaining an utilization level of ρ implies that the average number of jobs in process are the same both in the static and dynamic situation. Thus, our attention is on the average number of jobs in the backlog. Let $B(r)$ denote the random variable for the backlog level with the probability mass function

$$\Pr\{B(r)=n\} = \begin{cases} \sum_{j=-W_0}^{0} p^{(r)}(0,j), & n=0 \\ p^{(r)}(0,j), & n=j,\ j=1,2,\ldots,r-1, \\ p^{(r)}(l,j), & n=rl+j,\ l=1,2,\ldots,\ j=0,1,\ldots,r-1 \end{cases}$$

The larger the backlog level is, the higher the costs that the production unit has to face. Using the explicit derivations provided in the previous section we formulate the relationship between the static and the dynamic case average backlog levels as in the following proposition:

PROPOSITION 1. *The average number of jobs in the production backlog of a dynamic system is always larger than it is under a static system, and their relationship is given by*

$$E[B(r)] = E[B(\infty)]\,(1+\theta(\rho,\ r))\,, \tag{18}$$

where the average backlog level under static situation is $E[B(\infty)] = \frac{\rho^{W_0+1}}{1-\rho}$, *and*

$$\theta(\rho,\ r) = 2 \cdot \frac{1-\rho}{1+\rho} \cdot \frac{\rho^{r-1}}{1-\rho^{r-1}}$$

is a monotonically increasing function of the update frequency $\alpha = 1/r$ *for* $r = 2, 3, \ldots$.

Proof See Appendix C.

Proposition 1 has an intuitively appealing interpretation. We identify that updating the planned lead time increases the long-term average backlog level by a multiplicative term, which is a function of the utilization level and the update frequency. We name this term as *lead-time-update-effect*, and denote it by $\theta(\rho,\ r)$. When the planned lead time is increased, additional jobs are released increasing the backlog level until the planned lead time is decreased. On average, this results in a higher average backlog level although the utilization of the production department does not change. It is a fundamental intuition from Hopp and Spearman (2000) that inefficient increase in congestion is undesirable due to the increase in order processing, material handling and inventory holding costs. The monotonicity of $\theta(\rho,\ r)$ implies that the fixed planned lead time policy yields the lowest cost situation. As the production department becomes more sensitive to the changes in its workload level by more frequent updating of the planned lead time, more workload is generated.

Figure 5 illustrates the level of increase in the average number of jobs in the backlog as a function of the utilization for different update frequencies. For the highest update frequency, $\alpha = 1/2$, and the utilization level close to 1, the backlog level is twice as much as it is when the planned lead time is fixed. As the update frequency decreases, the deteriorating effect significantly decreases. The lead-time-update-effect is relatively small for low utilization levels, and increases as the utilization increases.

In addition, the delivery performance of the jobs are evaluated by considering the actual duration of time that the processed jobs spend between the moment that they are released and the moment that they are completed. The random variable for the job flow time is denoted by $C(r)$, and it is the flow time of a finished job (not canceled). As the average workload level increases in the dynamic case, we would also expect to see on average longer flow times. The result we have found is even stronger as provided in the following proposition:

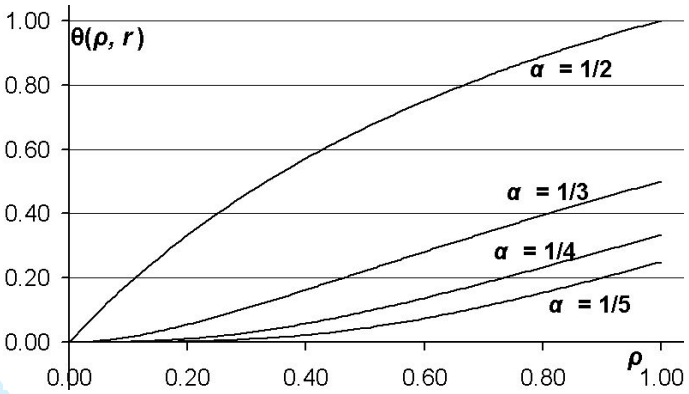


Figure 5 **The lead-time-update-effect,** $\theta(\rho,\ r)$**;** $0<\rho<1$ **and** $r=1/\alpha$**,** $\alpha=1/2,1/3,1/4,1/5$**.**

PROPOSITION 2. *Given that only the last job in the backlog queue can be canceled, the flow time of a processed job in the dynamic situation is stochastically greater than or equal to the flow time of a job in the static situation.*

$$Pr\{C(r)\geq l\}\geq Pr\{C(\infty)\geq l\}\,,$$

where $C(\infty)$ *denotes the random variable for the flow time of a job in the static situation.*

Proof See Appendix D.

This result is related to the increased variability in the release pattern of jobs created by updating the planned lead time. When the planned lead time is updated, jobs are released earlier than they would be under the static case. Some of the jobs are canceled. However, the net input rate does not differ from the static case, and on average, jobs spend more time in the system causing worse delivery performance of the production department. Further, Proposition 2 directly implies that a certain (e.g. 90^{th}) percentile of the flow time distribution of the completed jobs in the dynamic case is greater than or equal to it is in the static case. Thus, the service level in terms of the percentage of jobs completed in a certain time frame will never improve, once the production department starts to quote dynamic planned lead times.

In addition to the actual flow time, the mean and the variability of the planned lead time is also important to analyze. From the perspective of the inventory planner a very long and a highly variable lead time is undesirable. A long lead time increases the uncertainty, and an erratic lead time increases both the variability in inventory and the nervousness in the planning decisions. Let $L(r)$ denote the random variable for the planned lead time quoted by the production department, and its probability mass function is given by

$$\Pr\{L(r) = L_0 + n\} = \begin{cases} \sum_{j=-W_0}^{r-1} p^{(r)}(0,j), & n = 0 \\ \sum_{j=0}^{r-1} p^{(r)}(l,j), & n = l,\ l = 1,2,\ldots, \end{cases}$$

The first and second moment of the distribution of the planned lead time are derived based on the update frequency as follows:

PROPOSITION 3. *The first and second moment of the probability distribution of the lead time quoted by the manufacturer are given by*

$$E[L(r)] = L_0 + \frac{\rho^{W_0+1}}{1+\rho} \cdot \frac{2\rho^{r-1}}{1-\rho^{r-1}}, \tag{19}$$

$$E[L^2(r)] = L_0^2 + \frac{\rho^{W_0+1}}{1+\rho} \cdot \frac{2\rho^{r-1}}{1-\rho^{r-1}} \cdot \left(2L_0 + \frac{1+\rho^{r-1}}{1-\rho^{r-1}}\right). \tag{20}$$

Proof See Appendix E.

Let us denote the average increase over the minimum planned lead time due to updating the planned lead time as $\Delta L(r)$, which is given by

$$\Delta L(r) = \frac{\rho^{W_0+1}}{1+\rho} \cdot \frac{2\rho^{r-1}}{1-\rho^{r-1}}.$$

The relationship between the average number of jobs in the backlog and the average planned lead time directly follows from Equations (18) and (19). It gives an analytical description of the lead time syndrome as

$$E\left[B(r)\right] = \Delta L(r) \cdot \frac{1+\theta(\rho,\ r)}{\theta(\rho,\ r)}. \tag{21}$$

Equation (21) provides insight into the fundamental inventory replenishment behavior; as the estimated delivery time increases, the inventory position is increased by larger orders generating in return increased workload. As we exemplified in Section 1, the longer the planned lead time gets, the higher the production backlog is. Here, we emphasize the insight that the long-term average behavior of a physical phenomenon (number of jobs in the backlog) is expressed in terms of the long-term average behavior of a dynamic planning parameter (planned lead time) and a coefficient determined by its update policy (lead-time-update-effect).

The coefficient of variation of the planned lead time $CV\left[L(r)\right]$ follows from the Equations (19) and (20),

$$CV\left[L(r)\right] = \frac{\sqrt{\Delta L(r)\left(\frac{1+\rho^{r-1}}{1-\rho^{r-1}} - \Delta L(r)\right)}}{L_0 + \Delta L(r)}.$$

The lead time distribution depends on the design variables including, L_0, W_0 and r. For a specific situation with $L_0 = 1$ and $W_0 = 1$, $CV\,[L(r)]$ is depicted as a function of the utilization level for different update frequencies in Figure 6. Starting from zero for a utilization level close to zero, the coefficient of variation of the planned lead time increases as the utilization increases, and asymptotically approaches $\sqrt{2}$ as the utilization approaches to 1.

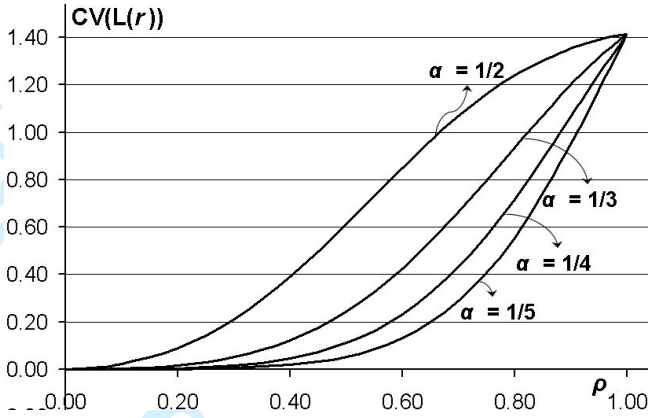


Figure 6 **Coefficient of variation of the lead time,** $CV\,[L(r)]$**;** $0 < \rho < 1$ **and** $r = 1/\alpha$**,** $\alpha = 1/2, 1/3, 1/4, 1/5$**.**

For a given utilization level, $CV\,[L(r)]$ increases as the update frequency increases. However, the relative effect of the update frequency depends on the utilization level. The update frequency is much influential for moderate values of ρ, i.e. $0.5 \leq \rho \leq 0.8$. The effect diminishes at the utilization boundaries. For a utilization level close to zero it is trivial that there will be no variation. For a utilization level close to 1, the marginal effect of the update frequency diminishes since the system possess inherently very high variability. As a result, for moderate levels of the utilization, the planned lead time should be updated less frequently in order to decrease the nervousness in planning decisions.

5. Conclusion

We have modeled a situation known as the lead time syndrome in the literature. Our study has been motivated by lack of analytical analysis for production-inventory systems utilizing dynamic, adaptive planned lead times based on the workload status. We have provided insightful results on the stability condition and the performance evaluation of systems with adaptive planned lead times. We are aware of the fact that the results depend on the specific policy by which the planned lead time is updated, and the planner's response to changes in the planned lead time. At the same time, we believe that our problem setting is realistic in the sense that it describes the phenomenon,

and the type of behavior described in this paper can be experienced in common planning tools such as MRP, linear programming or base-stock policies. In this study, for a stationary demand and a stationary production process, we have shown that:

- Production planning systems utilizing dynamic load dependent planned lead times can be modeled through a two-dimensional Markov process and solved explicitly using the matrix geometric analysis.
- The stability condition and utilization level are independent of whether or not the planned lead time is updated.
- When the planned lead time is updated in response to the changing workload level, the average number of jobs waiting in the production backlog increases by a multiplicative term, which is larger for a higher utilization level and higher update frequency.
- Although the utilization does not change, due to the increased variability in the workload, the flow time of the processed jobs increases as the lead time is updated.
- The variability of the planned lead time increases with the update frequency, with the exception that the effect diminishes at utilization boundaries.

The analysis and results provided in this study promote further improvements in modeling and analyzing dynamic, adaptive systems. The strength of our analysis lies in the modeling of the update frequency, which is an important design parameter for dynamic, adaptive planning tools. Update frequency is related to the sensitivity of the production department to changes in the workload status. There is also the response of inventory planner, modeled by $h(\lambda)$, which is related to the degree that the inventory planner is sensitive to changes in the planned lead time. In this respect, it would be an interesting extension to consider alternative strategies in setting $h(\lambda)$ and to evaluate the joint effect of $h(\lambda)$ and the update frequency. It is also promising to extend the analysis by considering multiple-stages of the supply chain.

Our results in the previous section implies that, under stationary conditions, the fixed planned lead time policy is preferable. However, it is unclear what the best policy is when the demand pattern is non-stationary. In that case, new policies may need to be developed (see Selçuk et al. (2007)). The analysis for non-stationary demand or production conditions is an interesting research direction.

Appendix A: Proof of Theorem 1

For $r=2,3$ the balance equations yield

$$\pi_{r-2}^{(r)}=\pi_0^{(r)}(1+\rho)^{r-2},\quad \pi_1^{(r)}=\pi_{r-1}^{(r)}\left(\frac{1+\rho}{\rho}\right)^{r-2},\ \text{and}\quad \pi_1^{(r)}=\pi_{r-2}^{(r)}\rho^{3-r}.$$

Hence,

$$\pi_{r-1}^{(r)}=\pi_0^{(r)}\rho,\qquad r=2,3, \tag{A.1}$$

For $r=4,5,\ldots$ we obtain, by balancing the flow out and into the set $\{0,1,\ldots,k\}$, $k=0,1,\ldots r-2$,

$$\pi_1^{(r)}=\pi_0^{(r)}(1+\rho),\quad \pi_{r-2}^{(r)}=\pi_{r-1}^{(r)}\left(\frac{1+\rho}{\rho}\right),$$

$$\pi_k^{(r)}=\pi_0^{(r)}+\pi_{k-1}^{(r)}\rho-\pi_{r-1}^{(r)}\rho,\qquad k=2,\ldots,r-2.$$

Hence,

$$\pi_{r-1}^{(r)}\left(\frac{1+\rho}{\rho}\right)=\pi_0^{(r)}\left(1+\rho+\ldots+\rho^{r-2}\right)-\pi_{r-1}^{(r)}\rho\left(1+\rho+\ldots+\rho^{r-4}\right),$$

which simplifies to

$$\pi_{r-1}^{(r)}=\pi_0^{(r)}\rho,\qquad r=4,5\ldots \tag{A.2}$$

Equation (A.1) and (A.2) together with Condition (2) show that the system is stable for every $r=2,3,\ldots$ as long as $\rho<1$. This completes the proof of Theorem 1.

Appendix B: Proof of Theorem 2

We apply the results of Ramaswami and Latouche (1986), and derive explicit solutions using matrix algebra. Given the Markov process with generator $Q^{(r)}$ is ergodic, the rate matrix $R^{(r)}$ that exactly solves the matrix quadratic equation (9) is given by

$$R^{(r)}=-A_0^{(r)}\left(A_1^{(r)}+A_0^{(r)}e^{(r)}\left(e_{r-2}^{(r)}\right)^T\right)^{-1}. \tag{B.1}$$

The structure of $A_0^{(r)}$ allows us to describe the rate matrix in more detail. First, let us denote the r-dimensional square matrix $\left(A_1^{(r)}+A_0^{(r)}e^{(r)}\left(e_{r-2}^{(r)}\right)^T\right)$ by $A_3^{(r)}$. Then, $R^{(r)}$ has rows of zero except the last one, and we can write its last row as;

$$R_{r-1}^{(r)}=-\lambda\cdot\left(\text{the second row of the matrix }\left(A_3^{(r)}\right)^{-1}\right). \tag{B.2}$$

Explicitly, the description of $A_3^{(r)}$ is as follows:

$$\begin{aligned}\left(A_3^{(r)}\right)_{i,i}&=-(\lambda+\mu),\qquad i=0,1,\ldots r-1,\\ \left(A_3^{(r)}\right)_{i,i+1}&=\lambda,\qquad i=0,1,\ldots r-2,\\ \left(A_3^{(r)}\right)_{i,i-1}&=\mu,\qquad i=1,2,\ldots r-2,\end{aligned}$$

$$\left(A_3^{(r)}\right)_{r-1,r-2} = (\lambda+\mu),$$

and all other elements of $A_3^{(r)}$ are zero. The second row of $\left(A_3^{(r)}\right)^{-1}$, denoted by $\left(a_{3(1)}^{(r)}\right)^{-1}$, solves the following matrix equation,

$$\left(A_3^{(r)}\right)^T \left(\left(a_{3(1)}^{(r)}\right)^{-1}\right)^T = e_1^{(r)}.$$

By exploiting the tri-diagonal structure of $A_3^{(r)}$ we find

$$\left(a_{3(1)}^{(r)}\right)^{-1} = \left(-1/\mu,\ -(1+\rho)/\mu,\ -\rho(1+\rho)/\mu,\ \ldots,\ -\rho^{r-3}(1+\rho)/\mu,\ -\rho^{r-2}/\mu\right).$$

Consequently, due to Equation (B.2), this derivation yields

$$R_{r-1}^{(r)} = \left(\rho,\ \rho(1+\rho),\ \rho^2(1+\rho),\ \ldots,\ \rho^{r-2}(1+\rho),\ \rho^{r-1}\right).$$

This completes the proof of Theorem 2.

Appendix C: Proof of Proposition 1

The expected number of jobs in the backlog is given by

$$E[B(r)] = r\sum_{l=1}^{\infty} l p_l^{(r)} e^{(r)} + \sum_{j=1}^{r-1} j \sum_{l=0}^{\infty} p_l^{(r)} e_j^{(r)},$$

and using the relationship we have found in Equation (14), and the derivation of $p_0^{(r)}$ in Equations (12) and (13), we further simplify this equation as:

$$E[B(r)] = \frac{r}{(1-\rho^{r-1})^2} \cdot p_0^{(r)} R^{(r)} e^{(r)} + \sum_{j=1}^{r-1} j p_0^{(r)} \left(I^{(r)} + \frac{R^{(r)}}{1-\rho^{r-1}}\right) e_j^{(r)}. \tag{C.1}$$

Let us denote the first and the second terms on the right-hand-side of Equation (C.1) by $B_r(1)$ and $B_r(2)$ respectively. We first write $B_r(1)$ in its explicit form, and then, write $B_r(2)$ similarly in order to provide an explicit expression for $E[B(r)]$.

From Equations (12) and (13),

$$p_0^{(r)} = p^{(r)}(0,0)\left(1,\ \rho,\ \rho^2,\ \ldots,\ \rho^{r-2},\ \frac{\rho^{r-1}}{1+\rho}\right). \tag{C.2}$$

Besides, Equation (15) implies that the term $\frac{R^{(r)}e^{(r)}}{1-\rho^{r-1}}$ is a column vector of zeros except the last row being equal to $\frac{2\rho}{1-\rho}$. Then, using Equation (C.2) and the explicit solution for $p^{(r)}(0,0)$, provided in Equation (16), we can rewrite $B_r(1)$ as:

$$B_r(1) = \frac{2r\rho^{W_0+1}\rho^{r-1}}{(1-\rho^{r-1})(1+\rho)}. \tag{C.3}$$

Similar derivations are performed to determine $B_r(2)$, and we rewrite it as:

$$B_r(2) = \frac{\rho^{W_0+1}}{1-\rho} - \frac{2(r-1)\rho^{W_0+1}\rho^{r-1}}{(1+\rho)(1-\rho^{r-1})}. \tag{C.4}$$

Then, the average number of jobs in the backlog given in Equation (C.1) can be written explicitly by

$$E[B(r)] = \frac{\rho^{W_0+1}}{1-\rho}\left(1+2\cdot\frac{1-\rho}{1+\rho}\cdot\frac{\rho^{r-1}}{1-\rho^{r-1}}\right). \tag{C.5}$$

Denote

$$\theta(\rho,\ r) = 2\cdot\frac{1-\rho}{1+\rho}\cdot\frac{\rho^{r-1}}{1-\rho^{r-1}}.$$

Then, for all $r = 2, 3, \ldots$

$$\theta(\rho,\ r+1) - \theta(\rho,\ r) = 2\cdot\frac{1-\rho}{1+\rho}\cdot\frac{\rho^{r}-\rho^{r-1}}{(1-\rho^{r})(1-\rho^{r-1})}.$$

From the stability condition, $\rho < 1$, it directly follows that

$$\theta(\rho,\ r+1) < \theta(\rho,\ r),$$

which implies that the lead time update effect $\theta(\rho,\ r)$ is monotonically increasing in the update frequency $\alpha = 1/r$.

Together with Equation (C.5) this completes the proof of Proposition 1.

Appendix D: Proof of Proposition 2

Let us assume that the static system is initially empty and consider a realization of arrival times and processing times of jobs; let a_j denote the j-th arrival time at the shop and let τ_j denote the j-th processing time in the shop. Now we couple the static and dynamic system by assuming that the dynamic system is also initially empty and fed by exactly the same stream of jobs; so a_j is the j-th (external) arrival time in the dynamic system. It is important to note that, in the dynamic system, additional jobs may be generated at the j-th arrival and additional jobs may leave (i.e., be canceled) when a job finished processing. Further we assume that the processing times of the jobs are exactly the same as in the static system, i.e., the j-th processing time in the dynamic system is τ_j.

Let $\hat{w}_r(t)$ denote the workload, i.e., number of jobs in the dynamic system at time t; the subscript refers to the update parameter r in the dynamic system. Similarly, $\hat{w}_\infty(t)$ denotes the workload in the static system at time t. Now we immediately have that $\hat{w}_r(t) \geq \hat{w}_\infty(t)$, or more precisely, for all $t \geq 0$,

$$\hat{w}_r(t) = \hat{w}_\infty(t) + \hat{L}_r(t), \tag{D.1}$$

where $\hat{L}_r(t)$ is the quoted lead time at time t minus L_0. In Figure 7 a sample path of $\hat{w}_\infty(t)$ and $\hat{w}_r(t)$ is shown for $r = 2$ and $W_0 = 1$. Note that, in both systems, processing of the j-th job starts at exactly the same point in time for all j.

The claim is that, for all $t \geq 0$, the j-th job in the queue of the static system arrived later (or at the same time) than the j-th job in the dynamic system. The claim is immediate from Equation (D.1): an arriving job in the static system will find at most as many jobs in the queue as the corresponding job in the dynamic system. Thus, if the job in the static system is the j-th one in the queue, then the j-th job in the dynamic

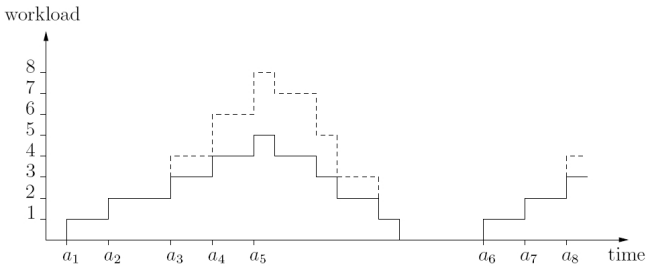


Figure 7 **Realization of the workload $\hat{w}_\infty(t)$ (solid line) and $\hat{w}_r(t)$ (dashed line) for $r=2$ and $W_0=1$.**

system was already present or arrived at the same time. Further, since the job cancelation policy in the dynamic system is to only cancel the last job in the queue, the position of both jobs in the queue will remain the same until they have been processed.

Since the flow time is the difference between the arrival and completion time, the claim above implies Proposition 2.

Appendix E: Proof of Proposition 3

Using the explicit derivation of the rate matrix in Equations (10) and (11), and $p_0^{(r)}$ given in (C.2), we rewrite the probability mass function of the lead time as follows:

$$\Pr\{L(r)=L_0+l\}=\begin{cases}1-2\rho^{r-1}\cdot\frac{\rho^{W_0+1}}{1+\rho}, & l=0\\ 2\rho^{r-1}\cdot\frac{\rho^{W_0+1}}{1+\rho}\cdot(1-\rho^{r-1})\left(\rho^{r-1}\right)^{l-1}, & l=1,2,\ldots.\end{cases}$$

The lead time quoted by the manufacturer to its customers is L_0 plus a random variable. We rewrite $L(r)$ in terms of its distribution characteristics as follows:

$$L(r)=L_0+\begin{cases}0, & \text{with probability } 1-2\rho^{r-1}\cdot\frac{\rho^{W_0+1}}{1+\rho}\\ 1+\tilde{L}_r, & \text{with probability } 2\rho^{r-1}\cdot\frac{\rho^{W_0+1}}{1+\rho},\end{cases}$$

where the random variable $\tilde{L}_r$ has a geometric distribution with success probability ρ^{r-1}.

$$\Pr\left\{\tilde{L}_r=l\right\}=\left(1-\rho^{r-1}\right)\left(\rho^{(r-1)}\right)^l, l=0,1,\ldots.$$

Consequently, the average lead time is given by

$$E\left[L(r)\right]=L_0+2\rho^{r-1}\cdot\frac{\rho^{W_0+1}}{1+\rho}\cdot\left(1+E\left[\tilde{L}_r\right]\right)=L_0+\frac{\rho^{W_0+1}}{1+\rho}\cdot\frac{2\rho^{r-1}}{1-\rho^{r-1}}. \tag{E.1}$$

Similarly, using the moments of the geometric random variable $\tilde{L}_r$, we solve for the second moment of $L(r)$.

$$\begin{aligned}E\left[L^2(r)\right]&=L_0^2\left(1-2\rho^{r-1}\cdot\frac{\rho^{W_0+1}}{1+\rho}\right)+\left(2\rho^{r-1}\cdot\frac{\rho^{W_0+1}}{1+\rho}\right)E\left[\left(L_0+1+\tilde{L}_r\right)^2\right]\\&=L_0^2+(2L_0+1)\cdot\frac{\rho^{W_0+1}}{1+\rho}\cdot\frac{2\rho^{r-1}}{1-\rho^{r-1}}+\frac{\rho^{W_0+1}}{1+\rho}\cdot\left(\frac{2\rho^{r-1}}{1+\rho^{r-1}}\right)^2,\end{aligned}$$

which simplifies to

$$E\left[L^2(r)\right] = L_0^2 + \frac{\rho^{W_0+1}}{1+\rho} \cdot \frac{2\rho^{r-1}}{1-\rho^{r-1}} \cdot \left(2L_0 + \frac{1+\rho^{r-1}}{1-\rho^{r-1}}\right). \quad \text{(E.2)}$$

Together with Equation (E.1) this completes the proof of Proposition 3.

References

Breithaupt, J.W., M. Land, P. Nyhuis. 2002. The workload control concept: Theory and practical extensions of load oriented order release. *Production Planning and Control* **13**(7) 625–638.

Buzacott, J.A., S.M. Price, J.G. Shanthikumar. 1992. Service level in multistage MRP and base stock controlled production systems. G. Fandel, T. Gulledge, A. Jones, eds., *New Directions for Operations Research in Manufacturing*. Springer-Verlag, New York, 445–463.

Chang, F.C.R. 1994. A study of factors affecting due-date predictability in a simulated dynamic job shop. *Journal of Manufacturing Systems* **13**(6) 389–406.

Conway, R.W., W.L. Maxwell, L.W. Miller. 1967. *Theory of scheduling*. Addison-Wesley, London.

De Kok, A.G., J.C. Fransoo. 2003. Planning supply chain operations: Definition and comparison of planning concepts. A.G. De Kok, S.C. Graves, eds., *Handbook in Operations Research and Management Science, Volume 11: Design and Analysis of Supply Chains*. Elsevier, Amsterdam, 597–675.

Enns, S.T. 2001. MRP performance effects due to lot size and planned lead time settings. *International Journal of Production Research* **39**(3) 461–480.

Enns, S.T., P. Suwanruji. 2004. Workload responsive adjustment of planned lead times. *Journal of Manufacturing Technology Management* **15**(1) 90–100.

Forrester, J.W. 1980. *Industrial dynamics*. 10th ed. The M.I.T Press, Massachusetts.

Homem-de-Mello, T., A. Shapiro, M. L. Spearman. 1999. Finding optimal material release times using simulation-based optimization. *Management Science* **45**(1) 86–102.

Hopp, W.J., M.L. Spearman. 2000. *Factory physics: Foundations of manufacturing management*. 2nd ed. McGraw-Hill, New York.

Hoyt, J. 1978. Dynamic lead times that fit today's dynamic planning (Q.U.O.A.T lead times). *Production and Inventory Management* **19** 63–72.

IBM. 1972. Manufacturing activity planning. *Communications Oriented Production Information and Control System, Volume 5*. IBM Corp., White Plains, New York.

Kanet, J.J. 1982. Towards understanding lead times in MRP systems. *Production and Inventory Management* **3**rd **Quarter** 1–14.

Kanet, J.J. 1986. Towards a better understanding of lead times in MRP systems. *Journal of Operations Management* **6** 305–316.

Karaesmen, F., J.A. Buzacott, Y. Dallery. 2002. Integrating advance order information in make-to-stock production. *IIE Transactions* **34**(8) 649–662.

Kingsman, B.G., I.P. Tatsiopoulos, L.C. Hendry. 1989. A structural methodology for managing manufacturing lead times in make-to-order companies. *European Journal of Operational Research* **40** 196–209.

Lambrecht, M.R., P.L. Ivens, N.J. Vandaele. 1998. ACLIPS: A capacity and lead time integrated procedure for scheduling. *Management Science* **44**(11) 1548–1561.

Lambrecht, M.R., J.A. Muckstadt, R. Luyten. 1984. Protective stocks in multi-stage production systems. *International Journal of Production Research* **22**(6) 1001–1025.

Liberopoulos, G., S. Koukoumialos. 2005. Tradeoffs between base stock levels, numbers of kanbans and production lead times in production-inventory systems with advance demand information. *International Journal of Production Economics* **96**(2) 213–232.

Mather, H., G.W. Plossl. 1978. Priority fixation versus throughput planning. *Production and Inventory Management* **3**rd **Quarter** 27–51.

Molinder, A. 1997. Joint optimization of lot-sizes, safety stocks and safety lead times in an MRP system. *International Journal of Production Research* **35**(4) 983–994.

Neuts, M.F. 1981. *Matrix-geometric solutions in stochastic models: An algorithmic approach*. The Johns Hopkins University Press, Baltimore.

Plossl, G.W. 1988. Throughput time control. *International Journal of Production Research* **26**(3) 493–499.

Ramaswami, V., G. Latouche. 1986. A general class of markov processes with explicit matrix-geometric solutions. *OR Spectrum* **8** 209–218.

Selçuk, B., J.C. Fransoo, A.G. De Kok. 2006. The effect of updating lead times on the performance of hierarchical planning systems. *International Journal of Production Economics* **104**(2) 427–440.

Selçuk, B., J.C. Fransoo, A.G. De Kok. 2007. Supply chain operations planning with load dependent planned lead times. Working paper, Technische Universiteit Eindhoven, The Netherlands.

Spitter, J.M., C.A.J. Hurkens, A.G. De Kok, E.G. Negenman, J.K. Lenstra. 2005. Linear programming models with planned lead times. *European Journal of Operational Research* **163** 706–720.

Tatsiopoulos, I.P., B.G. Kingsman. 1983. Lead time management. *European Journal of Operational Research* **14** 351–358.

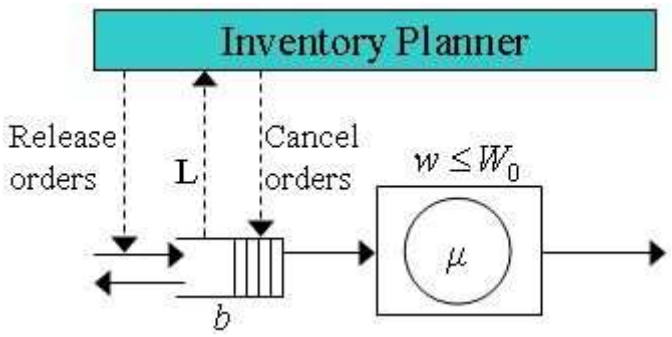
Van Leeuwaarden, J.S.H., E.M.M. Winands. 2005. Quasi-birth-and-death processes with an explicit rate matrix. To appear in Stochastic Models.

Vandaele, N.J., M.R. Lambrecht, N. De Schuyter, R. Cremmery. 2000. Spicer off-highway products division-brugge improves its lead-time and scheduling performance. *Interfaces* **30**(1) 83–95.

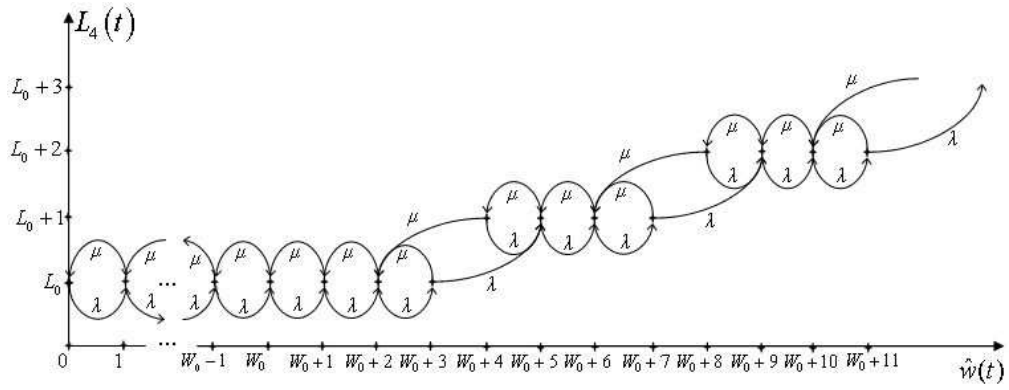
Yano, C.A. 1987. Setting planned lead times in serial production systems with tardiness costs. *Management Science* **33**(1) 95–106.

Zäpfel, G., H. Missbauer. 1993. New concepts for production planning and control. *European Journal of Operational Research* **67** 297–320.

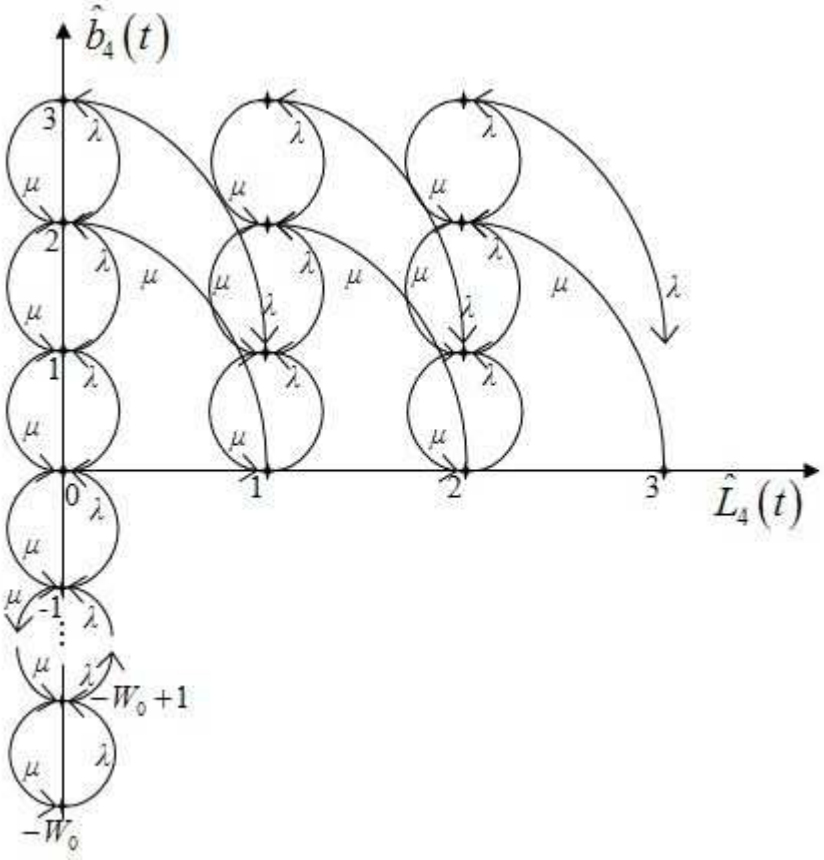
Zijm, W.H.M., R. Buitenhek. 1996. Capacity planning and lead time management. *International Journal of Production Economics* **46/47** 165–179.



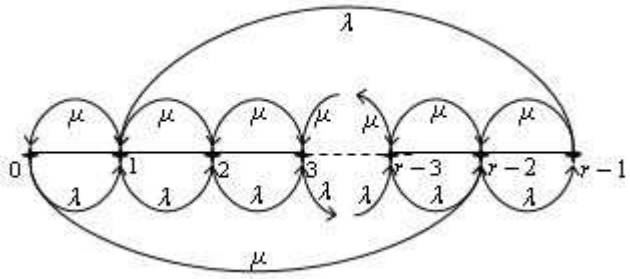
A single-server queueing system with adaptive planned lead time.
88x43mm (96 x 96 DPI)



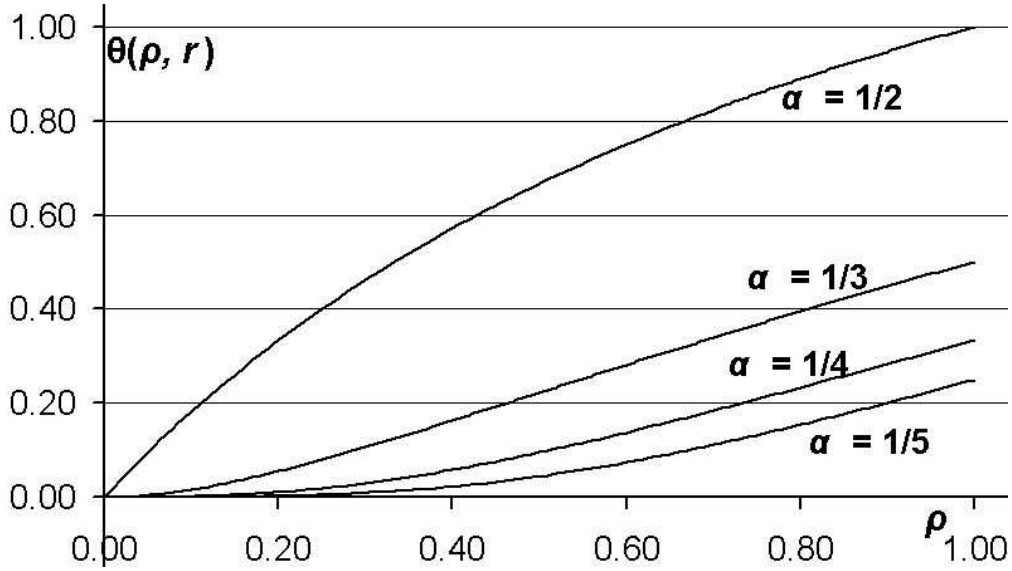
Transition rate diagram for the process $X_4(t), t \geq 0$.
200x77mm (96 x 96 DPI)



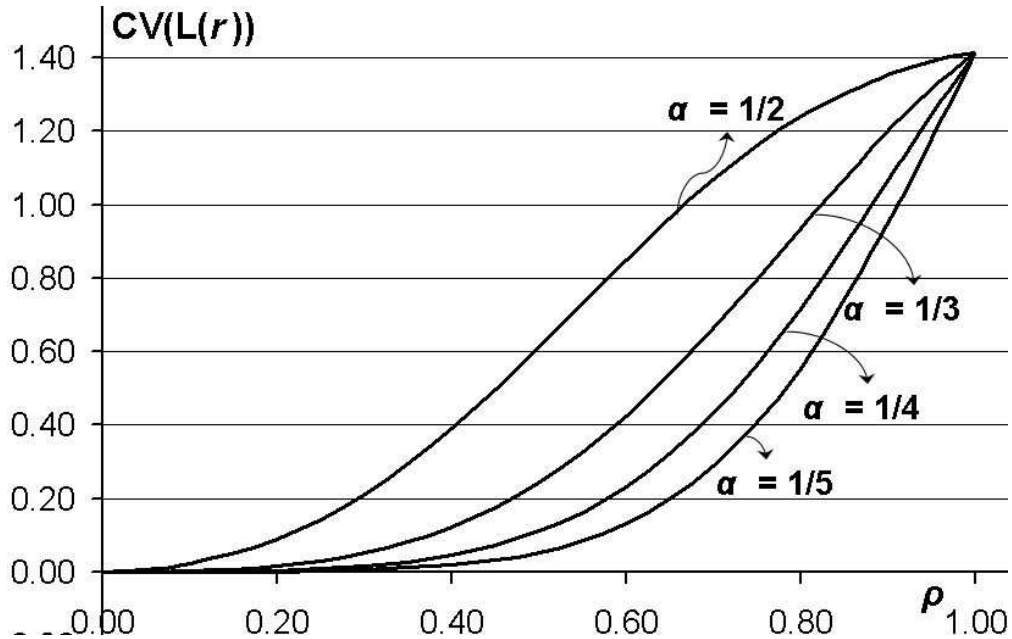
Transition rate diagram for the process $Y_4(t), t \geq 0$.
150x152mm (72 x 72 DPI)



Transition rate diagram for the Markov process with generator $A^{(r)}$.
83x37mm (96 x 96 DPI)



The lead-time-update-effect, θ(ρ,r); 0<ρ<1 and r=1/α, α=1/2,1/3,1/4,1/5.
210x118mm (96 x 96 DPI)



Coefficient of variation of the lead time, CV[L(r)]; 0<ρ<1 and r=1/α, α=1/2,1/3,1/4,1/5.
216x136mm (96 x 96 DPI)