EURANDOM PREPRINT SERIES
2011-009

The Effect of Workload Constraints in Linear Programming Models for Production Planning

M.M. Jansen, A.G. de Kok, I.J.B.F. Adan
ISSN 1389-2355

The Effect of Workload Constraints in Linear Programming Models for Production Planning

M.M. Jansen[†]
A.G. de Kok[‡]
I.J.B.F. Adan[§]

25-10-2011

Abstract: Linear programming (LP) models for production planning incorporate a model of the manufacturing system that is necessarily deterministic. Although these deterministic models are the current state-of-the-art, it should be recognized that they are used in an environment that is inherently stochastic. This fact should be kept in mind, both when making modeling choices and when setting the parameters of the model. In this paper we study the effect of a workload constraint on the efficiency of resource usage and on the reliability of production planning in a stochastic environment represented by a queueing model. The main novelty of the queueing model is the fact that jobs are admitted to the production facility periodically but are processed continuously. We show that there may not be an acceptable trade-off between efficiency and reliability if planned lead times are not explicitly modeled. The impact of production uncertainty on the design and parametrization of the LP model is demonstrated by numerical examples.

Keywords: Queueing Model, Production Planning, Rolling Schedule, Linear Programming

[†]Eindhoven University of Technology, Department of Industrial Engineering, Paviljoen E.14, Postbus 513, 5600 MB Eindhoven / EURANDOM
[‡]Eindhoven University of Technology, Department of Industrial Engineering, Paviljoen E.14, Postbus 513, 5600 MB Eindhoven
[§]EURANDOM, Postbus 513, 5600 MB Eindhoven

1. Introduction

With the emergence of Advanced Planning Systems [15], linear programming (LP) models for production planning [6, 28, 12, 26, 24, 27, 20] are becoming more commonplace in supporting firm's decision making processes. These models deal with the coordination of goods flows and the allocation of resource capacity in an integrated fashion. In order for them to remain computationally tractable, the dynamics of the manufacturing system are generally modeled as being entirely deterministic. In particular this implies that the planned throughput of a resource never exceeds the capacity of the resource. In reality though, throughput can be both higher and lower than the anticipated capacity. This fact is largely ignored in the literature on LP models for production planning and the choice of the capacity (or the production rate) parameter is left unaddressed [3]. Because there is uncertainty in the input data of LP models for production planning, they are typically applied in a rolling schedule context (see [5, 22]). This implies that the system state (WIP, stocks) is observed, and planning decisions are taken, at fixed equidistant points in time. These mark the starts of planning periods during which the execution of the production process takes place. Another implication of the rolling schedule context of production planning is the fact that decisions are implemented only for the first period in the planning horizon. Consequently, the capacity parameter forms an upper bound on the amount of work in the facility. Due to the periodic nature of production planning, it is not possible to observe or react to deviations of the state of the manufacturing system from the planned state until the start of the next planning period. Particularly if the throughput was higher than expected, this may lead to unplanned idleness of the resource. As such, loss in the efficiency of resource usage is inevitable due to the periodic nature of production planning.

It is intuitively clear that a poor choice of the capacity parameter or workload constraint leads to poor planning performance. However, the effect of the choice of workload constraints in LP models for production planning that are applied in a context where the production process is subject to uncertainty has not been extensively studied. In this paper we fill in this gap. We study the trade-off between the reliability of a production plan and the efficiency of resource usage. Here, reliability refers to the event that a production plan is executed without tardiness or corrective (planning) actions. We also show that it is often necessary to model capacity over more than one planning period in order to get a good trade-off between reliability and efficiency.

Planned lead times arise naturally as a modeling tool if capacity is considered over multiple planning periods. They turn out to be an important instrument to get a good trade-off between reliability and efficiency of the resource usage. Planned lead times are absent in most LP models proposed in the literature. In these models the decision variables are planned production quantities. It is argued that the lead time is a function of workload and capacity and need not be explicitly modeled. Some of the literature considers a minimum planned lead time that is not related to the capacity (for example [7]). In other literature [16, 19, 2], order releases are decoupled from the expected production quantities resulting in implicitly planned lead times, but these models do not account for the availability and consumption of materials in the production network. Explicitly planned lead times are, for example, found in [17, 11, 23, 25]. Rather than restricting production quantities to the period capacity, in models with explicit planned lead times the workload is constrained to the cumulative capacity in the planned lead time. Note that production planning models without planned lead times are special cases with a planned lead time of a single period, in which case the production quantity and order release quantity coincide.

Planned lead times are essential for the coordination of goods flows in the production network. If the planned lead time is to be reliable, limitation of the workload in the manufacturing system is inevitable. A workload constraint in the production planning model leads to smoothing of

the schedule of order releases. The effect of production smoothing is build-ahead inventory (cf. [**?**,]section 15.5]Hopp01). A tighter workload constraint results in a higher probability that the production plan is reliable but increases the build-ahead inventory. It is already shown in [21, chapter 5] that a higher throughput can be achieved on a resource (i.e. a resource can be utilized more efficiently) if a longer planned lead time is selected. In this paper we explore this effect further by explicitly relating the workload constraint and the planned lead time to the build-ahead inventory and the reliability of the production plan.

2. The model of the manufacturing system

In order to analyze the effect of a workload constraint on the reliability of the production plan and the efficiency of resource usage exactly, we consider the simplest stochastic manufacturing system possible: a facility with a single resource that processes jobs with identical processing characteristics. We represent it by the queueing model that is shown in Figure 1.

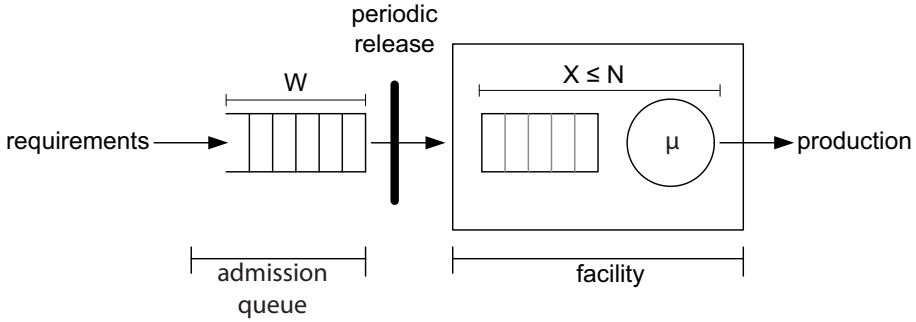


Figure 1. Model of the Manufacturing System

In this queueing model, jobs arrive to a queue outside the facility that we refer to as the *admission queue*. Jobs may arrive to this queue continuously (or at the period boundaries) but can enter the facility only at the start of each period. We refer to these instances as the release epochs. At any time, at most N jobs are allowed in the facility. Jobs that cannot be admitted into the facility at a release epoch due to the workload constraint wait in the admission queue until the next release epoch. The number of jobs that can be processed in a period is random. Clearly, only those jobs that are inside the facility at the start of a period can be processed. At the end of a period, jobs in the facility that have not been processed remain there in order to be processed in the next period. The number of jobs that can be admitted into the facility is reduced by the number of unprocessed jobs remaining in the facility.

Now we discuss how the queueing model relates to the LP controlled manufacturing system. The discrete time release epochs reflect that the order release decisions are taken only at the start of each production period and the restriction on the number of jobs in the facility represents the workload constraint. Job arrivals represent the period demands faced by the manufacturing system. We assume that future demands are known to the order release function (safety stocks may be kept to deal with the uncertain part of the demand) such that it is possible to produce build-ahead inventory. Note that it is not relevant for the queueing model that the requirements are known a priori, since the decisions (workload constraint and planned lead time) do not depend on the demands for individual periods but only on the long run distribution of the requirements. As such, the arrival process can be modeled as a random process in the queueing model. Note furthermore

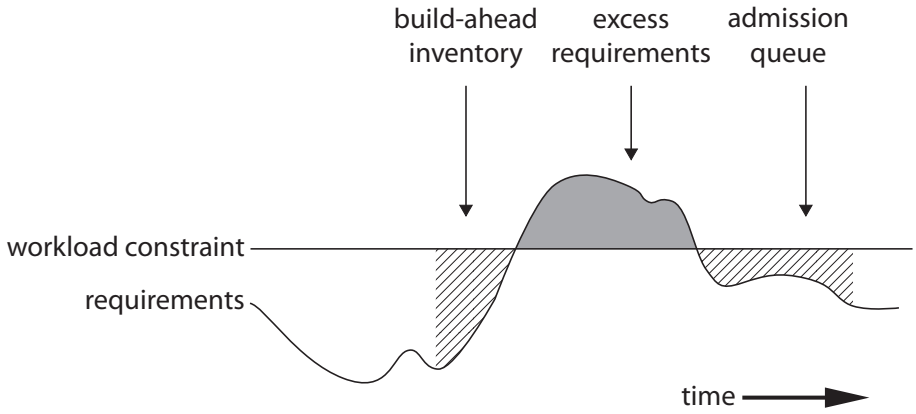


FIGURE 2. Illustration build-ahead inventory

that in the queueing model the releases are *delayed* due to the workload constraint whereas in the LP model, order releases typically *precede* the requirements in time due to the workload constraint (hence the term build-ahead inventory). The correspondence of the long run number of delayed releases and the long run build-ahead inventory is eminent from Figure 2 where the shaded and grey areas in this figure are of the same size.

We make the following assumptions about the manufacturing system:

(1) Both the characteristics of the production process and the workload constraint are constant over time.
(2) Periodic arrival quantities are independent and identically distributed.
(3) The number of jobs processed in the facility depends only on the WIP in the facility at the start of the period.
(4) Processing times are independent and identically distributed.

We note that assumptions three and four together are only consistent with exponential processing times. In case of non-exponential processing times, the state of the facility at the start of a period is not fully specified by the WIP, but also requires the specification of the residual processing time of the job in process. Hence, the state space will be two dimensional, with one component being discrete and the other continuous. Clearly, in this case the analysis would be much more complicated. Therefore, in the sequel we will assume exponential processing times, although this assumption is only explicitly used in the calculation of the sojourn time distribution in the facility.

We consider two measures for the efficiency of resource usage. The first measure is the maximum utilization level under which the manufacturing system remains stable. The second measure is the expected length of the admission queue. The reliability of the planned lead time is defined to be the probability that the sojourn time in the facility of a job does not exceed the planned lead time.

We now describe the model in mathematical terms. The total number of jobs waiting in the admission queue, the number residing in the facility and the total number in the manufacturing system (admission queue plus facility) are denoted by W_n, X_n and L_n respectively. The manufacturing system is observed at the release epochs that are indexed by $n = 1, 2, \ldots$. The variables W_n and X_n denote the state just after admissions at epoch n and they are related to the total number of jobs in the manufacturing system in the following way:

$$X_n = \min\{L_n, N\} \tag{1}$$

$$W_n = \left(L_n - N\right)^+ \tag{2}$$

We denote the number of arrivals in a period n by the random variable A_n and the long run average number of arrivals per period by λ. The long run average number of jobs that can be processed in the facility per period, provided there is ample WIP at the start of the period, is denoted by μ. $V_{x,n} = \min\{V_{\infty,n}, x\}$ denotes the throughput in period n conditioned on a workload level x at the start of the period, where $\{V_{\infty,n}\}_{n\in\mathbb{N}^+}$ are i.i.d. random variables with mean μ. Hence, the throughput in period n is $V_{X_n,n}$. The dynamics of the system are described by the following Lindley-type equation:

$$L_{n+1} = L_n - V_{X_n,n} + A_n, \tag{3}$$

We denote by L, V_X, A, X, and W the random variables that have the stationary distribution of L_n, $V_{X_n,n}$, A_n, X_n, and W_n respectively. Furthermore, we use the additional notation $\rho := \lambda/\mu$, $(x)^+ = \max\{0, x\}$, $(x)^- = \max\{0, -x\}$, and $|x| = (x)^+ + (x)^-$.

The remainder of this paper is organized as follows. First we briefly discuss literature on queueing models that are related to ours. We then present the stability condition for the system. Next we describe in detail the Markov chain corresponding to (3), derive the probability generating function (PGF) for the state distribution at the release epochs and the Laplace-Stieltjes transform (LST) of the sojourn times. We also give the CDF of the sojourn time distribution for the case where the processing times are exponentially distributed. Finally we use these results to compare various settings of the workload constraint and the planned lead time.

3. Literature

A model that is closely related to ours is the bulk service queue. In the bulk service queue, jobs enter service in batches of a maximum size. There typically is a fixed time to service completion after which all jobs in service depart together and a new batch can enter service. A seminal paper in the area of bulk service queues is Bailey [4]. Other important references include [18, 9]. Van Leeuwaarden [29] presents an extensive treatment of the discrete bulk service queue that is described by the Lindley equation $X(t+1) = (X(t) - N)^+ + A(t)$. Our model can be seen to correspond to a bulk service queue where the service time is equal to a planning period and the maximum batch size corresponds to the workload constraint. The main difference in our model, however, is the fact that only part of the batch may be processed and that the jobs in the batch that are not processed, enter the queue again.

One particular trait of our model is that the server may be idle even though there are jobs lining up in the queue, waiting to be processed. This leads to a loss of processing potential and possibly instability of the system. This trait is shared with fixed-cycle traffic light (FCTL) queues [10, 14, 30]. However, in the FCTL queue, the loss of processing potential is due to the fact that there are several queues that get allocated a fixed amount of service time that is independent of the queue length whereas in our model, processing potential is lost due to the fact that a finite number of jobs are selected from the queue that are eligible to receive service in a period.

The analysis in this paper is similar to that found both in the bulk service queue literature and in the FCTL queue literature. The key step is the characterization of the number of jobs at the end of a service time or a cycle. To this end, a probability generating function (PGF) is formulated that includes N unknowns where N is the maximum number of jobs that can pass in a green period. Solving for these unknowns involves complex root-finding for the denominator of the PGF.

A paper that requires separate mentioning is that of Wang [31]. Wang analyzes a queue where jobs can enter service only at fixed time intervals. There are c identical exponential servers and jobs arrive according to a Poisson process. Using techniques similar to [4], Wang characterizes the

steady state queue-length distribution by a PGF, and obtains closed-form expressions for the PGF of the steady state queue-length in the special cases $c = 1$ and $c = \infty$.

4. Stability Condition

The manufacturing system in Figure 1 is stable if the number of jobs in the admission queue does not grow to infinity in the long run. In most queueing models, the stability condition simply is the requirement that the long run number of arrivals does not exceed the maximum service rate. This condition does not depend on the control policy for the queue. Stability is achieved by the fact that the server is working continuously if there are many jobs in the queue. In the system of our study this is not the case. The stability of the manufacturing system depends on the parameter N. Due to the periodicity of order releases and the workload constraint, the server may idle even though there are many jobs in the admission queue. (Note that this event occurs if $V_{\infty,n} > N$ and $L_n > N$.) This phenomenon reduces the effective capacity of the system and therefore the stability condition changes. The stability condition of our system is given in the following proposition.

Proposition 1. *A necessary and sufficient condition for stability of the system is $\rho < \rho_{\max}$, where*

$$\rho_{\max} = 1 - \frac{\mu - N + \mathbb{E}[|V_\infty - N|]}{2\mu} \leq 1 \tag{4}$$

The inequality in (4) is strict if $N < \mu$.

Here, $|V_\infty - N|$ denotes the absolute difference between the random variable V_∞ and N, and V_∞ should be interpreted as the processing potential of the facility (i.e. the number of jobs processed if there is ample work in the facility).

Proof. It is clear to see that the stability condition for the system is $\mathbb{E}[A] < \mathbb{E}[V_N]$ or

$$\rho < \frac{\mathbb{E}[V_N]}{\mu}$$

The numerator at the right-hand side can be rewritten as

$$\mathbb{E}[V_N] = \mathbb{E}[\min\{V_\infty, N\}] = \mathbb{E}\left[V_\infty - (V_\infty - N)^+\right] = \mu - \mathbb{E}\left[(V_\infty - N)^+\right]$$

Using furthermore that

$$2(V_\infty - N)^+ = (V_\infty - N) + (V_\infty - N)^- + (V_\infty - N)^+ = (V_\infty - N) + |V_\infty - N|,$$

we have

$$\frac{\mathbb{E}[V_N]}{\mu} = 1 - \frac{\mathbb{E}[(V_\infty - N)^+]}{\mu} = 1 - \frac{\mu - N + \mathbb{E}[|V_\infty - N|]}{2\mu}.$$

It follows that a stability condition for the system is

$$\rho < \rho_{\max} = 1 - \frac{\mu - N + \mathbb{E}[|V_\infty - N|]}{2\mu}$$

By Jensen's inequality we have

$$\mu - N + \mathbb{E}[|V_\infty - N|] \geq \mu - N + |\mathbb{E}[V_\infty - N]| = 2(\mu - N)^+ \geq 0$$

which shows that $\rho_{\max} \leq 1$ and $\rho_{\max} < 1$ for $\mu > N$. □

A common probabilistic interpretation of parameters and variables in deterministic models of manufacturing systems is that they correspond to the first moment (mean) of the actual variables and parameters involved (cf. [**?**,]section 8.2.2]Hopp01). For LP models with $\tau = 1$ this implies that $N = \mu$. In this case, the formula for maximum utilization (4) simplifies to a function of the mean and MAD of V_∞. The MAD (Mean Absolute Deviation) is a measure of variability for a random variable, say X, and is defined as

$$\mathbb{E}\left[|X - \mathbb{E}\left[X\right]|\right]$$

Corollary 1. *The maximum utilization rate for a resource with a workload limit $N = \mu$ is*

$$\rho_{\max} = 1 - \frac{MAD[V_\infty]}{2\mu}$$

Proof. The proof follows readily from Proposition 1. □

The MAD is a measure of variability that is often used by practitioners. Figure 3 shows the maximum utilization rate for Poisson distributed V_∞ with mean $\mu = 5$, $\mu = 10$, and $\mu = 20$. The workload constraint is plotted on the horizontal axes as a multiple of μ and the maximum utilization level for that constraint is plotted on the vertical axes. Figure 3 shows that the maximum utilization is reduced substantially for $N = \mu$.

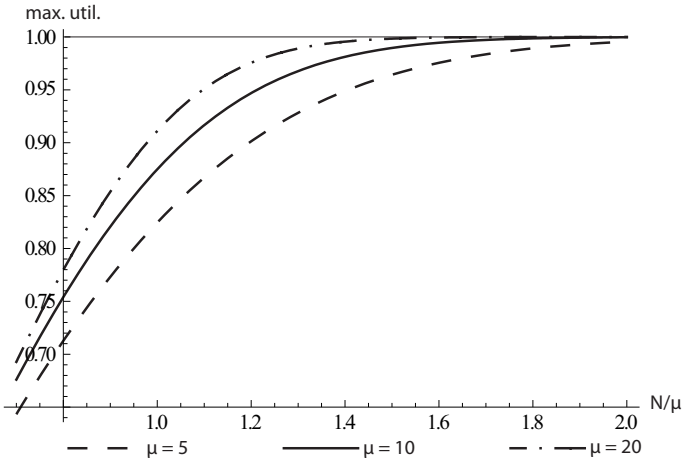


FIGURE 3. Maximum Utilization Level

It is well known that the queue-length (build-ahead inventory) explodes as the utilization rate reaches its maximum. In the next section we discuss how the queue-length for our model can be calculated if the utilization rate is less than its maximum ($\rho < \rho_{\max}$).

5. The Stationary Distributions

5.1. **A Discrete Time Markov Chain Representation.** We consider the model in Figure 1 at the release epochs. Since A_n are independent, and $V_{X_n,n}$ depends only on the state of the system at the n^{th} release epoch, the process $\{L_n\}_{n\in\mathbb{N}_+}$ forms a discrete time Markov Chain (DTMC) with transition matrix

$$\mathbf{P} = \begin{pmatrix} \beta_{00} & \beta_{01} & \beta_{02} & \beta_{03} & \cdots \\ \beta_{10} & \beta_{11} & \beta_{12} & \beta_{13} & \cdots \\ \beta_{20} & \beta_{21} & \beta_{22} & \beta_{23} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \beta_{N-1,0} & \beta_{N-1,1} & \beta_{N-1,2} & \beta_{N-1,3} & \cdots \\ \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 & \cdots \\ 0 & \alpha_0 & \alpha_1 & \alpha_2 & \cdots \\ 0 & 0 & \alpha_0 & \alpha_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \tag{5}$$

where

$$\beta_{ij} = \mathbb{P}\{A - V_i = j - i\}, \quad \text{for all } 0 \le i < N, j \ge 0, \tag{6}$$

$$\alpha_j = \mathbb{P}\{A - V_N = j - N\}, \quad \text{for all } j \ge 0, \tag{7}$$

The elements of the matrix P can be calculated as follows:

$$\beta_{ij} := \begin{cases} \sum_{k=0}^{j} \mathbb{P}\{A = k\}\, \mathbb{P}\{V_i = i - j + k\}, & \text{if } 0 \le j < i \\ \sum_{k=0}^{i} \mathbb{P}\{A = j - i + k\}\, \mathbb{P}\{V_i = k\}, & \text{if } j \ge i \end{cases}$$

and

$$\alpha_j := \begin{cases} \sum_{k=0}^{j} \mathbb{P}(A = k)\mathbb{P}(V_N = N - j + k) & \text{if } 0 \le j < N \\ \sum_{k=0}^{N} \mathbb{P}(A = j - N + k)\mathbb{P}(V_N = k) & \text{if } j \ge N \end{cases}$$

If the stability condition ($\rho < \rho_{\max}$) is satisfied, the DTMC is ergodic. We define the stationary probabilities $p_i := \lim_{n\to\infty} \mathbb{P}\{L_n = i\}$. We characterize the stationary distribution of the DTMC by its PGF. First consider the PGF's for the arrival and service processes:

$$G_A(z) := \mathbb{E}\left[z^A\right] = \sum_{k=0}^{\infty} \mathbb{P}\{A = k\}\, z^k \tag{8}$$

$$G_{V_i}(z) := \mathbb{E}\left[z^{V_i}\right] = \sum_{k=0}^{i} \mathbb{P}\{V_i = k\}\, z^k \tag{9}$$

$$= \sum_{k=0}^{i-1} \mathbb{P}\{V_\infty = k\}\, z^k + \mathbb{P}\{V_\infty \ge i\}\, z^i \tag{10}$$

The PGF for the limiting distribution of the DTMC is:

$$G_L(z) := \mathbb{E}\left[z^L\right] = \sum_{i=0}^{\infty} p_i\, z^i$$

From (3) we have:

$$\begin{aligned} G_L(z) &= \mathbb{E}\left[z^{L+A-V_{\min\{L,N\}}}\right] \\ &= \sum_{i=0}^{N-1} p_i\, z^i\, G_A(z) G_{V_i}(z^{-1}) + \sum_{i=N}^{\infty} p_i\, z^i\, G_A(z) G_{V_N}(z^{-1}) \\ &= \sum_{i=0}^{N-1} p_i\, z^i\, G_A(z)\big(G_{V_i}(z^{-1}) - G_{V_N}(z^{-1})\big) + G_L(z) G_A(z) G_{V_N}(z^{-1}) \end{aligned}$$

which reduces to

$$G_L(z) = \frac{G_A(z)\,\sum_{i=0}^{N-1} p_i\, z^{N+i}\left(G_{V_i}(z^{-1}) - G_{V_N}(z^{-1})\right)}{z^N - z^N G_A(z) G_{V_N}(z^{-1})} \tag{11}$$

The numerator of G_L has N unknowns that can be found by considering the roots $z_1, z_2, \ldots$ of the denominator within the unit circle in the complex plane. It can be shown using Rouché's theorem that there are exactly $N-1$ roots inside the unit circle [1] and these roots can routinely be found using computer packages such as Wolfram Mathematica and MATLAB. Since the PGF is bounded inside the unit circle, the numerator must also vanish for these roots. Substituting the roots in the numerator and adding the normalization equation $G_L(1) = 1$ gives a system of N linear equations in N unknowns. The following normalization equation is obtained by applying L'Hôpital's rule:

$$\sum_{i=0}^{N-1} p_i\Big(\mathbb{E}\left[V_N\right] - \mathbb{E}\left[V_i\right]\Big) = \mathbb{E}\left[V_N\right] - \mathbb{E}\left[A\right] \tag{12}$$

Note that equation (12) is precisely the equation that balances mean inflow and mean outflow:

$$\mathbb{E}\left[A\right] = \sum_{i=0}^{N-1} p_i \mathbb{E}\left[V_i\right] + \mathbb{P}\left\{L \geq N\right\} \mathbb{E}\left[V_N\right]$$

The system of equations that needs solving becomes:

$$\left(\mathbf{Z}\,\mathbf{V} - \mathbf{z^N}\,\bar{\mathbf{v}}^T\right)\mathbf{p} = \begin{pmatrix} \mathbf{0} \\ \mathbb{E}\left[V_N\right] - \mathbb{E}\left[A\right] \end{pmatrix}, \tag{13}$$

where

$$\mathbf{z^N} = \begin{pmatrix} z_1^N \\ z_2^N \\ \vdots \\ z_{N-1}^N \\ 0 \end{pmatrix}, \quad \mathbf{Z} = \begin{pmatrix} z_1^{N-1} & z_1^{N-2} & \ldots & z_1 & 1 \\ z_2^{N-1} & z_2^{N-2} & \ldots & z_2 & 1 \\ \vdots & \vdots & & \vdots & \vdots \\ z_{N-1}^{N-1} & z_{N-1}^{N-2} & \ldots & z_{N-1} & 1 \\ 1 & 2 & \ldots & N-1 & N \end{pmatrix},$$

$$\bar{\mathbf{v}} = \begin{pmatrix} \bar{v}_1 \\ \bar{v}_2 \\ \vdots \\ \bar{v}_{N-1} \\ \bar{v}_N \end{pmatrix}, \quad \mathbf{V} = \begin{pmatrix} v_1 & v_2 & \ldots & v_{N-1} & \bar{v}_N \\ v_2 & v_3 & \ldots & \bar{v}_N & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ v_{N-1} & \bar{v}_N & \ldots & 0 & 0 \\ \bar{v}_N & 0 & \ldots & 0 & 0 \end{pmatrix},$$

with $v_i = \mathbb{P}\{V_\infty = i\}$, and $\bar{v}_i = \mathbb{P}\{V_\infty \geq i\}$. Equation (13) is obtained from the numerator of (11) as follows. First we divide the numerator by $-G_A(z)$ and rewrite the result as

$$\begin{aligned}
& \sum_{i=0}^{N-1} p_i\, z^{N+i} \left(G_{V_N}(z^{-1}) - G_{V_i}(z^{-1})\right) \\
= & \sum_{i=0}^{N-1} z^{N+i} \left(\sum_{j=0}^{N-1} v_j z^{-j} - \sum_{j=0}^{i-1} v_j z^{-j} + \bar{v}_N z^{-N} - \bar{v}_i z^{-i} \right) p_i \\
= & \sum_{i=0}^{N-1} \left(\sum_{j=i}^{N-1} v_j z^{N+i-j} + \bar{v}_N z^i - \bar{v}_i z^N \right) p_i \\
= & \sum_{i=0}^{N-1} \left(\sum_{j=1}^{N-i-1} v_{j+i} z^{N-j} + \bar{v}_N z^i - \bar{v}_{i+1} z^N \right) p_i
\end{aligned}$$

Subsequently substituting $z = z_1, z_2, \ldots, z_{N-1}$ and equating to zero yields the first $N-1$ rows in (13). The last row is simply equation (12). With the solution $\mathbf{p} = (p_0, p_1, \ldots, p_{N-1})^T$, the PGF $G_L(z)$ is fully defined. For more information on finding the unknowns in PGF's, see for example chapter 2 of [29].
In B we present equations to recursively obtain the entire probability distribution and expressions for the first two moments of L, W, and X. In A we also describe an alternative, numerically stable method to obtain the stationary distribution of L by analyzing an embedded Markov chain.

5.2. **Sojourn times.** As we study the effect of the workload constraint on the reliability of the lead time, we are interested in the time that a job spends in the facility. We obtain the characteristics of the sojourn time in the facility by conditioning on the number of jobs that precede an arbitrary job entering the facility in processing. Recall that X_n is the number of jobs in the facility just after a release epoch. Let Y_n be the number of jobs left in the facility just before that release epoch:

$$Y_{n+1} = X_n - V_{X_n,n}$$

and let $Q_n = X_n - Y_n$. Furthermore, let Q, Y be the random variables that have the stationary distribution of Q_n and Y_n. $G_X(z)$, $G_Y(z)$, and $G_Q(z)$ are the PGF's of X, Y, and Q, defined in the usual way. Let q_i denote the probability that a released job finds i jobs in front of it, or in other words, q_i is the long-run fraction of jobs that find on the moment of release i jobs in front of them. Such a job will be released at the start of a planning period if and only if $Y_n \leq i$ and $X_n > i$. Hence, the long run expected number of such jobs released at the start of a planning period is $1 \cdot (P(Y_n \leq i, X_n > i) = P(Y \leq i) - P(X \leq i)$. Since the expected number of jobs released per period is $E(X) - E(Y) = \lambda$, we get for the long-run fraction (cf. Lemma 2.2.1 in [8]),

$$q_i = \frac{P(Y \leq i) - P(X \leq i)}{E(X) - E(Y)} = \frac{1}{\lambda}(P(Y \leq i) - P(X \leq i)). \tag{14}$$

Hence its PGF is

$$\begin{aligned}
G_Q(z) &= \sum_{i=0}^{\infty} q_i z^i = \frac{1}{\lambda} \sum_{i=0}^{\infty} \sum_{k=0}^{i} (P(Y=k) - P(X=k)) z^i \\
&= \frac{1}{\lambda} \sum_{k=0}^{\infty} (P(Y=k) - P(X=k)) \frac{z^k}{1-z} = \frac{G_Y(z) - G_X(z)}{\lambda(1-z)}\,, \qquad (15)
\end{aligned}$$

where we used that, in equilibrium, $\lambda = E(X) - E(Y)$. Let $B^*(s)$ denote the LST of a service time and let $T^*(s)$ denote the LST of the sojourn time. We have

$$\begin{aligned} T^*(s) &= E(e^{-sT}) = \sum_{i=0}^{\infty} p_i B^*(s)^{i+1} = B^*(s)\, G_Q\big(B^*(s)\big) \\ &= \frac{B^*(s)}{\lambda\big(1 - B^*(s)\big)} \left[G_Y\big(B^*(s)\big) - G_X\big(B^*(s)\big)\right] \end{aligned} \tag{16}$$

We can easily calculate the moments of T from $T^*(s)$. The first two moments are given in B. The CDF of T can be obtained by numerically inverting the $T^*(s)$. For exponentially distributed service times the

Proposition 2. *Let the service time be exponentially distributed with mean μ^{-1}. Then the CDF of the sojourn time is $F_T(t)$:*

$$\begin{aligned} F_T(t) = \frac{\mu}{\lambda}\Big[\Big(\mathbb{P}\{Y=0\} - \mathbb{P}\{X=0\}\Big)t \\ + \sum_{k=1}^{N} \Big(\mathbb{P}\{Y=k\} - \mathbb{P}\{X=k\}\Big)\Big(t\Gamma_{k,\mu}(t) - \frac{k}{\mu}\Gamma_{k+1,\mu}(t)\Big)\Big], \end{aligned} \tag{17}$$

where $\Gamma_{k,\mu}(t)$ is the CDF of the Gamma distribution with mean $k\,\mu^{-1}$ and variance $k\,\mu^{-2}$.

Proof. The PGF of the exponential distribution is

$$B^*(s) := \frac{s}{s+\mu}.$$

Then equation (16) becomes

$$T^*(s) = \frac{\mu}{\lambda} \sum_{k=0}^{N} \big(\mathbb{P}\{Y=k\} - \mathbb{P}\{X=k\}\big) \frac{1}{s}\Big(\frac{\mu}{s+\mu}\Big)^k$$

It can easily be verified that the Laplace transform of the function $\Gamma_{k,\mu}(t)$ is

$$\frac{1}{s}\Big(\frac{\mu}{s+\mu}\Big)^k.$$

Applying this to T^* gives the PDF of T:

$$f_T(t) = \frac{\mu}{\lambda}\Big[\mathbb{P}\{Y=0\} - \mathbb{P}\{X=0\} + \sum_{k=1}^{N} \big(\mathbb{P}\{Y=k\} - \mathbb{P}\{X=k\}\big)\Gamma_{k,\mu}(t)\Big] \tag{18}$$

Finally, using that

$$\int_{s=0}^{t} \Gamma_{k,\mu}(s)ds = t\Gamma_{k,\mu}(t) - \frac{k}{\mu}\Gamma_{k+1,\mu}(t)$$

yields the CDF of T. □

6. Numerical examples

In this section we present numerical results for the performance of the workload constrained manufacturing system with exponential processing times. First we compare the queue-lengths for various settings of the workload constraint. Next, we show the relation between the workload constraint and the lead-time reliability. For our numerical examples, we consider three cases corresponding

to a system with a production rate of $\mu = 20$, $\mu = 10$, and $\mu = 5$. These cases have $c_V^2 = 0.05$, $c_V^2 = 0.1$, and $c_V^2 = 0.2$ respectively, where $c_V^2 = \frac{1}{\mu}$ is the squared coefficient of variation of V_∞. Table 1 shows the expectation and variance of the number of jobs in the manufacturing system after a release epoch. The data is vertically organized according to the workload constraint expressed as a multiple of μ such that the figures may be easily compared. The data is horizontally organized according to the utilization level ρ. Table 2 shows information about the corresponding sojourn times T. Besides the mean and variance, also the probability that the sojourn time is less than a planned lead time of respectively 1, 2, and 3 periods is given.

TABLE 1. Effect of the Workload Constraint on Queue Lengths

			$\rho = 0.78$				$\rho = 0.82$				$\rho = 0.86$			
μ	N/μ	$\rho_{\max}$	$\mathbb{E}[W]$	$\sigma^2(W)$	$\mathbb{E}[X]$	$\sigma^2(X)$	$\mathbb{E}[W]$	$\sigma^2(W)$	$\mathbb{E}[X]$	$\sigma^2(X)$	$\mathbb{E}[W]$	$\sigma^2(W)$	$\mathbb{E}[X]$	$\sigma^2(X)$
	1	**0.911**	1.37	11.06	16.45	11.88	2.99	31.50	17.49	9.84	7.75	126.18	18.57	6.53
	1.2	**0.976**	0.38	2.93	16.80	18.66	0.85	8.14	18.03	19.04	1.96	23.83	19.38	18.17
	1.4	**0.996**	0.13	1.06	16.99	22.57	0.36	3.50	18.39	25.53	0.95	11.79	20.01	27.91
20	**1.6**	**0.999**	0.05	0.39	17.07	24.66	0.16	1.60	18.57	29.88	0.51	6.54	20.41	36.00
	1.8	**1.000**	0.02	0.15	17.10	25.68	0.07	0.73	18.66	32.54	0.28	3.63	20.64	42.14
	2	**1.000**	0.01	0.05	17.11	26.15	0.03	0.33	18.70	34.05	0.15	2.01	20.76	46.47
	3	**1.000**	0.00	0.00	17.12	26.51	0.00	0.01	18.73	35.72	0.01	0.10	20.91	53.68
	1	**0.875**	3.32	29.81	8.67	3.82	7.57	104.19	9.22	2.51	36.06	1544.96	9.79	0.77
	1.2	**0.947**	1.21	9.14	9.01	7.54	2.27	21.28	9.67	6.82	4.60	57.70	10.37	5.51
	1.4	**0.981**	0.64	4.82	9.30	10.66	1.24	11.38	10.08	10.68	2.49	28.78	10.93	10.02
10	**1.6**	**0.995**	0.37	2.88	9.50	13.27	0.79	7.39	10.40	14.27	1.69	19.83	11.42	14.58
	1.8	**0.999**	0.22	1.77	9.64	15.38	0.52	5.01	10.64	17.52	1.22	14.76	11.83	19.15
	2	**1.000**	0.14	1.08	9.73	17.00	0.35	3.42	10.81	20.32	0.90	11.13	12.14	23.56
	3	**1.000**	0.01	0.09	9.85	20.34	0.05	0.48	11.11	27.86	0.20	2.60	12.83	39.38
	1	**0.825**	10.06	147.32	4.68	0.75	119.99	14985.54	4.97	0.08	∞	∞	n/a	n/a
	1.2	**0.901**	2.93	23.55	4.94	2.32	5.50	60.66	5.28	1.75	∞	∞	n/a	n/a
	1.4	**0.949**	1.68	12.26	5.19	3.80	2.89	26.07	5.59	3.35	5.39	65.28	6.01	2.66
5	**1.6**	**0.976**	1.16	8.36	5.42	5.25	1.99	17.38	5.88	4.97	3.60	39.81	6.38	4.37
	1.8	**0.989**	0.86	6.28	5.62	6.68	1.52	13.38	6.15	6.64	2.79	30.63	6.74	6.21
	2	**0.996**	0.65	4.89	5.78	8.07	1.21	10.86	6.39	8.35	2.30	25.66	7.07	8.17
	3	**1.000**	0.19	1.48	6.24	13.59	0.44	4.27	7.13	16.37	1.06	12.91	8.24	19.02

TABLE 2. Effect of the Workload Constraint on Sojourn Times

			$\rho = 0.78$					$\rho = 0.82$					$\rho = 0.86$				
μ	N/μ	$\rho_{\max}$	$\mathbb{E}[T]$	$\sigma^2 T$	$F_T(1)$	$F_T(2)$	$F_T(3)$	$\mathbb{E}[T]$	$\sigma^2 T$	$F_T(1)$	$F_T(2)$	$F_T(3)$	$\mathbb{E}[T]$	$\sigma^2 T$	$F_T(1)$	$F_T(2)$	$F_T(3)$
20	**1**	**0.911**	0.47	0.09	0.95	1.00	1.00	0.50	0.10	0.93	1.00	1.00	0.53	0.10	0.92	1.00	1.00
	1.2	**0.976**	0.50	0.10	0.92	1.00	1.00	0.53	0.11	0.90	1.00	1.00	0.58	0.12	0.87	1.00	1.00
	1.4	**0.996**	0.51	0.11	0.91	1.00	1.00	0.56	0.13	0.88	1.00	1.00	0.61	0.14	0.84	1.00	1.00
	1.6	**0.999**	0.51	0.12	0.91	1.00	1.00	0.57	0.14	0.87	1.00	1.00	0.64	0.16	0.82	1.00	1.00
	1.8	**1.000**	0.51	0.12	0.90	1.00	1.00	0.57	0.14	0.86	1.00	1.00	0.65	0.18	0.80	1.00	1.00
	2	**1.000**	0.52	0.12	0.90	1.00	1.00	0.57	0.15	0.86	1.00	1.00	0.66	0.19	0.80	0.99	1.00
	3	**1.000**	0.52	0.12	0.90	1.00	1.00	0.58	0.15	0.86	1.00	1.00	0.67	0.21	0.80	0.99	1.00
10	**1**	**0.875**	0.54	0.13	0.89	1.00	1.00	0.56	0.13	0.88	1.00	1.00	0.59	0.13	0.86	1.00	1.00
	1.2	**0.947**	0.59	0.15	0.85	1.00	1.00	0.62	0.16	0.82	1.00	1.00	0.66	0.17	0.80	1.00	1.00
	1.4	**0.981**	0.62	0.18	0.81	1.00	1.00	0.67	0.19	0.78	0.99	1.00	0.73	0.21	0.74	0.99	1.00
	1.6	**0.995**	0.65	0.21	0.79	0.99	1.00	0.71	0.23	0.74	0.99	1.00	0.79	0.25	0.69	0.98	1.00
	1.8	**0.999**	0.67	0.23	0.78	0.99	1.00	0.74	0.26	0.72	0.98	1.00	0.83	0.29	0.65	0.97	1.00
	2	**1.000**	0.68	0.25	0.77	0.98	1.00	0.76	0.29	0.71	0.97	1.00	0.87	0.33	0.63	0.96	1.00
	3	**1.000**	0.69	0.28	0.77	0.97	1.00	0.80	0.37	0.70	0.95	0.99	0.95	0.50	0.61	0.91	0.99
5	**1**	**0.825**	0.63	0.20	0.81	0.99	1.00	0.65	0.20	0.80	0.99	1.00	n/a	n/a	n/a	n/a	n/a
	1.2	**0.901**	0.71	0.25	0.75	0.98	1.00	0.74	0.25	0.73	0.98	1.00	n/a	n/a	n/a	n/a	n/a
	1.4	**0.949**	0.78	0.30	0.70	0.97	1.00	0.82	0.31	0.67	0.97	1.00	0.86	0.31	0.64	0.96	1.00
	1.6	**0.976**	0.84	0.35	0.66	0.96	1.00	0.89	0.36	0.62	0.95	1.00	0.95	0.37	0.58	0.94	1.00
	1.8	**0.989**	0.89	0.41	0.63	0.94	1.00	0.96	0.43	0.58	0.93	0.99	1.03	0.44	0.53	0.91	0.99
	2	**0.996**	0.93	0.46	0.60	0.92	0.99	1.02	0.49	0.55	0.90	0.99	1.11	0.52	0.49	0.88	0.99
	3	**1.000**	1.05	0.70	0.57	0.86	0.97	1.20	0.83	0.50	0.81	0.95	1.38	0.96	0.42	0.74	0.93

As expected, the admission queue-length increases with both the utilization and the reciprocal of the workload limit. We can see that even for a moderately variable V_∞, $\mathbb{E}[W]$ grows rapidly as N is reduced to μ. The effect of the workload constraint on $\mathbb{E}[X]$ is relatively small but $\sigma^2 X$ is reduced substantially with a more restrictive workload constraint. The squared coefficient of variation of V_∞ appears to be an important factor for the sensitivity of the sojourn times. For $\mu = 20$ ($c_V^2 = 0.05$) the effect of the workload constraint on the sojourn time is relatively small. On the other hand, for $\mu = 5$ ($c_V^2 = 0.2$) reliable production is hardly possible unless a planned lead time of more than one period is selected.

We now turn to the trade-off between the efficiency of the resource usage and the reliability of the planned lead time. Figure 4 shows the maximum utilization rate of the manufacturing system for a given lead time reliability $\phi \in \{0.9, 0.98\}$ and planned lead time $\tau \in \{1, 2, 3\}$. The bold line shows the maximum utilization rate. Lead time reliability is defined as $\mathbb{P}\{T \leq \tau\} \geq \phi$. The workload constraint N is set out on the horizontal axis. The maximum utilization under which the system is stable and the planned lead time is reliable is set out on the vertical axis.

Figure 4 reveals two important insights. Firstly, we observe that the utilization of the manufacturing system is highly restricted if $\tau = 1$. Secondly, for $\tau > 1$, we see that the best choice for the workload constraint is not trivial. In fact, particularly for smaller τ the curve is rather sharp near the maximum. Furthermore, the seemingly obvious choice of setting $N = \tau\,\mu$ is clearly not the best in most cases. If the setting $N = \tau\,\mu$ is desirable (e.g. because it corresponds more closely to the available capacity over the planning horizon), Figure 4 can be used to find which settings of τ are feasible under the reliability constraint ϕ. For example, a system with $\mu = 10$ that is running at $\rho = 0.8$ must have $\tau \geq 2$ for $\phi = 0.9$.

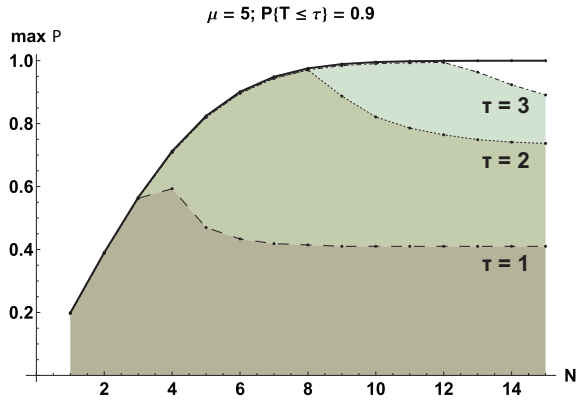
7. Summary and conclusions

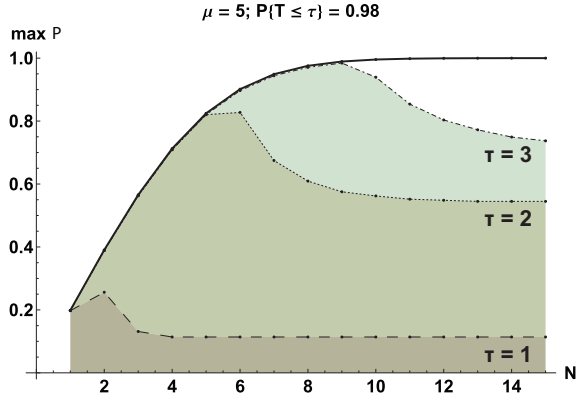
Linear programming (LP) models for production planning found in today's Advanced Planning Systems typically treat production capacity as a simple deterministic upper bound on the period throughput. Following the principles of rolling schedule planning, the amount of work that is released to the facility is limited by the choice of the capacity parameter. The obvious choice of the capacity parameter seems to be the average production rate. In this paper we show that this approach may substantially reduce the efficiency of the manufacturing system if the throughput is subject to uncertainty. We show that there is a simple relation between the maximum utilization rate of a manufacturing system, and the variability of the output. For the special case where the workload is restricted to the average production rate, we see that the MAD measure of variation naturally arises in this relation.

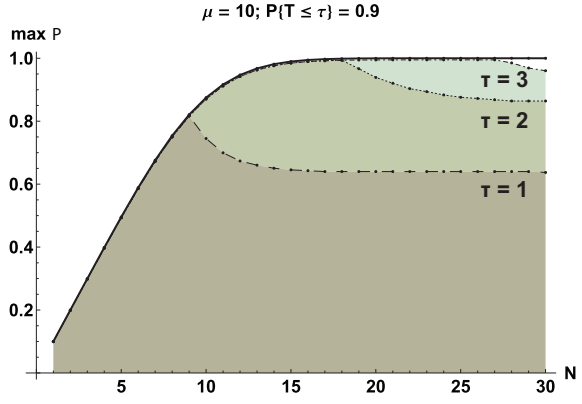
We also present expressions for the stationary queue-length and sojourn time distributions of the manufacturing system. In order to evaluate these expressions, we need the first N probability masses of the stationary queue-length distribution of the total number of jobs L in the manufacturing system. We propose numerical procedures to obtain these probabilities.

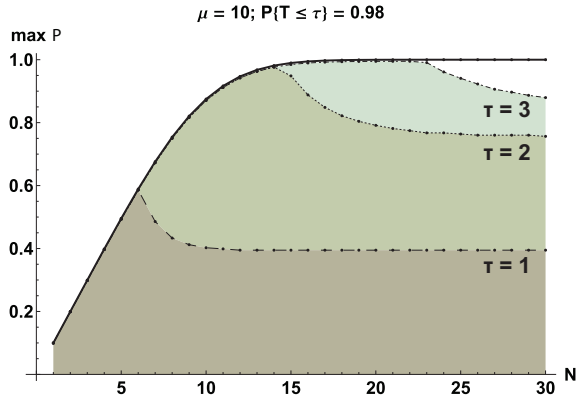
We use the results to study the trade-off between the efficiency of resource usage and the reliability of the planned lead time. The special case where lead times are not explicitly modeled is equivalent to setting $\tau = 1$. The efficiency is given by the maximum utilization rate of the system, and the amount of build-ahead inventory which is represented by the admission queue length.

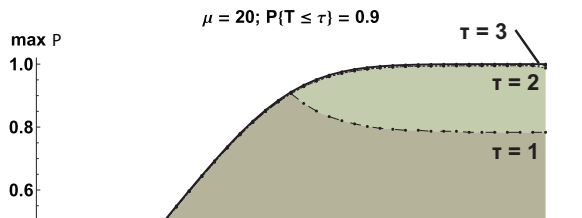
The numerical study shows that the seemingly obvious choice of setting the capacity parameter equal to the average production rate ($N = \mu$) leads to both poor reliability and poor efficiency. Whereas relaxing the workload constraint leads to deterioration of the reliability, restricting it further leads to a increase of the build-ahead inventory. A better trade-off between reliability and efficiency is obtained for higher values of the planned lead time parameter.


μ = 5; P{T ≤ τ} = 0.9
max P
1.0
0.8
0.6
0.4
0.2
τ = 3
τ = 2
τ = 1
2
4
6
8
10
12
14
N


μ = 5; P{T ≤ τ} = 0.98
max P
1.0
0.8
0.6
0.4
0.2
τ = 3
τ = 2
τ = 1
2
4
6
8
10
12
14
N


μ = 10; P{T ≤ τ} = 0.9
max P
1.0
0.8
0.6
0.4
0.2
τ = 3
τ = 2
τ = 1
5
10
15
20
25
30
N


μ = 10; P{T ≤ τ} = 0.98
max P
1.0
0.8
0.6
0.4
0.2
τ = 3
τ = 2
τ = 1
5
10
15
20
25
30
N


μ = 20; P{T ≤ τ} = 0.9
max P
1.0
0.8
0.6
τ = 3
τ = 2
τ = 1

References

[1] I.J.B.F. Adan, J.S.H. van Leeuwaarden, and E.M.M. Winands. On the application of Rouché's theorem in queueing theory. *Operations Research Letters*, 34(3):355–360, 2006.

[2] J. Asmundsson, R.L. Rardin, C.H. Turkseven, and R. Uzsoy. Production planning with resources subject to congestion. *Naval Research Logistics*, 56(2):142–157, 2009.

[3] H. Aytug, M.A. Lawley, K. McKay, S.Mohan, and R. Uzsoy. Executing production schedules in the face of uncertainties: A review and some future directions. *European Journal of Operational Research*, 161:86–110, 2005.

[4] N.T.J. Bailey. On queueing processes with bulk service. *Journal of the Royal Statistical Society. Series B (Methodological)*, 16(1):80–87, 1954.

[5] K.R. Baker. An experimental study of the effectiveness of rolling schedules in production planning. *Decision Sciences*, 8(1):19–27, 1977.

[6] K.R. Baker. Requirements planning. *Handbooks in Operations Research and Management Science*, 4:571–627, 1993.

[7] P.J. Billington, J.O. McClain, and L.J. Thomas. Mathematical programming approaches to capacity-constrained MRP systems: Review, formulation and problem reduction. *Management Science*, 29(10):1126–1141, 1983.

[8] S.C. Borst. *Polling systems*. PhD thesis, CWI, Amsterdam, 1996.

[9] M.L. Chaudry and J.G.C. Templeton. *A first course in bulk queues*. Wiley, New York, 1983.

[10] J.N. Darroch. On the traffic-light queue. *The Annals of Mathematical Statistics*, 35(1):380–388, 1964.

[11] A.G. de Kok and J.C. Fransoo. Planning supply chain operations: definition and comparison of planning concepts. *Handbooks in operations research and management science*, 11:597–676, 2003.

[12] A. Drexl and A. Kimms. Lot sizing and scheduling–Survey and extensions. *European Journal of Operational Research*, 99(2):221–235, 1997.

[13] W.J. Hopp and M.L. Spearman. *Factory Physics*. Irwin McGraw-Hill, 2001.

[14] D.R. McNeil. A solution to the fixed-cycle traffic light problem for compound Poisson arrivals. *Journal of Applied Probability*, 5(3):624–635, 1968.

[15] H. Meyr, M. Wagner, and J. Rohde. Structure of advanced planning systems. *Supply chain management and advanced planning*, pages 109–115, 2005.

[16] H. Missbauer. Aggregate order release planning for time-varying demand. *International Journal of Production Research*, 40(3):699–718, 2002.

[17] E.G. Negenman. *Material coordination under capacity constraints*. PhD thesis, Eindhoven University of Technology, 2000.

[18] M.F. Neuts. A general class of bulk queues with Poisson input. *The Annals of Mathematical Statistics*, 38(3):759–770, 1967.

[19] J. Pahl, S. Voß, and D.L. Woodruff. Production planning with load dependent lead times: an update of research. *Annals of Operations Research*, 153(1):297–345, 2007.

[20] Y. Pochet and L.A. Wolsey. *Production Planning by Mixed Integer Programming*. Springer, 2006.

[21] B. Selçuk. *Dynamic Performance of Hierarchical Planning Systems: Modeling and Evaluation with Dynamic Planned Lead Times*. PhD thesis, Eindhoven University of Technology, 2007.

[22] N.C. Simpson. Multiple level production planning in rolling horizon assembly environments. *European Journal of Operational Research*, 114(1):15–28, 1999.

[23] J.M. Spitter. *Rolling Schedule Approaches for Supply Chain Operations Planning*. PhD thesis, Eindhoven University of Technology, 2005.

[24] JM Spitter, AG De Kok, and NP Dellaert. Timing production in LP models in a rolling schedule. *International Journal of Production Economics*, 93:319–329, 2005.

[25] J.M. Spitter, C.A.J. Hurkens, A.G. De Kok, J.K. Lenstra, and E.G. Negenman. Linear programming models with planned lead times for supply chain operations planning. *European Journal of Operational Research*, 163:706–720, 2004.

[26] H. Stadtler. Multilevel lot sizing with setup times and multiple constrained resources: Internally rolling schedules with lot-sizing windows. *Operations Research*, 51(3):487–502, 2003.

[27] H. Tempelmeier. *Material-Logistik: Modelle und Algorithmen für die Produktionsplanung und-steuerung in Advanced Planning-Systemen*. Springer, 6th ed. edition, 2006.

[28] L.J. Thomas and J.O. McClain. An overview of production planning. *Handbooks in Operations Research and Management Science*, 4:333–370, 1993.

[29] J.S.H. van Leeuwaarden. *Queueing models for cable access networks*. PhD thesis, Ph. D. thesis, Eindhoven University of Technology, The Netherlands, 2005.

[30] J.S.H. van Leeuwaarden. Delay analysis for the fixed-cycle traffic-light queue. *Transportation Science*, 40(2):189–199, 2006.

[31] P. Patrick Wang. Markovian queueing models with periodic-review. *Computers & Operations Research*, 23(8):741 – 754, 1996.

Appendix A. An Embedded Markov Chain Approach for Obtaining the Unknowns of G_L

The transition matrix for the DTMC of L may be reduced to a finite Markov chain by embedding L on the states $0, \ldots, N-1$. That is, the Markov chain of L is only observed at the times n when $L_n < N$. There are two types of transitions in the embedded Markov chain. Firstly, there are the direct transitions between states $0, \ldots, N-1$. Secondly, there are the indirect transitions via states outside the embedded Markov chain. For these indirect transitions, we need to know the return probabilities. The return probability from state $i > N$ to state $j < N$ is the probability to return for the first time to the embedded set in state j, given that the Markov chain starts outside the embedded set in state i. The return probabilities can be determined as follows.
Suppose the DTMC is in state $n+m$, where $n \geq N, m \geq 0$. We define the return probability $b_{m,i}^{(k)}, i > 0, k \geq 0$ to be the probability that the first transition to some state smaller or equal to n will be to the state $n-i$ and will take at most k jumps. These probabilities can be calculated recursively:

$$\begin{aligned} b_{m,i}^{(0)} &= 0 \\ b_{m,i}^{(k)} &= \alpha_{N-(m+i)} + \sum_{j=0}^{\infty} \alpha_{N+(j-m)}\, b_{j,i}^{(k-1)}, \quad k > 0 \end{aligned}$$

Note that a jump to the right in k steps is of maximum size N such that we can restrict the above summation and obtain:

$$b_{m,i}^{(k)} = \alpha_{N-(m+i)} + \sum_{j=0}^{(k-1)N-i} \alpha_{N+(j-m)}\, b_{j,i}^{(k-1)}, \quad k > 0 \tag{19}$$

This sequence is increasing and bounded so the limit exists. We denote this limit by $b_{m,i}$:

$$b_{m,i} = \lim_{k \to \infty} b_{m,i}^{k} \tag{20}$$

We now define the Markov chain embedded on $\{0, 1, \ldots, N-1\}$ with transition matrix $Q = (q_{ij})$,

$$q_{ij} = \beta_{ij} + \sum_{m=0}^{\infty} \beta_{i,N+m}\, b_{m,N-i}, \quad , 0 \leq i, j < N \tag{21}$$

Let $\tilde{p}$ be a solution of the balance equations of the embedded Markov chain (i.e. $\tilde{p}\, Q = \tilde{p}$). Then we use (12) for normalization to find the original probabilities $p_i, i = 0, \ldots, N-1$:

$$p_i = \frac{1}{c}\, \tilde{p}_i, \tag{22}$$

where

$$c = \frac{\sum_{i=0}^{N-1} \tilde{p}_i \Big(\mathbb{E}\left[V_N\right] - \mathbb{E}\left[V_i\right] \Big)}{\mathbb{E}\left[V_N\right] - \mathbb{E}\left[A\right]}$$

To obtain the numerical results presented in this paper, the infinite sums in equations (19)-(21) need to be truncated. We use the following criterion to truncate the infinite sum in (21). For a given i, let $\hat{\beta}_i := \max\{j : \beta_{i,j} > \epsilon\}$, where ϵ is small. The summation in (21) is truncated at $m = \hat{\beta}_i - N$ and the limit in (20) is approximated by $b_{m,i}^{\hat{k}}$ where

$$\hat{k} = \inf\{k : \max_{m \leq \hat{\beta}_i - N} \{b_{i,m}^{(k)} - b_{i,m}^{(k-1)}\} < \epsilon\}.$$

Appendix B. More Details of the Stationary Probability Distributions

B.1. **The probability masses for states $i \geq N$.** The probability masses of the stationary distribution of L for the states $i \geq N$ can be found through the following balance equation for state i:

$$p_i = \frac{1}{\alpha_0}\Big(p_{i-N} - \sum_{j=N}^{i-1} \alpha_{i-j}\, p_j - \sum_{j=0}^{N-1} \beta_{j,i-N}\, p_j\Big), \quad 0 < n < N, \tag{23}$$

Equation (23) involves subtractions which may lead to numerical instabilities (i.e. with negative probabilities). Alternatively we may obtain the probabilities by extending the embedded Markov chain to include higher states. Given the return probabilities $b_{m,1}$ we can calculate the stationary probability p_i for $i = N, N+1, \ldots$ by considering the Markov chain embedded on states $\{0, 1, \ldots, i\}, i \geq N$. The balance equation for state i becomes:

$$\begin{aligned} p_i = \frac{1}{(1 - \alpha_N - \sum_{k=0}^{\infty} \alpha_{N+k+1}\, b_{k,1})} \times \\ \Big[\sum_{j=0}^{N-1} p_j \Big(\beta_{ji} + \sum_{k=0}^{\infty} \beta_{j,i+k+1}\, b_{k,1}\Big) \\ + \sum_{j=N}^{i-1} p_j \Big(\alpha_{N+(i-j)} + \sum_{k=0}^{\infty} \alpha_{N+(i-j)+k+1}\, b_{k,1}\Big)\Big] \end{aligned} \tag{24}$$

Note that this balance equation needs no further normalization since $p_0, p_1, \ldots, p_{N-1}$ are already properly normalized.

B.2. **Moments of the distribution of the number of jobs.** The moments of the distribution of the number of jobs in the system can be found by standard differentiation of the PGF in (11) and taking the limit $z \to 1$. Applying L'Hôpital's rule, the first two moments become:

$$\mathbb{E}[L] = \frac{\Delta_1 (\Theta_2 - \Theta_1) - \Theta_1 (\Delta_2 - \Delta_1)}{2\Delta_1^2} \tag{25}$$

$$\begin{aligned} \mathbb{E}\left[L^2\right] = \frac{1}{6\Delta_1^3}\Big[3 (\Delta_2 - \Delta_1) (\Theta_1 (\Delta_2 - \Delta_1) - \Delta_1 (\Theta_2 - \Theta_1)) \\ + 2\Delta_1 (\Delta_1 (\Theta_3 - 3\Theta_2 - 3\Theta_1) - \Theta_1 (\Delta_3 - 3\Delta_2 + 3\Delta_1))\Big] \end{aligned} \tag{26}$$

where

$$\begin{aligned} \Delta_k &= N^k - \mathbb{E}\left[J_N^k\right] \\ \Theta_k &= \sum_{i=0}^{N-1} \mathbb{E}\left[(N + J_i)^k\right] - \mathbb{E}\left[(i + J_N)^k\right] \end{aligned}$$

Although these equations look somewhat ugly, they are straightforward to calculate once the probabilities $p_0, \ldots, p_{N-1}$ are known. The moments of the distribution of the number of jobs waiting to be admitted (W) and the number of jobs in the production unit (X) are directly found through their relation to the total number of jobs:

$$\mathbb{E}[X] = \sum_{i<N} p_i\, i + N(1 - \sum_{i<N} p_i) \tag{27}$$

$$\mathbb{E}\left[X^2\right] = \sum_{i<N} p_i\, i^2 + N^2(1 - \sum_{i<N} p_i) \tag{28}$$

$$\mathbb{E}[W] = \sum_{i>N} (i-N)\, p_i = \mathbb{E}[L] - \mathbb{E}[X] \tag{29}$$

$$\mathbb{E}\left[W^2\right] = \sum_{i>N} (i-N)^2\, p_i = \mathbb{E}\left[L^2\right] - \mathbb{E}\left[X^2\right] - 2N(\mathbb{E}[L] - \mathbb{E}[X]) \tag{30}$$

B.3. **Moments of the Sojourn Time in the Facility.** The first two moments of the sojourn time are found by taking the derivative of the Laplace-Stieltjes transform $T^*(s)$ and letting $s \to 0$. The first two moments are:

$$\mathbb{E}[T] = \frac{\mathbb{E}[B]\left(\mathbb{E}[X] + \mathbb{E}\left[X^2\right] - \mathbb{E}[Y] - \mathbb{E}\left[Y^2\right]\right)}{2(\mathbb{E}[X] - \mathbb{E}[Y])} \tag{31}$$

$$\mathbb{E}\left[T^2\right] = \frac{3\,\mathbb{E}\left[B^2\right]\left(\mathbb{E}[X] + \mathbb{E}\left[X^2\right] - \mathbb{E}[Y] - \mathbb{E}\left[Y^2\right]\right)}{6\left(\mathbb{E}[X] - \mathbb{E}[Y]\right)} \tag{32}$$

$$\frac{-2\,\mathbb{E}[B]^2\left(\mathbb{E}[X] - \mathbb{E}\left[X^3\right] - \mathbb{E}[Y] + \mathbb{E}\left[Y^3\right]\right)}{6\left(\mathbb{E}[X] - \mathbb{E}[Y]\right)}$$

where

$$\mathbb{E}\left[Y^k\right] = \sum_{i=1}^{N-1} p_i\, \mathbb{E}\left[\left(i - V_i\right)^k\right] + \left(1 - \sum_{i=0}^{N-1} p_i\right) \mathbb{E}\left[\left(N - V_N\right)^k\right]$$