The effect of workload constraints in mathematical programming models for production planning

The effect of Workload Constraints in Mathematical Programming Models for Production Planning

Michiel Jansen, Ivo Adan, Ton de Kok

Beta Working Paper series 331

BETA publicatie	WP 331 (working paper)
ISBN	978-90-386-2386-3
ISSN	
NUR	982
Eindhoven	October 2010

Elsevier Editorial System(tm) for European Journal of Operational Research
Manuscript Draft

Manuscript Number:

Title: The Effect of Workload Constraints in Mathematical Programming Models for Production Planning

Article Type: Innovative Application of OR

Section/Category: Queueing

Keywords: Queueing, Production Planning, Rolling Schedules, Math Programming

Corresponding Author: Mr. Michiel Jansen, MSc.

Corresponding Author's Institution: Eindhoven University of Technology

First Author: Michiel Jansen, MSc.

Order of Authors: Michiel Jansen, MSc.; Ivo Adan, Prof.Dr.Ir.; Ton de Kok, Prof.Dr.

Abstract: Linear and mixed integer programming models for production planning incorporate a model of the manufacturing system that is necessarily deterministic. Although these deterministic models are the current-state-of-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 relation between workload constraints that reflect the finite capacity of the manufacturing system, and the use of planned lead times. It is a common practice in rolling schedule based production planning to limit the periodic output to the average production rate. If lead times are not modeled explicitly, this also implies a restricition on the periodic releases to the average production rate. We demonstrate that this common practice results in inefficient use of the production capacity and show that the use of planned lead times leads to a better trade-off between efficiency and reliability. We analyze a stylized model of a manufacturing system with a single exponential server and two queues in series: an admission queue and a work-in-progress (WIP) queue. The admission queue represents the pool of unreleased orders that is virtually present in the state variables of the planning model. Periodically, jobs from the admission queue are released to the WIP queue such that the number of jobs in WIP and in service does not exceed the workload constraint. We present a simple formula for the maximum utilization rate of such a system, characterize the stationary queue-length distribution by its generating function, and give the distribution of the sojourn time of a job. We use the results to compare various settings of the workload constraint and the planned lead time.

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

M.M. Jansen[a,b,*], A.G. de Kok[a], I.J.B.F. Adan[b]

[a] *Eindhoven University of Technology, Department of Industrial Engineering, Paviljoen E.14, Postbus 513, 5600 MB Eindhoven*
[b] *EURANDOM, Postbus 513, 5600 MB Eindhoven*

Abstract

Linear and mixed integer programming models for production planning incorporate a model of the manufacturing system that is necessarily deterministic. Although these deterministic models are the current-state-of-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 relation between workload constraints that reflect the finite capacity of the manufacturing system, and the use of planned lead times. It is a common practice in rolling schedule based production planning to limit the periodic output to the average production rate. If lead times are not modeled explicitly, this also implies a restricition on the periodic releases to the average production rate. We demonstrate that this common practice results in inefficient use of the production capacity and show that the use of planned lead times leads to a better trade-off between efficiency and reliability. We analyze a stylized model of a manufacturing system with a single exponential server and two queues in series: an admission queue and a work-in-progress (WIP) queue. The admission queue represents the pool of unreleased orders that is virtually present in the state variables of the planning model. Periodically, jobs from the admission queue are released to the WIP queue such that the number of jobs in WIP and in service does not exceed the workload constraint. We present a simple formula for the maximum utilization rate of such a system, characterize the stationary queue-length distribution by its generating function, and give the distribution of the sojourn time of a job. We use the results to compare various settings of the workload constraint and the planned lead time.

Keywords: Queueing, Production Planning, Rolling Schedules, Math Programming

*Corresponding author
Email addresses: m.m.jansen@tue.nl (M.M. Jansen), a.g.d.kok@tue.nl (A.G. de Kok), i.j.b.f.adan@tue.nl (I.J.B.F. Adan)

1. Introduction

With the emergence of Advanced Planning Systems [10], Mathematical Programming (MP) models for production planning [8, 12, 15, 17–19] are becoming more commonplace in supporting firm's decision making processes. Contrary to the Manufacturing Resources Planning (MRP-II) concept [22], these models deal with goods flow coordination and resource allocation in an integrated fashion. Periodic production quantities are constrained by some capacity parameter specifying the maximum possible throughput in a period. The maximum periodic throughput is treated as a deterministic variable in these MP models. Although it is recognized by most authors that in reality there is uncertainty in the maximum throughput, it is suggested that this should be dealt with by installing safety stock or safety time. Suggestions on how to set the capacity parameter in practice are generally absent. It seems to be an obvious choice to set this parameter equal to the average production rate. In this paper we show that this approach may seriously reduce the efficiency of resource utilization.

In practice, production planning is carried out in a rolling schedule based fashion. A plan is developed for multiple periods into the future but only the decisions for the first period are implemented. One period later, the planning model is updated with actual forecast and state information and a new plan is created. For most MP models that are proposed in the literature, these decisions are the production quantities. In other literature [5, 16] the order release decision is decoupled from the production quantities through the use of an explicitly defined planned lead time that is expressed as an integer number of planning periods. 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 explicit planned lead times are special cases with a planned lead time of a single period. It is argued in [13, 14] that optimal planned lead times may be larger than a single period. We shall provide another argument for the use of explicit planned lead times that are possibly larger than a single period.

The planned lead time is essential for the coordination of goods flows in a supply chain 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 (see section 15.5 of [6]). Hence, there is a trade off between restricting the workload to ensure a reliable planned lead time and relaxing the workload constraint to reduce the build-ahead inventory.

In this paper we study the effect of the workload constraint and the planned lead time parameters on the efficiency of the resource utilization. We study this effect in a stylized model of the manufacturing system. The manufacturing system consists of a facility in which at most N jobs are allowed (the workload constraint). In front of the facility is a queue that contains those jobs that cannot be admitted to the facility due to the workload constraint. This admission queue represents the build-ahead inventory. New requirements in a period are represented by arrivals of jobs to the admission queue. We consider two measures of efficiency. The first measure is the maximum utilization level under which the system remains stable. The second measure is the expected length of the admission queue. Reliability of the planned lead time is expressed as the probability that the sojourn time of a job in the facility exceeds the planned lead time.

A graphical representation of the model is shown in Figure 1. The total number of jobs in the manufacturing system, the number waiting in the admission queue, and the number residing in the facility are denoted by L_n, W_n, and X_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 are related to the total number of jobs in the manufacturing system in the following way

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

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

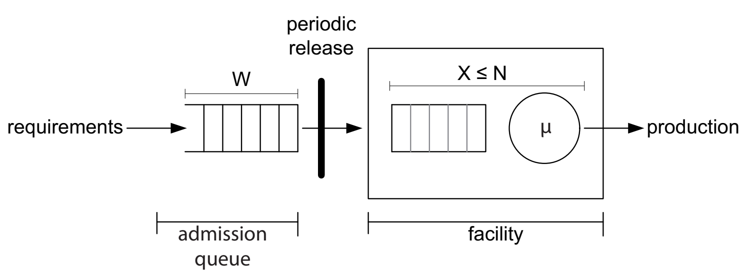

Figure 1: Model of the Manufacturing System

We assume that jobs arrive according to a (compound) Poisson process with mean λ and we denote the number of arrivals in a period that starts with epoch n by A_n. Jobs in the facility are processed by a single server with exponential service times with mean μ^{-1}. $V_{x,n} = \min\{V_{\infty,n}, x\}$ denotes the throughput in period n conditioned on a workload level x at the start of the period, where $V_{\infty,n}$ is a Poisson random variable with rate μ. In other words, the throughput in period n is $V_{X_n,n}$. Since $V_{x,n}$ are i.d.d. random variables, we ommit the index n whenever this is convenient (similarly for A_n). The dynamics of the total number of customers is described by the following Lindley type equation:

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

The remainder of his paper is organized as follows. First we briefly discuss some of the literature on queueing models that are related to ours. We then present a stability condition for the system. Next we describe the mathematical model and give the probability generating function (PGF) for the state distributions at the release epochs. We also derive the CDF of the sojourn time distribution. Finally we use these results to compare various settings of the workload constraint and the planned lead time.

Before we continue, we introduce some notation that we use throughout the paper. Let $V = \lim_{n\to\infty} V_{X_n,n}$, $L = \lim_{n\to\infty} L_n$, $W = \lim_{n\to\infty} W_n$, $X = \lim_{n\to\infty} X_n$. We use the additional notation $\rho := \lambda/\mu$, $(x)^+ = \max\{0, x\}$, $(x)^- = \max\{0, -x\}$, and $|x| = (x)^+ + (x)^-$.

2. Literature

The queueing model described in this paper is related to two streams of literature. The first stream is the literature on the bulk service queue. The analysis of the bulk service queue model is very similar to ours. In the bulk service queue, jobs enter into 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. 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 is the fact that there may be jobs left in the facility at the end of the period. A seminal paper in the area of bulk service queues is Bailey [2]. Other important references include [3, 11]. Van Leeuwaarden [20] presents an extensive treatment of the discrete bulk service queue that is described by the Lindley equation $X(t+1) = (X(t) - N)^+ + A(t)$.

The other stream of literature that is related to our model is the literature on the fixed-cycle traffic light (FCTL) queue. In the FTCL queue, there are cycles that consist of a red period during which jobs may arrive but are not served and a green period in which jobs both arrive and are being served. The length of a cycle and the red and green period is fixed. The similarity of the FTCL queue with the periodic order release model described in this paper is that capacity is lost due to the fact that the server may idle even though there are jobs queueing. Most traffic light queues assume a constant rate of departure and therefore there is a natural limit on the number of jobs that can be processed in the green period. Traffic light systems are for example discussed in [4, 9] and more recently in [21]. The key step in each of these papers is the characterization of the number of jobs at the end of a cycle. To this purpose, a probability generating function (PGF) is formulated that include 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 [23]. 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 [2], Wang characterizes the steady state queue-length distribution by a PGF. Wang obtains closed-form expressions for the cases $c = 1$ and $c = \infty$.

3. A 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 long run 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. 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. 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*

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

where the latter inequality is strict if $N < \mu$.

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 is this fraction 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 $\mu > N$.

Remark 1. Although we consider a facility with Poisson arrival and service processes, Proposition 1 holds for any independent and identical discretely distributed A and V_∞.

As we already mentioned in the introduction, it seems to be a natural choice to set the capacity parameter in MP models for production planning equal to the mean or expected throughput (i.e. $N = \mu$). In this case, the formula for maximum utilization simplifies as follows.

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

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

where MAD stands for the Mean Absolute Deviation.

Proof. The proof follows readily from Proposition 1.

The MAD is a measure of variability that is often used by practitioners. Figure 2 shows the maximum utilization rate for three different squared coefficients of variation (scv) of the maximum throughput. 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. The figure shows that the maximum utilization is strongly reduced for if the workload constraint is set equal to the expected maximum throughput.

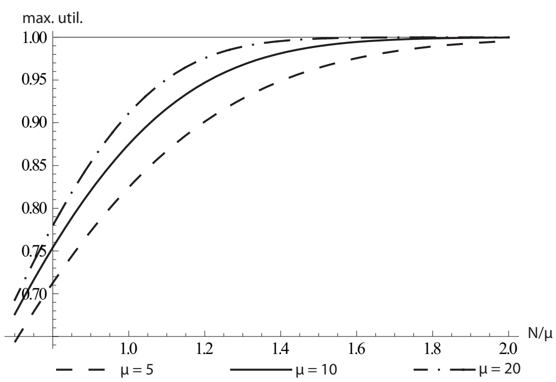

Figure 2: Maximum Utilization Level

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

4. The Stationary Distributions

4.1. 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

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

where

$$\beta_{ij} = \mathbb{P}\{A - V_i = j - i\}, \qquad \text{for all } 0 < i \leq N, j \geq 0 \tag{6}$$

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

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

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

and

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

If the stability condition 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:

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

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

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:

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

which reduces to

$$G_L(z) = \frac{G_A(z)\, \sum_{i=0}^{N-1} p_i\, z^{N+i}\, \big(G_{V_i}(z^{-1}) - G_{V_N}(z^{-1})\big)}{z^N - z^N G_A(z) G_{V_N}(z^{-1})} \tag{10}$$

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 $N-1$ such roots [1] and these roots can routinely be found using computer packages such as Wolfram Mathematica and MATLAB. Since the PGF is finite inside the unit circle, the numerator must be zero 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'Hopital's rule:

$$\sum_{i=0}^{N-1} p_i \Big(\mathbb{E}[V_N] - \mathbb{E}[V_i]\Big) = \mathbb{E}[V_N] - \mathbb{E}[A] \tag{11}$$

Note that equation (11) is precisely the equation that balances inflow and outflow:

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

The system of equations that needs solving becomes:

$$\big(\mathbf{Z}\,\mathbf{V} - \mathbf{z^N}\,\bar{\mathbf{v}}^T\big)\,\mathbf{p} = \begin{pmatrix} \mathbf{0} \\ \mathbb{E}[V_N] - \mathbb{E}[A] \end{pmatrix}, \tag{12}$$

where

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

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

with $v_i = \mathbb{P}\{V_\infty = i\}$, and $\bar{v}_i = \mathbb{P}\{V_\infty \geq i\}$. With the solution $\mathbf{p} = (p_0, p_1, \dots, p_{N-1})^T$, the PGF $G_L(z)$ is fully defined. For more information on finding the unknowns in PGF's, see for example chapter 2 of [20].

In Appendix Appendix B we give equations that can be used to obtain the entire probability distribution and expressions for the first two moments of L, W, and X. In Appendix Appendix A we also describe an alternative, numerically stable method for obtaining the stationary distribution of L by analyzing an embedded Markov Chain.

So far, we have not used the fact that arrivals and departures have exponentially distributed interarrival times. The result up to here hold for any discrete distribution of periodic arrivals and departures. The analysis can also be applied to facilities with load-dependent exponential servers (including the multi-server queue). For the determination of the sojourn times in the next sub-section, we do rely on the exponential distribution of the interarrival and service times.

4.2. Sojourn times

In queueing systems where jobs arrive in unit size according to a Poisson process, a distributional form of Little's Law applies (cf. [7]). Consider a job departing from the system. The number of jobs $\tilde{L}$ in the system after its departure is equal to the number of arrivals that occurred during its sojourn time S. By a level crossing argument and the PASTA property, the number of customers arriving during an arbitrary sojourn time is equal in distribution to the number of jobs in the system at an arbitrary point in time. If arrivals follow a Poisson process, the PGF of the arbitrary time number of jobs in the system $G_{\tilde{L}}$ is related to the LST of the time spent in the system S^* as follows:

$$\begin{aligned} G_{\tilde{L}}(z) &= \sum_{k=0}^{\infty} z^k \int_{t=0}^{\infty} \frac{e^{-\lambda t}\,(\lambda t)^k}{k!}\, d\mathbb{P}\{S < t\} \\ &= S^*\big(\lambda(1-z)\big) \end{aligned}$$

Hence, the LST of the time spent in the system becomes:

$$S^*(s) = G_{\tilde{L}}\Big(\frac{\lambda - s}{\lambda}\Big) \tag{13}$$

In a similar way, we can find the sojourn time in the admission queue using $G_{\tilde{W}}$. The PGF's $G_{\tilde{L}}$ and $G_{\tilde{W}}$ are derived in Appendix Appendix C.

As we study the trade-off between a workload constraint and lead time reliability in this paper, we are particularly interested in the time that a job spends in the facility. The PASTA property does not hold for the arrivals into the facility so it is not possible to follow the approach described above here. Instead, we condition the sojourn time in the facility on the number of jobs residing in the facility just before a new admission, and the size of the batch in which a job enters. Here we must take into account the inspection paradox: an arbitrary job is more likely to be part of a large admission batch. We use the following lemma for the calculation of the sojourn time of an arbitrary job:

Lemma 1. *Let (Y, Q) be the number of jobs in the manufacturing system just before a release epoch and the number of jobs subsequently admitted. Let $(\tilde{Y}, \tilde{Q})$ be the number of jobs prior to a release and the number of jobs admitted* as seen by an arbitrary job. *Then the joint probability distributions of these random variables are related in the following way:*

$$\mathbb{P}\left\{\tilde{Y} = y, \tilde{Q} = q\right\} = \frac{q\,\mathbb{P}\{Y = y, Q = q\}}{\mathbb{E}[Q]} \tag{14}$$

Proof. Take a large number of release epochs M and mark each job that enters by with the size q of the batch in which it enters. The fraction of batches of size q is $\mathbb{P}\{Q = q|Q > 0\}$, and the expected number of batches is $M\,\mathbb{P}\{Q = q|Q > 0\}$. We put all jobs together in a bin. By the law of large numbers, for large enough M, the number of jobs marked q in the bin is $q\,\mathbb{P}\{Q = q|Q > 0\}\ M$ and the total number of jobs in the bin is $\sum_{q=1}^{N} q\,M\,\mathbb{P}\{Q = q|Q > 0\}$. Letting M go to infinity and randomly picking one job from the bin, the probability of having picked a job marked q is:

$$\begin{aligned}
\mathbb{P}\left\{\tilde{Q} = q\right\} &= \lim_{M\to\infty} \frac{\binom{q\,\mathbb{P}\{Q = q|Q > 0\}\ M}{1}}{\binom{\sum_{q=1}^{N} u\,\mathbb{P}\{Q = q|Q > 0\}\ M}{1}} \\
&= \lim_{M\to\infty} \frac{q\,\mathbb{P}\{Q = q\}/\mathbb{P}\{Q > 0\}\ M}{\sum_{q=1}^{N} u\,\mathbb{P}\{Q = q\}/\mathbb{P}\{Q > 0\}\ M} \\
&= \frac{q\,\mathbb{P}\{Q = q\}}{\mathbb{E}[Q]}
\end{aligned}$$

Finally, we condition the number of jobs in the manufacturing system before admission on the admission batch size to get

$$\mathbb{P}\left\{\tilde{Y} = y, \tilde{Q} = q\right\} = \frac{q\,\mathbb{P}\{Q = q\}}{\mathbb{E}[Q]}\mathbb{P}\{Y = y|Q = q\} = \frac{q\,\mathbb{P}\{Y = y, Q = q\}}{\mathbb{E}[Q]}$$

Proposition 2. *Consider the workload constrained manufacturing system consisting of a single server with exponential service rate μ to which orders arrive with a rate λ. Let $G_Y(z) := \sum_{y=0}^{N} \mathbb{P}\{Y=y\}\, z^y$ and $G_X(z) := \sum_{x=0}^{N} \mathbb{P}\{X=x\}\, z^x$ be the PGF's of respectively the number of jobs in the manufacturing system just before and just after the release epoch. Then the LST of the sojourn time in the manufacturing system for an arbitrary job has a distribution with LST T^*:*

$$T^*(s) = \frac{\mu}{s\lambda}\left[G_Y\Big(\frac{\mu}{s+\mu}\Big) - G_X\Big(\frac{\mu}{s+\mu}\Big)\right] \tag{15}$$

The CDF of the sojourn time is $F_T(t)$:

$$\begin{aligned} F_T(t) = \frac{\mu}{\lambda}\Big[&\Big(\mathbb{P}\{Y=0\} - \mathbb{P}\{X=0\}\Big)t \\ &+ \sum_{k=1}^{N}\Big(\mathbb{P}\{Y=k\} - \mathbb{P}\{X=k\}\Big)\Big(t\Gamma_{k,\mu}(t) - \frac{k}{\mu}\Gamma_{k+1,\mu}(t)\Big)\Big], \end{aligned} \tag{16}$$

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

Proof. Let B denote a service time with $\mathbb{E}[B] = \mu^{-1}$ and denote the PGF

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

For a job entering in the j^{th} position of a batch into a system with y customers already present, the sojourn time T is the $(y+j)$-fold convolution of B. Let $\chi = \{(y,q) \in \mathbb{N}^2 : y \geq 0, q > 0, y+q \leq N\}$ and denote $p_{yq} = \mathbb{P}\{Y=y, Q=q\}$. Noting that the probability that a job is in the j^{th} place of a batch of size q is $\frac{1}{q}$ and using Lemma 1 the LST of T can be written as

$$\begin{aligned} T^*(s) &= \mathbb{E}\left[e^{-sT}\right] = \sum_{(y,q)\in\chi} \mathbb{P}\left\{\tilde{Y}=y, \tilde{Q}=q\right\} \sum_{j=1}^{q} \frac{1}{q}\mathbb{E}\left[e^{-sB^{(y+j)}}\right] \\ &= \sum_{(y,q)\in\chi} \frac{q\,p_{yq}}{\mathbb{E}[Q]} \sum_{j=1}^{q} \frac{1}{q}\big(B^*(s)\big)^{y+j} \\ &= \sum_{(y,q)\in\chi} p_{yq}\frac{\big(B^*(s)\big)^{y+1}}{\mathbb{E}[Q]}\frac{1-\big(B^*(s)\big)^{q}}{1-B^*(s)} \\ &= \frac{B^*(s)}{\mathbb{E}[Q]\big(1-B^*(s)\big)}\left[\sum_{y=0}^{N-1}\sum_{q=1}^{N-y} p_{yq}\big(B^*(s)\big)^{y} - \sum_{y=0}^{N-1}\sum_{q=1}^{N-y} p_{yq}\big(B^*(s)\big)^{y+q}\right] \end{aligned}$$

We are free to extend the summation ranges in the term between square brackets to include $q=0$ and $y=N$ since these terms cancel out. Since

$X(t) = Y(t) + Q(t)$ we have

$$\begin{aligned} T^*(s) &= \frac{B^*(s)}{\mathbb{E}[Q]\big(1-B^*(s)\big)}\left[\sum_{y=0}^{N}\sum_{q=0}^{N-y} p_{yq}\big(B^*(s)\big)^y - \sum_{y=0}^{N}\sum_{q=0}^{N-y} p_{yq}\big(B^*(s)\big)^{y+q}\right] \\ &= \frac{B^*(s)}{\mathbb{E}[Q]\big(1-B^*(s)\big)}\left[G_Y\big(B^*(s)\big) - G_X\big(B^*(s)\big)\right] \\ &= \frac{\mu}{s\lambda}\left[G_Y\big(\tfrac{\mu}{s+\mu}\big) - G_X\big(\tfrac{\mu}{s+\mu}\big)\right], \end{aligned}$$

where in the last step, we used that $\mathbb{E}[Q] = \lambda$.

Equation (15) can be written as

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

It can easily be verified that the Laplace transform in t 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:

$$f_T(t) = \frac{\mu}{\lambda}\Big[\mathbb{P}\{Y=0\} - \mathbb{P}\{X=0\} + \sum_{k=1}^{N}\big(\mathbb{P}\{Y=k\} - \mathbb{P}\{X=k\}\big)\Gamma_{k,\mu}(t)\Big] \quad (17)$$

Finally, using that

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

yields the CDF of T.

Using its LST, we can easily calculate the moments of T. The first two moments are given in Appendix Appendix B. The CDF of T can be obtained by numerically inverting the LST.

5. Numerical examples

In this section we present some numerical results for the performance of the workload constrained manufacturing system. 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 $\mu = 5$ (increasing coefficient of variation).

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 an variance, also the probability that the sojourn time is less than a planned lead time of 1, 2, and 3 periods respectively is given.

Table 1: Effect of the Workload Constraint on Queue Lengths

			0.78				0.82				0.86			
μ	**N/μ**	**ρ_{max}**	**E[W]**	**Var[W]**	**E[X]**	**Var[X]**	**E[W]**	**Var[W]**	**E[X]**	**Var[X]**	**E[W]**	**Var[W]**	**E[X]**	**Var[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

			0.78					0.82					0.86				
μ	**N/μ**	**ρ_{max}**	**E[T]**	**Var[T]**	**P{T<1}**	**P{T<2}**	**P{T<3}**	**E[T]**	**Var[T]**	**P{T<1}**	**P{T<2}**	**P{T<3}**	**E[T]**	**Var[T]**	**P{T<1}**	**P{T<2}**	**P{T<3}**
	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
20	**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
	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
10	**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
	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
5	**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 was 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 $\mathrm{Var}[V_\infty]$, $\mathbb{E}[W]$ grows rapidly as N is reduced to μ. The effect of the workload constraint on $\mathbb{E}[X]$ is relatively small but $\mathrm{Var}[X]$ is reduced substantially with a more restrictive workload constraint. The squared coefficient of variation (SCV) of V_∞ appears to be an important factor for the sensitivity of the sojourn times. For $\mu = 20$ ($SCV = 0.05$) the effect of the workload constraint on the sojourn time is relatively small. On the other hand, for $\mu = 5$ ($SCV = 0.2$) 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 utilization

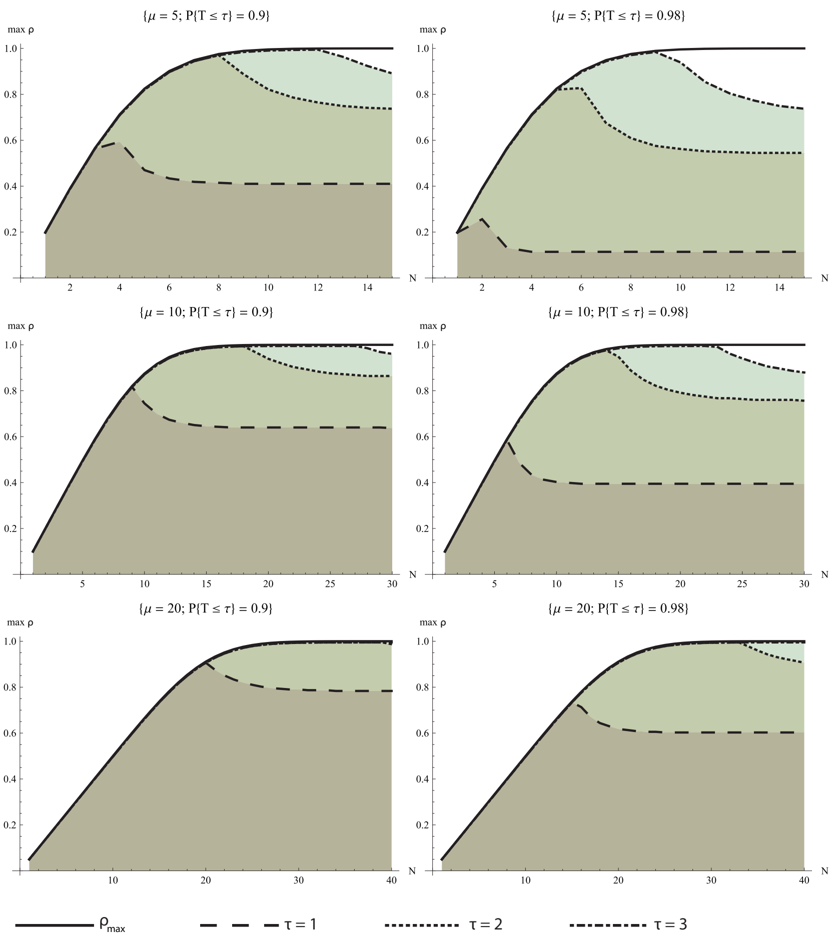

Figure 3: Maximum Utilization

and the reliability of the planned lead time. Figure 3 shows the maximum utilization rate of the manufacturing system for a given lead time reliability $\alpha \in \{0.9, 0.98\}$ and planned lead time $\tau \in \{1, 2, 3\}$. Lead time reliability is defined as $\mathbb{P}\{T \leq \tau\} \geq \alpha$. 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 3 gives 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 3 can be used to find which settings of τ are feasible under the reliability constraint α. For example, a system with $\mu = 10$ that is running at $\rho = 0.8$ must have $\tau \geq 2$ for $\alpha = 0.9$.

6. Conclusions

Mathematical programming (MP) 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 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 require 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 utilization 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 extend of the load smoothing effect that is the result of the workload constraint. The load smoothing effect is the average number of items in backlog or produced in advance of the requirement due to the workload constraint. This effect is reflected in the number of jobs waiting to be admitted to the manufacturing system.

The numerical study shows that the common practice of setting the capacity parameter in MP models for production planning equal to the average production rate ($N = \mu$) leads both to poor reliability and a poor efficiency. Whereas relaxing the workload constraint leads to deterioration of the reliability, restricting it further leads to a increase of the smoothing effect. A better trade-off between reliability and efficiency is obtained for higher values of the planned lead time parameter.

References

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

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

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

[5] A.G. de Kok and J.C. Fransoo. *Supply Chain Management: Design, Coordination and Operation*, volume 11, chapter Planning Supply Chain Operations: Definition and Comparison of Planning Concepts, pages 597–676. Elsevier, 2003.

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

[7] J. Keilson and L.D. Servi. A distribution form of Littles law. *Operations Research Letters*, 7(5):223–227, 1988.

[8] K.R. Baker. *Logistics of Production and Invetory*, volume 4 of *Handbooks in OR & MS*, chapter Requirements Planning, pages 571–627. North-Holland, 1993.

[9] 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.

[10] H. Meyr, M. Wagner, and J. Rohde. *Supply Chain Management and Advanced Planning - Concepts, Models, Software and Case Studies*, chapter Structure of Advanced Planning Systems, pages 75–77. Springer-Verlag Berlin, 1st edition, 2000.

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

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

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

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

[15] J.M. Spitter, A.G. de Kok, and N.P. Dellaert. Timing production in lp models in arolling schedule. *Int. J. Production Economics*, 93-94:319–329, 2005.

[16] 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.

[17] 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.

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

[19] L.J. Thomas and J.O. McClain. An overview of production planning. *Logistics of Production and Inventory*, pages 333–370, 1993.

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

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

[22] T.E. Vollmann, W.L. Berry, and D.C. Whybark. *Manufacturing Planning and Control Systems*. Homewood, Ill. : Dow Jones-Irwin, 1984.

[23] P. 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 on the states $0, \ldots, N-1$. That is, the Markov Chain is only observed at the times n where $L_n < N$. 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 require the return probabilities that are 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 a state $j \leq n$ 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:

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

This sequence is increasing and bounded so the limit exists. We now define the Markov Chain embedded on $\{0, 1, \ldots, N-1\}$ with transition matrix $Q = (q_{ij})$,

$$q_{ij} = \beta_{ij} + \sum_{m=0}^{\infty} \beta_{i,N+m}\, b_{m,N-i}, \quad , 0 \leq i, j < N \tag{A.2}$$

Let $\tilde{p}$ be the solution of the embedded Markov Chain (i.e. $\tilde{p}\, Q = \tilde{p}$). Then we use (11) for normalization to find the original probabilities $p_i, i = 0, \ldots, N-1$:

$$p_i = \frac{1}{c}\, \tilde{p}_i, \tag{A.3}$$

where

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

Remark 2. To obtain the numerical results presented in this paper, we use the following stopping criterion for the iteration in A.2. For a given i, let $\hat{\beta}_i := \max\{j : \beta_{i,j} > \epsilon\}$, where ϵ is small. Stop whenever $\max_{m \leq \hat{\beta}_i - N}\{b_{i,m}^{(k)} - b_{i,m}^{(k-1)}\} < \epsilon$.

Appendix B. More Details of the Stationary Probability Distributions

Appendix 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:

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

The balance equation in (B.1) 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:

$$\begin{aligned} p_i = {} & \frac{1}{(1 - \alpha_N - \sum_{k=0}^{\infty} \alpha_{N+k+1}\, b_{k,1})} \times \\ & \Big[\sum_{j=0}^{N-1} p_j \Big(\beta_{ji} + \sum_{k=0}^{\infty} \beta_{j,i+k+1}\, b_{k,1}\Big) \\ & + \sum_{j=N}^{i-1} p_j \Big(\alpha_{N+(i-j)} + \sum_{k=0}^{\infty} \alpha_{N+(i-j)+k+1}\, b_{k,1}\Big)\Big] \end{aligned} \tag{B.2}$$

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

Appendix 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 (10) and taking the limit $z \to 1$. Applying l'Hopital's rule, the first two moments become:

$$\mathbb{E}[L] = \frac{\Delta_1 (\Theta_2 - \Theta_1) - \Theta_1 (\Delta_2 - \Delta_1)}{2\Delta_1^2} \tag{B.3}$$

$$\begin{aligned} \mathbb{E}\left[L^2\right] = {} & \frac{1}{6\Delta_1^3}\Big[3(\Delta_2 - \Delta_1)(\Theta_1(\Delta_2 - \Delta_1) - \Delta_1(\Theta_2 - \Theta_1)) \\ & + 2\Delta_1(\Delta_1(\Theta_3 - 3\Theta_2 - 3\Theta_1) - \Theta_1(\Delta_3 - 3\Delta_2 + 3\Delta_1))\Big] \end{aligned} \tag{B.4}$$

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 $p_0, \dots, p_{N-1}$ 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:

$$\mathbb{E}[X] = \sum_{i<N} p_i\, i + N(1 - \sum_{i<N} p_i) \tag{B.5}$$

$$\mathbb{E}\left[X^2\right] = \sum_{i<N} p_i\, i^2 + N^2(1 - \sum_{i<N} p_i) \tag{B.6}$$

$$\mathbb{E}[W] = \sum_{i>N} (i-N)\, p_i = \mathbb{E}[L] - \mathbb{E}[X] \tag{B.7}$$

$$\mathbb{E}\left[W^2\right] = \sum_{i>N} (i-N)^2\, p_i = \mathbb{E}\left[L^2\right] - \mathbb{E}\left[X^2\right] - 2N(\mathbb{E}[L] - \mathbb{E}[X]) \tag{B.8}$$

Appendix 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:

$$\mathbb{E}[T] = \frac{\mathbb{E}[B]\left(\mathbb{E}[X] + \mathbb{E}\left[X^2\right] - \mathbb{E}[Y] - \mathbb{E}\left[Y^2\right]\right)}{2(\mathbb{E}[X] - \mathbb{E}[Y])} \tag{B.9}$$

$$\mathbb{E}\left[T^2\right] = \frac{3\,\mathbb{E}\left[B^2\right]\left(\mathbb{E}[X] + \mathbb{E}\left[X^2\right] - \mathbb{E}[Y] - \mathbb{E}\left[Y^2\right]\right)}{6\left(\mathbb{E}[X] - \mathbb{E}[Y]\right)} \frac{-2\,\mathbb{E}[B]^2\left(\mathbb{E}[X] - \mathbb{E}\left[X^3\right] - \mathbb{E}[Y] + \mathbb{E}\left[Y^3\right]\right)}{6\left(\mathbb{E}[X] - \mathbb{E}[Y]\right)} \tag{B.10}$$

where

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

Appendix C. Distribution of the Stationary Queue Lengths at Arbitrary Time

In order to obtain the time that a job spends in the admission queue and in the whole manufacturing system, we require the stationary distributions of the admission queue length and the total number of jobs in the system *at arbitrary time*. Without loss of generality we assume that the length of a period is one. Let $A_n(t)$ and $V_{i,n}(t)$ be the number of arrivals and jobs processed in the interval $[r_n, r_n + t]$, $\quad 0 \le t < 1$, where , r_n is the time of the n^{th} release epoch.

Furthermore, let $L_n(t)$ be the number of jobs at time $r_n + t$. The queue-length processes at arbitrary times are given by:

$$L_n(t) = L_n + A_n(t) - V_{X_n,n}(t) \tag{C.1}$$

The stationary distribution of $L_n(t)$ is denoted by $L(t)$. We define the PGF's $G_A(z,t)$, $G_{V_i}(z,t)$, $G_L(z,t)$, where

$$G_L(z,t) := \sum_{k=0}^{\infty} z^k \, \mathbb{P}\{L(t) = k\}$$

and the others are defined similarly.

From (C.1) we obtain expressions for the PGF defined above:

$$\begin{aligned} G_L(z,t) &= \sum_{i=0}^{N-1} p_i z^i G_A(z,t)\, G_{V_i}(z^{-1},t) \\ &\quad + \Big(1 - \sum_{i=0}^{N-1} p_i\Big) z^N G_A(z,t) G_{V_N}(z^{-1},t) \end{aligned} \tag{C.2}$$

We use the notation $\tilde{L}$ to denote the arbitrary-time variant of L. The PGF of $\tilde{L}$ is

$$G_{\tilde{L}}(z) := \int_0^{1-} G_L(z,t)\, d\tau \tag{C.3}$$

$$\tag{C.4}$$

For a compound Poisson arrival process and a Poisson departure process we have:

$$\begin{aligned} G_A(z,t) &= e^{-t\lambda\big(1-G_D(z)\big)}, \\ G_{V_i}(z,t) &= e^{-t(1-z)/\mu}\, \bar{\Gamma}_{i,\mu}(z\,t) + z^i\, \Gamma_{i,\mu}(t), \end{aligned}$$

where $G_D(z)$ is the PGF of the size of individual arrivals, $\Gamma_{i,\mu}$ is the CDF of a Erlang/Gamma random variable with mean $i\,\mu^{-1}$ and variance $i\,\mu^{-2}$ whose complement is denoted by $\bar{\Gamma}_{i,\mu}$.

For notational convenience we introduce $G_{\tilde{J}_i}(z) = \int_0^{1-} z^i G_A(z,t) G_{V_i}(z^{-1},t)dt$. Furthermore, let $r := \lambda\big(1 - G_D(z)\big)$, $s := \mu(1 - z^{-1})$, and $v := \mu/(r + \mu)$. With some algebra we obtain

$$\begin{aligned} G_{\tilde{J}_i}(z) &= \int_0^{1-} z^i G_A(z,t) G_{V_i}(z^{-1},t) \\ &= \frac{1}{r+s}\Big[z^i\big(1 - G_A(z) G_{V_i}(z^{-1})\big) + \frac{s}{r}\, v^i\,\big(1 - e^{-r}\, G_{V_i}(v^{-1})\big)\Big] \end{aligned} \tag{C.5}$$

so that

$$G_{\tilde{L}}(z) = \sum_{i=0}^{N-1} p_i G_{\tilde{J}_i}(z) + \Big(1 - \sum_{i=0}^{N-1} p_i\Big) G_{\tilde{J}_N}(z) \tag{C.6}$$

Working Papers Beta 2009 - 2010

nr.	Year	Title	Author(s)
331	2010	The Effect of Workload Constraints in Mathematical Programming Models for Production Planning	M.M. Jansen, A.G. de Kok, I.J.B.F. Adan
330	2010	Using pipeline information in a multi-echelon spare parts inventory system	Christian Howard, Ingrid Reijnen, Johan Marklund, Tarkan Tan
329	2010	Reducing costs of repairable spare parts supply systems via dynamic scheduling	H.G.H. Tiemessen, G.J. van Houtum
328	2010	Identification of Employment Concentration and Specialization Areas: Theory and Application	F.P. van den Heuvel, P.W. de Langen, K.H. van Donselaar, J.C. Fransoo
327	2010	A combinatorial approach to multi-skill workforce scheduling	Murat Firat, Cor Hurkens
326	2010	Stability in multi-skill workforce scheduling	Murat Firat, Cor Hurkens, Alexandre Laugier
325	2010	Maintenance spare parts planning and control: A framework for control and agenda for future research	M.A. Driessen, J.J. Arts, G.J. v. Houtum, W.D. Rustenburg, B. Huisman
324	2010	Near-optimal heuristics to set base stock levels in a two-echelon distribution network	R.J.I. Basten, G.J. van Houtum
323	2010	Inventory reduction in spare part networks by selective throughput time reduction	M.C. van der Heijden, E.M. Alvarez, J.M.J. Schutten
322	2010	The selective use of emergency shipments for service-contract differentiation	E.M. Alvarez, M.C. van der Heijden, W.H. Zijm
321	2010	Heuristics for Multi-Item Two-Echelon Spare Parts Inventory Control Problem with Batch Ordering in the Central Warehouse	B. Walrave, K. v. Oorschot, A.G.L. Romme
320	2010	Preventing or escaping the suppression mechanism: intervention conditions	Nico Dellaert, Jully Jeunet.

319	2010	Hospital admission planning to optimize major resources utilization under uncertainty	
318	2010	Minimal Protocol Adaptors for Interacting Services	R. Seguel, R. Eshuis, P. Grefen.
317	2010	Teaching Retail Operations in Business and Engineering Schools	Tom Van Woensel, Marshall L. Fisher, Jan C. Fransoo.
316	2010	Design for Availability: Creating Value for Manufacturers and Customers	Lydie P.M. Smets, Geert-Jan van Houtum, Fred Langerak.
315	2010	Transforming Process Models: executable rewrite rules versus a formalized Java program	Pieter van Gorp, Rik Eshuis.
314	2010	Ambidexterity and getting trapped in the suppression of exploration: a simulation model	Bob Walrave, Kim E. van Oorschot, A. Georges L. Romme
313	2010	A Dynamic Programming Approach to Multi-Objective Time-Dependent Capacitated Single Vehicle Routing Problems with Time Windows	S. Dabia, T. van Woensel, A.G. de Kok
312	2010	Tales of a So(u)rcerer: Optimal Sourcing Decisions Under Alternative Capacitated Suppliers and General Cost Structures	Osman Alp, Tarkan Tan
311	2010	In-store replenishment procedures for perishable inventory in a retail environment with handling costs and storage constraints	R.A.C.M. Broekmeulen, C.H.M. Bakx
310	2010	The state of the art of innovation-driven business models in the financial services industry	E. Lüftenegger, S. Angelov, E. van der Linden, P. Grefen
309	2010	Design of Complex Architectures Using a Three Dimension Approach: the CrossWork Case	R. Seguel, P. Grefen, R. Eshuis
308	2010	Effect of carbon emission regulations on transport mode selection in supply chains	K.M.R. Hoen, T. Tan, J.C. Fransoo, G.J. van Houtum
307	2010	Interaction between intelligent agent strategies for real-time transportation planning	Martijn Mes, Matthieu van der Heijden, Peter Schuur
306	2010	Internal Slackening Scoring Methods	Marco Slikker, Peter Borm, René van den Brink
305	2010	Vehicle Routing with Traffic Congestion and Drivers' Driving and Working Rules	A.L. Kok, E.W. Hans, J.M.J. Schutten, W.H.M. Zijm
304	2010	Practical extensions to the level of repair analysis	R.J.I. Basten, M.C. van der Heijden, J.M.J. Schutten
303	2010	Ocean Container Transport: An Underestimated and Critical Link in Global Supply Chain Performance	Jan C. Fransoo, Chung-Yee Lee
302	2010	Capacity reservation and utilization for a manufacturer with uncertain capacity and demand	Y. Boulaksil; J.C. Fransoo; T. Tan
300	2009	Spare parts inventory pooling games	F.J.P. Karsten; M. Slikker; G.J. van Houtum
299	2009	Capacity flexibility allocation in an outsourced supply chain with reservation	Y. Boulaksil, M. Grunow, J.C. Fransoo

298	2010	An optimal approach for the joint problem of level of repair analysis and spare parts stocking	R.J.I. Basten, M.C. van der Heijden, J.M.J. Schutten
297	2009	Responding to the Lehman Wave: Sales Forecasting and Supply Management during the Credit Crisis	Robert Peels, Maximiliano Udenio, Jan C. Fransoo, Marcel Wolfs, Tom Hendrikx
296	2009	An exact approach for relating recovering surgical patient workload to the master surgical schedule	Peter T. Vanberkel, Richard J. Boucherie, Erwin W. Hans, Johann L. Hurink, Wineke A.M. van Lent, Wim H. van Harten
295	2009	An iterative method for the simultaneous optimization of repair decisions and spare parts stocks	R.J.I. Basten, M.C. van der Heijden, J.M.J. Schutten
294	2009	Fujaba hits the Wall(-e)	Pieter van Gorp, Ruben Jubeh, Bernhard Grusie, Anne Keller
293	2009	Implementation of a Healthcare Process in Four Different Workflow Systems	R.S. Mans, W.M.P. van der Aalst, N.C. Russell, P.J.M. Bakker
292	2009	Business Process Model Repositories - Framework and Survey	Zhiqiang Yan, Remco Dijkman, Paul Grefen
291	2009	Efficient Optimization of the Dual-Index Policy Using Markov Chains	Joachim Arts, Marcel van Vuuren, Gudrun Kiesmuller
290	2009	Hierarchical Knowledge-Gradient for Sequential Sampling	Martijn R.K. Mes; Warren B. Powell; Peter I. Frazier
289	2009	Analyzing combined vehicle routing and break scheduling from a distributed decision making perspective	C.M. Meyer; A.L. Kok; H. Kopfer; J.M.J. Schutten
288	2009	Anticipation of lead time performance in Supply Chain Operations Planning	Michiel Jansen; Ton G. de Kok; Jan C. Fransoo
287	2009	Inventory Models with Lateral Transshipments: A Review	Colin Paterson; Gudrun Kiesmuller; Ruud Teunter; Kevin Glazebrook
286	2009	Efficiency evaluation for pooling resources in health care	P.T. Vanberkel; R.J. Boucherie; E.W. Hans; J.L. Hurink; N. Litvak
285	2009	A Survey of Health Care Models that Encompass Multiple Departments	P.T. Vanberkel; R.J. Boucherie; E.W. Hans; J.L. Hurink; N. Litvak
284	2009	Supporting Process Control in Business Collaborations	S. Angelov; K. Vidyasankar; J. Vonk; P. Grefen
283	2009	Inventory Control with Partial Batch Ordering	O. Alp; W.T. Huh; T. Tan
282	2009	Translating Safe Petri Nets to Statecharts in a Structure-Preserving Way	R. Eshuis
281	2009	The link between product data model and process model	J.J.C.L. Vogelaar; H.A. Reijers
280	2009	Inventory planning for spare parts networks with delivery time requirements	I.C. Reijnen; T. Tan; G.J. van Houtum
279	2009	Co-Evolution of Demand and Supply under Competition	B. Vermeulen; A.G. de Kok
			B. Vermeulen, A.G. de Kok

278	2010	Toward Meso-level Product-Market Network Indices for Strategic Product Selection and (Re)Design Guidelines over the Product Life-Cycle	R. Seguel, R. Eshuis, P. Grefen
277	2009	An Efficient Method to Construct Minimal Protocol Adaptors	
276	2009	Coordinating Supply Chains: a Bilevel Programming Approach	Ton G. de Kok, Gabriella Muratore
275	2009	Inventory redistribution for fashion products under demand parameter update	G.P. Kiesmuller, S. Minner
274	2009	Comparing Markov chains: Combining aggregation and precedence relations applied to sets of states	A. Busic, I.M.H. Vliegen, A. Scheller-Wolf
273	2009	Separate tools or tool kits: an exploratory study of engineers' preferences	I.M.H. Vliegen, P.A.M. Kleingeld, G.J. van Houtum
272	2009	An Exact Solution Procedure for Multi-Item Two-Echelon Spare Parts Inventory Control Problem with Batch Ordering	Engin Topan, Z. Pelin Bayindir, Tarkan Tan
271	2009	Distributed Decision Making in Combined Vehicle Routing and Break Scheduling	C.M. Meyer, H. Kopfer, A.L. Kok, M. Schutten
270	2009	Dynamic Programming Algorithm for the Vehicle Routing Problem with Time Windows and EC Social Legislation	A.L. Kok, C.M. Meyer, H. Kopfer, J.M.J. Schutten
269	2009	Similarity of Business Process Models: Metics and Evaluation	Remco Dijkman, Marlon Dumas, Boudewijn van Dongen, Reina Kaarik, Jan Mendling
267	2009	Vehicle routing under time-dependent travel times: the impact of congestion avoidance	A.L. Kok, E.W. Hans, J.M.J. Schutten
266	2009	Restricted dynamic programming: a flexible framework for solving realistic VRPs	J. Gromicho; J.J. van Hoorn; A.L. Kok; J.M.J. Schutten;

Working Papers published before 2009 see: http://beta.ieis.tue.nl