Setting planned lead times for a make-to-order production system with master schedule smoothing

Teo, Chee-Chong; Bhatnagar, Rohit; Graves, Stephen C.

2011

Teo, C.-C., Bhatnagar, R., & Graves, S. C. (2011). Setting planned lead times for a make-to-order production system with master schedule smoothing. IIE Transactions, 43(6), 399-414. doi:10.1080/0740817X.2010.523765

https://hdl.handle.net/10356/89757

https://doi.org/10.1080/0740817X.2010.523765

Setting Planned Lead-times for a Make-To-Order Production System with Master Schedule Smoothing

Chee-Chong Teo[1,2], Rohit Bhatnagar[1,3], Stephen C. Graves[1,4]

[1]Singapore-MIT Alliance, Nanyang Technological University, Singapore

[2]Corresponding author
School of Civil & Environmental Engineering
Nanyang Technological University
Nanyang Avenue, Singapore 639798
Phone: +65 6790 4887 Fax: +65 6791 0676
Email: teocc@ntu.edu.sg

[3]Nanyang Business School
Nanyang Technological University
Nanyang Avenue, Singapore 639798
Email: arbhatnagar@ntu.edu.sg

[4]Sloan School of Management
Massachusetts Institute of Technology, MA 02139-4307
Email: sgraves@mit.edu

Abstract

We consider a make-to-order manufacturing environment with fixed guaranteed delivery lead-times and multiple product families, each with a stochastic demand process. The primary challenge in this environment is how to meet the quoted delivery times subject to fluctuating workload and capacity limits. We consider the tactical planning parameters, namely the planning windows and planned lead-times. We model the planning process in which the demand represents a dynamic input into the master production schedule. A planning window for each product family controls how the schedule of each product family is translated into job release. The planning window is the slack that exists when the fixed quoted delivery lead-time is longer than the total planned production lead-time. Further, the planned lead-time of each station regulates the workflow within a multi-station shop. The model has underlying discrete time periods to allow the modeling of the planning process that is typically defined in time buckets; within each time period, we model the intra-period workflow that permits multiple job movements within the time period. Our model characterizes key performance measures for the shop as functions of the planning windows and planned lead-times. We formulate an optimization model to determine the values of these planning parameters that minimize the relevant production-related costs.

Supplementary materials are available for this article. Go to the publisher's online edition of *IIE Transactions* for additional results of the simulation study in Section 6 of this manuscript.

Keywords: Make-To-Order; Production Smoothing; Master Production Schedule; Material Requirements Planning; Planned Lead-time; Planning Window.

1. Introduction

Managing customer lead-times in the face of demand uncertainty is a critical challenge faced by make-to-order (MTO) firms that, unlike make-to-stock firms, cannot use finished goods inventory to buffer against demand variations. Some firms continually adjust the delivery times quoted to customers depending on capacity load; others quote fixed (or a tight range of) lead-times to their customers, while varying the production rate in response to demand fluctuations. The fixed quoted delivery time represents a lead-time guarantee to customers that the manufacturer strives to meet with high likelihood. Rao et al. (2005) highlight that the fixed lead-time strategy enables the firm "to set clear expectations to customers, and thereby improves customer satisfaction". For example, Toyota's mass customized Camry Solaras cars have a five-day delivery lead-time after they are ordered by dealers (Robinson 1999). Weng (1999) reports an industrial application in a make-to-order lighting fixture manufacturer wherein the use of a 10-day fixed lead-time policy significantly improved the on-time performance of deliveries. The fixed lead-time strategy has been utilized in industries where lead-times are largely dictated by the market and when predictability is valued by customers, e.g. customized printed circuit boards, for which customers need to coordinate the concurrent availability of components to build an assembly.

If the MTO manufacturer adopts a fixed lead-time strategy, it has to first set the fixed lead-time, which involves complex tradeoffs encompassing decisions at the strategic and tactical level. The manufacturer must consider its market position in terms of its pricing, competitors' lead-times, market requirements as well as its own production capability in meeting the lead-time. The setting of the fixed quoted delivery lead-time is a cross-functional decision, involving marketing, sales and production departments. After determining the fixed lead-time, the manufacturer must make tactical and operational decisions on how best to meet this guaranteed fixed lead-time. The key tactical and operational planning decisions include how the firm should utilize its production resources to meet the fixed quoted delivery lead-time with high probability. In this paper we assume that the firm has adopted a fixed lead-time strategy and has set the delivery lead-time, and we focus herein on the tactical production decisions needed to support such a strategy. To meet the fixed lead-time, the MTO manufacturer must be able to flex its production capability. For instance, the manufacturer might utilize overtime or subcontracting in periods of high demand; alternatively, it might plan to operate at a lower capacity utilization level so as to have regular-time capacity available to meet high demand. In either case there is a cost for this flexibility, due to under-utilized capacity or premium production costs from overtime and subcontracting.

Typically, the two core functions at the tactical level needed to support a fixed lead-time policy are the material requirements planning (MRP) and master production scheduling (MPS). We consider the two sets of planning parameters that have the most influence on regulating the flow of work in such an environment, namely the planned lead-times in the MRP and the planning window in the MPS. The planned lead-time (PLT) is a key input for MRP and is used to project how long each job will spend in each production step or workstation. There is a tradeoff involved in setting the PLT; a short PLT leads to a more variable workload at the workstation, while a long PLT results in more work-in-process (WIP) at the station.

The planning window is the difference between the quoted delivery lead-time and the total planned production lead-time for a job. We assume the MPS is smoothed over the planning window (if a slack exists) while preserving the integrity of the fixed quoted delivery lead-time. There is prior evidence that regulating the release of work to a shop provides benefits. A longer planning window allows greater smoothing of the MPS, which leads to a less variable job release into the shop and subsequently fewer occurrences of "spikes" in capacity loading. The idea is to avoid releasing surges of customer orders that would otherwise cause capacity saturation with no significant increase in throughput. In addition, Vollmann et al. (2005, p. 321) highlight the growing importance of production smoothing in non-repetitive and make-to-order production, as firms increasingly need to respond to customer pressure for greater flexibility in volume and product mix. It is also well documented in *just-in-time* literature that a stable MPS facilitates process and quality improvements. Without a stable production environment, resources tend to be channeled to react to changes in production output instead of achieving process improvements. Further, a smoothed MPS leads to more stable requirements for raw materials or components, which results in lower component safety stock and/or a more stable delivery schedules for suppliers.

The key challenge in developing the model is how to incorporate both the planning system decisions and the workflow control decisions for a production network. On the one hand, most MPS and MRP systems use time buckets of equal intervals (e.g., a day, a week or a work shift) as the underlying time unit for the key manufacturing decisions; the planned production output, inventory levels and job releases are specified for each time bucket in the planning horizon. On the other hand, job flow occurs at much shorter time intervals (essentially in continuous time) and it is important to capture the dynamics of these intra-period job movements. The problem is further complicated by the presence of multiple products in the context of MTO manufacturing, where we need to capture the planning decisions and job routings of each individual product.

To connect the discrete-time planning system and the intra-period job flows, we extend the Tactical Planning Model (TPM) of Graves (1986). The TPM is a discrete-time model in which all transitions within the model are governed by an underlying time period. The model assumes that all job movements from one station to another occur *only* at the start of each time period, which corresponds to the planning time bucket in the context of this paper. This assumption is restrictive in many situations where jobs can make multiple moves between stations within a single time bucket. We address this limitation and extend the model by allowing work to flow between workstations throughout each period, and not only at the start of the period. To our knowledge, work on intra-period job movement in multi-station systems has not been previously reported in literature. Due to the structure of the TPM-based workflow model, the workflow of each product family can be modeled separately. Hence the TPM extension is able to capture the key planning and workflow attributes of each product family, including the demand process, release planning, job routing, planning windows and planned lead-time at each station.

We model how the stochastic demand process is translated into the MPS, where its smoothness is controlled with the length of the planning window. We then model the workflow in a multi-station production network, in which the MPS determines the job release and the PLTs control the production at each station. In this fashion, we model the linkage between the MPS planning process and the workflow in the production network. Our model determines the variance of production requirements (a measure of capacity requirements) and the expected WIP level at each workstation. This allows us to characterize the tradeoff between the capacity requirements and WIP inventory as it depends on the planning windows and the PLTs. We embed the model in an optimization model to determine the PLTs and planning windows.

In the next section, we review related work. In Section 3, we describe the problem setting, and explain the roles of the planning window and PLT in production smoothing. In Section 4, we develop the model that permits multiple job movements in each discrete-time bucket, and subsequently develop a workflow model for a single product family. We then discuss how we can extend the model to characterize the production of multiple product families in Section 5. In Section 6, we discuss the results of a simulation study to validate the model. In Section 7, we formulate an optimization model to set the planning windows and PLTs. Section 8 presents an example that illustrates how the model can be used to support tactical planning of a MTO production system. We conclude in Section 9 with a discussion of future research opportunities.

2. Related Work

Much of the literature on managing customer lead-times in MTO manufacturing is on due date management. Earlier work encompasses setting of due dates and sequencing of orders (see surveys in Cheng and Gupta 1980, and Baker 1984). The tradeoff in this problem is that tight due dates reflect better responsiveness to customer demand but may be difficult or costly to achieve due to conflicts with sequencing objectives. Subsequent work includes the additional consideration of order acceptance decisions, i.e., the effect of quoted due date towards customer orders. More recent work considers the additional dimension of how pricing affects customer demand, often together with order acceptance decisions. The reader is referred to the review by Keskinocak and Tayur (2004) for details of the various aspects of the due date management policies.

The literature on due date setting focuses on a single-stage aggregated system, for which the firm has some flexibility to vary the quoted delivery lead-time. Our research differs in that we consider the production planning aspects of how best to satisfy a fixed quoted delivery lead-time for a multi-stage, multi-product system. The due date setting models support real-time operational decisions for the variable lead-time strategy. In contrast, our model is not intended for real-time decisions. Rather, we focus on the production-related tactics of setting the internal lead-times in order to support the fixed lead-time strategy.

There is a limited amount of research that deals with MTO manufacturers who quote a fixed delivery lead-time. Cruickshanks et al. (1984) consider production smoothing in a single-stage manufacturing process and introduce the concept of the planning window. We also utilize the concept of the planning window to smooth the MPS; however, in addition, we model production smoothing over a multi-stage production network. So and Song (1998) consider a single product, single stage queueing system with a fixed quoted delivery lead-time. They characterize the lead-time, price and capacity expansion decisions with the objective to maximize profits subject to a service constraint. Rao et al. (2005) also consider a single stage problem with a fixed quoted delivery lead-time with similar decision variables (except that it excludes capacity decisions). Rather than a service constraint, their model assumes a completely reliable delivery lead-time achieved by outsourcing in times of capacity shortfall. Our model is similar to theirs in that we also assume the delivery lead-time is always met through subcontracting or overtime, if necessary. Both So and Song (1998) and Rao et al. (2005) focus on the strategic decision of determining the fixed delivery lead-time to maximize profits for a single-stage system that produces an aggregate product. In contrast, our model addresses the tactical decisions associated with master production scheduling, and setting the release process and workflow control policies in a multi-station, multi-product MTO environment.

Another related research area encompasses analytical models that determine the optimal PLTs. Weeks (1978), Kanet (1986), Matsuura and Tsubone (1993) and Matsuura et al. (1996) develop single-stage models while Yano (1987) and Gong et al. (1994) formulate serial multi-echelon models. Our paper differs in that we model a general network of workstations with no restriction on process routes by focusing on tactical planning rather than detailed scheduling, from which we achieve much tractability. The prior work assumes an exogenous job arrival process but we determine work releases that are generated from the demand process, and explicitly model how the discrete-time planning process affects workflow and behavior of individual workstations. For a job shop, Graves (1986) presents the Tactical Planning Model (TPM) that highlights the tradeoff between production smoothness and WIP in selecting the PLT at each workstation. Other work that is related to the TPM includes Graves (1988a, 1988b, 1988c), Fine and Graves (1989), Graves et al. (1998), Graves and Hollywood (2001), Hollywood (2005), Teo (2006) and Teo et al. (2009). In this paper, we extend the TPM in a number of ways. First, we explicitly model the release process for a multi-product MTO environment. Second, we model the dynamics of job flows within each time bucket, allowing us to characterize how the planning process affects workflow. Third, we use the model to formulate an optimization model to support the setting of planning windows and PLTs.

Our extension of the TPM is an effort to address the gap between discrete-time planning models and continuous-time workflow models; Karmarkar (1993) states, "*There is a need for an approach that will work at the planning level, so that facility loading under seasonal demand can be analyzed. The technical problem here is that planning models tend to be in discrete time while queuing and scheduling models are in continuous time*." Planning models involve dividing the planning horizon into discrete time periods that correspond to the time buckets; the models determine the optimal production quantities and capacity plans in each bucket that satisfy aggregate resource constraints. However, to characterize the specific shop workflow and capacity loading requires the capturing of the intra-period processing and movement of jobs.

Work on clearing functions (see e.g., Karmarkar 1989, Asmundsson et al. 2006 and Selcuk et al. 2008) is similar in intent to our TPM extension, since this work also attempts to connect discrete-time planning with continuous-time shop behavior. A nonlinear clearing function approximates the queue congestion due to capacity loading by specifying the amount of work "cleared" in a given time period by a resource as a nonlinear function of WIP; the function is embedded in optimization models for aggregate production planning. Our model also assumes a WIP dependent production output (as will be seen) but differs in that the production output in each time period is to preserve the integrity of planned lead-times.

Further, our intent is more to determine the planning tactics (i.e., the PLTs and planning windows) rather than detailed production plans.

3. Problem Setting

We begin with a summary of notations in Table 1.

Table 1. Notations used in the paper.

t :	index for time period (i.e., planning time bucket)
i :	index for workstations, $i = 1, 2, \ldots, m$. ($i = 0$ represents the dummy station)
k :	index for product family
s_i :	index for sub-period at workstation i, $s_i = 1, 2, \ldots, p_i$
Δ_i :	length of each sub-period s_i at workstation i ($\Delta_i = 1/p_i$)
DLT_k :	fixed quoted delivery lead-time for product family k
$PPLT_k$:	product planned lead-time for product family k
W_k :	length of planning window for product family k
c_i :	penalty cost per hour of capacity shortfall at workstation i
h_{ik} :	unit WIP inventory holding cost at workstation i for product family k
n_{ik} :	station planned lead-time at workstation i for product family k
ω_{ik} :	number of times each job from product family k visits workstation i
ϕ_{ij} :	average amount of work input to station i generated by a unit of output at station j
P_{it} :	production requirements (in workhours) at workstation i in period t
M_i :	nominal capacity at workstation i (in workhours) per period t
Q_{ikt} :	queue length (in workhours) for product family k at workstation i at start of period t
A_{it} :	amount of workload that arrives at workstation i in period t
D_{kt} :	demand for product family k in period t
ξ_{it} :	zero-mean noise term at workstation i in period t
$Y(\Delta_i, s_i)$:	production requirement (in workhours) at workstation i in sub-period s_i of length Δ_i
$X(\Delta_i, s_i)$:	queue length (in workhours) at start of sub-period s_i of length Δ_i
$\boldsymbol{P_t}$:	column vector $\{P_{0t}, P_{1t}, \ldots, P_{mt}\}'$
$\boldsymbol{Q_t}$:	column vector $\{Q_{0t}, Q_{1t}, \ldots, Q_{mt}\}'$
$\boldsymbol{A_t}$:	column vector $\{A_{0t}, A_{1t}, \ldots, A_{mt}\}'$
$\boldsymbol{\Phi}$:	square matrix with elements ϕ_{ij}
$\boldsymbol{\zeta_t}$:	column vector $\{D_{t-1}, \xi_{1t,\ldots\ldots,} \xi_{mt}\}'$
$\boldsymbol{\mu}$:	expectation of vector $\boldsymbol{\zeta_t}$

We consider a MTO production system that produces a variety of products. We assume that end items can be grouped into product families that consist of products with similar designs and production processes. We define the MPS at the product family level (see Vollmann et al. 2005 for defining a group of end items as the MPS unit). Each product family k has a fixed quoted delivery lead-time, denoted as DLT_k; for instance, if a product family has a fixed DLT_k of (say) one week, then the manufacturer would commit to deliver each order exactly one week after the receipt of the order. Each workstation is assigned a *station planned lead-time* for each product family, denoted as n_{ik}. It is the *intended* amount of time required for each job from product family k to complete its processing at the workstation, including both processing and waiting time. We associate a *product planned lead-time* $PPLT_k$ with each product family, which is the total planned time required for an item of product family k to be completed by the shop. Each product family has predetermined distinct routings that defines the sequence of workstations that a product must visit for processing; we permit re-entry in which a routing can re-visit any workstation. If the product family k has a single routing, $PPLT_k$ is the planned time required for a job from product family k to move through the route, which we express as:

$$PPLT_k = \sum_i \omega_{ik} n_{ik} \tag{1}$$

where ω_{ik} is the number of times each job from product family k visits workstation i. A product family may have multiple routings, e.g., if the end item requires assembly of multiple parts. In this case, each routing indexed by $r(k)$ has to satisfy (1) and in addition, $\omega_{i,r(k)}$ replaces ω_{ik}, which denotes the number of times each job in routing $r(k)$ visits workstation i.

3.1 Master Production Scheduling

The MPS can be smoothed over the planning window. The relationship between DLT_k, $PPLT_k$ and W_k is illustrated in Figure 1.

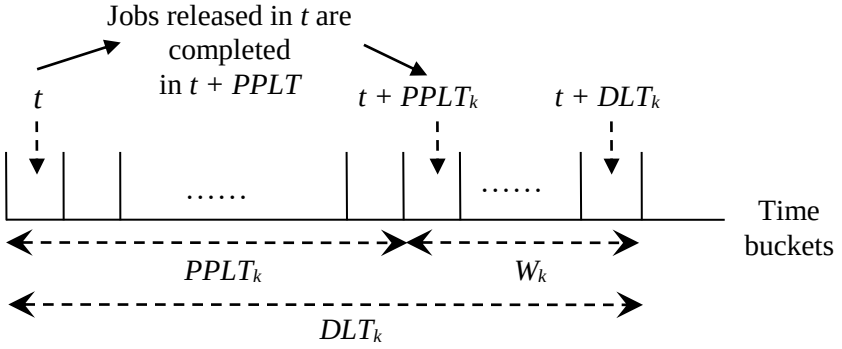

Figure 1. Relationship between DLT_k, $PPLT_k$ and W_k

Jobs to be completed in period $t + PPLT_k$ need to be released in period t. In any period t, to determine the release quantity, the production level to be completed in $t + PPLT_k$ is set with knowledge of all contracted orders for delivery over the planning window encompassing period $t + PPLT_k$ to $t + DLT_k$. We express the length of the planning window as:

$$W_k = DLT_k - PPLT_k + 1. \quad (2)$$

We assume the production over the planning window is leveled, where a larger W_k allows a greater extent of smoothing. We illustrate the MPS smoothing with an example in the *Appendix*. We note that if $DLT_k = PPLT_k$, then $W_k = 1$: this corresponds to no smoothing as the releases in each period t must correspond exactly to the orders due in time $t + PPLT_k$.

The smoothing of the MPS results in early production (and release) of some orders. We do not explicitly consider the inventory of finished products that result from the orders that are built ahead of time. There is very little risk in holding this inventory, as it has been made to order. Furthermore, orders that are built in advance can be delivered to customers who accept early deliveries. Alternatively, the built ahead inventory might be shipped by a more economical mode given the additional time for shipping before the due date. In both cases, there is an opportunity to provide additional value to the customer. Therefore we argue that the finished goods inventory level is not a crucial planning consideration in this context.

3.2 Production at Workstations

The workload at each workstation varies over time due to both varying job arrivals and inherent production variability. Sources of the inherent production variability include dissimilar processing requirements of jobs, machine breakdowns, setups and yields. In face of the workload variability, a larger n_{ik} allows smoother production (as will be seen) and a higher likelihood for the workstation to meet the station planned lead-time. However, with a larger n_{ik}, jobs stay longer at the workstation, which leads to a higher WIP inventory.

We let P_{it} be the production requirement in workhours at workstation i in time period t. We define production requirement P_{it} as the production output *required* at workstation i in period t in order to ensure that the work-in-queue satisties the station planned lead-times. The motivation for this specification is to model current planning practice. In MRP systems, the MPS for end items, after offsetting for the station planned lead-times, is translated into production requirements at each workstation. To achieve coordination of production with scheduled receipts of both produced and procured components, the station

planned lead-time is used to determine job start times at each workstation using the well known backward scheduling logic of MRP.

We assume the production requirements at each workstation are controlled by its station planned lead-times. Specifically, the production schedule is regulated by the job start times whereby a job is not required to be produced earlier even if capacity is available. Effectively, each workstation produces a quantity sufficient to fulfill the station planned lead-time of each product family in each period. We do not consider the procured components in our model and assume the scheduled receipts of purchased components are always reliable. Finally, we like to stress that our intent is to capture the tactical level issues, i.e., how best to set the internal lead-times to meet the fixed quoted delivery lead-time in normal operating circumstances. Thus we do not take into account the actual detailed operational decisions that might deviate from the tactics deployed in the model, e.g., expediting of a particular order from an important customer.

3.3 Problems in Production Planning

There are several planning decisions to make in the MTO environment:

- For each product family, given the fixed DLT_k and the constraint $PPLT_k \leq DLT_k$, how does one assign n_{ik} to each workstation? This decision must account for the tradeoff between capacity requirements in the face of variable workload and WIP level at each workstation.
- The planner must consider the tradeoff between the smoothing of the MPS and the time allowed at the workstations. Increasing the $PPLT_k$ means longer n_{ik} for job completion at the workstations; however, a longer $PPLT_k$ results in a shorter planning window, yielding a more variable job release pattern over time and thus greater variability in job arrivals to the workstations. A larger $PPLT_k$ also leads to longer time spent at the workstations, and thus higher WIP levels.
- The planner must consider the interdependence of workflow between the workstations. In addition, each station can process jobs from multiple product families and the production planner must take into account the interactions between the various product families.

Our model helps the planner in setting n_{ik} for each workstation, which determines the $PPLT_k$ and W_k. In considering the tradeoff between capacity requirements and WIP levels, we must consider the sum of expected total penalty cost due to capacity shortfall and expected WIP inventory holding cost.

We let c_i be the penalty cost per workhour of capacity shortfall for workstation i; h_{ik} be the unit WIP inventory holding cost (per workhour per period) at workstation i for product family k; M_i denotes

the nominal capacity for workstation i in workhours per period; Q_{ikt} be the queue length in workhours for product family k at station i at start of t. We express the total cost as:

$$\sum_i \left[c_i E\left[P_{it} - M_i\right]^+ + \sum_k h_{ik} E\left[Q_{ikt}\right] \right]. \tag{3}$$

where $x^+ = \max(x, 0)$. The term $c_i E[P_{it} - M_i]^+$ denotes the expected penalty cost incurred due to capacity shortfall, i.e., if $P_{it} > M_i$. A higher variability of P_{it} leads to a higher expected penalty cost and its variability is influenced by $PPLT_k$, n_{ik} and W_k as described earlier.

We interpret the expected penalty cost as follows. We assume that each workstation can always output the production requirement; that is, there are no rigid capacity constraints but a nominal capacity limit M_i. When P_{it} exceeds M_i, we assume that the workstation is able to accomplish the production requirement but that it incurs an additional cost c_i per hour for all production in excess of the nominal capacity. For instance, we might assume that in times of high workload, a workstation takes expediting actions, e.g., overtime and subcontracting, to produce the jobs within the station planned lead-time. The workstation would then incur an expediting or subcontracting cost $c_i(P_{it} - M_i)$ for its production that is beyond its nominal capacity. (Refer to Holt et al. 1960, which is one of the works that establishes the similar approach of using convex cost functions in modeling production systems.)

The key to this research is to characterize P_{it} to enable us to compute the expected penalty cost. In essence, we determine the first two moments of P_{it} wherein the variance of P_{it} represents a measure of variability of the production requirements. By further assuming P_{it} is normally distributed, we are able to compute the capacity costs. In addition, we determine the expected WIP level $E[Q_{it}]$. This allows us to develop an expression for (3) that we use as an objective function in an optimization model to support the setting of W_k and n_{ik}.

4. Single Family Model

In this section, we develop the characterization of P_{it} for a single aggregate product family. We first extend the TPM of Graves (1986) to allow multiple work movements within a time period to characterize the intra-period transfer of production requirements. We then present how we convert the stochastic demand for a single family into a smoother MPS and release. Next we model the workflow in a general network of workstations. In Section 5, we build upon this single family model to incorporate multiple families.

4.1 Multiple Job Movements in Discrete Time Bucket

Since we are modeling a single product family, we drop the subscript k for notational convenience. We start with the discrete time period t with length of one time unit, corresponding to that for the planning time bucket (e.g., a day or a week). We sub-divide each period t into p_i equal subintervals, and use index s_i, where $s_i = 1, 2, \ldots, p_i$, to denote the sub-periods at workstation i. We define the length of each sub-period as $\Delta_i = 1/p_i$.

We assume that work arrives at the start of each sub-period and the workflow is measured in units of work content at a station (e.g. hours) rather than jobs. We assume that the production requirement at workstation i in sub-period s_i of length Δ_i, denoted by $Y(\Delta_i, s_i)$, is a linear function of its queue length at the start of sub-period s_i, denoted by $X(\Delta_i, s_i)$; this is given by:

$$Y(\Delta_i, s_i) = (\Delta_i / n_i) X(\Delta_i, s_i) \qquad \text{for } s_i = 1, 2, \ldots, p_i, \tag{4}$$

where $n_i \geq \Delta_i$. Equation (4) is similar to the linear control rule of the TPM in Graves (1986) but differs in that (4) is expressed as a function of the size of the time grid; as such, we allow arrivals to the workstation throughout each period.

We interpret (4) as follows: in order to realize the station planned lead-time, the workstation i must not allow work to wait for more than n_i periods and thus we approximate that it must process close to (Δ_i / n_i) of the work-in-queue in each sub-period. We note that the production requirements vary according to the work-in-queue to assure that the actual lead-time for nearly all of jobs in queue matches the station planned lead-time. Further, when a relatively high workload arrives to join the queue, the workload would be "spread" over the station planned lead-time. Thus a larger n_{ik} permits a relatively smoother production requirement as it permits more workload leveling. (Note that the production-workload relationship in (4) aligns with empirical findings in labor psychology that show that productivity, and thus production rate, in some human operated production systems increases with workload because of the effect of work pressure on human performance; see Bertrand and Van Ooijen 2002 for a review.)

We further assume that the arrival of work to the workstation in period t, which we denote as A_{it}, occurs uniformly over period t. In particular, we assume that the arrival amount at the start of each sub-period is equal to A_{it} / p_i. The validity of this assumption depends upon the characteristics of the arrival stream as well as the length of the time period. If the arrival stream is highly varying and consists of only a few arrivals in each period, then this assumption may be less reasonable. Similarly, this assumption will be more valid if the time period is long compared to the inter-arrival times, as this will result in a more even spread of the workload within each period. Nevertheless, although we recognize that the variability

of inter-arrival times affects the expected queuing time, we note that the purpose of our model is for tactical planning; therefore we are more concerned with the aggregate random arrivals $\{A_{it}\}$ in each time bucket, rather than the detailed random variations within the time period.

We define Q_{it} as the queue length at the start of period t prior to any arrival. We have the boundary condition for the queue length at the first sub-period within each time period:

$$X(\Delta_i, s_i = 1) = Q_{it} + A_{it}/p_i. \tag{5}$$

For $s_i = 2, \ldots, p_i$, we model the queue length at the start of sub-period s_i by the balance equation:

$$X(\Delta_i, s_i) = X(\Delta_i, s_i - 1) - Y(\Delta_i, s_i - 1) + A_{it}/p_i. \tag{6}$$

By substituting (4) into (6), we obtain for $s_i = 2, \ldots, p_i$:

$$X(\Delta_i, s_i) = (1 - (\Delta_i/n_i))\, X(\Delta_i, s_i - 1) + A_{it}/p_i. \tag{7}$$

We express P_{it} as the sum of the production requirements over the sub-periods:

$$P_{it} = \sum_{s_i=1}^{p_i} Y(\Delta_i, s_i) = (\Delta_i / n_i) \sum_{s_i=1}^{p_i} X(\Delta_i, s_i). \tag{8}$$

We use (5) and (7) to find

$$\sum_{s_i=1}^{p_i} X(\Delta_i, s_i) = Q_{it} + [1 - (\Delta_i / n_i)] \sum_{s_i=1}^{p_i - 1} X(\Delta_i, s_i) + A_{it},$$

from which we find:

$$(\Delta_i / n_i) \sum_{s_i=1}^{p_i} X(\Delta_i, s_i) = Q_{it} + A_{it} - [1 - (\Delta_i / n_i)] X(\Delta_i, p_i). \tag{9}$$

We now substitute (9) into (8) to get

$$P_{it} = Q_{it} + A_{it} - [1 - (\Delta_i / n_i)] X(\Delta_i, p_i). \tag{10}$$

From (10), in order to get an expression for P_{it}, we need to find $X(\Delta_i, p_i)$. From (7) and repeated substitution, we can write

$$X(\Delta_i, p_i) = [1 - (\Delta_i / n_i)]^{p_i - 1} \times Q_{it} + \left(1 + [1 - (\Delta_i / n_i)] + \ldots + [1 - (\Delta_i / n_i)]^{p_i - 1}\right) \times \frac{A_{it}}{p_i}. \tag{11}$$

We can use (11) to re-write (10) as

$$P_{it} = \beta_i(\Delta_i) Q_{it} + \gamma_i(\Delta_i) A_{it}, \tag{12}$$

where

$$\beta_i(\Delta_i) = 1 - [1 - (\Delta_i / n_i)]^{p_i}$$

and

$$\gamma_i(\Delta_i) = 1 - \left(\frac{1-(\Delta_i / n_i)}{(1/n_i)} \right) \left(1 - [1-(\Delta_i / n_i)]^{p_i}\right)$$

$$= 1 - \beta_i(\Delta_i) \left(\frac{1-(\Delta_i / n_i)}{(1/n_i)} \right).$$

Therefore the production in each time period is a linear function of both the work-in-queue at the start of the period and the arrivals during the period. The coefficients for the linear function depend on the size of the subintervals.

By letting the length of the sub-period go to zero, we obtain the continuous time limits for $\beta_i(\Delta_i)$ and $\gamma_i(\Delta_i)$. In effect, we assume that (4) holds at every instant in time, which corresponds to a fluid-like workflow. Since $\Delta_i = 1/p_i$ by definition, we obtain the continuous time limits of $\beta_i(\Delta_i)$ and $\gamma(\Delta_i)$ as:

$$\beta_i = 1 - e^{-1/n_i}$$

and

$$\gamma_i = 1 - n_i \beta_i .$$

We can now restate (12) for the continuous-time function as:

$$P_{it} = \beta_i Q_{it} + \gamma_i A_{it}, \tag{13}$$

where β_i and γ_i are given above. The validity of the fluid flow assumption yielding (13) depends on the frequency of job movements in each time bucket; a higher frequency would suggest a more fluid work-flow. However, in situations where job flow rates are low, we expect the sub-period coefficients $\beta_i(\Delta_i)$ and $\gamma_i(\Delta_i)$ to be more appropriate. In such cases, Δ is set to a value on the order of the average inter-arrival time. For notational consistency, we will use β_i and γ_i for the subsequent model development.

The balance equation for workstation *i* at the start of each time period is given by

$$Q_{it} = Q_{i,t-1} - P_{i,t-1} + A_{i,t-1}. \tag{14}$$

We substitute (13) into (14) and perform repeated substitution to obtain

$$P_{it} = \beta_i (1-\gamma_i) \sum_{r=0}^{\infty} (1-\beta_i)^r A_{i,t-1-r} + \gamma_i A_{it} . \tag{15}$$

If the arrival stream $\{A_{it}\}$ to the workstation were independent and identically distributed (i.i.d.) with variance σ^2, then we find the variance of production requirements in (15) is

$$Var(P_{it}) = \left(\frac{\beta_i}{2-\beta_i} (1-\gamma_i)^2 + \gamma_i^2 \right) \sigma^2 \quad . \tag{16}$$

(The arrival stream to a station from upstream stations is generally not i.i.d. but correlated over time. We consider correlated arrivals in Section 4.3.) In Figure 2 we compare the standard deviation of P_{it}, denoted

by $\sigma(P_{it})$, obtained using the coefficients β_i and γ_i with that computed with the sub-period coefficients $\beta_i(\Delta_i)$ and $\gamma_i(\Delta_i)$.

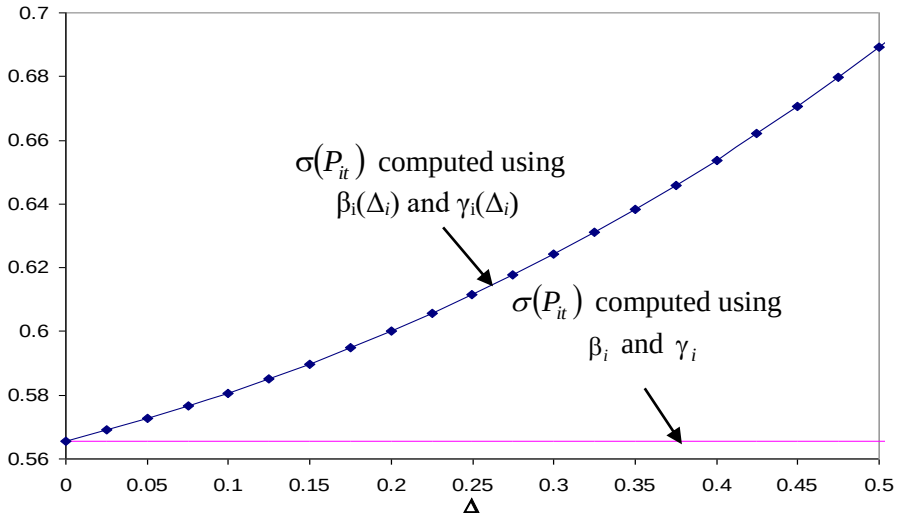

Figure 2. $\sigma(P_{it})$ computed for different sub-period lengths (with $\sigma = 1$)

We observe that the results are close over a reasonable range of flow rates; at $\Delta_t = 0.1$ and $\Delta_t = 0.05$ (i.e., corresponding to 10 and 20 arrivals per period, respectively), the percentage differences are 2.6%, and 1.2% respectively. Thus we expect (13) to be a reasonable approximation. In Section 6, we discuss the suitability of the coefficients for different job flow rates based on a simulation study.

We have expressed (13) in terms of the parameters β_i and γ_i, both of which are functions of n_i. To see how (13) behaves for different values of n_i, we graph $Var(P_{it})$, β_i and γ_i as functions of n_i in Figure 3. We observe that $Var(P_{it})$ decreases with larger values of n_i, i.e., longer station planned lead-time leads to a smoother production.

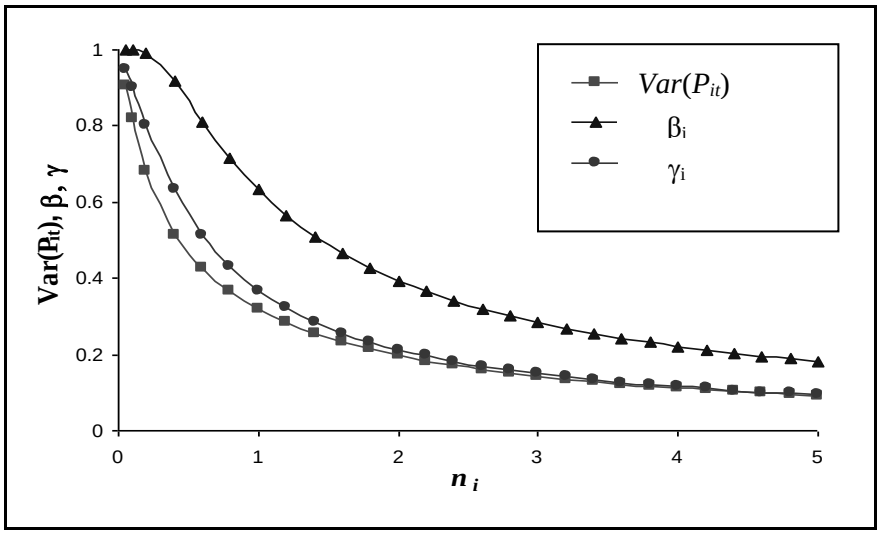

Figure 3: $Var(P_{it})$, β_i and γ_i as functions of n_i (with $\sigma = 1$)

4.2 Model for MPS Smoothing

We denote the demand in period t by D_t and assume that the demand process is i.i.d. over time with mean $E[D_t]$. As discussed in Section 3.1, the production planner smoothes the MPS by considering the orders not yet released into production as a demand queue, waiting to be released into the shop.

To model the work release, we introduce a dummy workstation that is the first station that a job "visits" before it is released to the shop. The queue at the dummy station represents the order backlog of work waiting to be released. In each period, the job arrival to the dummy station is the demand D_t and the production output of the dummy station corresponds to the job release in that period. Denoting station 0 as the dummy station, we model the MPS smoothing by

$$P_{0t} = \frac{Q_{0t}}{W}, \tag{17}$$

where $W \geq 1$. Equation (17) is equivalent to (4) and the linear control rule of the TPM. In (17), we assume that in order to ensure an order in the input queue does not wait for more than W periods, the shop attempts to release $1/W$ of the on-hand orders in each period. Further, (17) models the smoothing of the release by spreading the on-hand orders evenly over the planning window.

The material balance equation for the dummy station is

$$Q_{0t} = Q_{0,t-1} - P_{0,t-1} + D_{t-1}. \tag{18}$$

By substituting (17) into (18) to replace Q_{0t} in (18), we obtain

$$P_{0t} = (1 - 1/W)P_{0,t-1} + (1/W)D_{t-1},$$

namely a first-order exponential smoothing model with 1/*W* as the smoothing parameter. A larger value of *W* results in a smoother job release into the shop.

4.3 Workflow Model

We model the work arrivals to workstation *i* by

$$A_{it} = \sum_j \phi_{ij} P_{jt} \,, \tag{19}$$

where ϕ_{ij} is a positive scalar. We assume that each time unit of production requirement at workstation *j* generates ϕ_{ij} time units of input to station *i* on average. When station *j* represents the dummy station, we assume that each order unit triggers ϕ_{ij} time units of work, on average, for station *i*. We note that $\phi_{0j} = 0$ for all *j* and that ϕ_{j0} is the average amount of work that starts at workstation *j* for each new job.

To analyze the workflow model, we combine the production functions in (13) and (17), and express the result in matrix notation as:

$$\boldsymbol{P_t} = \boldsymbol{F}\boldsymbol{Q_t} + \boldsymbol{G}\boldsymbol{A_t} \tag{20}$$

where $\boldsymbol{P_t} = \{P_{0t}, P_{1t}, \ldots, P_{mt}\}'$, $\boldsymbol{Q_t} = \{Q_{0t}, Q_{1t}, \ldots, Q_{mt}\}'$ and $\boldsymbol{A_t} = \{A_{0t}, A_{1t}, \ldots, A_{mt}\}'$ are column vectors and *m* is the number of workstations. $\boldsymbol{F}$ is a diagonal matrix with first diagonal element equal to 1/*W* and the remaining diagonal elements set to β_i (i = 1, 2, …, m). $\boldsymbol{G}$ is a diagonal matrix with zero as the top element and γ_i (i = 1, 2, …, m) as the next *m* diagonal elements.

We express the work arrivals to the workstations given in (19) in matrix form as:

$$\boldsymbol{A_t} = \boldsymbol{\Phi}\boldsymbol{P_t} \,, \tag{21}$$

where $\boldsymbol{\Phi}$ is a square matrix with elements ϕ_{ij}. By substituting (21) into (20), we obtain

$$\boldsymbol{Q_t} = \boldsymbol{F}^{-1}(\boldsymbol{I} - \boldsymbol{G}\boldsymbol{\Phi})\boldsymbol{P_t} \,, \tag{22}$$

where $\boldsymbol{I}$ is an identity matrix. Note that $\boldsymbol{F}^{-1}$ exists since it is a diagonal matrix, with each diagonal element being positive. We combine the balance equations for the workstations in (14) and for the dummy stations in (18), and express them in matrix form as:

$$\boldsymbol{Q_t} = \boldsymbol{Q_{t-1}} - \boldsymbol{P_{t-1}} + \boldsymbol{A_{t-1}} + \boldsymbol{\zeta_t} \,. \tag{23}$$

The term $\boldsymbol{\zeta_t}$ is a column random vector that combines both demand arrivals to the dummy station and the noise arrivals to the workstations; it has D_{t-1} in the first row and $\{\xi_{1t,\ldots\ldots,}\xi_{mt}\}$ in the next *m* rows. The term ξ_{it} is a zero-mean noise term that represents any inherent production variability in the arrival

stream to station *i*. The noise signifies any deviation between actual and expected workload, based on the upstream production levels. We assume that the amount of noise is independent of the work arrivals. We can use ξ_{it} to model the sources of variability due to machine failures, setups, production yields or any dissimilarity in processing requirements within the product family. Here we adopt the concept of *effective processing time*. The effective processing time is the adjusted processing time that incorporates the variability of processing requirements as well as the time that is unavailable for processing due to random events. See Hopp and Spearman (2000) for a more detailed explanation.

Now by substituting (21) and (22) into (23), we have

$$\boldsymbol{P}_t = \boldsymbol{B}\boldsymbol{P}_{t-1} + \boldsymbol{H}\boldsymbol{\zeta}_t \,, \tag{24}$$

where $\boldsymbol{H} = (\boldsymbol{I} - \boldsymbol{G}\boldsymbol{\Phi})^{-1}\boldsymbol{F}$ and $\boldsymbol{B} = \boldsymbol{I} - \boldsymbol{H}(\boldsymbol{I} - \boldsymbol{\Phi})$. We can show that the matrix $(\boldsymbol{I} - \boldsymbol{G}\boldsymbol{\Phi})$ is invertible as long as the spectral radius of $\boldsymbol{\Phi}$ is less than one, which is a necessary condition for the system to reach steady state, as will be seen. Assuming an infinite history for the system, we obtain by repeated iteration:

$$\boldsymbol{P}_t = \sum_{s=0}^{\infty} \boldsymbol{B}^s \boldsymbol{H}\boldsymbol{\zeta}_{t-s} \,. \tag{25}$$

We denote the expectation of vector $\boldsymbol{\zeta}_t$ by the vector $\boldsymbol{\mu}$, a column vector with $E[D_t]$ in the first row as the only non-zero element, since by definition, the noise terms $\{\xi_{it}\}$ have zero mean. From (25), we obtain the expectation of the production requirements vector:

$$\boldsymbol{E}[\boldsymbol{P}] = (\boldsymbol{I} - \boldsymbol{\Phi})^{-1}\boldsymbol{\mu} \,. \tag{26}$$

Note that $\boldsymbol{E}[\boldsymbol{P}]$ does not depend on the planning parameters but depends only on the workflow matrix $\boldsymbol{\Phi}$ and the vector $\boldsymbol{\mu}$. The variance of the production vector is given by

$$\boldsymbol{Var}(\boldsymbol{P}) = \sum_{s=0}^{\infty} \boldsymbol{B}^s \boldsymbol{H}\boldsymbol{Var}(\boldsymbol{\zeta}_t)\boldsymbol{H}'\boldsymbol{B}'^s \,, \tag{27}$$

where $\boldsymbol{Var}(\boldsymbol{\zeta}_t)$ is a covariance matrix that consists of the variance of D_t and the covariances of the noise terms ξ_{it}. We note that (27) provides the production variances for each station as well as the covariance for each pair of workstations. These covariances are useful for understanding the interdependence among workloads at different workstations.

We can compute the power series in (27) using the methods in Graves (1986). We show in Teo (2006) that the series in (24) converges if and only if the spectral radius of $\boldsymbol{\Phi}$ is less than one. This condition assures that a unit of work processed at one station cannot eventually generate more than one unit of work for the same workstation; otherwise the system will not reach a steady state.

Now we compute the expected queue length at each station from (22):

$$\boldsymbol{E}[\boldsymbol{Q}] = \boldsymbol{F}^{-1}(\boldsymbol{I} - \boldsymbol{G}\boldsymbol{\Phi})\boldsymbol{E}[\boldsymbol{P}] = \boldsymbol{F}^{-1}(\boldsymbol{I} - \boldsymbol{G}\boldsymbol{\Phi})(\boldsymbol{I} - \boldsymbol{\Phi})^{-1}\boldsymbol{\mu} \tag{28}$$

5. Extension to Multi-Family Model

The single family model assumes a single workflow matrix $\boldsymbol{\Phi}$ and assumes a single planned lead-time for each workstation. In the multi-product setting, this implies each station has to have a common station planned lead-time across all product families and all product families have same workflow and routing through the shop. In this section we show how to relax these restrictions and extend the single family model for the multi-product setting.

We consider a manufacturing system with multiple product families and where each workstation processes jobs from one or more families. We assume that each workstation *i* has a station planned lead-time n_{ik} for each product family that visit the workstation. Furthermore, each product family is allowed to have its own DLT_k, $PPLT_k$ and planning windows W_k.

We first analyze the production for each individual product family. We model each product family as in the single-family model, i.e., each product family *k* has a dummy station and has column vector $\boldsymbol{\zeta}_{\boldsymbol{kt}}$ and matrix Φ_k to model each family's demand, noise levels and routing. Furthermore, we have to satisfy (1) and (2) for the DLT_k, $PPLT_k$ and planning window W_k of each product family. We use the single-family model to characterize the production requirements and queue lengths for product family *k*, i.e., $\boldsymbol{E}[\boldsymbol{P}_{\boldsymbol{k}}]$, $\boldsymbol{Var}(\boldsymbol{P}_{\boldsymbol{k}})$ and $\boldsymbol{E}[\boldsymbol{Q}_{\boldsymbol{k}}]$.

Similar to the single family model, we assume an infinite capacity (albeit with a finite nominal capacity M_i) at each workstation. Hence each workstation can always meet the production requirements for all product families processed at the station. By further assuming that demand is independent between the product families, we obtain the mean and variance of the total production requirements and the total queue lengths by aggregating across all product families:

$$\boldsymbol{E}[\boldsymbol{P}_{\boldsymbol{total}}] = \sum_{\boldsymbol{k}} \boldsymbol{E}[\boldsymbol{P}_{\boldsymbol{k}}], \tag{29}$$

$$\boldsymbol{Var}(\boldsymbol{P}_{\boldsymbol{total}}) = \sum_{\boldsymbol{k}} \boldsymbol{Var}(\boldsymbol{P}_{\boldsymbol{k}}), \tag{30}$$

$$\boldsymbol{E}[\boldsymbol{Q}_{\boldsymbol{Total}}] = \sum_{\boldsymbol{k}} \boldsymbol{E}[\boldsymbol{Q}_{\boldsymbol{k}}]. \tag{31}$$

The above results require independence between the demands of the product families. We can relax this assumption to incorporate correlated demands by building a single linear-systems model in which the first

k terms of the noise vector ζ_{kt} correspond to the demand processes for the *k* product families. We can then incorporate the demand correlations into the covariance matrix for the noise vector ζ_{kt}. See Teo (2006) for details.

6. Simulation Study

We perform a simulation study to test the accuracy of (27). Specifically, we aim to validate the following workflow assumptions in deriving (27):

- We permit the production requirements to vary within each period according to (4); thus, if jobs move instantly downstream upon completion, the arrival process to the downstream stations will also vary, which contradicts the uniform-arrival assumption.
- We model flow of discrete jobs but assume workflow is measured in units of work content (e.g., hours).
- In deriving the continuous-time coefficients β_i and γ_i, we assume that each job flows continuously like a fluid and that the production output satisfies (4) at every instant of time. We assume that the sub-period coefficients $\beta_i(\Delta_i)$ and $\gamma_i(\Delta_i)$ can depict job flow of lower frequencies.

We simulate a six-station serial flow system with a stationary work input at the first station. Processing times are fixed and equal at all stations. The motive for such a simplified simulation model is to attain a controlled experiment and to test whether the accuracy of (27) deteriorates for the downstream workstations. In the simulation model, discrete jobs are processed and jobs move immediately to the next station upon completion, instead of flowing downstream uniformly over each time period as assumed in the model. For the work input with mean of 80 workhours and standard deviation of 20 workhours, we experiment with different processing times (1, 2, 4, 8 and 16 hours) to vary the fluidity of workflow; the longer the processing time, the less fluid (i.e., more lumpy) the workflow. In addition, we assume the station planned lead-times are equal for all stations, and we experiment with different values (1, 2 and 3 periods). Note that the utilization levels are not an experimental factor because we assume flexible capacity. A more detailed discussion of the simulation setup and results are presented in the *Online Appendix*.

The results are encouraging as the study shows that (27) computed using β_ι and γ_i is reasonably accurate for a wide range of job fluidity. The average percentage error in the standard deviation of P_{it} for processing times 1, 2 4 and 8 hours is 2.3% and the maximum error is 6.5%. As expected, the errors computed for the less fluid 16-hour jobs using β_ι and γ_i are relatively higher, with the largest percentage error at 17.6%. The sub-period coefficients $\beta_\iota(\Delta_\iota)$ and $\gamma_\iota(\Delta_\iota)$ yield a reasonably accurate approximation for the

16-hour jobs with a percentage error of 2%. We observe no obvious trends in percentage errors between the different stations and the station planned lead-times. We also test the model's robustness to production variability by experimenting with different standard deviations of processing times, namely 0.4, 0.6, 0.8 and 1.0 hour for a mean processing time of 4 hours. The average percentage error is 3.0% for all test problems.

The simulation model consists of a serial flow system that processes a single product family. We offer a few comments to how we can extend the simulation results to the multi-family, general network setting. Since the multi-family model involves modeling each product family independently using family specific parameters, the validity of the multi-family model depends on the model for each individual product family. Therefore, we focus our attention on the generalized routing of a single product family, wherein each workstation can receive job stream from multiple stations and its output can be dispatched to more than one workstation. One limitation arises if the arrival streams at a workstation (of the same product family) are of vastly different level of fluidity. For instance, a workstation receives two arrival streams, one with lumpy jobs at low frequencies and the other with high fluidity; as such, there is no appropriate set of coefficients that is able to match the fluidity of both streams. We note that this limitation does not occur if the arrival streams are of suitably high fluidity, in which case the coefficients β_i and γ_i would match both streams. We expect this could be a probable scenario because as stated earlier, the simulation results show that β_i and γ_i give accurate computations for a reasonably wide range of job fluidity. Likewise, if the streams are from different product families, the limitation will not happen because a family-specific set of coefficients can be employed to match the fluidity of each product family.

7. Optimization

We formulate a nonlinear optimization program for the multi-family model. Our objective is to minimize the total expected expediting cost (e.g., overtime and subcontracting) plus total WIP inventory holding cost. The decision variables are W_k and n_{ik}. We assume the delivery lead-time DLT_k is exogenously determined for each product family k.

$$Min \quad \sum_i \left[c_i E\left[P_{it} - M_i\right]^+ + \sum_k h_{ik} E\left[Q_{ikt}\right] \right]$$

$$s.t. \quad \sum_i \omega_{i,r(k)} n_{ik} + W_k - 1 = DLT_k, \quad \forall k, r(k)$$

$$W_k \geq a_k, \quad \forall k$$

$$n_{ik} \geq b_i, \quad \forall i, k.$$

The objective function is the total expected cost in (3), wherein P_{it} is a function of both W_k and n_{ik}, and Q_{ikt} is a function of n_{ik}. We assume that the production requirement P_{it} is normally distributed. We note that P_{it} is normally distributed if the demands and noise terms are normal random variables. We use the mean and variance of P_{it} from (29) and (30) to evaluate the first term of the objective function by the normal linear loss integral.

The first set of constraints combines (1) and (2) for product family k, defining the relationship between the planning windows, station planned lead-times and *DLT*s for every routing r_k. The second and third sets of constraints assure a lower bound of at least a_k and b_i on the planning windows and n_{ik} of each workstation i, respectively. A smaller n_{ik} would imply a need for frequent monitoring at the workstation to track the job progress. Likewise, a short planning window would require more efforts in monitoring the MPS. Hence both a_k and b_i are managerial inputs that must be set appropriately in order to avoid complications in production control.

We have not established the convexity of the objective function. As a consequence we cannot assert that we obtain the global optimum to this nonlinear optimization program. Nevertheless, from our computational experience with test problems, we observe that standard nonlinear optimization algorithms perform very reliably on this problem: from a wide range of starting points, for each test problem we always obtain the same solution. In the next section we report on one computational study. Teo et al. (2009) report another computational study based on an application of the model to an oil-rig manufacturer.

8. Numerical Example

We use a real-world setting to illustrate the use of the model and to draw some insights. We consider a production shop that is part of a manufacturing facility of an oil-rig builder. Teo et al. (2009) present a case study on another larger facility of the same oil-rig builder in which the model's predictive capability is validated. The data presented in this section has been altered to protect proprietary information. However the resulting qualitative relationships and insights drawn from this example are the same as they would be from using the actual data.

The shop processes and cuts steel plates into required dimensions, which are then used for downstream production stages to construct the necessary steel assemblies. The shop operates in a MTO environment because the internal customer orders are highly customized, resulting in numerous possible cutting dimensions and plate thicknesses. The delivery lead-times for the internal orders are fixed to facilitate planning in the downstream processes. The production control is based on planned lead-times in which the production schedules synchronizes with the "scheduled receipts" of engineering drawings for the cut-

ting dimensions. In addition, due to the large physical size of the plates (each weighing 0.5 to 10 tons) and the space constraints in the facilities, the stations usually produce a quantity that is just sufficient to meet the station planned lead-times, so as to avoid taking up the downstream shop space unnecessarily.

As part of ongoing improvement efforts, management was examining how best to group the numerous production options into product families and how to set planned lead-times appropriately. In this example, we categorize the customer orders according to the plate thickness into two product families: *Thick Plates* and *Thin Plates*. The shop consists of 4 workstations, namely *Blasting*, *NC* (Numerical Control) *Gas Cut*, *NC Plasma Cut* and *Manual Cut*. The process flow map, including the dummy stations to model the MPS smoothing, is shown in Figure 4.

1. Dummy station for *Thick Plates*
2. Dummy station for *Thin Plates*
3. Blasting
4. NC Gas Cut
5. NC Plasma Cu
6. Manual Cut

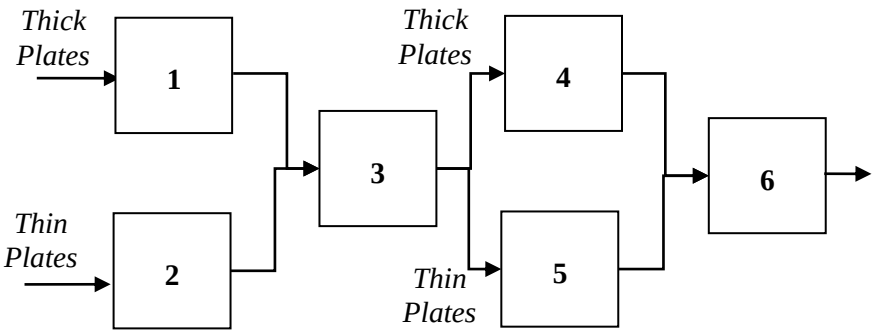

Figure 4. Process Flow Map

Raw steel plates (both *Thick Plates* and *Thin Plates*) are first sent to the *Blasting* station where ball grids are splashed onto the plate surfaces to remove impurities, followed by an application of a corrosion preventive paint. The blasted *Thick Plates* are then transferred to the *NC Gas Cut* station, which is capable of cutting the *Thick Plates*. The blasted *Thin Plates* are moved to the *NC Plasma Cut* station, which can only cut the *Thin Plates*. The steel plates are then sent to the *Manual Cut* station where the final cutting is done manually using hand-held tools.

The demands for both product families have been found to be i.i.d. over time and are approximately normally distributed. The daily demand for the *Thick Plates* has a mean of 20 plates and standard deviation of 10 plates, while the demand for *Thin Plates* has a mean of 26 plates and standard deviation of 12 plates. The demands for the product families are uncorrelated. The fixed delivery lead-times for the *Thick Plates* and *Thin Plates* are 9 days and 8 days respectively. We show in Table 2 the expectation and standard deviation of the effective processing times at each workstation. The standard deviation of the effective processing time is mainly due to the different processing requirements for the different plates. Table 2 shows the nominal capacity available per day in workhours. Each workstation has flexibility to expand its

capacity in each day by subcontracting its production to nearby shops. The subcontracting cost per hour for outstanding work at each workstation is also shown in the Table 2.

Table 2. Data of workstations

	Processing time per plate (hour)				Capacity (workhours per day)	Subcontract Cost ($ per workhour)	Holding Cost ($ per workhour per day)
	Mean		Standard deviation				
	Thick	Thin	Thick	Thin			
Blasting[1]	0.55	0.55	0.35	0.35	28	550	0.72
NC Gas Cut	1.69	-	1.96	-	43	368	0.61
NC Plasma Cut	-	1.34	-	1.46	49	441	0.77
Manual Cut	3.50	1.07	2.55	0.86	128	788	0.74

[1]The expectation and standard deviation of processing times at the *Blasting* station are identical for both product families

We set up the workflow matrices for *Thick Plates* and *Thin Plates*, denoted by $\mathbf{\Phi}_{Thick}$ and $\mathbf{\Phi}_{Thin}$ respectively, as:

$$\mathbf{\Phi}_{Thick} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0.55 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3.08 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2.07 & 0 & 0 \end{bmatrix} \quad \mathbf{\Phi}_{Thin} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.55 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2.43 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.80 & 0 \end{bmatrix}$$

For example, each workhour at *Blasting* generates on average 1.69/0.55 = 3.08 workhours for the *NC Gas Cut* station, i.e. $\phi_{43} = 3.08$. To set up the variance of the vector $\boldsymbol{\zeta}_t$, we note that the off-diagonal terms of the matrix are all equal zero as there is assumed to be no correlation between the demands or the random variability at the stations. We set the diagonal elements for the dummy stations (first two rows) equal to the variance of the daily demand; the diagonal elements for the workstations (bottom four rows) represent the variance of the noise terms $Var(\xi_{it})$ due to the variability in the processing times, which we set to the variance of the workload for the mean number of jobs processed at the station per day. For example, at the *Blasting* station, the mean number of *Thick Plates* processed per day is 20 and the standard deviation of the processing time is 0.35 hour, which $Var(\xi_{it}) = 20(0.35)^2 = 2.45$ hour^2. We assume that the noise term at each station is normally distributed. The matrices for the variances of $\boldsymbol{\zeta}_t$ for the two product families are:

$$\boldsymbol{Var}(\zeta_{Thick,t}) = \begin{bmatrix} 100 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2.45 & 0 & 0 & 0 \\ 0 & 0 & 0 & 75.8 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 129.7 \end{bmatrix} \quad \boldsymbol{Var}(\zeta_{Thin,t}) = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 144 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3.19 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 55.0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 19.2 \end{bmatrix}$$

Base case

In the base case, the raw plates are "pushed" into production with planning window W_k = 1 (i.e., no smoothing of the MPS). Each station has a common station planned lead-time n_{ik} for both product families. The station planned lead-times assigned to the stations are roughly equal. Table 3 shows the values of W_k and n_i. The table also shows the mean and standard deviation of the production requirement as well as the queue length at each station computed using the model. Due to the normality assumption for the demands and variability at each station, the production requirements at the stations are also normally distributed; this allows us to compute the probability of subcontracting. We also compute the expected subcontracting cost and holding cost per day, which are shown in the same table.

Table 3. Base case

	W_k or n_i	$E[P_{it}]$	$\sigma(P_{it})$	$E[Q_{it}]$	Probability of subcontracting in a day	Expected subcontract cost per day ($)	Expected holding cost per day ($)
Release (Thick)	1	20.00	10.00	-	-	-	-
Release (Thin)	1	26.00	12.00	-	-	-	-
Blasting	3	25.30	3.38	75.90	0.21	223.4	55.20
NC Gas Cut	3	33.84	6.14	101.6	0.07	68.31	62.49
NC Plasma Cut	2	34.88	6.33	69.50	0.01	11.72	53.87
Manual Cut	3	97.85	12.19	293.8	0.01	21.37	216.5
					Total ($):	324.80	388.06
						Total Cost ($) = 712.86	

The release into the production shop is highly variable as there is no smoothing of the MPS. At the *Blasting* station, there is a relatively high probability of 0.21 that the shop has to subcontract the outstanding work. The expected subcontracting cost of the *Blasting* station is $223.40 per day. The other workstations have much lower probabilities of insufficient nominal capacity. The queue length at the

Manual Cut station is the highest, with 293.8 hours of work-in-queue, resulting in a daily holding cost of $216.50.

Smoothing of MPS

We increase the planning windows of each product family to 3 days. To keep the delivery lead-time fixed, this increase in the planning window must be compensated by a reduction of the station planned lead-times at the stations. We observe in the base case that the *Manual Cut* station has the lowest expected subcontracting cost and the highest inventory holding cost. Hence we choose to reduce the station planned lead-time of the *Manual Cut* station from 3 days to 1 day.

Table 4. Smoothing of the MPS

	W_k or n_i	$E[P_{it}]$	$\sigma(P_{it})$	$E[Q_{it}]$	Probability of subcontracting in a day	Expected subcontract cost per day ($)	Expected holding cost per day ($)
Release (Thick)	3	20.00	4.47	-	-	-	-
Release (Thin)	3	26.00	5.37	-	-	-	-
Blasting	3	25.30	2.76	75.90	0.16	119.6	55.20
NC Gas Cut	3	33.84	5.77	101.6	0.06	48.36	62.49
NC Plasma Cut	2	34.88	5.87	69.5	0.01	6.55	53.87
Manual Cut	1	97.85	14.58	97.9	0.02	77.14	72.12
					Total ($):	251.65	243.68
						Total Cost ($) = 495.33	

Table 4 shows that the releases for both product families become less variable. The standard deviation of the release for *Thick* Plates falls from 10 to 4.47 plates and that for the *Thin Plates* from 12 to 5.37 plates. The smoother release results in less variable job arrivals at the workstations which in turn causes smoother production requirements. For example, the standard deviation of production requirements for the *Blasting* station decreases from 3.38 hours in the base case to 2.76 hours. This leads to a reduction in the probability of subcontracting at the *Blasting* station from 0.21 to 0.16, with a corresponding fall in expected subcontracting cost from $223.36 to $119.60. The *NC Gas Cut* and the *NC Plasma Cut* also undergo a reduction in expected subcontracting costs due to the less variable work arrival from the *Blasting* station; this illustrates how production variability at an upstream station can affect the downstream stations.

The *Manual Cut* station experiences two opposite effects on its production variability. On the one hand, the longer planning windows lead to smoother job arrivals to the station, but it now has a shorter station planned lead-time to smooth the arrival variability. The combined effect causes an increase in production variability at the workstation. We note that its expected subcontracting cost rises from $21.37 to $77.14. But its queue length decreases from 293.79 hours to 97.9 hours with a decline in expected holding cost from $216.53 to $72.12. Overall, the expected total cost for the shop falls from $712.86 in the base case to $495.33.

Optimal solution

We obtain the optimal planning windows and station planned lead-times using the optimization model in Section 7, where we set $a_k = 1$ for all k and $b_i = 1$ for all i. The optimal solution and associated results are shown in Table 5. The optimal solution involves further reducing the station planned lead-times of the *Blasting* station and *NC Plasma Cut* station, and to a lesser extent the *NC Gas Cut* station; in turn, the planning windows are increased. Even though the total expected subcontracting cost increases, the decrease in the total expected holding cost more than offsets this increase; thus, it leads to a reduction in the total cost. The minimum expected total cost is $464.09, which translates into a 34.9% cost savings over the base case.

Table 5. Optimal solution.

	W_k or n_i	$E[P_{it}]$	$\sigma(P_{it})$	$E[Q_{it}]$	Probability of Subcontracting in a day	Expected subcontract cost per day ($)	Expected holding cost per day ($)
Release (Thick)	4.16	20.00	3.70	-	-	-	-
Release (Thin)	5.06	26.00	3.97	-	-	-	-
Blasting	1.94	25.30	2.73	49.02	0.16	115.6	35.65
NC Gas Cut	2.90	33.84	5.83	98.16	0.06	50.78	60.35
NC Plasma Cut	1	34.88	6.84	34.88	0.02	20.59	27.04
Manual Cut	1	97.85	14.71	97.85	0.02	81.96	72.12
					Total ($):	268.93	195.16
						Total Cost ($) = 464.09	

In addition, we can also use the model for various "what-if" analyses. For example: what if the delivery lead-time is reduced? What if the nominal capacity at a station is increased? We can also utilize the model to determine the impact of unplanned changes: what if the demand increases? If the subcontracting costs

are raised, should we acquire more capacity? The model can be employed to determine the impact of these changes on the expected total cost to support tactical planning.

9. Conclusions

In this paper, we model a MTO manufacturing environment that has a fixed quoted delivery lead-time. We focus on the setting of the planning windows and station planned lead-times, which are two key tactical planning parameters. We develop for a single aggregate product a workflow model that allows intra-period workflows. We show how to characterize the production requirements as well as the expected queue lengths for each workstation. We then extend the single-family model to systems with multiple families. We embed the model into a nonlinear optimization program to find the optimal values of the planning parameters. We illustrate the use of the model with a small case example from a shop that processes steel panels for oil-rig construction.

One possible extension to this research is to model the stability of the MPS when it is subject to changes in customer orders. In our model, we assume that there is no change to the order quantity after the customer order is received. However in many MTO firms, it is common for customers to adjust or modify their orders within the delivery lead-time. Thus an important enhancement to the model would be to incorporate the dynamics of the MPS and its impact on the production flow, especially in light of the techniques that are used to achieve a stable MPS, e.g., firm planned orders, frozen time periods and time fencing (see Vollmann et al. 2005).

It would also be worthwhile to explore how the model can be employed to provide feedback in supporting the setting of the fixed quoted delivery lead-time. The feedback mechanism would be equivalent to the approach in hierarchical planning, wherein our model serves as a disaggregation model that provides feedback to the more aggregate models (e.g., So and Song 1998 and Rao et al. 2005) that determine the fixed quoted delivery lead-time. It would facilitate the capturing of the problem at different detail levels, i.e., aggregated data for setting the fixed lead-time and more detailed production data for determining the planning windows and station planned lead-times.

Acknowledgements

This research has been supported by the Singapore-MIT Alliance (SMA) program. The authors thank the Departmental Editor, Associate Editor and three anonymous referees for their valuable comments which greatly improved the paper.

Appendix

We consider a product family with $DLT_k = 4$, $PPLT = 2$ and planning window of length = 4 – 2 + 1 = 3. The current period is $t = 0$. The production planner has to decide on the MPS and the planned releases over the planning horizon that ranges from $t = 0$ to $t = 4$. To simplify the example, there are no orders after period $t = 4$ and there is no planned inventory at $t = 1$. We illustrate the smoothing of the MPS in Table A1.

Table A1: Smoothing of the MPS

Period t	0	1	2	3	4
Orders	-	-	10	20	30
MPS	-	-	20	20	20
Planned Inventory	-	0	10	10	0
Planned Release	20	20	20	-	-

The second row shows orders contracted to be delivered in each period, from $t = 2$ to 4. We do not state the order quantities for period $t = 0$ and $t = 1$, as they are already in process and are thus inconsequential. In this example, we smooth the MPS by spreading the orders as evenly as possible over the planning window, which ranges from $t = 2$ to $t = 4$; in particular, we level the MPS at 20 units for each period of the planning window. Since the order quantity in $t = 2$ is only 10 units, the MPS results in a planned inventory of 10 units in period $t = 2$ that is carried over to period $t = 3$ and is finally applied to the order in $t = 4$. Furthermore, since the $PPLT_k = 2$, we release jobs two periods earlier to meet the corresponding MPS. For instance, we plan to release 20 units at $t = 0$ to meet the MPS in period $t = 2$.

References

Asmundsson, J. M., R. L. Rardin and R. Uzsoy. (2006) Tractable Nonlinear Production Planning Models for Semiconductor Wafer Fabrication Facilities. *IEEE Transactions on Semiconductor Manufacturing* **19**: 95-111.

Baker, K. R. (1984) Sequencing Rules and Due-Date Assignments in a Job Shop. *Management Sci*., **30**, 1093 – 1104.

Bertrand, J.W.M. and Ooijen, H.P.G. van. (2002) Workload based order release and productivity: a missing link. *Production Planning and Control*, 13, 7, 665-678.

Cruickshanks, A. B., R. D. Drescher and S. C. Graves. (1984) A Study of Production Smoothing in a Job Shop Environment. *Management Sci.*, **30**, 36-42.

Cheng, T.C.E. and Gupta, M.C. (1989) Survey of scheduling research involving due date determination decisions. *European Journal of Operational Research*, 38, pp. 156-166.

Fine, C. H. and S. C. Graves. (1989) A Tactical Planning Model for Manufacturing Subcomponents in Mainframe Computers. *J. Manuf. and Opns. Mgmt*., **2**, 1, pp 4-34.

Gong, L., T. d. Kok., and J. Ding. (1995) Optimal Leadtimes Planning in a Serial Production System. *Management Sci.,* **40**, 5, 629-632.

Graves, S. C. (1986). A Tactical Planning Model for a Job Shop. *Oper. Res.*, **34**, 4, pp 522-533.

Graves, S. C. (1988a) Safety Stocks in Manufacturing Systems. *J. Manuf. and Opns. Mgmt*. **1**, 1, 67-101.

Graves, S. C. (1988b) Determining the Spares and Staffing Levels for a Repair Depot. *J. Manuf. and Opns. Mgmt*. **1**, 2, 227-241.

Graves, S. C. (1988c) Extensions to a Tactical Planning Model for a Job Shop. *Proceedings of the 27th IEEE Conference on Decision and Control*, Austin, Texas.

Graves, S. C., D. B. Kletter, and W. B. Hetzel. (1998) A Dynamic Model for Requirements Planning with Application to Supply Chain Optimization. *Oper. Res.*, **46**, 3, 35 – 49.

Graves, S. C. and J. S. Hollywood. (2001) revised March 2004, January 2006. A Constant-Inventory Tactical Planning Model for a Job Shop. *Working paper*, 36 pp.

Hopp, W. J., and M. L. Spearman. (2001) Factory Physics: Foundations of Manufacturing Management. 2nd edition. Irwin McGraw Hill, Boston.

Hollywood, J. S. (2005) An Approximate Planning Model for Distributed Computing Networks. *Naval Res. Logistics*. **52**, 6, 590-605.

Holt, C.C., F. Modigliani, J.F. Muth and H.A. Simon. (1960) Planning Production, Inventories and Work Force. Englewood Cliffs, NJ, Prentice Hall.

Kanet, J. J. (1986) Toward a better understanding of lead-times in MRP systems. *J. of Oper. Mgt*., **6**, 3, 305 – 315.

Karmarkar, U. S. (1989) Capacity Loading and Release Planning with Work-in-Progress (WIP) and Lead-times. *Journal of Manufacturing and Operations Management*. **2**, 105-123.

Karmarkar, U. S. (1993) Manufacturing Lead-times, Order Release and Capacity Loading. In *Handbooks in Operations Research and Management Science, Vol. 4., Logistics of Production and Inventory*, S. C. Graves, A. H. Rinnooy Kan, P. H. Zipkin (eds.), North-Holland, Amsterdam, pp. 287-329.

Keskinocak, P., R. Ravi, S. Tayur. (2001) Scheduling and reliable lead-time quotation for orders with availability intervals and lead-time sensitive revenues. *Management Sci*., 47, 264–279.

Keskinocak, P. and S. Tayur. (2004) Due Date Management Policies. In *Handbook of Quantitative Supply Chain Analysis; Modeling in the E-Business Era.* D. Simchi-Levi, S-D. Wu and Z-M Shen (eds), Kluwer Academic Publishers, pp. 485-554.

Matsuura, H. and H. Tsubone. (1993) Setting Planned Lead-times in Capacity Requirements Planning. *Journal of Oper. Res. Society*., **44**, 8, 809-816.

Matsuura, H., H. Tsubone and M. Kanezashi. (1996) Setting Planned Lead-times for Multi Operation Jobs. *European Journal of Oper. Res.*, **88**, 287-303.

Rao, U.S., J.M. Swaminathan and J. Zhang. (2005) Demand and Production with Uniform Guaranteed Lead-time. *Production and Operations Management*, **14**, 4, 400-412.

Robinson, A. (1999) Toyota to test quick turn on Solara. Automotive News. 9 August. www.automotivenews.com

So, K. C., J.-S. Song. (1998) Price, delivery time guarantees and capacity selection. *Eur. J. Oper. Res.*, **111**, 28 – 49.

Selcuk, B., J. C. Fransoo and A. G. de Kok. (2008) Work in Process Clearing in Supply Chain Operations Planning. *IIE Transactions*., **40**, 3, 206-220.

Teo, C. C. (2006) A Tactical Planning Model for Make-To-Order Environment under Demand Uncertainty. Ph.D. thesis, Singapore-MIT Alliance, Nanyang Technological University.

Teo, C.C., R. Bhatnagar and S.C. Graves. (2009) An Application of Master Schedule Smoothing and Planned Lead-times Control. *Working Paper*, Singapore-MIT Alliance, Singapore.

Vollmann, T. E., W. L. Berry, D. C. Whybark, and F. R. Jacobs. (2005) *Manufacturing Planning and Control for Supply Chain Management*, 5th edition, McGraw-Hill.

Weng, K.Z. (1999) Strategies for integrating lead-time and customer-order decisions. *IIE Transactions*. **31**, 2, 161-171.

Weeks, J. K. (1978) Optimizing planned lead-times and delivery date. *Proceedings of the 21st American Production and Inventory Control Society Annual Meeting*, 177 – 188.

Yano, C. A. (1987) Setting Planning Leadtimes in Serial Production Systems with Earliness Costs. *Management Sci.*, **33**, 1, 95-106.