

# **A Tactical Planning Model for a Production Network with Continuous-Time Control**

**Chee-Chong Teo<sup>1,2</sup>, Rohit Bhatnagar<sup>1,2</sup>, Stephen C. Graves<sup>1,3</sup>**

<sup>1</sup>Singapore-MIT Alliance, Nanyang Technological University, Singapore

<sup>2</sup>Nanyang Technological University, Singapore

<sup>3</sup>Sloan School of Management, Massachusetts Institute of Technology, USA

5 January 2007

We develop an extension to the tactical planning model (TPM) for a job shop by Graves (1986). The TPM is a discrete-time model in which all transitions occur at the start of each time period. The time period must be defined appropriately in order for the model to be meaningful. On the one hand, each period must be short enough so that a job is unlikely to travel through more than one station in a single period. On the other hand, in order for the TPM to be useful for production planning, the time period needs to be long enough to coincide with the underlying time buckets of planning systems. We overcome this restriction of period sizing by building an extension to the TPM for a single production station with a continuous-time production function. We then determine the first two moments of the production requirement and queue length based on the derived continuous-time model. We show how to use this single-station model to build a multi-station model for a general stochastic flow network.

*Subject classifications:* Inventory/production: tactical planning that considers tradeoffs between production smoothness, planned lead time and work-in-process inventory.

*Area of review:* Manufacturing, Service and Supply Chain Operations

# A Tactical Planning Model for a Production Network with Continuous-Time Control

5 January 2007

---

We develop an extension to the tactical planning model (TPM) for a job shop by Graves (1986). The TPM is a discrete-time model in which all transitions occur at the start of each time period. The time period must be defined appropriately in order for the model to be meaningful. On the one hand, each period must be short enough so that a job is unlikely to travel through more than one station in a single period. On the other hand, in order for the TPM to be useful for production planning, the time period needs to be long enough to coincide with the underlying time buckets of planning systems. We overcome this restriction of period sizing by building an extension to the TPM for a single production station with a continuous-time production function. We then determine the first two moments of the production requirement and queue length based on the derived continuous-time model. We show how to use this single-station model to build a multi-station model for a general stochastic flow network.

*Subject classifications:* Inventory/production: tactical planning for tradeoffs between production smoothness, planned lead time and work-in-process inventory.

---

## 1. Introduction

Graves (1986) introduces a tactical planning model (TPM) for a stochastic flow network, as a model for a general production system such as a job shop. The intent of the TPM is to characterize the tactical tradeoffs for a general production network, namely the tradeoffs among capacity, inventory and time. The TPM, and variants of it, have been successfully applied in several manufacturing contexts (Cruickshanks et al., 1984; Graves 1988a; Fine and Graves, 1989; Graves et al., 1998; Graves and Hollywood, 2001; Teo, 2006).

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 occur at the start of each time period. As such, one must set the time period to be short enough so that it is unlikely for one job in the actual system to travel through two successive stations in a single period. However, for the TPM to be useful as an analytical tool for production planning, we require the

TPM planning period to match the underlying time buckets of planning systems. But in many situations, the restriction of period sizing prevents the setting of a period long enough to match the time bucket. Moreover, a short period size does not correspond to the model's assumption of Markovian job movements. The intent of this note is to address this limitation of the model; in particular, we develop a continuous-time version of the single-station model so as to remove the dependence on the choice of time period.

We begin with a review of the TPM. Central to the TPM is the planned lead time of each station, which is a key control parameter in most planning systems. There is a capacity-inventory tradeoff involved in setting the planned lead time. If the planned lead time were set too short, it might be difficult to smooth a fluctuating workload and consequently, the capacity might not be sufficient to meet the "peaks" of the varying workload. Conversely, if the planned lead time were set too long, it may lead to higher work-in-process (WIP) inventory.

The TPM computes the mean and variance of the production requirement and queue length at each station, both of which are functions of the planned lead time. In particular, the variance of the production requirement provides a measure of the production smoothness while the mean queue length signifies the expected WIP level. Thus the TPM provides a simple way to analyze the fundamental tradeoff between the three elements of lead time, capacity and inventory.

Furthermore, with additional distribution assumptions on the work arrivals, one obtains complete characterizations of the requirement distributions from the first two moments. These distributions can be used for performance evaluations that support the analysis of the aforementioned tradeoffs, e.g. the probability that the production requirement exceeds the capacity and the expected backlog. Fine and Graves (1989) apply the TPM to analyze these tradeoffs in the tacti-

cal planning for an actual shop that produces computer components. Graves (1988a) develops a TPM-based model to characterize the tradeoff between labor and inventory requirements in a repair depot.

Other work on the TPM includes Graves (1988b), which extends the model to incorporate machine failures, lot-sizing and a capacity constraint. Graves and Hollywood (2001) develop a constant-inventory TPM in which the job release is regulated to maintain a constant inventory level; they illustrate the model with an application to a machine shop. Graves et al. (1998) develop an adaptation of the TPM for modeling the dynamic requirements in a multi-stage production-inventory context, and apply the model to a Kodak internal supply chain. Hollywood (2005) develops an approximate model based on the TPM for a distributed computing network. Teo (2006) extends the TPM for a make-to-order environment for engineered products, and tests it on a production system for oil rigs.

We now give a summary of the model's assumptions and formulations. The TPM is a continuous flow model in which we model the workload of jobs at each station, rather than the number of jobs. The workflow is assumed to have a Markov property, i.e. the processing requirements at a downstream station depend only on the upstream station from which the work arrives. The key assumption of the TPM is the linear control rule, which is stated as

$$P_{it} = \alpha_i Q_{it} \tag{1}$$

where  $P_{it}$  is the production at station  $i$  in time period  $t$ ,  $Q_{it}$  is the queue level (or WIP level) at the start of period  $t$ , and the parameter  $\alpha_i$ ,  $0 < \alpha_i \leq 1$ , is a smoothing parameter. Both  $P_{it}$  and  $Q_{it}$  are measured in units of the workload at station  $i$ , e.g. hours of work. This rule states that the production  $P_{it}$  at station  $i$  is a fixed portion  $\alpha_i$  of the queue  $Q_{it}$ . Here we interpret  $1/\alpha_i$  as the planned lead time. In particular, station  $i$  must process  $\alpha_i$  of the work-in-queue on average in each period

in order to realize the planned lead time; this is approximated by (1) in which the production requirement is assumed to be *precisely*  $\alpha_i$  of the queue. Equation (1) is analogous to the approximate single-stage smoothing model in Cruickshanks et al. (1984); they demonstrate through simulation that the approximate model is indicative of the behavior of their proposed smoothing procedure.

The TPM assumes that there is no hard capacity constraint and that each station produces according to the production rule (1). This assumption is reasonable in settings where expediting actions (e.g. overtime) are taken in periods of high workload. Alternatively, for contexts with identifiable constraints, we use the model to characterize the production requirements for a given setting of the planned lead times; knowledge of these requirement levels is useful for assessing the feasibility of the production plan, e.g. for setting the planned lead times (and WIP level) to be consistent with the capacity planning.

The queue level  $Q_{it}$  satisfies the standard inventory balance equation

$$Q_{it} = Q_{i,t-1} - P_{i,t-1} + A_{it} \quad (2)$$

where  $A_{it}$  is the amount of work that arrives at station  $i$  at the start of period  $t$ . By substituting (1) into (2), we obtain a first-order smoothing equation with  $\alpha_i$  as the smoothing parameter:

$$P_{it} = (1 - \alpha_i)P_{i,t-1} + \alpha_i A_{it} \quad (3)$$

By repeated substitution and assuming an infinite history of arrivals, we find that

$$P_{it} = \sum_{s=0}^{\infty} \alpha_i (1 - \alpha_i)^s A_{i,t-s} \quad (4)$$

If we assume that the arrival stream  $\{A_{it}\}$  to station  $i$  is independent and identically distributed (i.i.d.) with mean  $\mu$  and variance  $\sigma^2$ , then we obtain

$$E[P_{it}] = \mu \quad (5)$$

$$Var(P_{it}) = \frac{\alpha_i \sigma^2}{2 - \alpha_i} \quad (6)$$

We combine (1) and (4) to obtain the first two moments of  $Q_{it}$ :

$$E[Q_{it}] = \frac{\mu}{\alpha_i} \quad (7)$$

$$Var(Q_{it}) = \frac{\sigma^2}{2\alpha_i - \alpha_i^2} \quad (8)$$

If we consider a network of production stations, the arrival stream to a station from upstream stations are generally not i.i.d. but are correlated over time.

Each station can receive two types of arrivals: exogenous arrivals consisting of new jobs and endogenous arrivals consisting of jobs from other stations. We model the work arrival to station  $i$  by

$$A_{it} = \sum_j \phi_{ij} P_{j,t-1} + \varepsilon_{it} \quad (9)$$

The first term in (9) represents the total workflow arriving at station  $i$  from other stations. We assume that one unit (e.g. hour) of work at station  $j$  will trigger, on average,  $\phi_{ij}$  time units of work that arrives next at station  $i$ . The second term in (9) is a random variable  $\varepsilon_{it}$  that represents the work content from new jobs arriving at station  $i$ , as well as noise from both the exogenous and endogenous flows. By assumption, the time series  $\varepsilon_{it}$  is i.i.d. over time.

We can restate (3) and (9) in matrix-vector form:

$$\mathbf{P}_t = (\mathbf{I} - \mathbf{D})\mathbf{P}_{t-1} + \mathbf{D}\mathbf{A}_t \quad (10)$$

$$\mathbf{A}_t = \mathbf{\Phi}\mathbf{P}_{t-1} + \boldsymbol{\varepsilon}_t \quad (11)$$

where  $\mathbf{P}_t = \{P_{1t}, \dots, P_{nt}\}'$ ,  $\mathbf{A}_t = \{A_{1t}, \dots, A_{nt}\}'$ , and  $\boldsymbol{\varepsilon}_t = \{\varepsilon_{1t}, \dots, \varepsilon_{nt}\}'$  are column vectors of random variables,  $n$  is the number of stations,  $\mathbf{I}$  is the identity matrix,  $\mathbf{D}$  is a diagonal matrix with  $\{\alpha_1, \dots, \alpha_n\}$  on the diagonal, and  $\boldsymbol{\Phi}$  is an  $n$ -by- $n$  matrix with elements  $\phi_{ij}$ . By substituting (11) into (10) and by iterating the resulting equation, we obtain  $\mathbf{P}_t$  as an infinite series

$$\mathbf{P}_t = \sum_{s=0}^{\infty} (\mathbf{I} - \mathbf{D} + \mathbf{D}\boldsymbol{\Phi})^s \mathbf{D}\boldsymbol{\varepsilon}_{t-s}$$

The infinite series converge provided that  $\rho(\boldsymbol{\Phi}) < 1$ , where  $\rho(\boldsymbol{\Phi})$  denotes the spectral radius of  $\boldsymbol{\Phi}$  (see Graves 1986). The mean and the covariance for the noise vector  $\boldsymbol{\varepsilon}_t$  are denoted by  $\boldsymbol{\mu} = \{\mu_1, \dots, \mu_n\}'$ , and  $\boldsymbol{\Sigma} = \{\sigma_{ij}\}$  respectively. If we assume the noise vector  $\boldsymbol{\varepsilon}_t$  to be a normally-distributed random vector, then the production vector  $\mathbf{P}_t$  is also normally distributed. The first two moments of  $\mathbf{P}_t$  are given by

$$\begin{aligned} E[\mathbf{P}_t] &= \sum_{s=0}^{\infty} (\mathbf{I} - \mathbf{D} + \mathbf{D}\boldsymbol{\Phi})^s \mathbf{D}\boldsymbol{\mu} \\ &= (\mathbf{I} - \boldsymbol{\Phi})^{-1} \boldsymbol{\mu} \end{aligned} \tag{12}$$

$$\mathbf{S} = \text{Var}(\mathbf{P}_t) = \sum_{s=0}^{\infty} \mathbf{B}^s \mathbf{D}\boldsymbol{\Sigma}\mathbf{D}\mathbf{B}^{s'}$$

where  $\mathbf{B} = \mathbf{I} - \mathbf{D} + \mathbf{D}\boldsymbol{\Phi}$ . We note that  $\mathbf{S}$  provides the variance of the production requirement for each station, as well as the covariance for each pair of stations. In addition, we determine the first two moments of the queue levels. From (1), we note that

$$\mathbf{Q}_t = \mathbf{D}^{-1} \mathbf{P}_t$$

Therefore we have

$$E[\mathbf{Q}_t] = \mathbf{D}^{-1} E[\mathbf{P}_t]$$

$$\text{Var}(Q_t) = D^{-1}SD^{-1}$$

In the next section, we discuss the limitations of the TPM due to the period sizing restriction. In Section 3, we present a continuous-time model with a linear control rule and determine the first two moments of the production requirements and queue lengths. We present this model for a single-station setting so as to highlight the key insights. We conclude in Section 4 with some discussion on the model and describe how this development extends to a multi-stage system.

## 2. Limitations of TPM

In the TPM, work flows between the stations at the start of each time period; thus, in the TPM, jobs visit at most one workstation in each period. In order to model the workflow in an actual shop, we need to set the period length to be short enough so that it is highly improbable for one job to travel through more than one station in a single period. The setting of the period length is even more complex for a production network with many stations. One must set the period length short relative to the frequency of job movements between each pair of stations. Furthermore, the smoothing parameter  $\alpha_j$  in the TPM are constrained to be  $0 < \alpha_j \leq 1$ , which entails that the planned lead time must be at least one time period; this is restrictive in the multi-station setting, as the time period has to be set less than the minimum planned lead time of all the stations in the network.

In contrast, the TPM's time period needs to be long to justify the assumption of Markovian job movements. For Markovian workflow, we assume that each station processes a relatively stable mix of jobs in each time period, so that the subsequent flow to downstream stations is stable as well. The validity of this assumption depends on the period length. If only a few jobs

are completed in each period, then it is unlikely that the output is very stable. Therefore this assumption would be more valid if we were able to set a longer time period to allow more jobs to be processed. A longer time period, which results in more jobs processed per period by each station, also allows us to invoke the Central Limit Theorem as justification for modeling the production quantity in each period as a normally-distributed random variable.

We also need to consider how the time period for the model matches that for the planning system. Most production systems utilize time buckets of equal intervals for their time-phased planning systems; a time bucket represents the underlying time unit for the key planning and control decisions. For instance production quantities and production capacity are measured in units per time bucket. The decisions on job release are made for each time bucket, and re-planning is done on the same time interval. The choice of time bucket should correspond to the frequency with which the production plan changes, due to control decisions, job movements or stochastic events. The length of the time bucket is typically a day or a week, but can be as short as a work shift in some contexts.

The time period of the TPM should equal the planning time bucket for the model to be useful for production planning. There are three reasons for this. First, with the matching of both time intervals, the model output would be more relevant to production planning; for instance, the probability of capacity shortfall over one-day time buckets would be more meaningful to the production planner than say, over one-hour periods. Second, from our experience in applying the model, the matching of the model time period to the planning time bucket will greatly facilitate the characterization of the model inputs using historical production data, due to the simple fact that most production data corresponds to the planning time bucket. Third, having the same time periods opens up opportunities to extend the TPM to incorporate key aspects of production plan-

ning into the workflow model; for example, this permits consideration of the demand planning and master scheduling processes, which are typically defined in time buckets.

### 3. Model

In this section, we develop a single-station model to overcome the period-sizing limitations of the TPM. We first derive a linear production function that accommodates more frequent arrivals to the station; as a consequence, we can now permit work to flow through more than one station within a time period. We then find the first two moments of the production requirement and the queue length using the derived production function.

Without loss of generality, we start with a discrete time period, indexed by  $t$ , with length of one time unit. The length of this time period would equal that for the planning time bucket. We sub-divide each time period  $t$  into  $m$  equal subintervals, and will use index  $s$ , where  $s = 1, 2, \dots, m$ , to denote the sub-periods. We define  $m = 1/\Delta$ , where  $\Delta$  is the length of each sub-period.

We assume that work can arrive at the start of each sub-period. We also assume that we set the production in each sub-period according to a linear control rule. Thus, we control the production according to a finer time grid, and we allow for more fluid arrivals to the station.

We restate the control rule (1) for each sub-period  $s$  as

$$Y(\Delta, s) = \alpha\Delta X(\Delta, s) \quad \text{for } s = 1, 2, \dots, m \quad (13)$$

where  $Y(\Delta, s)$  is the production in sub-period  $s$  of length  $\Delta$ ,  $X(\Delta, s)$  is the queue length at start of sub-period  $s$  and  $\alpha$  is the smoothing parameter. Similar to the TPM, we interpret  $1/\alpha$  as the planned lead time; however, we now permit  $\alpha$  to assume any positive value, and thus, we permit the planned lead time to be less than one time period.

We now proceed to develop an expression for  $P_t$  in terms of  $Q_t$  and  $A_t$ . These variables have the same definition as in TPM:  $P_t$  is the production in period  $t$ ,  $Q_t$  is the queue length at the

start of period  $t$ , and  $A_t$  is the arrival of work to the station in period  $t$ . However, we now assume that  $A_t$  does not arrive at the start of the period, but rather arrives uniformly over period  $t$ . In particular, we assume that the arrival amount at the start of each sub-period is equal to  $A_t/m$ . 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 we define the time period to be long compared to the order of the inter-arrival times, as this will result in a smoother spread of the workload within each period.

We have the following boundary condition for the queue length at the start of the first sub-period within each time period:

$$X(\Delta, s = 1) = Q_t + A_t/m \quad (14)$$

For  $s = 2, \dots, m$ , we model the queue length at the start of sub-period  $s$  by the standard balance equation

$$X(\Delta, s) = X(\Delta, s - 1) - Y(\Delta, s - 1) + A_t/m \quad (15)$$

By substituting (13) into (15), we obtain for  $s = 2, \dots, m$ :

$$X(\Delta, s) = (1 - \alpha\Delta) X(\Delta, s - 1) + A_t/m \quad (16)$$

Since the production in a time period is the sum of the production over the sub-periods, we note that

$$P_t = \sum_{s=1}^m Y(\Delta, s) = \alpha\Delta \sum_{s=1}^m X(\Delta, s) \quad (17)$$

We use (14) and (16) to find

$$\sum_{s=1}^m X(\Delta, s) = Q_t + (1 - \alpha\Delta) \sum_{s=1}^{m-1} X(\Delta, s) + A_t$$

from which we find:

$$\alpha\Delta \sum_{s=1}^m X(\Delta, s) = Q_t + A_t - (1 - \alpha\Delta)X(\Delta, m) \quad (18)$$

We now substitute (18) into (17) to get

$$P_t = Q_t + A_t - (1 - \alpha\Delta)X(\Delta, m) \quad (19)$$

From (19), in order to get an expression for  $P_t$ , we need to find  $X(\Delta, m)$ . From (16) and repeated substitution, we can write

$$X(\Delta, m) = (1 - \alpha\Delta)^{m-1} \times Q_t + \left(1 + (1 - \alpha\Delta) + \dots + (1 - \alpha\Delta)^{m-1}\right) \times \frac{A_t}{m} \quad (20)$$

We can use (20) to re-write (19) as

$$P_t = \beta(\Delta)Q_t + \gamma(\Delta)A_t \quad (21)$$

where

$$\beta(\Delta) = 1 - (1 - \alpha\Delta)^m$$

and

$$\gamma(\Delta) = 1 - \left(\frac{1 - \alpha\Delta}{\alpha}\right) \left(1 - (1 - \alpha\Delta)^m\right) = 1 - \beta(\Delta) \left(\frac{1 - \alpha\Delta}{\alpha}\right)$$

Thus, we have a linear control rule for the production in each time period, where now the production depends on both the work queue at the start of the period plus the arrivals during the period. The coefficients for the linear rule depend on the size of the subintervals.

To get some insight into the structure and behavior of this model, we examine the continuous-time limits for  $\beta(\Delta)$  and  $\gamma(\Delta)$  as the length of the sub-period goes to zero. This corresponds to a continuous-time control that is independent of  $\Delta$ ; in effect, we assume that (13) holds at every instant in time. We obtain the continuous-time limit of  $\beta(\Delta)$ :

$$\begin{aligned}
\beta &= \lim_{\Delta \rightarrow 0} \beta(\Delta) = \lim_{\Delta \rightarrow 0} [1 - (1 - \alpha\Delta)^m] \\
&= \lim_{\Delta \rightarrow 0} [1 - (1 - \alpha\Delta)^{1/\Delta}] \\
&= 1 - e^{-\alpha}
\end{aligned}$$

For  $\gamma(\Delta)$ , we find that

$$\begin{aligned}
\gamma &= \lim_{\Delta \rightarrow 0} \gamma(\Delta) = \lim_{\Delta \rightarrow 0} \left\{ 1 - \left( \frac{1 - \alpha\Delta}{\alpha} \right) (1 - (1 - \alpha\Delta)^m) \right\} \\
&= 1 - \lim_{\Delta \rightarrow 0} \left( \frac{1 - \alpha\Delta}{\alpha} \right) + \lim_{\Delta \rightarrow 0} \frac{(1 - \alpha\Delta)(1 - \alpha\Delta)^m}{\alpha} \\
&= 1 - \frac{1}{\alpha} + \frac{1}{\alpha} e^{-\alpha} \\
&= 1 - \frac{1}{\alpha} (1 - e^{-\alpha}) = 1 - \frac{\beta}{\alpha}
\end{aligned}$$

We can now restate (21) for the continuous-time control as:

$$P_t = \beta Q_t + \gamma A_t \quad (22)$$

where  $\beta$  and  $\gamma$  are given above.

The balance equation for the queue length for the single station is now given by:

$$Q_t = Q_{t-1} - P_{t-1} + A_{t-1} \quad (23)$$

Equation (23) differs from the balance equation in (2) of the TPM because of the new assumption that arrivals occur continuously throughout a period. Hence, we define  $Q_t$  to be the queue length at the start of period  $t$ , *prior* to any arrivals in period  $t$ ; in contrast, for the TPM, we had defined  $Q_t$  to be the queue length *after* all arrivals in that period. By substituting (22) into (23) and repeated substitution, we obtain:

$$Q_t = (1 - \gamma) \sum_{i=1}^{\infty} (1 - \beta)^{i-1} A_{t-i} \quad (24)$$

If we assume that the arrivals are i.i.d. with mean  $\mu$  and variance  $\sigma^2$ , then we find the two moments for the queue length from (24):

$$E[Q_t] = \left( \frac{1-\gamma}{\beta} \right) \mu = \frac{\mu}{\alpha} \quad (25)$$

$$Var(Q_t) = \frac{(1-\gamma)^2 \sigma^2}{2\beta - \beta^2} \quad (26)$$

Similarly, we obtain the two moments for the production variable:

$$E[P_t] = \mu \quad (27)$$

$$Var(P_t) = \left( \frac{\beta}{2-\beta} (1-\gamma)^2 + \gamma^2 \right) \sigma^2 \quad (28)$$

We note that the expected queue length (25) and expected production (27) for our continuous-time model are identical to that of the TPM (5 and 7). Furthermore, we see by Little's Law that the expected lead time corresponds to the planned lead time, namely  $1/\alpha$ . The variances (26 and 28), however, are different from the TPM's variances (6 and 8). The queue variances differ due to the different definitions of  $Q_t$ , while the production variances are different due to our uniform workflow assumption; we will discuss the production variances of both models in the next section.

## 4. Discussion

We have determined the production function that enables us to define a time period that corresponds to the time buckets in typical planning systems. In particular, we have derived a continuous-time production function that allows work to travel through more than one station within a single time period. With the relaxation of the period sizing limitation, we are able to match the time period with the planning time bucket. As a result, we are able to extend the TPM to include

the important parameters of production planning; Teo (2006) employs the continuous-time function to model the planning process of converting customer orders into a master schedule, and subsequently into job releases for a multistage shop. We review and discuss the proposed model in the rest of this section.

We have expressed (22) in terms of the parameters  $\beta$  and  $\gamma$ , both of which are functions of  $\alpha$ . For small values of  $\alpha$  (i.e., long planned lead times), we see that  $\beta$  approaches  $\alpha$  and  $\gamma$  approaches zero. Hence for small values of  $\alpha$ , the continuous-time production rule in (22) becomes the TPM linear rule in (1). However, it is less clear how the rule behaves for large values of  $\alpha$  (i.e., short planned lead times). To get some insights into this, we graph  $\beta$ ,  $\gamma$ , and  $\text{Var}(P_t)$  as functions of the smoothing parameter  $\alpha$  in Figure 1 (with  $\sigma = 1$ ). We observe that all three values increase with  $\alpha$ . The coefficients  $\beta$  and  $\gamma$  correspond to the fraction of the current queue and arrivals, respectively, that are produced in the current period. We see that both of these approach one as we increase  $\alpha$ , or equivalently as we reduce the planned lead time. The increase in  $\text{Var}(P_t)$  signifies that the production becomes more variable as we perform less smoothing with shorter planned lead times. Nevertheless, there is still substantial smoothing over this range of  $\alpha$ , as no smoothing would equate  $\text{Var}(P_t) = 1$ .

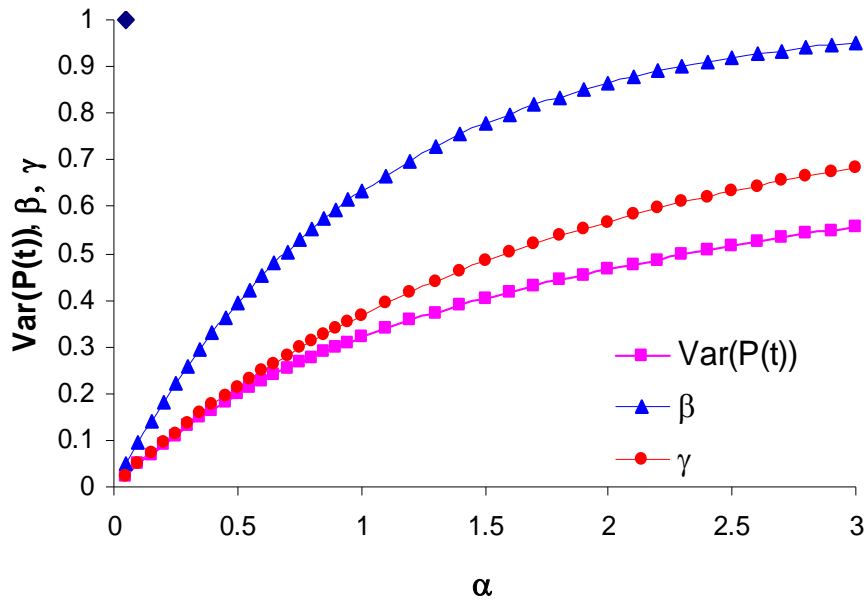


Figure 1.  $\beta$ ,  $\gamma$ , and  $\text{Var}(P_t)$  as functions of  $\alpha$  (with  $\sigma = 1$ )

We illustrate the tradeoff between production smoothness and inventory in Figure 2, in which we graph both  $\text{Var}(P_t)$  and  $E[Q_t]$  for the arrival parameters  $\mu = 1$  and  $\sigma = 1$ . We observe that as we reduce  $\alpha$ ,  $\text{Var}(P_t)$  decreases but the expected queue length  $E[Q_t]$  grows. For given arrival variability, as we reduce the planned lead time, the production requirement becomes more variable and more capacity is required but it results in less WIP inventory; alternatively, as we smooth production by increasing the planned lead time, we would need less capacity but it leads to more WIP inventory. These insights are the same as for the TPM. But the model given here is more general in that we allow continuous arrivals to the station and have no restrictions on the size of the planned lead time.

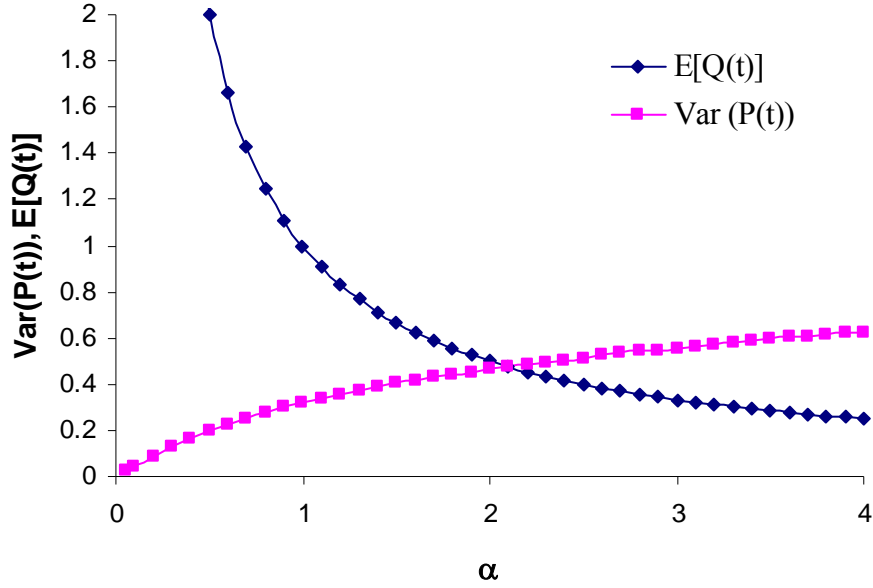


Figure 2.  $\text{Var}(P_t)$  and  $E[Q_t]$  as functions of  $\alpha$  ( $\mu = 1, \sigma = 1$ )

The continuous-time model assumes continual arrivals to the work station as well as instantaneous adjustments of the production output; in contrast, the TPM assumes arrivals and the setting of the production level only at the start of each period. In order to gain insight into how each model captures the production smoothing, we compare the production variance  $\text{Var}(P_t)$  of the continuous-time model with that of the TPM for a single-station setting. Since the TPM requires  $0 < \alpha \leq 1$ , we only perform the comparison for this range of  $\alpha$ . We graph  $\text{Var}(P_t)$  of the two models for  $\sigma = 1$  in Figure 3. We observe that  $\text{Var}(P_t)$  of the continuous-time model is always smaller than that of the TPM. We also see that the difference is small when the planned lead time is large, say greater than three time periods; but for shorter planned lead times, there is an increasing difference between the two models.

Why does  $\text{Var}(P_t)$  for the continuous-time model grow less quickly relative to that for the TPM as we shorten the planned lead time? The control rule of the TPM in (1) defines the produc-

tion requirement in a period as a fixed proportion of the queue length *after* all work arrives at the start of each period. In contrast, in the continuous-time model, we assume that work arrives uniformly within each period even though work arrivals can vary between periods. Since the instantaneous production requirement is a fixed proportion of the queue length (13), the production requirement changes gradually within the period in response to a change in the amount of work arrivals. Therefore the production requirement of the continuous-time model is less sensitive to variability in the work arrival for a single station.

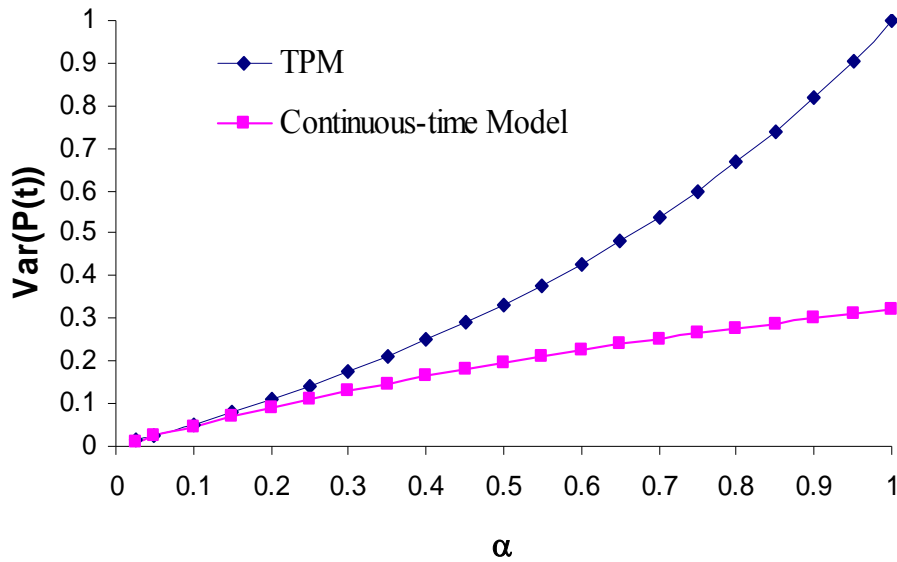


Figure 3. Comparison of  $\text{Var}(P_t)$  between both models

We can extend the single-station model to a multi-station network for which each work-station produces according to the continuous-time production rule in (22). We first restate (22) and (23) in matrix form:

$$P_t = FQ_t + GA_t \quad (29)$$

$$Q_t = Q_{t-1} - P_{t-1} + A_{t-1} \quad (30)$$

where matrix  $F$  and  $G$  are diagonal matrices with  $\beta_i$  and  $\gamma_i$  on the diagonal. To model the arrival stream at each station, we note the assumption of continuous work arrival implies that the arrival stream in the current period is generated by upstream production in the same period; consequently, we model the arrival stream to station  $i$  by  $A_{it} = \sum_j \phi_{ij} P_{jt} + \varepsilon_{it}$ ; we state its matrix form as

$$A_t = \Phi P_t + \varepsilon_t \quad (31)$$

From (29) and (31), we find that

$$\begin{aligned} P_t &= FQ_t + G \times (\Phi P_t + \varepsilon_t) \\ (I - G\Phi)P_t &= FQ_t + G\varepsilon_t \\ P_t &= (I - G\Phi)^{-1}(FQ_t + G\varepsilon_t) \end{aligned} \quad (32)$$

We substitute (31) into (30) and obtain

$$Q_t = Q_{t-1} - (I - \Phi)P_{t-1} + \varepsilon_{t-1}$$

By substituting (32) for  $P_{t-1}$ , we find that

$$\begin{aligned} Q_t &= Q_{t-1} - (I - \Phi)(I - G\Phi)^{-1}(FQ_{t-1} + G\varepsilon_{t-1}) + \varepsilon_{t-1} \\ &= \left( I - (I - \Phi)(I - G\Phi)^{-1}F \right) Q_{t-1} + \left( I - (I - \Phi)(I - G\Phi)^{-1}G \right) \varepsilon_{t-1} \\ &= BQ_{t-1} + H\varepsilon_{t-1} \end{aligned} \quad (33)$$

where  $B$  and  $H$  are as defined. By iterating (33), we obtain an expression for  $Q_t$  in terms of  $\varepsilon_t$ :

$$Q_t = \sum_{j=1}^{\infty} B^{j-1} H \varepsilon_{t-j}$$

We prove in Teo (2006) that the spectral radius of  $\Phi$  needs to be less than one for the series to converge. The expected queue length is given by

$$\begin{aligned}
E[Q_t] &= (I - B)^{-1} H \mu \\
&= F^{-1} (I - G\Phi)(I - \Phi)^{-1} \left( I - (I - \Phi)(I - G\Phi)^{-1} G \right) \mu \\
&= F^{-1} (I - G\Phi)(I - \Phi)^{-1} \mu - F^{-1} G \mu \\
&= F^{-1} (I - G)(I - \Phi)^{-1} \mu
\end{aligned} \tag{34}$$

For the last expression, we use the fact that

$$\begin{aligned}
(I - G\Phi)(I - \Phi)^{-1} &= (I - G\Phi) \sum_{j=0}^{\infty} \Phi^j \\
&= \sum_{j=0}^{\infty} \Phi^j - G \sum_{j=1}^{\infty} \Phi^j \\
&= \sum_{j=0}^{\infty} \Phi^j - G \sum_{j=0}^{\infty} \Phi^j + G \\
&= (I - G)(I - \Phi)^{-1} + G
\end{aligned}$$

We determine the variance of queue length by

$$R = \text{Var}(Q_t) = \sum_{j=1}^{\infty} B^{j-1} H \Sigma B^{j-1} H'$$

We now determine the first two moments for the production vector. We use (34) to get the expectation of the production vector (32)

$$\begin{aligned}
E[P_t] &= (I - G\Phi)^{-1} (B E[Q_t] + G \mu) \\
&= (I - G\Phi)^{-1} \left( B B^{-1} (I - G)(I - \Phi)^{-1} + G \right) \mu \\
&= (I - G\Phi)^{-1} \left( (I - G)(I - \Phi)^{-1} + G \right) \mu \\
&= (I - G\Phi)^{-1} (I - G\Phi)(I - \Phi)^{-1} \mu \\
&= (I - \Phi)^{-1} \mu
\end{aligned}$$

We note that the expected production vector  $E[\mathbf{P}_t]$  is identical to that of the TPM in (12). We find the second moments by taking the variance of production requirement in (32); we find that

$$\mathbf{S} = \text{Var}(\mathbf{P}_t) = \mathbf{M}\mathbf{R}\mathbf{M}' + \mathbf{N}\mathbf{\Sigma}\mathbf{N}' \quad (35)$$

where  $\mathbf{M} = (\mathbf{I} - \mathbf{G}\mathbf{\Phi})^{-1}\mathbf{F}$  and  $\mathbf{N} = (\mathbf{I} - \mathbf{G}\mathbf{\Phi})^{-1}\mathbf{G}$ . The power series  $\mathbf{S}$  can be computed analytically or numerically using the methods for the TPM that are discussed in Graves (1986).

We assume in our model that the arrivals within each period are uniform in order to achieve a significantly more tractable model. In particular, we assume that the arrival stream is not highly varying within each period and the time period is long compared to the inter-arrival times. But in some production systems, the arrival of workload is “lumpy”, e.g. in systems with diverse batch sizes, and as such, the arrivals cannot be made more uniform by any reasonable period length. Therefore an important enhancement to the model would be to model the flow variability within each time period.

## Acknowledgments

This research has been supported by the Singapore-MIT Alliance, an engineering education and research collaboration among the National University of Singapore, Nanyang Technological University, and MIT.

## References

- Cruickshanks, A. B., R. D. Drescher and S. C. Graves. 1984. A Study of Production Smoothing in a Job Shop Environment. *Mgmt. Sci.*, **30**, 368-380.
- Fine, C. H. and S. C. Graves. 1989. A Tactical Planning Model for Manufacturing Subcomponents of Mainframe Computers. *J. Manuf. and Opns. Mgmt.* **2**, 1, 4-34.

- Graves, S. C. 1986. A Tactical Planning Model for a job shop. *Opns. Res.* **34**, 4, 522-533.
- Graves, S. C. 1988a. Determining the Spares and Staffing Levels for a Repair Depot. *J. Manuf. and Opns. Mgmt.* **1**, 2, 227-241.
- Graves, S. C. 1988b. Extensions to a Tactical Planning Model for a Job Shop. *Proceedings of the 27<sup>th</sup> IEEE Conference on Decision and Control*, Austin, Texas, December.
- Graves, S. C., D. B. Kletter, and W. B. Hetzel. 1998. A Dynamic Model for Requirements Planning with Application to Supply Chain Optimization. *Opns. Res.* **46**, 3, 35-49.
- Graves, S. C. and J. S. Hollywood. 2001. A Constant-Inventory Tactical Planning Model for a Job Shop. Working paper, January, revised March 2004, 35 pp.
- Hollywood, J. S. 2005. An Approximate Planning Model for Distributed Computing Networks. *Naval Res. Logistics.* **52**, 6, 590-605.
- 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, Singapore.