A CONSTANT-INVENTORY TACTICAL PLANNING MODEL FOR A JOB SHOP

Stephen C. Graves
Massachusetts Institute of Technology, Cambridge MA 02139, sgraves@mit.edu

John S. Hollywood
RAND Corporation, Arlington VA 22202, johnsh@rand.org

January 2001, revised March 2004, January 2006

Abstract: We develop a constant-inventory tactical planning model for a generic manufacturing system, such as a job shop. The model is based on a discrete-time, continuous flow model developed by the first author. In the model, we assume that we can regulate the release of work to the shop to maintain the constant-inventory constraint. We characterize the conditions for which the shop production levels converge to a steady-state distribution. We then determine the first two moments for the production random vector. We illustrate the use of the model with an industry application.

Subject classifications: performance evaluation and planning for a job shop; constant WIP control; work-flow smoothing
Area of review: manufacturing, service and supply chain operations

1. Introduction

This paper discusses an analytic model for a manufacturing system that is comprised of multiple workstations and produces multiple products or services. Each product or service typically requires processing at a series of stations, and different products will have different process routes through the shop; there is no dominant flow path, and there is often significant uncertainty in process routing and workloads. A distinguishing characteristic of this generic manufacturing system is its inherent flexibility, in terms of its ability to accommodate a wide range of products and processes. However, this creates challenges for how to plan and schedule the workflow through the shop.

In this paper we extend the tactical planning model in Graves (1986) to incorporate an additional control policy. In particular we add a constraint on the total amount of work-in-process inventory in the manufacturing system. Thus, the control policy regulates the release of work to the system so as to maintain this inventory constraint.

The intent of this paper is similar to that of the Graves (1986). We wish to develop a model that can be used to understand the interrelationship and interplay of variability, capacity and work-in-process inventory in a generic manufacturing system, such as a job shop. The intent is to help managers with tactical decisions in the design, improvement and operation of a complex manufacturing system. The key tactical challenge is to find the most efficient balance of resources, in terms of capacity and work-in-process inventory, needed to satisfy customer requirements, specified in terms of volume of work and flow times. Beyond Graves (1986), we now consider the tactic of regulating the input flow on work released to the shop; we wish to understand how this tactic can help to improve the operation of a shop.

We expect this model to be used in an exploratory fashion to identify high leverage points for improving the performance of a shop, and for addressing various what if questions. For instance, what is the impact on the manufacturing system of imposing a limit on the work-in-process inventory, and how best should one set this limit? What is the impact from increasing capacity at one workstation in the shop? What additional capacity is needed to reduce the flow times through the shop? These are the types of questions we hope to explore with the model.

But our intent is to develop a simple model that is easy to implement and to use. In this respect we make simplifying assumptions so as to keep the model analytically tractable and parsimonious in terms of required data. As such, the model provides first-order analyses and answers to the questions posed above, and we expect the model to be used, in a complementary fashion, with more detailed models or simulations that would verify the findings and results from our first-order model.

Keeping the inventory in a shop constant is an increasingly popular strategy to control inventory and work-in-progress levels, along with production variability, and has been studied in detail. Relevant work on constant work-in-progress rules includes that of that of Spearman, Hopp and Woodruff (1989, 1990), Duenyas and Hopp (1993), Duenyas, Hopp and Spearman (1993), Gstettner and Kuhn (1996), Herer and Masin (1997), Hopp and Roof (1998), Dar-El, Herer and Masin (1999), Ryan, Baynat and Choobineh (2000), and Ryan and Choobineh (2003). Framinan, Gonzalez and Ruiz-Usano (2003) provide a comprehensive review of the research literature on CONWIP systems.

For the purposes of tactical planning, a common approach is to model a job shop as a network of queues. Jackson (1957, 1963) developed the basic models for open queuing networks, now known as Jackson networks. Gordon and Newell (1967) extended the analysis of

Jackson networks to a closed queuing system, whereby the number of jobs in the system remains constant. Kelly (1979) provides a comprehensive treatment and generalization. Conway and Georganas (1989) discuss a number of queuing network models, and numerical algorithms to calculate their values. Most of these models assume that arrivals and/or processing times are exponentially distributed, and assume that workstations process work at a constant rate.

The tactical planning model (TPM) in Graves (1986) is a discrete-time linear-systems model of a job shop that determines the first two moments of the production random variables, given a set of planned lead-times for the workstations. Rather than track individual jobs (as in a Jackson network), the model tracks workloads at each station, and assumes that work arrivals to a station are linear functions of work completed at upstream stations. Fine and Graves (1989) apply the TPM to an IBM manufacturing system that produced components for mainframe computers, and show how to modify the model to account for a control policy on the input mix to the shop. Leong (1987) shows how to adapt the TPM to permit a Kanban control policy in which downstream stations pull work from upstream stations. Graves (1988a) extends the one-station model to permit workstation failures, to incorporate variability due to lot-sizing, and to include an explicit capacity constraint. Mihara (1988) extends the TPM to multiple-station systems in which each station fails in accordance with a Bernoulli process. Hollywood (2005) develops the TPM for the case of non-linear production rules for modeling distributed computing networks. He employs a linearization approach to obtain second-order and first-order approximations of the expectations and of the variances, respectively, for the production and queue length random variables. He then shows the effectiveness of these approximations compared to a simulation.

In this paper, we show how to extend the TPM to incorporate a constraint on the work-in-process in the shop, and show how this extension significantly reduces production variability in the shop. In section 2, we review the TPM. We then present and analyze the constant-inventory model in section 3. We report in section 4 on an application to illustrate how the constant-inventory model can be used. In section 5 we show how to set the parameters for the constant-inventory model so that it has the same expected throughput and work-in-process as a given TPM. In section 6 we test the constant-inventory model on an example job shop to demonstrate the advantages of the model over both the TPM and a constant-release model where we keep the input to a station constant. We conclude in section 7 with a few ideas about next steps.

2. Review of the Tactical Planning Model

The tactical planning model (TPM) is a discrete-time model of the work-flow through a network of workstations, where there are no restrictions on the flow structure. We assume an underlying time period for the model and express the arrival of work per period in terms of time units (e.g. hours) of work rather than individual jobs. We model production per period at a workstation as the amount of work performed rather than as the number of jobs completed. Individual jobs have no identity in the model.

Each station uses a linear control rule to determine the amount of work to perform each period. The rule is:

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

where P_{it} is the production of work station i in time period t, Q_{it} is the work-in-process or work-in-queue at the start of period t, and the parameter $\alpha_i, 0 < \alpha_i \leq 1$, is a smoothing parameter. This rule states that the production at workstation i is a fixed portion (α_i) of the queue of work

remaining at the start of the period. We discuss the motivation for this control rule at the end of the section. We model the queue level Q_{it} by a standard inventory balance equation:

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

where A_{it} is the amount of work that arrives at workstation i at the start of period t. By using the control rule (1) to replace Q_{it} in the balance equation (2), we get a simple smoothing equation:

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

The next step in the model development is to characterize the work arrivals. A workstation receives two types of arrivals: new jobs that have their first processing step at the station, and in-process jobs that have just completed processing at an upstream station. We model the arrivals to a station from another station by the equation:

$$A_{ijt} = \phi_{ij} P_{j,t-1} + \varepsilon_{ijt}\,. \tag{4}$$

A_{ijt} is the amount of work arriving to station i from station j at the start of period t, ϕ_{ij} is a positive scalar, and ε_{ijt} is a random variable. We assume that one unit (e.g. hour) of work at station j generates ϕ_{ij} time units of work at station i, on average. ε_{ijt} is a noise term that introduces uncertainty into the relationship between production at j and arrivals to i; we assume this term is a serially i.i.d. random variable with zero mean and a known variance.

Then, the arrival stream to station i is given by:

$$A_{it} = \sum_j A_{ijt} + N_{it}\,, \tag{5}$$

where N_{it} is an i.i.d. random variable for the workload from new jobs that enter the shop at station i at time t. Substituting for A_{ijt}, we find:

$$A_{it} = \sum_j \phi_{ij} P_{j,t-1} + \varepsilon_{it}, \text{ where } \varepsilon_{it} = N_{it} + \sum_j \varepsilon_{ijt} . \quad (6)$$

Note that ε_{it} represents arrivals that are not predictable from the production levels of the previous period, and consists of work from new jobs and noise in the flow. By assumption, the time series ε_{it} is independent and identically distributed over time.

To present the analysis for the TPM, we rewrite the equations for production (3) and for arriving work (6) in matrix-vector form:

$$\mathbf{P_t} = \mathbf{(I - D)P_{t-1} + DA_t} , \quad (7)$$

$$\mathbf{A_t} = \mathbf{\Phi\, P_{t-1} + \varepsilon_t} . \quad (8)$$

where $\mathbf{P_t} = \{P_{1t},...,P_{nt}\}'$, $\mathbf{A_t} = \{A_{1t},...,A_{nt}\}'$, and $\boldsymbol{\varepsilon_t} = \{\varepsilon_{1t},...,\varepsilon_{nt}\}'$ are column vectors of random variables, n is the number of workstations, **I** is the identity matrix, **D** is a diagonal matrix with $\{\alpha_1,\ldots,\alpha_n\}$ on the diagonal, and $\mathbf{\Phi}$ is an n-by-n matrix with elements ϕ_{ij}. By substituting equation (8) into equation (7), we obtain a recursion:

$$\mathbf{P_t = (I - D + D\Phi)P_{t-1} + D\varepsilon_t} . \quad (9)$$

By iterating this equation and assuming an infinite history of the system, we rewrite $\mathbf{P_t}$ as an infinite series:

$$\mathbf{P_t} = \sum_{s=0}^{\infty} \mathbf{(I - D + D\Phi)^s D\varepsilon_{t-s}} . \quad (10)$$

We denote the mean and the covariance for the noise vector $\boldsymbol{\varepsilon_t}$ by $\boldsymbol{\mu} = \{\mu_1,\ldots,\mu_n\}'$, and $\mathbf{\Sigma} = \{\sigma_{ij}\}$, respectively. The first two moments of $\mathbf{P_t}$ are given by:

$$E[\mathbf{P_t}] = \sum_{s=0}^{\infty} (\mathbf{I - D + D\Phi})^{\mathbf{s}} \mathbf{D\mu} = (\mathbf{I - \Phi})^{-\mathbf{1}} \boldsymbol{\mu}, \tag{11}$$

and $$\mathbf{S} = \text{var}(\mathbf{P_t}) = \sum_{s=0}^{\infty} \mathbf{B}^{\mathbf{s}} \mathbf{D\Sigma D B'}^{\mathbf{s}}, \text{ where } \mathbf{B = I - D + D\Phi}. \tag{12}$$

provided that $\rho(\mathbf{\Phi}) < 1$, where $\rho(\mathbf{\Phi})$ denotes the spectral radius of $\mathbf{\Phi}$ (see Graves 1986). We note that **S** provides the production variance for each station, as well as the covariance for each pair of workstations.

Discussion: The assumption of a linear control rule (1) merits some discussion and justification. In the following we offer several points to clarify our intent and provide support for this assumption. Graves (1986) provides additional motivation for this control rule.

First, we intend for a simple model that is tractable and easy to apply. As summarized above, the linear control rule permits a fairly general model that entails just linear algebra and is readily implemented with MATLAB. We obtain the first two moments for the random variables for production and queue length at each work station with no distributional assumptions and no assumptions on the flow matrix other than $\rho(\mathbf{\Phi}) < 1$.

Second, we find that this rule is somewhat descriptive of actual behavior in many shops. Work output often does increase as the work-in-process increases in manufacturing and service systems. For example, Fine and Graves (1989) found a statistically significant linear relationship between station work-in-process and production at an IBM mainframe subcomponents plant; more recently, Holly (2006) finds the same at various work stations in an Intel semiconductor fabrication facility. In white-collar work systems, such as product development, people are the limiting resources, and can and do adjust their output rate as their work backlog grows. In machine shops, when the work-in-process grows too large, a manager might adjust the output

rate by various means: adding overtime, postponing preventative maintenance and training, running larger batch sizes. When the backlog drops, the manager cuts overtime and reinstates non-productive activities and events. Nevertheless, we would be hard pressed to argue that output is strictly linear over the entire range of possibilities. However, as with any model, we make simplifying assumptions so as to get a first-order understanding of the major effects under consideration.

Third, we find that many operations, both in manufacturing and white-collar settings, continue to operate with planned lead times. That is, they plan the flow of work through the operation by first decomposing the flow into unit steps each of which has a planned lead time. In effect, each unit step is allocated a fixed amount of time to transform its inputs into outputs. Questions arise as to how these planned lead times should be set, and what amount of capacity is required to assure that these times can be met reliably. The TPM provides an approach to this, as we interpret $(1/\alpha_i)$ to be the planned lead time for station i.

Fourth, related to the prior comment, we need to consider how we intend to use the model. The analysis does not explicitly account for any capacity constraints on the output of a workstation. Rather, we find the moments for the production random variable for each work station, e.g., equations (11) and (12), as a function of the smoothing parameter for the linear production rule. Given the nominal capacity of each workstation, the intent is to set the smoothing parameters, or equivalently the planned lead times, so that the resulting production random variables are consistent with the nominal capacity. Thus, the intent is to implicitly incorporate capacity limits into the model by means of the planned lead times. We illustrate this in the later sections of this paper.

Fifth, a linear control rule has desirable theoretical properties. Hollywood (2000) demonstrates that the optimal control rule for minimizing a weighted sum of production and work-in-process variances over a multi-period planning horizon (including infinite horizon) is a linear function of the work-in-process at multiple stations.

Finally, we have found that linear rules, particularly for production smoothing, have been helpful in practice. We have reported our efforts with using linear rules in a number of industrial projects: Cruickshanks et al. (1984), Graves et al. (1986), Graves (1988b), Fine and Graves (1989), and Graves et al. (1998). Our experience is that such models can be useful at highlighting the key tradeoffs and for evaluating counter measures for handling variability.

3. The Constant Inventory Model

In this section, we develop the constant-inventory tactical planning model (CIM). For this model we require that the weighted inventory of the shop is equal to some constant, W:

$$\sum_i w_i Q_{it} = W, \forall t . \tag{13}$$

Here, w_i is a non-negative weight for the work-in-process at station i. The units for the work-in-process Q_{it} are the time units of work at station i; thus, the weights reflect how to combine workloads at different stations into a common measure of inventory or workload for the shop.

We can use (13) to achieve an additional degree of control on the workflow through a shop. To enforce the constraint we need to assume that we can control the work release into the shop. There is extensive prior evidence that regulating the release of work to a shop provides great benefits in terms of either reduced and less variable flow times for a given throughput level, or increased throughput for a fixed level of work-in-process. For instance, we can use (13)

to incorporate into the TPM a constant work-in–process, or CONWIP, control policy, as proposed by Spearman et al. (1989, 1990); see also Hopp and Spearman (2001), who provide both a compelling justification for its use and strong evidence of its merits. We can also use (13) to model work-regulating control schemes, as developed by Wein (1988, 1990). Finally, we will show how to embed (13) within an optimization to determine the optimal weights for minimizing a measure of production variability.

Since we assume a linear control rule (1) for setting production, we can also use (13) to model any linear constraint on the production rates, e. g.,

$$\sum_i v_i P_{it} = V, \forall t .$$

For instance, if the shop has a single bottleneck, we might impose a simple constraint that specifies the bottleneck's production rate and then use the model to characterize the workflow throughout the shop. Another application is when v_i is the labor requirements per unit of work at station i, and there is a finite pool of shared labor V.

To adapt the TPM to include the constraint (13), we assume that the initial work-in-queue Q_{i0} satisfies (13). Then to assure (13) for all time periods, we must have that

$$\sum_i w_i \left(Q_{it} - Q_{i,t-1} \right) = 0, \forall t . \tag{14}$$

Recall from (2) that the differences in queue levels equal arrivals of new work less completed work. Then we can rewrite the above equation as:

$$\sum_i w_i \left(A_{it} - P_{i,t-1} \right) = 0, \forall t . \tag{15}$$

In order to achieve constant inventory, as specified by (13), we must assume that we can control the release of work to the shop in some way. Let R_{it} be the work that we release to station i at time t to maintain the constant-inventory constraint. We assume that we can set the work releases R_{it} immediately after observing the realization of the random vector $\boldsymbol{\varepsilon_t}$, which corresponds to the exogenous arrivals to the system plus random noise in the workflow between stations. The arrival random variable to workstation i at the start of time t is now:

$$A_{it} = \sum_j \phi_{ij} P_{j,t-1} + N_{it} + \sum_j \varepsilon_{ijt} + R_{it}, \tag{16}$$

where the variables have the same definitions as for the original TPM. We can substitute (16) into (15) to obtain:

$$\sum_i w_i \left(\sum_j \phi_{ij} P_{j,t-1} + N_{it} + \sum_j \varepsilon_{ijt} + R_{it} - P_{i,t-1} \right) = 0, \forall t\,. \tag{17}$$

We can rewrite (17) in matrix-vector form as follows:

$$\mathbf{w'R_t} = \mathbf{w'}\left(\left(\mathbf{I - \Phi}\right)\mathbf{P_{t-1}} - \boldsymbol{\varepsilon_t}\right), \tag{18}$$

where $\mathbf{w} = \{w_1, \ldots, w_n\}'$, $\mathbf{R_t} = \{R_{1t}, \ldots, R_{nt}\}'$, and the other vectors and matrices are the same as in the TPM model. Thus, in order to satisfy (13), we must determine the work release vector $\mathbf{R_t}$ so as to satisfy (18) in each period. Unfortunately, we have an under-specified requirement: n unknowns and only one equation. For the purpose of analysis, we introduce two simplifying assumptions.

Proportional Releases: We assume that in each period we set $\mathbf{R_t}$ as follows:

$$\mathbf{R_t} = r_t \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix} = r_t \boldsymbol{\beta}. \tag{19}$$

We interpret the scalar r_t as the control variable, equal to the amount of work that is released to the shop at the start of time period t in order to maintain the constant-inventory constraint. Then the vector $\boldsymbol{\beta}$ is a template that prescribes how to divide this aggregate work release across the workstations, where $\boldsymbol{\beta} \geq 0$; specifically, station i receives $R_{it} = r_t \beta_i$ in time period t.

Controllability: We assume that $\mathbf{w'\beta} > 0$, where $\mathbf{w} = \{w_1, \ldots, w_n\}'$, and $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_n\}'$. We cannot control the system if we can only release work to stations whose inventories do not contribute to the constant-inventory constraint, as will be seen.

We substitute (19) into (18):

$$\mathbf{w'R_t} = r_t(\mathbf{w'\beta}) = \mathbf{w'}\left(\left(\mathbf{I} - \boldsymbol{\Phi}\right)\mathbf{P_{t-1}} - \boldsymbol{\varepsilon_t}\right). \tag{20}$$

Without loss of generality, given that we assume $\mathbf{w'\beta} > 0$, then we can rescale $\boldsymbol{\beta}$ so that $\mathbf{w'\beta} = 1$. Thus, we find the work-release control variable to be:

$$r_t = \mathbf{w'}\left(\left(\mathbf{I} - \boldsymbol{\Phi}\right)\mathbf{P_{t-1}} - \boldsymbol{\varepsilon_t}\right) \tag{21}$$

Now, if we substitute (19) and (21) into (16), we find in matrix form:

$$\begin{aligned} \mathbf{A_t} &= \boldsymbol{\Phi}\mathbf{P_{t-1}} + \boldsymbol{\varepsilon_t} + \mathbf{R_t}, \\ &= \boldsymbol{\Phi}\mathbf{P_{t-1}} + \boldsymbol{\varepsilon_t} + \mathbf{w'}\left(\left(\mathbf{I} - \boldsymbol{\Phi}\right)\mathbf{P_{t-1}} - \boldsymbol{\varepsilon_t}\right)\boldsymbol{\beta} \\ &= \left(\boldsymbol{\Phi} + \boldsymbol{\beta}\mathbf{w'}\left(\mathbf{I} - \boldsymbol{\Phi}\right)\right)\mathbf{P_{t-1}} + \left(\mathbf{I} - \boldsymbol{\beta}\mathbf{w'}\right)\boldsymbol{\varepsilon_t} \\ &= \mathbf{F}\mathbf{P_{t-1}} + \mathbf{G}\boldsymbol{\varepsilon_t}. \end{aligned} \tag{22}$$

where $$\mathbf{F} = \mathbf{\Phi} + \boldsymbol{\beta}\mathbf{w}'\left(\mathbf{I} - \mathbf{\Phi}\right) \text{ and } \mathbf{G} = \mathbf{I} - \boldsymbol{\beta}\mathbf{w}'. \tag{23}$$

As in the TPM, the production vector at time *t* is given by the following equation:

$$\begin{aligned} \mathbf{P_t} &= (\mathbf{I} - \mathbf{D})\mathbf{P_{t-1}} + \mathbf{DA_t}, \\ &= (\mathbf{I} - \mathbf{D})\mathbf{P_{t-1}} + \mathbf{D}\left(\mathbf{FP_{t-1}} + \mathbf{G}\boldsymbol{\varepsilon}_\mathbf{t}\right) \\ &= \left(\mathbf{I} - \mathbf{D} + \mathbf{DF}\right)\mathbf{P_{t-1}} + \mathbf{DG}\boldsymbol{\varepsilon}_\mathbf{t}. \end{aligned} \tag{24}$$

We note that (24) has the same form as (9), the recursion for the TPM. By iterating this equation, we rewrite (24) for $\mathbf{P_t}$ as the series:

$$\mathbf{P_t} = (\mathbf{I} - \mathbf{D} + \mathbf{DF})^t \mathbf{DQ_0} + \sum_{s=0}^{t-1} (\mathbf{I} - \mathbf{D} + \mathbf{DF})^s \mathbf{DG}\boldsymbol{\varepsilon}_{\mathbf{t-s}} \tag{25}$$

where $\mathbf{P_0} = \mathbf{DQ_0}$ for $\mathbf{Q_0}$ being the initial work-in-process that satisfies (13). In the following propositions we show that if the spectral radius of **F** (denoted by $\rho(\mathbf{F})$) equals one, then the first two moments of $\mathbf{P_t}$ converge as $t \to \infty$.

Proposition 1. *We assume the spectral radius of matrix* $\mathbf{\Phi}$ is less than one (i. e., $\rho(\mathbf{\Phi}) < 1$). *Suppose we choose* **w** *such that* $\mathbf{w}'\mathbf{\Phi} \le \mathbf{w}'$. *Then* $\rho(\mathbf{F}) = 1$.

Proof. We first observe that $\rho(\mathbf{F}) \ge 1$: from (23) and the assumption that $\mathbf{w}'\boldsymbol{\beta} = 1$, we see that $\mathbf{w}'\mathbf{F} = \mathbf{w}'$. Thus, **w** is the left eigenvector of **F** with a corresponding eigenvalue of 1.

To show that $\rho(\mathbf{F}) \le 1$, we note that **F** is a positive matrix:

$$\mathbf{F} = \mathbf{\Phi} + \boldsymbol{\beta}\mathbf{w}'\left(\mathbf{I} - \mathbf{\Phi}\right) = \mathbf{\Phi} + \boldsymbol{\beta}\left(\mathbf{w}' - \mathbf{w}'\mathbf{\Phi}\right) \ge \mathbf{\Phi} \ge \mathbf{0}$$

where the first inequality is due to the supposition that $\mathbf{w}'\mathbf{\Phi} \le \mathbf{w}'$ and the assumption that $\boldsymbol{\beta} \ge 0$, and the second inequality is due to the assumption that $\mathbf{\Phi}$ is a positive matrix.

Suppose that $\rho(\mathbf{F}) > 1$, and let $\mathbf{y}$ be the maximal left eigenvector of $\mathbf{F}$ such that $\mathbf{y'F} = \lambda\,\mathbf{y'}$ for $\lambda > 1$. We consider two cases.

Case 1: Suppose $\mathbf{y'\beta} = 0$. Then consider

$$\mathbf{y'F} = \mathbf{y'\Phi} + \mathbf{y'\beta w'}\left(\mathbf{I} - \mathbf{\Phi}\right) = \mathbf{y'\Phi} = \lambda\mathbf{y'}$$

Thus, $\mathbf{y}$ is a left eigenvector $\mathbf{\Phi}$ with eigenvalue $\lambda > 1$; this contradicts the assumption that $\rho(\mathbf{\Phi}) < 1$.

Case 2: Suppose $\mathbf{y'\beta} \neq 0$. Without loss of generality we can re-scale $\mathbf{y}$ so that $\mathbf{y'\beta} = 1$. Then consider

$$\mathbf{y'F} = \mathbf{y'\Phi} + \mathbf{y'\beta w'}\left(\mathbf{I} - \mathbf{\Phi}\right) = \mathbf{y'\Phi} + \mathbf{w'}\left(\mathbf{I} - \mathbf{\Phi}\right) = \lambda\mathbf{y'}$$

By rearranging terms we get

$$\left(\mathbf{y'} - \mathbf{w'}\right)\mathbf{\Phi} = \lambda\mathbf{y'} - \mathbf{w'} \geq \mathbf{y'} - \mathbf{w'},$$

where the inequality is due to the supposition that $\lambda > 1$ and the fact that $\mathbf{y} \geq 0$ since $\mathbf{F}$ is a positive matrix. This contradicts the assumption that $\rho(\mathbf{\Phi}) < 1$.

Thus, if $\rho(\mathbf{\Phi}) < 1$, we conclude that $\rho(\mathbf{F}) \leq 1$.

As we have shown that $\rho(\mathbf{F}) \leq 1$ and $\rho(\mathbf{F}) \geq 1$, we conclude that $\rho(\mathbf{F}) = 1$. ∵

Proposition 2. *Assume that* $\rho(\mathbf{F}) = 1$. *The expectation of* $\mathbf{P_t}$ *converges to a finite value as* $t \to \infty$.

Proof. Define $\mathbf{B} = \mathbf{(I - D + DF)}$. $\mathbf{B}$'s largest eigenvalue equals the largest eigenvalue of $\mathbf{F}$ (the argument is identical to that given in Graves, 1986), so $\rho(\mathbf{B}) = 1$. Further, we see by substitution that $\mathbf{w'D^{-1}}$ is the left eigenvector of $\mathbf{B}$ corresponding to $\mathbf{B}$'s maximal eigenvalue:

$$\mathbf{w'D^{-1}B} = \mathbf{w'D^{-1}(I - D + DF)} = \mathbf{w'D^{-1}} - \mathbf{w'} \;+\; \mathbf{w'F} = \mathbf{w'D^{-1}}.$$

Define $\mathbf{v_1}$ to be the corresponding right eigenvector of $\mathbf{B}$, scaled so that $\mathbf{w'D^{-1}v_1} = 1$. We then have the following decomposition of $\mathbf{B^s}$:

$$\mathbf{B^s} = 1^{\mathbf{s}}\mathbf{v_1 w'D^{-1}} + \mathbf{E^s}.$$

Here, 1 is the largest eigenvalue of $\mathbf{B}$, and $\mathbf{E}$ is an orthogonal matrix such that $\mathbf{v_1 w'D^{-1}E} = \mathbf{0}$. Further, $\mathbf{E}$ has a spectral radius strictly less than one. By substituting the expression for $\mathbf{B^s}$ into (25), we find:

$$\mathbf{P_t} = \mathbf{B}^t\mathbf{DQ_0} + \sum_{s=0}^{t-1}\left(1^{\mathbf{s}}\mathbf{v_1 w'D^{-1}DG\varepsilon_{t-s}} + \mathbf{E^s DG\varepsilon_{t-s}}\right).$$

We know that $\mathbf{B}^t\mathbf{DQ_0}$ converges to a finite value as $t \to \infty$, since the spectral radius of $\mathbf{B}$ is one. This leaves the summation. From (23), we note that $\mathbf{w'G} = \mathbf{0}$, so the first term in the summation is zero. But then, since the spectral radius of $\mathbf{E}$ is less than one, the second term must converge to a finite value.

Consequently, the expectation of $\mathbf{P_t}$ converges to the following:

$$\begin{aligned} E\left[\lim_{t\to\infty}\mathbf{P_t}\right] &= \mathbf{B}^{\infty}\mathbf{DQ_0} + \sum_{s=0}^{\infty}\mathbf{E^s DG\mu} \\ &= \mathbf{v_1 w'Q_0} + \mathbf{(I-E)^{-1}DG\mu}, \end{aligned},$$

where $\boldsymbol{\mu} = \{\mu_1, \ldots, \mu_n\}'$ is the expectation of the noise vector $\boldsymbol{\varepsilon_t}$. Since by (13) $\mathbf{w'Q_0} = W$, we can simplify the above expression:

$$E\left[\lim_{t\to\infty}\mathbf{P_t}\right] = W\mathbf{v_1} + \mathbf{(I-E)^{-1}DG\mu}. \because \tag{26}$$

Proposition 3. *Assume that* $\rho(\mathbf{F}) = 1$. *The covariance matrix of* $\mathbf{P_t}$ *converges to a finite value as* $t \to \infty$.

Proof. Define $\mathbf{B} = (\mathbf{I} - \mathbf{D} + \mathbf{DF})$. We denote the covariance matrix for $\mathbf{P_t}$ as $t \to \infty$ by **S,** given by:

$$\begin{aligned} \mathbf{S} &= \text{var}\left[\sum_{s=0}^{\infty} \mathbf{B^s DG}\varepsilon_{\mathbf{t-s}}\right], \\ &= \text{var}\left[\sum_{s=0}^{\infty} \mathbf{E^s DG}\varepsilon_{\mathbf{t-s}}\right], \text{ as shown in Proposition 2,} \\ &= \sum_{s=0}^{\infty} \mathbf{E^s DG\Sigma G'DE'^s}. \end{aligned} \tag{27}$$

Here, $\mathbf{\Sigma}$ is the covariance matrix of the noise vector. This summation converges, as the maximal eigenvalue of **E** is less than one. ∵

To understand the convergence conditions, we compare (22) with (8): the matrix **F** for the constant-inventory model is analogous to the workflow matrix $\mathbf{\Phi}$ for the TPM. For the constant-inventory model, $\mathbf{FP_{t-1}}$ is the vector of arrivals at the start of period *t* that are due to the production in period *t*-1. For the TPM, we require $\rho(\mathbf{\Phi}) < 1$ for convergence. For the constant-inventory model, this condition becomes $\rho(\mathbf{F}) = 1$, as the workflow matrix **F** accounts for the endogenous workflow, given by $\mathbf{\Phi}$**,** and the induced workflow required to assure the inventory constraint, given by $\boldsymbol{\beta}\mathbf{w'(I - \Phi)}$.

Sufficient conditions for $\rho(\mathbf{F}) = 1$ are $\rho(\mathbf{\Phi}) < 1$, and $\mathbf{w'(I - \Phi)} \geq 0$. To understand the latter condition, we interpret the (*i, j*) element of the matrix $\mathbf{(I - \Phi)}$ as the work-in-process reduction at station *i* per unit production at station *j*. Then the j^{th} element of the vector $\mathbf{w'(I - \Phi)}$ is the reduction to the weighted work-in-process in the shop per unit production at station *j*; the

condition states that a unit of production at station *j* does not increase the weighted work-in-process in the shop. This condition is similar to the convergence condition for the TPM ($\rho(\mathbf{\Phi}) < 1$), whereby a unit of production at any workstation cannot ultimately generate more than one unit of return work for that station.

4. Illustrative Example

To illustrate how one might use the constant-inventory model we present an example based on an application for the Synchronous Machine Division at ABB. Synchronous Machines engineers and manufactures custom-designed electric generators and motors for a variety of industrial markets.

We model the production of the stators, which are one of two major subassemblies required for producing an engineered-to-order electric motor. The production of the subassembly entails three production stages in series: stator body assembly, stator winding and impregnation, and stator final assembly. In Table 1 we show the expectation and standard deviation of the workload per job for each stage. In Table 1 we also show the nominal capacity available each day for each stage. The nominal capacity represents the typical staffing level for each stage of the process; however, each stage has some amount of flexibility by which it can expand its capacity on a daily basis, through use of overtime and shared resources with other operations. The data in this example have been disguised for proprietary reasons, but remain representative of the actual data from the application.

As each order requires extensive customization, ABB cannot produce to stock; rather ABB maintains an order backlog and schedules production at a pre-determined rate based on the capacity of the bottleneck for the manufacturing system, which is the assembly and test operation

for the electric motors and generators. The stator subassembly is an input to these downstream operations. At the time of study, ABB had set a production rate of 2 jobs per day, or 10 jobs per week.

	E[hours per job]	Std Dev{hrs per job]	Capacity (hrs per day)
Stator body assembly	55.6	47	150
Stator winding and impregnation	165.9	135.9	385
Stator final assembly	30.2	27.9	85

Table 1: Sample Data for Three-Stage Shop

As part of an improvement project, ABB had decided to implement a CONWIP control policy for the production of the subassembly. That is, they would maintain a constant work-in-process of *W* jobs in the three-stage system for manufacturing the subassembly. The question at hand was how to set *W*. The primary tradeoff is that a smaller value of *W* results in a lower inventory and a shorter lead-time for producing the subassembly; however, with smaller values of *W*, more capacity is required to satisfy the production rate of 2 jobs per day.

To examine this question, we built a simple spreadsheet model that could inform the implementation of the CONWIP policy, in particular the choice of *W*. The inputs for the model are as follows:

Workflow matrix $\mathbf{\Phi}$: There are only two non-zero entries, $\phi_{21} = 2.98$ and $\phi_{32} = 0.18$. That is, from Table 1 each hour of work at stator body assembly generates on average 165.9/55.6 = 2.98 hours of work at stator winding and impregnation. Similarly, we find that each hour at stator winding and impregnation yields 0.18 hours of work at stator final assembly.

Mean and Covariance of noise vector $\boldsymbol{\varepsilon}_t$: There are no exogenous arrivals, as all job releases are controlled by the CONWIP policy. Hence $E[\boldsymbol{\varepsilon}_t] = \boldsymbol{\mu} = \mathbf{0}$. We use the covariance of the noise vector to model the workload uncertainty. In particular we have:

$$\boldsymbol{\Sigma} = \mathrm{cov}(\boldsymbol{\varepsilon}_t) = \begin{pmatrix} 4418 & & \\ & 36938 & \\ & & 1556 \end{pmatrix},$$

and we assume the off-diagonal elements are zero. The diagonal elements represent the variance of the workload that arrives at each stage each day. As the production rate for the shop is 2 jobs per day, we set the diagonal elements to be the variance of the workload for 2 independent jobs. In reality, these variances are larger, as there is also variability in the daily arrivals.

Release template $\boldsymbol{\beta}$: All new jobs enter the system at stage 1, so $\boldsymbol{\beta} = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}'$.

Weight vector **w**: The CONWIP policy creates a limit on the number of jobs in the system. Thus, we set the weight vector to convert the workload at each stage, measured in hours, into jobs: the weights are the inverses of the expected hours per job, given in Table 1. We re-scale the weight vector so that $\mathbf{w}'\boldsymbol{\beta} = 1$, to get: $\mathbf{w} = \begin{pmatrix} 1 & 0.34 & 1.84 \end{pmatrix}'$.

For each stage *i*, the model requires a smoothing parameter α_i or equivalently, a planned lead-time, which we denote by $\lambda_i = 1/\alpha_i$. Rather than specify these parameters a priori, we formulate an optimization problem for which the planned lead-times (λ_i) are the decision variables. In the optimization we minimize the capacity required to meet the production rate target with a lead-time constraint.

$$\begin{aligned} Min \ & \sum_{i=1}^{3} c_i \sigma\left(P_{it}\right) \\ s.t. \quad & \lambda_i \geq 1, \forall i \\ & \sum_{i=1}^{3} \lambda_i \leq L \\ & E\left[P_{it}\right] \geq 2h_i, \forall i \end{aligned} \tag{28}$$

The decision variables for the optimization are the planned lead-times (λ_i) and the inventory limit *W*. The objective in (28) is to minimize the weighted sum of the standard deviations of the production random variables. We find the standard deviations, $\sigma(P_{it})$, from the

covariance matrix **S** in (27). The objective function weights c_i signify the relative cost for capacity at the different stages. We use the objective function as a surrogate for the cost of the required capacity.

By the first set of constraints, we assure that the planned lead-times are at least one day for each stage, or equivalently that the smoothing parameters are no greater than one. The second constraint establishes an upper bound L on the total lead-time for the subassembly. The third set of constraints states that the expected workload at each stage, as given by (26) is at least that required for two jobs per day, where h_i is the expected hours per job at stage i.

We solve (28), using a generalized reduced gradient solver in Excel, over different values for the lead-time target L to explore the trade-off between the requirements for capacity and inventory (or equivalently lead-time). In Table 2, we report illustrative results for the model, where we use a weight vector $\mathbf{c} = (\mathbf{1} \quad \mathbf{1} \quad \mathbf{2})'$ in the objective function. Each row corresponds to a solution to (28) for the specified value of the lead-time L. We state the capacity requirements for each stage as being $E(P_{it}) + k\sigma(P_{it})$, where we use $k = 1.3$ for these computations. Thus, if the noise vector is normally distributed, then the actual production is within the capacity for 90% of the days; for the remaining 10% of the days, we assume each stage relies on its flexibility options (e. g., overtime) to expand the available capacity to cover the production.

As expected, we see that the capacity requirements increase as we reduce the lead-time L and equivalently W, the number of CONWIP jobs. The current staffing plans provide a nominal capacity of 150 hours/day, 385 hours per day, and 85 hours per/day for the three stages; then one might set $W = 28$ jobs. Furthermore, one can see what increases in capacity are necessary to reduce W. We can also use the model to evaluate the benefits from reducing variability in the process times at each stage, as well as the impact from changing the production rate.

L	*W*	Stage 1 Capacity	Stage 1 Lead-Time	Stage 2 Capacity	Stage 2 Lead-Time	Stage 3 Capacity	Stage 3 Lead-Time
8	16	166	2.5	415	3.6	93	1.8
9	18	162	2.8	406	4.1	90	2.0
10	20	159	3.1	400	4.7	88	2.3
11	22	156	3.3	395	5.2	86	2.5
12	24	153	3.6	390	5.7	85	2.7
14	28	150	4.2	384	6.7	82	3.1
16	32	147	4.7	379	7.7	81	3.6

Table 2: Model Results for Three-Stage Shop

5. Creating Comparable TPM and Constant Inventory Models

To appreciate the merits of a constant-inventory control policy, we will compare its performance to that for the TPM. In effect, the constant-inventory model corresponds to a closed shop or closed queuing network, whereas the TPM corresponds to an open shop or open queuing network. In this section we show how to create comparable TPMs and constant-inventory models. We first show how to parameterize the constant-inventory model so that it will have the same expected production and queue levels, and the same expected inputs as a given TPM. We then show how to parameterize the TPM so that it will have the same expected production and queue levels, and the same expected inputs as a given constant-inventory model.

5.1 Creating a Constant-Inventory Model from a TPM

In this section, we create a constant-inventory model from a TPM, where we use subscripts *CI* and *TPM* to denote the respective models.

Proposition 4. Suppose we have a TPM with parameters $\mathbf{\Phi}$, $\mathbf{D}$, $\mathbf{\Sigma}$ and μ_{TPM}. Consider a constant-inventory model with parameters $\mathbf{\Phi}$, $\mathbf{D}$, $\mathbf{\Sigma}$, $\mathbf{w}$, β and μ_{CI}, where $\mu_{TPM} \neq \mu_{CI}$, and where we set the release template β as follows:

$$\boldsymbol{\beta} = \frac{\boldsymbol{\mu}_{\mathbf{TPM}} - \boldsymbol{\mu}_{\mathbf{CI}}}{\mathbf{w}'\left(\boldsymbol{\mu}_{\mathbf{TPM}} - \boldsymbol{\mu}_{\mathbf{CI}}\right)}. \tag{29}$$

Then, the two models have the same expected production vectors, namely $\mathbf{E[P_{CI}] = E[P_{TPM}]}$.

Proof. To show $\mathbf{E[P_{CI}] = E[P_{TPM}]}$, we note from (25) that $\mathbf{E[P_{CI}]}$ satisfies the following recursion:

$$\mathbf{E[P_{CI}]} = (\mathbf{I} - \mathbf{D} + \mathbf{DF})\mathbf{E[P_{CI}]} + \mathbf{DG}\boldsymbol{\mu}_{\mathbf{CI}}, \tag{30}$$

which we can simplify to

$$(\mathbf{I} - \mathbf{F})\mathbf{E[P_{CI}]} = \mathbf{G}\boldsymbol{\mu}_{\mathbf{CI}}.$$

We use (23) to substitute for **F** and **G**, and can rewrite this recursion as:

$$(\mathbf{I} - \boldsymbol{\Phi})\mathbf{E}\left[\mathbf{P_{CI}}\right] - \boldsymbol{\mu}_{\mathbf{CI}} = \boldsymbol{\beta}\mathbf{w}'\left[(\mathbf{I} - \boldsymbol{\Phi})\mathbf{E[P_{CI}]} - \boldsymbol{\mu}_{\mathbf{CI}}\right].$$

Now we suppose that $\boldsymbol{\mu}_{\mathbf{TPM}} \neq \boldsymbol{\mu}_{\mathbf{CI}}$ and we select $\boldsymbol{\beta}$ as given by (29); then $\mathbf{E[P_{CI}]}$ must satisfy the following recursion:

$$(\mathbf{I} - \boldsymbol{\Phi})\mathbf{E}\left[\mathbf{P_{CI}}\right] - \boldsymbol{\mu}_{\mathbf{CI}} = \left[\boldsymbol{\mu}_{\mathbf{TPM}} - \boldsymbol{\mu}_{\mathbf{CI}}\right]\frac{\mathbf{w}'\left[(\mathbf{I} - \boldsymbol{\Phi})\mathbf{E[P_{CI}]} - \boldsymbol{\mu}_{\mathbf{CI}}\right]}{\mathbf{w}'\left[\boldsymbol{\mu}_{\mathbf{TPM}} - \boldsymbol{\mu}_{\mathbf{CI}}\right]}.$$

By substitution, we easily see that $\mathbf{E[P_{CI}]} = (\mathbf{I} - \boldsymbol{\Phi})^{-1}\boldsymbol{\mu}_{\mathbf{TPM}}$ is a solution, and that, by proposition 2, this is the unique solution for the recursion. Thus, by using (29) to set $\boldsymbol{\beta}$, we have shown that $\mathbf{E[P_{CI}]} = (\mathbf{I} - \boldsymbol{\Phi})^{-1}\boldsymbol{\mu}_{\mathbf{TPM}} = \mathbf{E[P_{TPM}]}$. ∵

From (19) and (21), we can use the results of proposition 4 to find the expectation for the releases for the constant-inventory model:

$$\mathbf{E}\left[\mathbf{R_{CI}}\right] = \boldsymbol{\beta}\mathbf{w}'\left((\mathbf{I} - \boldsymbol{\Phi})\mathbf{E[P_{CI}]} - \boldsymbol{\mu}_{\mathbf{CI}}\right) = \boldsymbol{\beta}\mathbf{w}'(\boldsymbol{\mu}_{\mathbf{TPM}} - \boldsymbol{\mu}_{\mathbf{CI}}) = \boldsymbol{\mu}_{\mathbf{TPM}} - \boldsymbol{\mu}_{\mathbf{CI}}$$

Thus, the expected inputs into the constant-inventory shop, $\mathbf{E[R_{CI}]} + \boldsymbol{\mu}_{\mathbf{CI}}$, equal the inputs into the TPM shop, $\boldsymbol{\mu}_{\mathbf{TPM}}$.

5.2 Creating a TPM from a Constant-Inventory Model

Now we suppose that we are given a constant-inventory model, and wish to create a comparable TPM.

Proposition 5. Suppose we have a constant-inventory model with parameters $\mathbf{\Phi}$, $\mathbf{D}$, $\mathbf{\Sigma}$, $\mathbf{w}$, β and μ_{CI}. Consider a TPM with parameters $\mathbf{\Phi}$, $\mathbf{D}$, $\mathbf{\Sigma}$, and μ_{TPM}, where $\mu_{TPM} \neq \mu_{CI}$ is set as follows:

$$\mu_{TPM} = \mu_{CI} + \beta \mathbf{w}'\left(\left(\mathbf{I} - \mathbf{\Phi}\right)\mathbf{E}\left[\mathbf{P}_{CI}\right] - \mu_{CI}\right) \tag{31}$$

Then, the two models have the same production vectors, namely $\mathbf{E[P_{CI}] = E[P_{TPM}]}$.

Proof. We note that (31) equates the expected inputs for the TPM to that for the constant inventory model, i.e., $\mu_{TPM} = \mu_{CI} + \mathbf{E}\left[\mathbf{R}_{CI}\right]$. To show that $\mathbf{E[P_{CI}] = E[P_{TPM}]}$, we know that $\mathbf{E[P_{CI}]}$ satisfies recursion (30). By substituting (23) for $\mathbf{F}$ and $\mathbf{G}$, and rearranging terms, we can rewrite (30) as:

$$\mathbf{E[P_{CI}]} = \mathbf{(I - D + D\Phi)E[P_{CI}]} + \mathbf{D}\mu_{CI} + \mathbf{D}\beta\mathbf{w}'\left(\left(\mathbf{I} - \mathbf{\Phi}\right)\mathbf{E}\left[\mathbf{P}_{CI}\right] - \mu_{CI}\right). \tag{32}$$

But (32) has the same form as the expectation of the TPM recursion (9), with $\mu_{TPM} = \mu_{CI} + \beta \mathbf{w}'\left(\left(\mathbf{I} - \mathbf{\Phi}\right)\mathbf{E}\left[\mathbf{P}_{CI}\right] - \mu_{CI}\right)$. Thus, $\mathbf{E[P_{TPM}]}$ also satisfies (32), and we conclude that $\mathbf{E[P_{CI}] = E[P_{TPM}]}$. ∵

6. A Computational Experiment

In this section we perform a computational experiment to illustrate both the use of the constant-inventory model and the superiority of a constant-inventory control policy relative to alternative controls. For this experiment, we use the example given in Graves (1986), which corresponds to a job shop that produces spindle components for grinding machines. The job shop consists of ten

stations, and the workflow is described by the matrix $\mathbf{\Phi}$ given in Table 3. The only station that receives work from outside the network is station 1 (the lathe); it receives 4 hours of work, on average, each time period. We note that each column of $\mathbf{\Phi}$ represents the expected amount of work generated for subsequent stations per unit of work at the current station.

To test the performance of the constant inventory model, we perform a 2^k design test, as follows. We assume that the covariance matrix $\mathbf{\Sigma}$ for the noise vector is a diagonal matrix. We consider pairs of stations (1 and 2, 3 and 4, etc.), and we alternate between giving the pairs "high" variances and "low" variances. Each "high" station *i* has a variance of 4, and a lead-time of 4 (i. e., $\sigma_i^2 = 4$ and $1/\alpha_i = 4$). Each "low" station has a variance of 0.05, and a lead-time of 2. There are 32 test cases formed by choosing all ways to assign "high" or "low" to the station pairs. Table 4 shows the variances and lead-times for each test.

	From Work Station									
To Work Station	1	2	3	4	5	6	7	8	9	10
1 (lathe)			0.11		0.68					
2 (copy lathe)	0.15									
3 (drill press)	0.04	0.01		0.71		0.6			0.07	
4 (milling)	0.01	0.41								
5 (rough grinder)	0.03	0.37	1.36							
6 (internal grinder)	0.24				0.15				0.13	
7 (thread cutting)					0.10					
8 (hole abrading)	0.01					0.22	1.00			
9 (precision grider)								3.43		
10 (ultra-precision grinder)									1.16	

Table 3: Workflow Matrix

	Station Input Variances					Station Lead Times						Station Input Variances					Station Lead Times				
Test	1,2	3,4	5,6	7,8	9,10	1,2	3,4	5,6	7,8	9,10	Test	1,2	3,4	5,6	7,8	9,10	1,2	3,4	5,6	7,8	9,10
1	0.05	0.05	0.05	0.05	0.05	2	2	2	2	2	17	4.0	0.05	0.05	0.05	0.05	4	2	2	2	2
2	0.05	0.05	0.05	0.05	4.0	2	2	2	2	4	18	4.0	0.05	0.05	0.05	4.0	4	2	2	2	4
3	0.05	0.05	0.05	4.0	0.05	2	2	2	4	2	19	4.0	0.05	0.05	4.0	0.05	4	2	2	4	2
4	0.05	0.05	0.05	4.0	4.0	2	2	2	4	4	20	4.0	0.05	0.05	4.0	4.0	4	2	2	4	4
5	0.05	0.05	4.0	0.05	0.05	2	2	4	2	2	21	4.0	0.05	4.0	0.05	0.05	4	2	4	2	2
6	0.05	0.05	4.0	0.05	4.0	2	2	4	2	4	22	4.0	0.05	4.0	0.05	4.0	4	2	4	2	4
7	0.05	0.05	4.0	4.0	0.05	2	2	4	4	2	23	4.0	0.05	4.0	4.0	0.05	4	2	4	4	2
8	0.05	0.05	4.0	4.0	4.0	2	2	4	4	4	24	4.0	0.05	4.0	4.0	4.0	4	2	4	4	4
9	0.05	4.0	0.05	0.05	0.05	2	4	2	2	2	25	4.0	4.0	0.05	0.05	0.05	4	4	2	2	2
10	0.05	4.0	0.05	0.05	4.0	2	4	2	2	4	26	4.0	4.0	0.05	0.05	4.0	4	4	2	2	4
11	0.05	4.0	0.05	4.0	0.05	2	4	2	4	2	27	4.0	4.0	0.05	4.0	0.05	4	4	2	4	2
12	0.05	4.0	0.05	4.0	4.0	2	4	2	4	4	28	4.0	4.0	0.05	4.0	4.0	4	4	2	4	4
13	0.05	4.0	4.0	0.05	0.05	2	4	4	2	2	29	4.0	4.0	4.0	0.05	0.05	4	4	4	2	2
14	0.05	4.0	4.0	0.05	4.0	2	4	4	2	4	30	4.0	4.0	4.0	0.05	4.0	4	4	4	2	4
15	0.05	4.0	4.0	4.0	0.05	2	4	4	4	2	31	4.0	4.0	4.0	4.0	0.05	4	4	4	4	2
16	0.05	4.0	4.0	4.0	4.0	2	4	4	4	4	32	4.0	4.0	4.0	4.0	4.0	4	4	4	4	4

Table 4: Variances and Lead-Times for Test Cases

We use the sum of the production standard deviations, that is, $\sum_{i=1}^{10}\sigma(P_{it})$, as a measure of the capacity requirements for the shop. For each test case, we evaluate this measure for three control policies: the TPM, the TPM with constant release, and the constant-inventory model.

- For the TPM, the variance of work arrivals to station 1 is 0.05 or 4, depending on whether the station is set to "low" or "high".
- For the TPM with constant release, the variance of work arrivals to station 1 is 0. That is, we assume that exactly four hours of work arrives each period to station 1.
- For the constant-inventory model, we regulate the work release to station 1 to maintain a constant weighted-inventory in the job shop. We solve a nonlinear program (see appendix) to find the weights and exogenous inputs that minimize the performance metric and assure equivalence to the TPM.

Table 5 gives the results of the 32 tests. For each test, we give the performance metric for each of the three models, and a percentage comparison between the models. The table shows that the

constant-inventory policy outperforms the other two policies. In comparison with the TPM, the sum of production standard deviations is 13% lower, on average, for the constant-inventory model; the percentage reduction ranges from about 5% to over 25% for the test cases. This performance is much better than the constant-release TPM, which only manages decreases from 0.2% to 12%, with the average being about 3%. The constant-release TPM is particularly ineffective for cases in which most of the variability is far downstream from the first station (cases 1 to 16). The constant-inventory model usually produces a significant decrease, even for cases in which most of the variability is far downstream from the first station.

Test	Sum of Production Standard Deviations			Percentage Reductions		
	TPM	TPM with Constant Release	Constant Inventory Model	C. Release Over TPM	C. Inventory over TPM	C. Inventory over C. Release
1	2.2781	2.2202	2.0465	2.54	10.12	7.82
2	3.2663	3.2119	3.0289	1.67	7.27	5.70
3	9.6651	9.6328	9.2429	0.33	4.37	4.05
4	8.9121	8.8788	8.5060	0.37	4.56	4.20
5	4.5684	4.5490	3.8723	0.42	15.24	14.88
6	5.1785	5.1594	4.5272	0.37	12.58	12.25
7	10.8417	10.8242	10.1488	0.16	6.39	6.24
8	10.1183	10.1008	9.4424	0.17	6.68	6.52
9	6.2609	6.1937	4.7082	0.21	24.14	23.98
10	6.8003	6.7874	5.4662	0.19	19.62	19.47
11	12.4466	12.4347	11.2520	0.10	9.60	9.51
12	11.7091	11.6971	10.5329	0.10	10.04	9.95
13	7.1819	7.1703	5.8271	0.16	18.86	18.73
14	7.6384	7.6269	6.3927	0.15	16.31	16.18
15	12.9572	12.9462	11.7935	0.08	8.98	8.90
16	12.2348	12.2239	11.0984	0.09	9.29	9.21
17	4.5344	3.8336	3.3131	15.46	26.93	13.58
18	5.3642	4.7063	4.2355	12.26	21.04	10.00
19	11.3964	10.8365	10.3013	4.91	9.61	4.94
20	10.6553	10.0920	9.5699	5.29	10.19	5.17
21	6.0995	5.5940	4.8906	8.29	19.82	12.57
22	6.6659	6.1757	5.5181	7.35	17.22	10.65
23	12.2109	11.7582	11.0923	3.71	9.16	5.66
24	11.4892	11.0355	10.3864	3.95	9.60	5.88
25	7.2972	6.8656	5.5114	5.91	24.47	19.72
26	7.8661	7.4519	6.2460	5.27	20.60	16.18
27	13.4416	13.0654	11.9967	2.80	10.75	8.18
28	12.7043	12.3272	11.2785	2.97	11.22	8.51
29	8.1720	7.7969	6.5855	4.59	19.41	15.54
30	8.6165	8.2508	7.1400	4.24	17.14	13.46
31	13.8974	13.5530	12.5217	2.48	9.90	7.61
32	13.1746	12.8299	11.8283	2.61	10.22	7.81

Table 5: Test Results

The implication of this computational experiment is a reaffirmation of the benefits from regulating the workflow into a shop. In particular, we show that the constant-inventory model has less production variability in comparison with a TPM with equivalent expected throughput and work-in-process for a job shop. The reduction in production variability translates into less required capacity and/or a lower production cost.

7. Concluding Remarks

In this paper, we show how to add a weighted-inventory constraint to the TPM, and to find the first two moments for the production and work-in-process random vectors. The algorithmic complexity for determining these moments is comparable to that for the TPM. To the extent that the TPM is a discrete-time model for an open queuing network, then we might interpret the constant-inventory model as the analog to a closed queuing network. We describe an application of the use of the model for sizing a CONWIP control policy. We also show how to parameterize a constant-inventory model to be comparable to a TPM, and vice versa. We report on a computational experiment that establishes the benefit of a constant-inventory control policy, over the TPM, in terms of reducing production variability.

We note some opportunities for future research on the constant-inventory model. First, one might explore further the stability of these models. In proposition 1, the condition on the inventory weights that guarantees convergence, is a sufficient but not necessary condition. Experimental tests suggest that models whose weights violate the proposition are able to converge if (1) the resulting total weighted arrivals are less than the total weighted production, or (2) if station lead-times are lengthened sufficiently. These phenomena warrant additional study.

Second, one might extend the analysis to permit multiple constraints. With multiple constraints we could split the shop into separate sub-sections and apply a constant-inventory constraint to each sub-section. For instance this would permit the analysis of a shop with multiple CONWIP loops. We expect that by having independent control on sub-sections of the shop would allow us to reduce production variations more effectively.

A third possible extension is to combine constant WIP constraints with production rules that are linear functions of the WIP at multiple stations. The optimal control rules minimizing

production and WIP variances in Hollywood (2000) were of this form; it may be possible to make significant reductions in production variability (and thus, capacity requirements) by allowing multi-station production rules.

A final area of study is to relax the linear control assumption for each station. Hollywood (2005) has developed a fairly accurate approximation for a TPM with nonlinear control rules, and has shown how to apply this to model distributed computing networks. It may be possible to extend this approximation to include a weighted-inventory constraint.

Appendix

Proposition 4 has shown how to set the parameters for the constant–inventory model so as to make it comparable to a given TPM. In particular, for a given TPM with expected arrivals μ_{TPM}, we set $\mathbf{w}$, β and μ_{CI} for the constant-inventory problem to satisfy (29). Thus, we can define several comparable constant-inventory models.

Suppose we wish to find the "best" from this set of comparable constant-inventory models. We can do this by optimization, subject to how we define "best." For the computational experiment in section 6, we seek the comparable constant-inventory model that minimizes the sum of the production standard deviations. We formulate an optimization whose decision variables are the inventory weight vector $\mathbf{w}$ and the expectation of the exogenous input vector $\boldsymbol{\mu}_{\mathbf{CI}}$. For a given value of μ_{TPM}, we solve the following problem:

$$\begin{aligned} &Min \sum_{i} \sigma\left(P_{it}\right) \\ &s.t. \quad \beta = \frac{\boldsymbol{\mu}_{\mathbf{TPM}} - \boldsymbol{\mu}_{\mathbf{CI}}}{\mathbf{w}'\left(\boldsymbol{\mu}_{\mathbf{TPM}} - \boldsymbol{\mu}_{\mathbf{CI}}\right)}. \\ &\boldsymbol{\mu}_{\mathbf{CI}} \leq \boldsymbol{\mu}_{\mathbf{TPM}} \\ &\boldsymbol{\mu}_{\mathbf{CI}}, \mathbf{w} \geq \mathbf{0} \end{aligned} \tag{A1}$$

The objective function is the sum of the standard deviations of the production random variable; we obtain this from the diagonal elements of the covariance matrix $\mathbf{S}$, as found from (27). The first constraint sets the release template β so that the resulting constant-inventory model is comparable to the TPM. In addition we require the decision variables $\mathbf{w}$ and $\boldsymbol{\mu}_{\mathbf{CI}}$ to be non-negative, and we bound $\boldsymbol{\mu}_{\mathbf{CI}}$ by the given $\boldsymbol{\mu}_{\mathbf{TPM}}$.

The objective function in (A1) is a continuous and continuously differentiable function in the decision variables, and is fairly easy to minimize. To find the optimal weights for the

examples in section 6, we employed the Two-Metric Projection Method (see Bertsekas, 1982) with diagonally scaled steepest descent iterations (taking numerical approximations of derivatives). Using this approach, an SGI workstation found the optimal weights in well under a minute of computing time.

Acknowledgment: This research has been supported in part by the MIT Leaders for Manufacturing Program, a partnership between MIT and major global manufacturing firms. The first co-author also acknowledges the support provided by the Singapore-MIT Alliance in completing this work. The authors also wish to acknowledge and thank Esko Savukoski, Senior Consultant, ABB Corporate Research Oy, Vaasa, Finland, for providing us with the application in Section 4 and for granting permission to use it in this paper.

References

Bertsekas, D. P., 1982. Projected Newton Methods for Optimization Problems with Simple Constraints. *SIAM J. on Control and Optimization*, **20**, 221-246.

Conway, Adrian E., and Nicolas D. Georganas. *Queuing Networks – Exact Computational Alogirthms: A Unified Theory Based on Decomposition and Aggregation*. MIT Press, Cambridge, MA.

Cruickshanks, A. B., R.D. Drescher, and S. C. Graves. 1984. A Study of Production Smoothing in a Job Shop Environment, *Management Science*, **30**, 368-380.

Dar-El, E. M., Y. T. Herer and M. Masin. 1999. CONWIP-based production lines with multiple bottlenecks: Performance and design implications, *IIE Transactions*, Vol. 31, No. 2; pp. 99 – 111.

Duenyas, Izak, and Wallace Hopp. 1993. Estimating the Throughput of an Exponential CONWIP Assembly System. *Queueing Systems*, **14**, 135-157.

Duenyas, Izak, Wallace Hopp, and Mark Spearman. 1993. Characterizing the Output Process of a CONWIP Line with Deterministic Processing and Random Outages. *Management Science*, **39**, 975-988.

Fine, Charles H., and Stephen C. Graves. 1989. A Tactical Planning Model for Manufacturing Subcomponents of Mainframe Computers. *J. Mfg. Oper. Mgt.*, **2**, 4-34.

Framinan, Jose M., Pedro L. Gonzalez and Rafael Ruiz-Usano. 2003. The CONWIP Production Control System: Review and Research Issues. *Production Planning & Control*, Vol. 14, No. 3, pp. 255-265.

Gordon, K. D. and G. F. Newell. 1967. Closed Queueing Systems with Exponential Servers. *Operations Research*, **15**, 254-265.

Graves, S. C., H.C. Meal, S. Dasu, and Y. Qiu. 1986. Two-Stage Production Planning in a Dynamic Environment, in Lecture Notes in Economics and Mathematical Systems, *Multi-Stage Production Planning and Inventory Control,* edited by S. Axsater, Ch. Schneeweiss, and E. Silver, Springer-Verlag, Berlin, **266**, 9-43.

Graves, Stephen C. 1986. A Tactical Planning Model for a Job Shop. *Operations Research*, **34**, 522-533.

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

Graves, Stephen C. 1988b. Determining the Spares and Staffing Level for a Repair Depot, *Journal of Manufacturing and Operations Management*, **1**, 227-241.

Graves, S. C., D. B. Kletter and W. B. Hetzel. 1998. A Dynamic Model for Requirements Planning with Application to Supply Chain Optimization, *Operations Research*, **46**, S35-S49.

Gstettner, S., and H. Kuhn. 1996. Analysis of Production Control Systems Kanban and CONWIP. *International Journal of Production Research*, **34**, 3253-3273.

Herer, Y.T., and M. Masin. 1997. Mathematical Programming Formulation of CONWIP based Production Lines; and Relationships to MRP. *International Journal of Production Research*, **35**, 1067-1076.

Holly, Sean M. 2006. "Lean Manufacturing in a Semiconductor Environment: Use of Variation Analysis to Focus Continuous Improvement Efforts," S.M. Thesis, Leaders for Manufacturing Program, MIT, Cambridge MA.

Hollywood, John S. 2000. Performance Evaluation and Optimization Models for Processing Networks with Queue-Dependent Processing Quantities. Ph.D. Thesis, Operations Research Center, MIT, Cambridge MA, June.

Hollywood, John S. 2005. "An Approximate Planning Model for Distributed Computing Networks," *Naval Research Logistics*, Vol. 52, No. 6, pp. 590-605.

Hopp, Wallace J. and M. L. Roof. 1998. "Setting WIP levels with Statistical Throughput Control (STC) in CONWIP Production Lines," *Int. J. Prod. Res.*, Vol. 36, No. 4 pp. 867-882.

Hopp, Wallace J., and Mark L. Spearman. 2001. *Factory Physics: Foundations of Manufacturing Management*. 2nd edition. Irwin McGraw Hill, Boston.

Kelly, Frank P. 1979. *Reversibility and Stochastic Networks*. Wiley, New York.

Jackson, J.R. 1957. "Networks of Waiting Lines." *Operations Research*, **5**, 518-521.

Jackson, J.R. 1963. "Jobshop-like Queuing Systems." *Management Science*, **10**, 131-142.

Leong, Thin-Yin. 1987. A Tactical Planning Model for a Mixed Push and Pull System. Ph.D. program second year paper, Sloan School of Management, Massachusetts Institute of Technology, July.

Mihara, S. 1988. A Tactical Planning Model for a Job Shop with Unreliable Work Stations and Capacity Constraints. S.M. Thesis, Operations Research Center, MIT, Cambridge MA, January.

Ryan, Sarah M., Bruno Baynat and Fred Choobineh. 2000. "Determining Inventory Levels in a CONWIP Controlled Job Shop," *IIE Transactions*, Vol. 32, pp. 105 – 114.

Ryan, Sarah M., and Fred Choobineh. 2003. "Total WIP and WIP Mix for a CONWIP Controlled Job Shop," *IIE Transactions*, Vol. 35, pp. 405 – 418.

Solberg, James J. 1976. A Mathematical Model of Computerized Manufacturing Systems. *Proceedings of the 4th International Conference on Production Research*. Tokyo, Japan.

Spearman, Mark L., Wallace J. Hopp, and David L. Woodruff. 1989. A Hierarchical Control Architecture for Constant Work-in-Progress (CONWIP) Production Systems. *Journal of Manufacturing and Operations Management*, **21**, 147-171.

Spearman, Mark L., David L. Woodruff, and Wallace J. Hopp 1990. CONWIP: A Pull Alternative to Kanban. *International Journal of Production Research*, **28**, 879-994.

Wein, Lawrence M. 1988. Scheduling Semiconductor Wafer Fabrication, *IEEE Transactions of Semiconductor Manufacturing* **1**, 115-130.

Wein, Lawrence M. 1990. Scheduling Networks of Queues: Heavy Traffic Analysis of a Two-Station Network With Controllable Inputs, *Operations Research* **38**, 1065-1078.