A Forecast-driven Tactical Planning Model for a Serial Manufacturing System

by

Pallav Chhaochhria, Sun Trading LLC, 100 South Wacker Drive,Suite 300

Chicago, IL 60606, USA, pallavc@yahoo.com

Stephen C. Graves, MIT, E62-579, 77 Massachusetts Avenue, Cambridge MA, 02139, USA,

sgraves@mit.edu

September 30, 2012, Revised September 28, 2013

Abstract

Our work is motivated by real-world planning challenges faced by a manufacturer of industrial products. We study a multi-product serial-flow production line that operates in a low-volume, long lead-time environment. The objective is to minimize variable operating costs, in the face of forecast uncertainty, raw material arrival uncertainty and in-process failure. We develop a dynamic-programming-based tactical model to capture these key uncertainties and trade-offs, and to determine the minimum-cost operating tactics. The tactics include smoothing production to reduce production-related costs, and segmenting the serial-flow line with decoupling buffers to protect against variance propagation. For each segment, we specify a work release policy and a production control policy to control both the work-in-process inventory within the segment and the inventory in the downstream buffer. We also optimize the raw material ordering policy with fixed ordering times, long lead-times and staggered deliveries. We test the model on both hypothetical and actual factory scenarios. The results confirm our intuition and provide new managerial insights on the application of these operating tactics. Moreover, we compare the performance predictions from the tactical model to a simulation of the factory, and find the predictions to be within 10% of simulation results, thus validating the model.

Key Words: production planning; production flow analysis; forecasting; flow shop

1. INTRODUCTION AND MOTIVATION

Our work is motivated by real-world planning challenges faced by a manufacturer of industrial products and systems in a make-to-forecast setting. The manufacturer has serial-flow systems, assembly systems and even job-shops for the different products and systems it manufactures. In this paper we develop a model for determining the optimal production tactics for a serial-flow system that produces discrete parts in a low-demand, long lead-time production environment.

The manufacturer orders raw materials and builds products based on forecasts of customer orders. However the forecast is dynamic: new orders come in while existing orders may change; some orders may have their due-dates advanced, some may be postponed, and some orders may be cancelled. Moreover, the lead-time for raw material is uncertain, as purchase orders may be delayed. This presents challenges in setting a raw material ordering policy, and in planning production and inventory so as to not delay customer orders.

The manufacturer has some flexibility in production planning as production capacity is elastic, i.e., the manufacturer can use overtime beyond the normal production shift to increase production capacity. However, the overtime comes at a higher labor cost than the regular production shift. Moreover, the amount of overtime in a day is limited due to labor union rules. Such overtime conditions are common in most production settings, allowing for some flexibility but at a cost.

The challenge for the manufacturer is to determine policies for ordering raw material and for planning production and inventory that minimize the relevant costs while meeting customer service requirements. The manufacturer needs to find the optimal mix of raw material, work-in-process and finished goods inventory to hold, balancing this inventory cost with the use of limited overtime to expand its production capacity when needed. Our research focuses on the tactical planning decisions that attempt to achieve the right balance of inventory and planned overtime to meet the customers' requirements.

One of the tactics we consider is the line configuration, namely how to divide the serial production line into segments. These segments are separated by decoupling buffer inventories, so that each segment can operate independently without blockage from the downstream segment or starvation from the upstream segment. Thus, each segment can be treated approximately as a separate or autonomous production line with its own raw material and finished goods inventories. Each segment sees a demand forecast from the downstream segment or from customers (if it is the customer-facing segment), and sets its production target in every period and releases work accordingly.

Another tactic that we examine is the use of production smoothing in each segment. Production smoothing involves the setting of production targets and inventory levels to meet variable demand with an efficient utilization of production resources. The aim of production smoothing is to find the best way to convert a highly variable demand stream into a less variable production output stream. A production smoothing policy necessitates the holding of additional inventory, either as work-in-process or finished goods or a combination thereof, in order to meet customer service requirements. Thus, the production plan for each segment becomes less variable than the demand forecast that it sees, and this in turn reduces the amount of production overtime that is performed in the segment.

We develop a model to find the best locations for these decoupling buffers while also suggesting the optimal production and inventory policies in each segment to minimize total system cost while facing a dynamic forecast process. We determine the optimal level of production smoothing and work-in-process inventory to have in each segment, as well as the safety stock needed in each decoupling buffer.

1.1 Related Work

Very few papers have studied tactical production planning under forecast uncertainty. However, there has been significant interest in the two research streams that are related to our work: production-inventory systems with dynamic forecast updates, and tactical level production planning. We examine the work done in each of these streams.

The earliest forecast evolution models were developed by Hausman and Peterson (1972). They developed a multi-period production-inventory model for style goods with capacity limitations and log-normal forecast revisions. They showed that threshold type inventory policies are optimal under those assumptions, and presented heuristic solutions.

Graves, Meal, Dasu and Qiu (1986) introduced a forecast-driven production planning model for a multi-product two-stage system. Their forecast process assumed additive forecast revisions with zero mean and finite variance, and used production smoothing as the main planning tactic. The paper also looked at how to disaggregate an aggregate plan in the two-stage context.

Heath and Jackson (1994) independently developed the same forecast process and highlighted its martingale properties, terming it the Martingale Model of Forecast Evolution (MMFE). They used the forecast process in a simulation model to analyze safety stock levels in a multi-product production-distribution system.

Graves, Kletter and Hetzel (1998) developed a forecast-driven model for requirements planning in a multistage production-inventory system. They assume an MMFE process, and built a single-stage model that determines the optimal level of production smoothness and the inventory requirements for that stage. They then use the single-stage model as a building block for modeling a network of stages. Their model was applied to a production system at Eastman Kodak.

Toktay and Wein (2001) studied a capacitated production system that produces a single item in a make-to-stock setting. They modeled the production stage as a single-server, discrete-time continuous-state queue and the demand forecast process as an MMFE process. They used heavy-traffic approximations to prove the optimality of the base-stock inventory policy. Kaminsky and Swaminathan (2001) introduce a forecast process where the forecast of future demand is uniformly distributed in a band. This forecast band gets narrower as the forecast gets closer to realization. They show that threshold-type inventory policies are optimal for a single-product in a capacitated production system using this forecast process. Gallego and Ozer (2001) assume advanced demand information and show that state-dependent (s,S) and base stock policies are optimal for a single-item periodic review inventory problem with and without a fixed ordering cost, respectively. Gallego and Ozer (2003) extends these results to a serial inventory system, proving the optimality of state-dependent echelon base stock policies. Iida and Zipkin (2006) consider a single-item, periodic-review inventory system with a deterministic lead-time subject to a MMFE process. They showed that a forecast-dependent base stock policy is optimal, and developed bounds on the optimal base-stock levels. Lu, Song and Regan (2006) also considered the same problem and devised heuristic solutions, as well as cost-error bounds. Altug and Muharremoglu (2010) show that state-dependent base-stock policies are optimal for a single-item, single stage inventory system with advance supply information modeled as an MMFE process, and where the replenishment from an upstream external source is capacitated and stochastic.

These papers (except Graves et al. 1986 and Graves et al. 1998) focus primarily on finding optimal inventory policies, and easily computable heuristics for the production-inventory systems under forecast uncertainty. They do not consider the interplay between employing production planning tactics and the inventory policies.

The second relevant stream of literature includes research on production smoothing and the placement of decoupling buffers in manufacturing systems. The production smoothing problem has been studied since the 1950’s. Simon (1952) used control theory (called servomechanism theory in the paper) to analyze the problem of setting the production rate in a single product manufacturing system. His key intuition was to use a low-pass filter on the demand forecast to set the production rate, essentially filtering out the high-frequency variations while capturing the low frequency demand changes. Modigliani and Hohn (1955) use Lagrange multipliers and calculus-based methods to study the same problem. They proved that using constant production rates is optimal when the inventory holding cost is low relative to the cost of production change, and setting the production rate equal to the expected demand in a period is optimal when the inventory holding cost is high relative to the cost of production change. They also characterize the length of the planning horizon needed to make the optimal decision in the current period.

Further work on this problem introduced the use of methods like linear programming (Hoffman and Jacobs 1954, Johnson and Dantzig 1955, Manne 1957), dynamic programming (Beckmann 1961, Symonds 1962) and queuing (Gaver 1961). These papers differ based on the assumed cost functions for production rate changes and on the knowledge and/or distribution of future demand.

Bertrand (1986) studied experimentally the behavior of production and inventory levels in a multi-product, multi-stage production system. His findings include that production smoothing in a stage led to increased inventory variability in the downstream buffer, and that inventory variability increased with the cumulative production lead-time. Thus, production smoothing would be better in upstream stages than in downstream stages.

More recent developments have studied the smoothing of replenishment orders for inventories in a multi-echelon environment to reduce the bullwhip effect. Balakrishnan et al. (2004) propose a coordinated inventory replenishment policy among the supply chain partners. They show that order smoothing rules dampen the variance propagation to upstream suppliers, and reduce total system costs.

Graves (1986) develops a different approach to production smoothing systems, related to determining planned lead-times for the individual work-stations in a job-shop. Longer planned lead-times for a work-station are effectively the result of more production smoothing at the work-station. This approach was developed further in Hollywood (2000) and Teo (2006).

We find no previous research on the placement of decoupling buffers in production systems. The closet work seems to be that on safety stock placement in supply chains e.g., Simpson (1958), Graves and Willems (2000). However, the safety stock placement models assume a fixed replenishment lead-time for each stage of the supply chain, irrespective of order quantity. This is equivalent to assuming unlimited production capacity at each stage. This is a fundamental difference from manufacturing system models where production capacity is usually highly constrained, causing load-dependent lead-times. These constraints make the planning and control of the manufacturing system much more complex, as starvation and blockage effects get propagated throughout the system.

The production control literature predominantly focuses on a single tactic rather than a combination of tactics which involves modeling their trade-offs. Moreover, these papers usually do not explicitly consider a forecast process that drives the manufacturing system.

Our work builds on the ideas proposed in Graves et al. (1986) and Graves et al. (1998). We develop models for a single production stage (or segment) that use production smoothing to minimize the total inventory-holding and production overtime costs subject to customer service requirements. We then use the single-stage model as a building block to represent a production system as a network of stages. We can then search over the choice of these stages so as to find the optimal configuration, equivalently the optimal locations for decoupling buffers in the system.

2. PROBLEM DESCRIPTION & SOLUTION STRATEGY

We examine a low-volume, long lead-time serial-flow manufacturing system for high-value component products produced by the research sponsor. The research is specific to the context studied, and the model especially addresses those concerns. However, the key ideas presented in the model apply to a large number of manufacturing systems, and so the model can be used for other systems with minor modifications. Moreover, the operating insights generated from the model are applicable to production planning in general. We first describe the problem and the solution strategy and then develop the model. We will present the model assuming a single product for ease of presentation; the actual case involves multiple products, and the specifics on how the model extends are given in Chhaochhria (2011).

2.1 System Features

We define a serial-flow manufacturing system as a production line where a work-piece requires a sequence of process steps to become a finished product. These process steps may involve any type of work except assembly with other work-pieces or parts. Example processes are casting, cutting, grinding, milling, painting, heat-treatment, and inspection.

Each process step is performed either in-factory or outsourced to a vendor or subcontractor. For the purposes of this research we do not examine in detail the process steps that are outsourced. Rather, we will treat these as a fixed time delay, corresponding to the planned lead-time for each step.

Each in-factory process step occurs at a work station. Each work station may have dedicated resources (machines and/or workers) assigned to it, or may share resources with other work stations. Some work stations might even entail dual resource constraints; for example, a worker and a machine need both to be available in order to process a work-piece. Furthermore, a work station might consist of multiple parallel resources (workers and/or machines). In this case, the processing of a work piece could happen at any of the parallel resources. The processing at a workstation can be serial (one unit a time) or in batches (concurrent processing of multiple units).

A work piece might revisit a work station as part of its process route, or possibly for re-work due to a failure at an inspection step. Whereas there can be some work stations that process a work piece multiple times, the vast majority of the work stations perform a single process step on each work piece.

The manufacturing system produces multiple products, each with a serial flow. The process route for each product need not be exactly the same, but there is a high level of commonality between the process routes for these products. That is, the set of work stations that one product visits, as well as the sequence, are almost identical to that for each of the other products. When switching from one product to another, a work station might require a different setup, thus incurring changeover times or costs.

The system has a large number of processes (on the order of 100) which leads to a long production lead-time (on the order of months) to complete a work-piece. There is also a long lead time for procuring the raw materials needed to initiate production. This necessitates the holding of inventory (raw, WIP and finished goods) in order to meet customer orders that have a delivery lead-time shorter than the production lead-time. Moreover, the production strategies that we propose are better suited for low-volume, long lead-time production lines.

The production system operates on a multi-shift schedule with resources operating one or two shifts per day. Additional production capacity is available by working overtime after the regular shift hours. However, there are limitations on the amount of overtime due to labor union and other constraints.

2.2 System Uncertainties

The manufacturing system is driven by a forecast that is updated monthly. The forecast horizon is on the order of 24 to 48 months. New orders, both regular and special, arrive each month. Regular orders have fixed quantities for each product type but special orders can be for any number of pieces. Regular orders have long delivery lead-times of between 18 and 36 months, but special orders have shorter lead-times of 3 to 6 months. Also, each month, there may be changes to the orders in the forecast; the delivery date of an order may be advanced or postponed by a number of months, or even cancelled. Such changes in the forecast are common as the sales department updates its list of confirmed and potential orders.

The system also faces raw material procurement uncertainty. The raw material has fixed ordering times with long delivery lead-times and staggered deliveries. Furthermore, the raw material delivery can have significant delays.

Finally, there is yield uncertainty in the production line. After inspection processes, some work-pieces do not meet quality specifications and are sent for rework or scrapped.

2.3 System Costs

At the tactical planning level, many of the production costs such as labor, machine operating costs, production facility-related costs are fixed. The controllable costs for purposes of tactical planning include: the inventory holding costs for raw material, work-in-process and finished goods; the production overtime cost when processes work beyond the regular hours of a shift; and a penalty cost that is incurred for each day that an order is delivered late beyond its due date. Our objective is to find operating policies that minimize the expectation of these controllable costs for the system.

2.4 Solution Strategy

We consider two main tactics to control the system. One is the placement of decoupling buffers in the serial-flow production line to create production segments. These buffers reduce variance propagation, both from the upstream raw material uncertainty and from the downstream uncertainty in the forecast process. They also divide the system into sub-systems that can operate with different production control policies due to differences in their production costs and capacity levels. The second tactic is to determine how to set the production level and work releases in each segment. We intend to do this to balance the overtime costs and inventory costs in each segment. How we level or smooth production in a segment can also affect the performance of other segments, by dampening the variance propagation to upstream segments.

Production smoothing tactics have been suggested in the academic literature and are commonly employed in practice. Decoupling buffers have not been studied in manufacturing systems, though they do occur in practice. There does not seem to be a comprehensive model that incorporates both these tactics simultaneously. The intent of this research is to develop operating tactics for production systems.

Figure 1: Insert here

In figure 1 we divide the serial-flow system into 3 segments by placing 2 decoupling buffers. Each segment j will experience its own demand variability; production smoothing in each segment dampens its demand variability, resulting in less variable production releases. As the release from downstream segment j+1 is the demand on the upstream segment j, smoothing dampens the demand variability as it is passed upstream. The key questions we want to answer with this model are:

- Where do we place these decoupling buffers and how much safety stock is needed?
- How much work-in-process inventory is needed in each segment?
- How should work be released into each segment?
- What production rate control policy should be used for each segment?
- What should be the raw material inventory ordering policy?

3. SINGLE PRODUCTION SEGMENT MODEL

We first develop a model for a single production segment. For a given set of inputs, this model will find the optimal control parameters that minimize the operating cost of the segment. We use this model to evaluate all possible segments. We assume that we can place a decoupling buffer between any two process steps. Then, for a system with N process steps, we have $\frac{(N+1)!N!}{2}$ possible production segments for the serial system.

Once we have evaluated the optimal operating costs for all possible segments, we then can determine the optimal set of segments. To do this, we develop a dynamic program that will find the optimal combination of segments to span the production line with the least cost. In effect, the dynamic program determines the locations of decoupling buffers and their safety stocks, and the optimal work release and production rate policies for each segment.

Figure 2: Insert here

A production segment consists of a set of contiguous process steps that are fed by a decoupling buffer and that are followed by a decoupling buffer. The initial buffer provides the starting material for the segment, while the final buffer holds the output from the segment. For example, in Figure 2, a segment (i,j) consists of process step (i+1) to j. Decoupling buffer #1 holds the input material, and finished work-pieces are stored in decoupling buffer #2. Each segment sees a forecast of its demand. This may be the finished goods demand forecast if the segment contains the last process step, or this may be a forecast of the work release from the downstream segment. For each segment we desire to meet its demand with a high level of service at the least cost. We develop this single segment model given the location of the decoupling buffers, and for specified demand variance and production variance. We want this model to be computationally efficient as we will use it to compute costs for all possible segments in a given serial-flow system, and for given ranges of the demand variances and the production variances.

If the segment includes the first process step, then it will need to set raw material inventory policies and associated costs. If the segment includes the last process step, then it will need to set finished goods inventory policies and estimate order delay penalty costs. Details on how exactly this is done are given in Chhaochhria (2011).

We first state the assumptions and then introduce the operational rules for work release and production rate control. We will establish some properties of these rules that make the analysis and computation involved in the segment evaluation tractable. Using these control rules and inventory balance equations, we can calculate the buffer inventory and production variances. We then use these variances to set the inventory level and estimate the production overtime. Finally, we will show how the operating costs of the segment are calculated, and how the control parameters are optimized.

3.1 Assumptions

A1. There is an underlying time period for planning. Demand forecasts get updated at this frequency, and we make production control decisions for each segment at this frequency.

A2. The forecast process can be modeled by the martingale model of forecast evolution (MMFE), with a constant demand mean and variance over our planning horizon.

A3. The decoupling buffer before a segment holds enough inventory so that the desired release rate into the segment is always possible. (Effectively, for the purposes of the analysis, we ignore any shortfalls in the availability of the input inventory for the segment.)

A4. The desired production rate in each period can be achieved. There is no absolute constraint on production and there is sufficient work-in-process to accomplish the desired rate.

The sequence of events in each segment in each time period as follows:

i. At the start of the period, the forecast is updated based on changes to confirmed and potential orders, including realization of the demand for the period.

ii. Demand for current period is filled from the output inventory

iii. The work release is determined for current and future periods

iv. The production rate for current and future periods is set, including overtime if necessary

v. Production occurs during the period as specified by the production rate, and goes into the output inventory at the end of the period

We find A2 to be a very reasonable assumption as the MMFE can be used to model a wide variety of demand types, including non-stationary and correlated demand as shown in Schoenmeyr and Graves (2009). Moreover, for the purposes of tactical planning, the demand mean and variance are unlikely to change in the short term.

A3 and A4 are strong assumptions but they greatly simplify the development of the single segment model. We will show through simulation tests that these assumptions hold for a reasonable range of parameter values.

3.2 Martingale Model of Forecast Evolution

In period t we denote the forecast for period $t+i$ as $f_t(t+i)$ for i=1,2,..,H where H is the forecast horizon. By convention we set $f_t(t)=D_t$ where D_t is the demand in period t. We assume that in each period t we make an initial forecast for the demand in period $t+H$, that is $f_t(t+H)$; and in each period we revise the nearer-term forecasts, where we define the *forecast revision* as:

$$\Delta f_t(t+i) = f_t(t+i) - f_{t-1}(t+i) \quad \text{for} \quad i=0,1,...,H$$

where we assume $f_t(t+i) = \mu$ (the mean demand rate) for all i>H. We can express demand as follows:

$$f_t(t) = D_t = \mu + \sum_{i=0}^{H} \Delta f_{t-H+i}(t) .$$

We let $\underline{\Delta f_t}$ be the vector of *H+1* forecast revisions. We assume that $\underline{\Delta f_t}$ is a random, iid vector with $E[\Delta f_t(j)] = 0$ for all *t* and *j*. With these assumptions Graves et al. (1986), Heath and Jackson (1994) and Graves et al. (1998) have established several properties for this forecast evolution model:

- $f_t(t+i)$ is a martingale and is an unbiased estimate of D_{t+i}
- The variance of the forecast error $\{D_{t+i} - f_t(t+i)\}$ increases (weakly) in *i*
- The random variable D_t has mean μ, and its variance is the trace of the covariance matrix for $\underline{\Delta f_t}$ which we denote by Σ_f.

The above properties do not require any distributional assumptions for the forecast revisions, beyond being iid with zero mean and covariance Σ_f. However, we will see that the distribution of the production rate and inventory random variables will depend on the distribution of forecast revisions. We will assume that the production rate and inventory random variables are normally distributed, which will be the case if we assume that the forecast revisions are normally distributed.

The MMFE process is descriptive of the planning process at many firms (see Graves et al. 1986, 1998) Abhyankar and Graves, 2001). In these cases, the initial forecast conveys the progress at identifying customers and in securing advanced orders. Subsequently, the forecast revisions correspond to changes to the master schedule, which reflect the success at converting the forecast (open orders) into demand (confirmed orders).

3.3 Work Release Rule

We assume the following linear release rule:

(1) $$r_t(t) = \alpha \times f_t(t) + (1-\alpha) \times r_{t-1}(t-1)$$

where

- $r_t(s)$ is the planned release rate in units or pieces for period s, determined in period t, s>t; when s=t it is the actual release rate for period t.
- $f_t(s)$ is the demand forecast for the segment for period s, determined in period t, s>t; when s=t it is the realized or actual demand for period t.
- α is a smoothing parameter, lying in (0, 1]. It is a decision variable to be determined for the segment.

In each period, we release at a rate that is a combination of two quantities: the first is the current demand rate, reflecting a desire to respond to the demand; the second is the previous release rate, reflecting a desire to keep the work release constant or near constant. The α parameter is a smoothing parameter, analogous to that for an exponential smoothing model: the smaller it is, the less responsive will be the work release rate, and the smoother will be the release process; the larger it is, the more responsive will be the release process, but it will then be more variable.

This rule permits some level of tractability. We can infer from (1) the planned release process, $r_t(s)$, for s > t. This is important as the planned release process for one segment is the forecast of the demand process for its upstream segment. Furthermore, we show next that the rule preserves the structure of the forecast process and dampens downstream variability as it gets passed upstream.

Result 1: If $f_t(s)$ is a MMFE process, then $r_t(s)$ will also be a MMFE process, albeit with a different covariance matrix.

Proof:

$$
\begin{aligned}
&r_t(t) = \alpha \times f_t(t) + (1-\alpha) \times r_{t-1}(t-1) \\
&= \sum_{i=0}^{t-1} \alpha (1-\alpha)^i f_{t-i}(t-i) + (1-\alpha)^t r_0(0) \\
&= \sum_{i=0}^{t-1} \alpha (1-\alpha)^i \left(\sum_{j=0}^{t-i-1} \Delta f_{t-i-j}(t-i) + f_0(t-i) \right) + (1-\alpha)^t r_0(0) \\
&= \sum_{i=0}^{t-1} \sum_{k=i}^{t-1} \alpha (1-\alpha)^i \Delta f_{t-k}(t-i) + \sum_{i=0}^{t-1} \alpha (1-\alpha)^i f_0(t-i) + (1-\alpha)^t r_0(0) \\
&= \sum_{k=0}^{t-1} \sum_{i=0}^{k} \alpha (1-\alpha)^i \Delta f_{t-k}(t-i) + r_0(t) \\
&= \sum_{k=0}^{t-1} \Delta r_{t-k}(t) + r_0(t)
\end{aligned}
$$

where we have defined:

$$r_0(0) = f_0(0)$$

$$r_0(t) = \sum_{i=0}^{t-1} \alpha (1-\alpha)^i f_0(t-i) + (1-\alpha)^t f_0(0)$$

$$\Delta r_{t-k}(t) = \sum_{i=0}^{k} \alpha (1-\alpha)^i \Delta f_{t-k}(t-i) \text{ for } 0 \le k \le t-1$$

We find it helpful to rewrite the revision above in equivalent form as:

$$\Delta r_t(t+k) = \sum_{i=0}^{k} \alpha (1-\alpha)^i \Delta f_t(t+k-i) \text{ for } t \ge 1, k \ge 0$$

From this we observe that if f$_t$(s) is a MMFE process, then the forecast revision vector $\underline{\Delta r_t}$ is iid, with zero mean. As we can write: $r_t\left(t\right)=r_0\left(t\right)+\sum_{k=0}^{t-1}\Delta r_{t-k}\left(t\right)$, we conclude that the planned release process r$_t$(s) is a MMFE process. □

We note that there is no forecast horizon for the planned releases, as derived above. When we revise the demand forecast in period t (i.e., when we observe Δf$_t$(t+i), for i = 0, 1,…H), all future planned releases are affected: that is, the release revision, Δr$_t$(t+k), is a linear combination of the forecast revisions, Δf$_t$(t+m) for m=0,1,..,H for all k≥0. This will be problematic for implementation, as the revision vector for releases $\underline{\Delta r_t}$ is of infinite dimension. Hence, we limit the revision vector to the forecast horizon H by redefining Δr$_t$(t+H):

$$\Delta r_t\left(t+H\right)=\sum_{i=H}^{\infty}\Delta r_t\left(t+i\right)\text{ for }t\geq 1$$

With this modification, we can express the revisions in period t in matrix form as $\underline{\Delta r_t}=M_1\underline{\Delta f_t}$ where M$_1$ is the (H+1)x(H+1) weight matrix. The ith column of M$_1$ has zeros above the diagonal and then α on the diagonal, followed by geometric weights $\alpha\left(1-\alpha\right)^{j-i}$ for row j, for i<j<H+1. We set the weights in the last row to be: $\left(1-\alpha\right)^{H},\left(1-\alpha\right)^{H-1},\ldots\left(1-\alpha\right),1$. We note that the weights in each column sum to one. Then the covariance matrix for the release process, r$_t$(t), is given by $\Sigma_r=M_1\Sigma_f M_1^T$.

For an MMFE process f$_t$(s), the variance of the demand, namely f$_t$(t), is given by the trace of the covariance matrix for the forecast revisions, namely, $tr\left(\Sigma_f\right)$. Similarly, the variance of the release process, r$_t$(t), is given by, $tr\left(\Sigma_f\right)=tr\left(M_1\Sigma_f M_1^T\right)$. We find that this release variance can be approximated well by the following:

$$\begin{aligned}Var\left(r_t\left(t\right)\right)&=tr\left(M_1\Sigma_f M_1^T\right)\approx\left(\frac{\alpha}{2-\alpha}\right)\left(\alpha\times tr\left(\Sigma_f\right)+\left(1-\alpha\right)\times sum\left(\Sigma_f\right)\right)\\&=\left(\frac{\alpha}{2-\alpha}\right)\left(\alpha\times Var\left(f_t\left(t\right)\right)+\left(1-\alpha\right)\times sum\left(\Sigma_f\right)\right)\end{aligned}\tag{2}$$

where $sum\left(\Sigma\right)$ represents the sum of all of the elements in the matrix Σ. The expression (2) is asymptotically exact as $H\rightarrow\infty$ when Σ_f is diagonal or tridiagonal, as is shown in Chhaochhria (2011).

Additionally, we find it to be a good approximation provided the non-diagonal terms of Σ_f are small relative to the diagonal terms, as is expected. Furthermore, we can show that:

$$sum\left(\Sigma_r\right) = sum\left(M_1 \Sigma_f M_1^T\right) = sum\left(\Sigma_f\right)$$

Hence the second term of (2) remains a constant, as we move from one segment to another. This is significant as we now have an easy way to model how the demand variability in a segment gets passed upstream to the next segment. In particular, let the variance of the demand process for a segment be denoted as $Var\left(f_t\left(t\right)\right) = \sigma_{OUT}^2$; then for a specified smoothing parameter α, we model the release variance of the segment $Var\left(r_t\left(t\right)\right) = \sigma_{IN}^2$, by:

(3) $$\sigma_{IN}^2 = \left(\frac{\alpha}{2-\alpha}\right)\left(\alpha \times \sigma_{OUT}^2 + \left(1-\alpha\right) \times sum\left(\Sigma_f\right)\right)$$

3.4 Production Rate Rule

We set the production rate by the linear control rule:

(4) $$p_t\left(t\right) = \mu + \beta \times \left(X - x_t\left(t\right)\right)$$

where

- $p_t(s)$ is the planned production rate for period s, as of period t, s>t; when s=t it is the production rate for period t, equal to the amount that completes processing in t.
- $x_t(s)$ is the planned buffer inventory for the segment for period s, determined in period t, s>t; when s=t it is the buffer inventory at the start of period t. Note that this is the inventory level *after* demand for the current period has been filled.
- μ is a mean production rate which is equal to the mean demand rate for the system.
- X is a target inventory level for the segment after the current period's demand has been filled. X is the safety stock and will be determined for the segment.
- β is a smoothing parameter, lying in (0, 1]. It is a decision variable to be determined for the segment.

In each period, we set the production at a rate that is a combination of two quantities: the mean production rate, reflecting a desire to keep the production at a constant level, and the deviation from the target inventory level, reflecting a desire to respond to the demand. The β parameter allows for smoothing; the smaller it is, the less responsive will be the production rate to demand, and vice-versa. The inventory dynamics are given by the following balance equation:

(5) $$x_t\left(t\right) = x_{t-1}\left(t-1\right) + p_{t-1}\left(t-1\right) - f_t\left(t\right)$$

Using the inventory balance equation with the production control rule, we observe that if the demand forecast is an MMFE process, then the production process is also a MMFE process:

Result 2

If $f_t(s)$ is a MMFE process, then $p_t(s)$ will also be a MMFE process.

Proof:

$$p_t\left(t\right) = \mu + \beta \times \left(X - x_t\left(t\right)\right)$$

$$= \mu + \beta \times \left(X - \left(X - \sum_{i=0}^{t-1}\sum_{k=0}^{i}\left(1-\beta\right)^k \times \Delta f_{t-i}\left(t-k\right)\right)\right)$$

$$= \mu + \beta \times \sum_{i=0}^{t-1}\sum_{k=0}^{i}\left(1-\beta\right)^k \times \Delta f_{t-i}\left(t-k\right)$$

$$= p_0\left(t\right) + \sum_{i=0}^{t-1}\Delta p_{t-i}\left(t\right)$$

Where we have defined:

$$p_0\left(t\right) = \mu$$

$$\Delta p_{t-i}\left(t\right) = \beta \times \sum_{k=0}^{i}\left(1-\beta\right)^k \times \Delta f_{t-i}\left(t-k\right)$$

$$or\ \Delta p_t\left(t+i\right) = \beta \times \sum_{k=0}^{i}\left(1-\beta\right)^k \times \Delta f_t\left(t+i-k\right)$$

Again, we note that under the assumptions of the MMFE model, $\Delta f_t(t+H+i)=0$ for $i>0$. Thus, the production revision vector, $\underline{\Delta p_t}$, is a linear combination of the forecast revision vector, $\underline{\Delta f_t}$. Since the revisions to the forecast, $\underline{\Delta f_t}$, are iid and have zero mean, the revisions to the planned production, $\underline{\Delta p_t}$, are iid and have zero mean also. Hence the production process $p_t(t+i)$ is an MMFE process.□

To estimate the average amount of overtime for each process step in the segment, we need to determine the variance of the production rate: $Var\left(p_t\left(t\right)\right) = \beta^2 Var\left(x_t\left(t\right)\right)$. We find this by calculating the variance of the buffer inventory.

3.5 Buffer Inventory Analysis

By replacing the previous period production term in (5) using the production rate rule (4), and simplifying, gives us:

$$
\begin{aligned}
x(t) &= x(t-1)+\left(\mu+\beta\times\left(X-x(t-1)\right)\right)-f_t(t)\\
&=(1-\beta)\times x(t-1)+\left(\mu-f_t(t)\right)+\beta X\\
&=(1-\beta)\times\left((1-\beta)\times x(t-2)+\left(\mu-f_{t-1}(t-1)\right)+\beta X\right)+\left(\mu-f_t(t)\right)+\beta X\\
&=\sum_{k=0}^{t-1}(1-\beta)^k\times\left(\mu-f_{t-k}(t-k)\right)+(1-\beta)^t\times x(0)+\sum_{k=0}^{t-1}(1-\beta)^k\times\beta X\\
&=\sum_{k=0}^{t-1}(1-\beta)^k\times\left(\mu-f_{t-k}(t-k)\right)+X
\end{aligned}
$$

where we assume $x(0)=X$, the target buffer inventory level. Given that the demand process is a MMFE process, we show next that the inventory process is also a MMFE process. This will enable us to calculate the inventory variance.

Result 3

If $f_t(s)$ is a MMFE process, then $x_t(s)$ will also be a MMFE process, albeit with a different covariance matrix.

Proof:

$$
\begin{aligned}
x(t) &= \sum_{k=0}^{t-1}(1-\beta)^k\times\left(\mu-f_{t-k}(t-k)\right)+X\\
&= X+\sum_{k=0}^{t-1}(1-\beta)^k\times\left(\mu-\sum_{j=0}^{t-k-1}\Delta f_{t-k-j}(t-k)-f_0(t-k)\right)\\
&= X-\sum_{k=0}^{t-1}(1-\beta)^k\times\left(\sum_{j=0}^{t-k-1}\Delta f_{t-k-j}(t-k)\right)\\
&= X-\sum_{j=0}^{t-1}\sum_{k=0}^{t-1-j}(1-\beta)^k\times\Delta f_{t-k-j}(t-k)\\
&= X-\sum_{i=0}^{t-1}\sum_{k=0}^{i}(1-\beta)^k\times\Delta f_{t-i}(t-k)\\
&= x_0(t)+\sum_{i=0}^{t-1}\Delta x_{t-i}(t)
\end{aligned}
$$

Where we define:

$$f_0(t-k)=\mu \quad \forall \text{k}$$

$$x_0(t)=X$$

$$\Delta x_{t-i}(t)=-\sum_{k=0}^{i}(1-\beta)^k\times\Delta f_{t-i}(t-k)$$

We find it helpful to rewrite the revision above in equivalent form as:

$$\Delta x_t(t+k)=-\sum_{i=0}^{k}(1-\beta)^i\,\Delta f_t(t+k-i) \text{ for } t\geq 1, k\geq 0$$

Similar to the prior results, there is no forecast horizon for the revisions to the planned inventory. To facilitate implementation, we limit the revisions to the forecast Horizon H, by redefining $\Delta x_t(t+H)$: $\Delta x_t\left(t+H\right)=\sum_{i=H}^{\infty}\Delta x_t\left(t+i\right)$ for $t\geq 1$.

We now can express the revisions in period t in matrix form as $\underline{\Delta x_t}=-M_2\underline{\Delta f_t}$ where M_2 is the (H+1)x(H+1) weight matrix. The ith column of M_2 has zeros above the diagonal and then 1 on the diagonal, followed by geometric weights $\left(1-\beta\right)^{j-i}$ for row j, i<j<H+1. We set the weights in the last row to be: $\left(1-\beta\right)^H/\beta,\left(1-\beta\right)^{H-1}/\beta,\ldots\left(1-\beta\right)^1/\beta,1/\beta$.

The covariance matrix for the MMFE inventory process is given by $\Sigma_x=M_2\Sigma_f M_2^T$. Since the revisions to the forecast, $\underline{\Delta f_t}$, are iid and have zero mean, the revisions to the planned inventory, $\underline{\Delta x_t}$, are iid and have zero mean. Hence the buffer inventory $x_t(t+i)$ is an MMFE process.□

Similar to the analysis of the release policy, we find that the inventory variance can be approximated well by the following expression:

$$\begin{aligned}(6)\qquad & Var\left(x_t\left(t\right)\right)=tr\left(M_2\Sigma_f M_2^T\right)\\ &\approx\frac{1}{2\beta-\beta^2}\times\left(\beta\times tr\left(\Sigma_f\right)+\left(1-\beta\right)\times sum\left(\Sigma_f\right)\right)\\ &=\frac{1}{2\beta-\beta^2}\times\left(\beta\times\sigma_{OUT}^2+\left(1-\beta\right)\times sum\left(\Sigma_f\right)\right)\end{aligned}$$

As in the case of the release variance, this expression is asymptotically exact when Σ_f is diagonal, and is a good approximation while the non-diagonal terms of Σ_f are small relative to the diagonal terms.

The inventory variance expression reflects the variability due to the responsiveness of the smoothing rule. The greater the smoothing (i.e, smaller β), the more variable will be the inventory, as it takes longer to respond to any variation in the demand.

We suggest setting the inventory target $X=z\times\sqrt{Var\left(x_t\left(t\right)\right)}$ where z is a safety factor set to achieve some service level target. In the implementation of the model, we assume that the inventory is normally distributed and set z accordingly. This would be the case if the forecast revisions were normally distributed.

3.6 Work-in-Process Inventory Analysis

There are two parts to the work-in-process inventory in this system. The first part is the minimum planned inventory at each process; this is the inventory due to the batch-forming, queuing times and service times of the process. The second part of the WIP is due to the difference in the work release and production rates in each segment; for example, if the work release is very responsive to demand ($\alpha \sim 1$), and the production is very smooth ($\beta \sim 0$) in a segment, then we need to hold a large amount of WIP so that even in periods of low demand, we are able to meet the desired production rate. This additional WIP is zero if the work release and production rates were synchronized.

We estimate the minimum planned inventory before each process by using Little's Law and queuing model approximations, as described in Hopp and Spearman (2007). We estimate the total expected waiting time for a work-piece at a process step by considering the time to form a batch (for batch processes), the waiting time in queue, and the actual service time. We then multiply it by the arrival rate to get the expected planned inventory attributable to the total waiting time. See Chhaochhria (2011) for the details.

However, this minimum WIP at each process step assumes that the work release and production rates are the same for the segment. If these rates differ from period to period, then there can be a shortfall in the WIP needed to achieve the desired production. Thus, we may need to hold additional quantities of WIP in order to meet our desired production rate when the work release and production rates are not synchronized and can differ. We assume that we can model the additional WIP at the end of the period as:

$$w_t(t) = w_{t-1}(t-1) + r_t(t) - p_t(t)$$

We can write the planned WIP in terms of the planned releases and production:

$$w_t(t+k) = w_t(t+k-1) + r_t(t+k) - p_t(t+k) \qquad \forall \mathrm{k} \ge 1$$

We can express the WIP revisions in terms of the planned release and production revisions as:

$$\begin{aligned}
&\Delta w_t(t+k) = w_t(t+k) - w_{t-1}(t+k) \qquad \forall \mathrm{k} \ge 0 \\
&= \Delta w_t(t+k-1) + \Delta r_t(t+k) - \Delta p_t(t+k) \\
&\Rightarrow \Delta w_t(t+k) = \sum_{j=0}^{k} \left(\Delta r_t(t+j) - \Delta p_t(t+j) \right)
\end{aligned}$$

We can substitute for the planned release and production revisions, using results developed the proofs of Results 1 and 2, and we get:

$$\begin{aligned}\Delta w_t(t+k) &= \sum_{j=0}^{k}\left(\Delta r_t(t+j)-\Delta p_t(t+j)\right)\\ &= \sum_{j=0}^{k}\sum_{i=0}^{j}\left(\alpha(1-\alpha)^i-\beta(1-\beta)^i\right)\Delta f_t(t+j-i)\\ &= \sum_{m=0}^{k}\sum_{n=0}^{k-m}\left(\alpha(1-\alpha)^n-\beta(1-\beta)^n\right)\Delta f_t(t+m)\\ &= \sum_{m=0}^{k}\left((1-\alpha)^{k-m+1}-(1-\beta)^{k-m+1}\right)\Delta f_t(t+m)\end{aligned}$$

To approximate the variance of the additional WIP, we will assume that Σ_f is diagonal; i.e., we assume each forecast revision in a period is independent of the other revisions in the same period. Then we can express the variance of the WIP revision as:

$$Var\left[\Delta w_t(t+k)\right]=\sum_{m=0}^{k}\left((1-\alpha)^{k-m+1}-(1-\beta)^{k-m+1}\right)^2 Var\left[\Delta f_t(t+m)\right]$$

We can now compute the variance of the WIP random variable as shown below:

$$\begin{aligned}&w_t(t)=\sum_{k=1}^{t}\Delta w_k(t)\\ &\Rightarrow Var\left[w_t(t)\right]=\sum_{k=1}^{t}Var\left[\Delta w_k(t)\right]=\sum_{k=0}^{t-1}Var\left[\Delta w_k(t+k)\right]\end{aligned}$$

As $t\to\infty$, we can simplify and write the above expression as :

$$\begin{aligned}Var\left[w_t(t)\right]&=\sum_{k=0}^{\infty}Var\left[\Delta w_k(t+k)\right]\\ &=\sum_{k=0}^{\infty}\sum_{m=0}^{k}\left((1-\alpha)^{k-m+1}-(1-\beta)^{k-m+1}\right)^2 Var\left[\Delta f_t(t+m)\right]\\ &=\sum_{m=0}^{\infty}\sum_{j=1}^{\infty}\left((1-\alpha)^{j}-(1-\beta)^{j}\right)^2 Var\left[\Delta f_t(t+m)\right]\\ &=\left(\frac{(1-\alpha)^2}{1-(1-\alpha)^2}-\frac{2(1-\alpha)(1-\beta)}{1-(1-\alpha)(1-\beta)}+\frac{(1-\beta)^2}{1-(1-\beta)^2}\right)\sum_{m=0}^{\infty}Var\left[\Delta f_t(t+m)\right]\\ &=\left(\frac{(1-\alpha)^2}{1-(1-\alpha)^2}-\frac{2(1-\alpha)(1-\beta)}{1-(1-\alpha)(1-\beta)}+\frac{(1-\beta)^2}{1-(1-\beta)^2}\right)Var\left[f_t(t)\right]\end{aligned} \tag{7}$$

We set the additional WIP level as: $W=z\times\sqrt{Var\left(w_t(t)\right)}$ where z is a safety factor set so that the probability of stocking out of this additional WIP is low. In the implementation of the model, we use the above approximation for Var(w_t(t)) and assume that this additional WIP is normally distributed and set z accordingly.

3.7 Production Overtime Cost Calculation

From (4) and (6), we can calculate the production rate variance:

(8)
$$\begin{aligned} Var\left(p_t(t)\right) &= \beta^2 Var\left(x_t(t)\right) \\ &= \frac{\beta}{2-\beta}\times\left(\beta\times\sigma_{OUT}^2+(1-\beta)\times sum\left(\Sigma_f\right)\right) \end{aligned}$$

We now model the processing time (Ω) on the bottleneck process step in the segment as a random variable. Namely, we have:

$$E[\Omega]=w\,E\left[p_t(t)\right]=w\mu,\quad Var[\Omega]=w^2 Var\left[p_t(t)\right]$$

where w is the processing time per work-piece at the bottleneck. We denote the nominal capacity for the bottleneck as χ , which is the amount of regular time available each period. Then the expected overtime per period for the segment is:

(9)
$$E\left[(\Omega-\chi)^+\right]$$

where we assume Ω is normally distributed. Thus the expression for overtime is just the expectation from a normal partial loss function.

3.8 Raw Material Ordering and Inventory Policy

The raw material ordering process at the research sponsor required specialized modeling due to its peculiarities. Raw material orders are placed N times a year, evenly spaced over the year. Each order has staggered deliveries: if the order is placed at month t, the first installment of raw material is scheduled to be delivered in month (t+L), the next installment delivered in month t+L+1, and so on, with the last delivery in month [t+L+(12/N)-1], where L is the lead time. However, each order installment (and all subsequent installments) can be delayed, with some known probability. We refer the reader to Chhachhria (2011) for the detailed development.

3.9 Finished Goods Inventory Analysis

The finished goods inventory policy at the research sponsor also required specialized modeling due to its peculiarities. Orders that are due in the current period are filled from the finished goods inventory at the start of the period. The order is late if it cannot be completely filled; for instance, if 49 pieces are available for an order requiring 50 units, the order is late until the 50th piece is done. Furthermore, a penalty cost is applied to the entire order size for the entire duration of its delay. For example, if the penalty cost is $100 per piece per day, an order for 50 pieces that is filled 2 days late incurs a penalty cost of $10,000 ($100 x 50 pieces x 2 days). See Chhachhria (2011) for the detailed development.

3.10 Segment cost computation & control parameter optimization

We now have the elements we need to evaluate the operating cost of a segment. We define $C\left(i, j, \sigma_{IN}, \sigma_{OUT}\right)$ to be the operating cost for a segment from process step i+1 to j, with demand variance, Var(f_t(t)) = σ_{OUT}^2, and with release variance, Var(r_t(t)) = σ_{IN}^2. We do this with the following steps:

For given values of $\left(\sigma_{IN}, \sigma_{OUT}\right)$ we obtain the smoothing parameter α by solving the following:

$$Var\left(r_t\left(t\right)\right) = \sigma_{IN}^2 = \left(\frac{\alpha}{2-\alpha}\right)\left(\alpha \times \sigma_{OUT}^2 + \left(1-\alpha\right) \times sum\left(\Sigma_f\right)\right)$$

where the $sum\left(\Sigma_f\right)$ is a constant for all segments. We note that the value of α is obtained by solving a quadratic equation, which can have two roots. In the rare event that both roots are in (0,1], we choose the larger root as it implies less production smoothing and leads to smaller inventory variance, and hence lower safety stocks.

We set the parameter β to minimize the operating cost of the segment, equal to the holding cost for the buffer inventory and the additional WIP, plus the expected overtime cost:

BufferInvCost = HoldingCost × (AvgCycleStock + SafetyStock)

$$= h \cdot \left(\frac{\mu}{2} + z \times \sqrt{Var\left(x_t\left(t\right)\right)}\right)$$

WIPCost = HoldingCost × Additional-WIP

$$= h \times z \times \sqrt{Var\left(w_t\left(t\right)\right)}$$

OT_Cost = OvertimeCost/hr × Expected_Overtime

$$= ot \cdot E\left[\left(\Omega - \chi\right)^+\right]$$

We find β by a line search over the range (0, 1]. For the demand-facing segment we include the expected penalty cost in the operating costs. For the most upstream segment, we include the raw material costs.

4. Dynamic Programming Model

We use the single segment model within a dynamic program to find the optimal combination of segments that span the production line; this provides the locations of the decoupling buffers and the optimal operating policies for each segment. The dynamic programming model is given by:

$$G\left(i, \sigma_{IN}\right) = \min_{j, \sigma_{OUT}} \left\{C\left(i, j, \sigma_{IN}, \sigma_{OUT}\right) + G\left(j, \sigma_{OUT}\right)\right\}$$

where

- $G\left(i,\sigma_{IN}\right)$ is the optimal operating cost for the production system downstream of process step i, assuming that there is a decoupling buffer after process step i, and the release variability from this buffer is given by σ_{IN} .
- $C\left(i, j,\sigma_{IN},\sigma_{OUT}\right)$ is the operating cost for a segment from process step i+1 to j, with demand variability given by σ_{OUT} and with release variability given by σ_{IN} . We define this only for $\sigma_{OUT} \geq \sigma_{IN}$. That is, a segment's releases are no more variable than its demand.
- The minimization is over $\left\{j,\sigma_{OUT} \mid j = i+1,...N,\sigma_{IN} \leq \sigma_{OUT} \leq \sigma_{Demand}\right\}$, i.e., j is the last process in the segment (i,j) after which the next decoupling buffer is located, and the release variability of this buffer is σ_{OUT} .

We have a boundary condition:

$$G\left(N,\sigma\right)=\begin{cases}0 \ \ for \ \ \sigma=\sigma_{Demand} \\ \infty \quad otherwise\end{cases}$$

where σ_{Demand}^{2} = Var(f_t(t)) = Trace(Σ_f) is the given variability of the external demand process. The optimal solution is given by:

$$G\left(i=0\right)=\min_{\sigma}\left\{G\left(i=0,\sigma\right)\right\}$$

where the minimization is over $\left\{\sigma \mid 0<\sigma\leq\sigma_{Demand}\right\}$, i.e., we choose the least cost system configuration G(i=0,σ) from the start of the production line (i=0) over all possible choices for the release variability from the raw material buffer

To solve this dynamic program numerically in reasonable time, we discretize the state space. We restrict the demand and release variability of a segment, σ_{OUT} and σ_{IN} respectively, to only take values from a finite set. This restriction may cause some sub-optimality in the solution but is a necessary trade-off in order to have reasonable computation times. However, we can control the optimality gap by having a larger set of allowable values with more granularity.

We calculate the operating costs of all possible segment-states given by the combination of (i, j, σ_{IN} , σ_{OUT}). This step involves finding the optimal value of the production smoothing parameter, β, for each segment-state. We then use the dynamic programming model to find the optimal combination of segment-states to span the production line while satisfying the constraints on the work release and demand variability between adjoining segment-states. We now establish the computational complexity of this method.

Let N be the number of process steps and R be the cardinality of the set of possible release and demand variances for each segment. Then the number of segment-states $\left(i, j, \sigma_{IN}, \sigma_{OUT}\right)$ is bounded by $\frac{N\left(N+1\right)}{2}\frac{R\left(R+1\right)}{2}$, which corresponds to the number of cost computations. For instance, the number of segment-states for a 30-process serial-flow system with 5 possible values for production and demand variances is bounded by 6,975. For each segment state, the largest computational effort is the calculation of the optimal production smoothing parameter, β, which is done be a line search over (0, 1]. The solution of the dynamic program is comparable in complexity to solving an acyclic shortest path with NR nodes, that is one node for every state in the state space of the cost-to-go function G(i, σ_{IN}).

To get a sense of the computation time, we solved test cases with N=30 and R=5; on a Windows laptop each problem instance took about 10 seconds to solve, with over 90% of the time used to calculate the costs of all possible segment-states in the system.

5. MODEL TESTING

In order to verify and validate the results of the tactical model, we developed a software prototype of the model using C#. This software takes as inputs the process flow information of the serial-flow system, the product demand statistics, the raw material ordering constraints, and user parameter settings. It then sets up and runs the dynamic program to find the optimal locations for decoupling buffers and the optimal operating policies for each of the segments. The software computes and reports all of the relevant costs, as well as the statistics for the inventory and production levels. Users can restrict the optimization by limiting the options for decoupling buffer locations, or by setting a maximum on the number of processes allowed in a segment, or by specifying a permissible range for the value of each control parameter.

We also developed a discrete-event simulation for a serial-flow system using C# (Huang, 2010). The purpose of the simulation is to test the accuracy of the tactical model, and in particular to assess the impact of key assumptions and approximations in the tactical model. The simulation takes as input the process flow information of the serial-flow system, the product demand statistics, the locations of the decoupling buffers, and operating policies for each segment. It then simulates all the operations of the factory under the process flow and the control rules for a number of years and records the inventory, production, order fulfillment and other outcomes for each day. For each simulation run, it computes statistical results for costs, order fulfillment and other performance metrics. We can compare these with the tactical model predictions for model validation.

We first tested the model on a hypothetical shop. These experiments are reported in Chhaochhria (2011). Here we report on numerical experiments on a real manufacturing system under a variety of scenarios. These tests help us validate the tactical model, its assumptions and cost predictions.

5.1 Manufacturing System

We obtained demand, process flow, and cost information for a real manufacturing system from our research sponsor. This serial-flow manufacturing system produces three families of discrete parts in a make-to-forecast setting. It has 94 processes including cutting, grinding, heat-treatment, coating, inspection, among others; these processes are shared by all of the product families. The production resources associated with each process step can entail both parallel machines and shared workers. Some process steps are batch processes, while a few others are outsourced to a subcontractor. The system includes several inspection steps, which can result in rework and repeat processes. Each process step can have setup times to switch between product types. Production is scheduled for two shifts a day, with an option for working overtime as needed. The system operates in a low volume, long lead time, production environment with total monthly demand of about 350 pieces and standard deviation of about 170. The forecast horizon is 24 months. We illustrate in Figure 3 how the forecast accuracy evolves over the horizon. The production lead-time is about 4 months. Two of the products share a common raw material, while the third product has its own raw material. The two raw materials are ordered twice a year, and have a six month lead-time. Deliveries are staggered monthly from the seventh to the twelfth month after order placement, and each delivery may be delayed by up to one month. Chhaochhria (2011) provides some additional details characterizing the process flow, including information on batch size, number of machines, and capacity utilization for each process step.

Insert Figure 3 here

We ran the tactical model software using the actual factory data. We restricted both σ_{IN} , σ_{OUT} to the set $\left\{a\sigma_{Demand}, \textit{for } a = 0.1, 0.2, \ldots 1\right\}$ for each product. The optimal solution for the base case entailed a single segment for the production line. The optimal operating parameter values and normalized costs are shown in Table 1. The upper section of the table shows the breakdown of costs by type of cost. The costs are normalized by dividing by the total cost of the one-segment solution, so as to add up to 100. The lower section of the table shows operating costs and parameters for each segment.

Insert Table 1 here

For this one-segment solution, there is some production smoothing (β=0.6) common to all products, and the work release is smoothed as well (α=0.61 for each product). Interestingly, the optimal α and β are so close that the expected variance in the WIP is almost negligible, thus requiring very little additional WIP and suggesting a constant WIP (CONWIP) policy. The WIP for each product is about 4 months of demand. This is in line with expectations given the production lead-time. The optimal FGI safety factor differs for each product due to the different inventory holding and penalty costs. Thus, the buffer safety stock and average inventory level are also significantly different for each of the products.

In practice there may be organizational reasons to limit the size of a segment. When we limit the segment to at most 50 processes, the optimal solution has two segments with a decoupling buffer between processes 45 and 46. The optimal operating parameter values, normalized inventory levels and costs (normalized to the one-segment solution total cost) are shown in Table 2.

Insert Table 2 here

The upstream segment has no production smoothing or work release smoothing (β=1.0, and α=1.0 for all three products), but the downstream segment has significant production and work release smoothing (β=0.5, and α=0.49 for all three products). This reduces the expected overtime and raw material inventory for this solution relative to the one-segment solution, but leads to higher WIP and buffer inventory levels. The buffer inventory cost is significantly higher than the one-segment solution due to the decoupling buffer that leads to more cycle stock being maintained in the system. The total cost is higher than the one-segment solution by 11.5%. We also considered tighter constraints on the segment size to produce solutions with 3 and 4 segments; these results are in Chhaochhria (2011).

5.2 Simulation Results

In this section, we validate the tactical model and verify the results shown in the previous section. We do this by comparing the tactical model's cost and inventory predictions with simulation results of factory performance under the tactical model's optimal operating policies.

We ran the simulation software using the actual factory data on demand, processes and costs. The simulation explicitly modeled the machines and workers at each process step; that is, both the availability of machine and a worker are required at certain steps in order to process a work-piece. The software also simulated the demand and forecast processes: at the start of a month, the software generated both new orders that are added to the list of orders, as well as changes in due dates to existing orders and order cancellations. The simulated forecast reflects the current list of orders, the expected new orders that will arrive, and expected future changes to these orders (including cancellations). Then orders that are currently due are filled from inventory, or backlogged if there is inadequate inventory.

The simulation replicated the policies as set by the tactical model. Decoupling buffers were located as suggested by the tactical model. For each segment, the work release and production rates are determined based on the rules given earlier, using the parameters given by the tactical model. Work is released from upstream buffers into segments, and backlogged if the upstream buffer has inadequate inventory. Completed work-pieces from each segment go into the downstream buffer of the segment. The backlogged orders and releases are then filled at the earliest possible opportunity as inventory is replenished by production. Overtime is scheduled as needed, based on the production rule. Performance metrics such as inventory at each process and in buffers, and overtime at each process, are recorded at the end of each day in the simulation.

The software simulated the factory operations for 30 years, and computed the average daily cost, overtime and inventory results for the run. In order to reduce the impact of variability in the simulation results, we performed 20 simulation runs and compared the median simulation performance with the tactical model results. These tests help us validate the tactical model's assumptions and cost predictions, and also verify its optimality.

We first compare the one-segment production line performance. The average inventory normalized to months of demand) and cost of operations (normalized by the one-segment tactical model solution) from the simulation along with the tactical model predicted inventory and costs are shown in the Table 3.

Insert Table 3 here

The raw material and WIP levels in the simulation are very similar to those in the tactical model, but finished goods level is somewhat higher in the simulation for all three products. The inventory holding costs for raw material and WIP are similar in the simulation and the tactical model, but the finished goods inventory and overtime costs are significantly different. This can be explained by the difference in operations between the simulation and the tactical model. In the tactical model, we assume that the production quantity in a period is equal to the calculated target for the period, even if it means not using available capacity. For example, if the production rule sets the production rate to 6 hours in an 8-hour day, then the tactical model assumes that this is exactly what is produced in each day. However, the simulation, reflecting real factory conditions, continues producing more parts even after the production target of 6 hours is reached if capacity and material are available, up to the length of a normal day (say, 8 hours). This effect pushes the WIP in the segment to the downstream buffer. So the simulation's buffer inventory level is almost double the tactical model's buffer inventory. Having a larger buffer inventory leads to lower delay penalties and less overtime in the simulation, bringing the total cost in line with the tactical model.

We did a similar comparison for the two-segment production line performance. The results are shown in the Table 4.

Insert Table 4 here

Again, the raw material levels (and costs) in the simulation are very similar to those in the tactical model, but the WIP is lower and buffer inventory is somewhat higher in the simulation for all three products. The reason for these differences is the same operational difference between the simulation and the tactical model as explained in the previous case. In the simulation, more of the segment WIP is processed and sent to the two buffers (intermediate buffer and finished goods), increasing the buffer inventory while reducing the WIP remaining in the segment. This in turn reduces the overtime and order delay penalty costs in the simulation. However, the total inventory of raw material, WIP and buffer stocks is very similar in the simulation and the tactical model. The total cost is again within 10% of the tactical model prediction.

In Chhaochhria (2011), we report results for a three-segment and four-segment production line performances. In both cases, we see the same pattern of inventory and cost differences; the total costs are again within 10% of the tactical model prediction.

6. CONCLUSIONS

The simulation results help us validate the tactical model: the approximations used in the tactical model analysis seem to work well as evidenced by the closely matching simulation results. The assumptions made in the tactical model, particularly about the decoupling buffers always meeting the desired work release requirement (A3) and the desired production target always being met (A4), seem to be reasonable with the buffers rarely stocking out and production targets almost always being satisfied in the simulation. Since the simulation modeled the actual factory in significant detail (in terms of the machines, workers, process-flow, costs, product demand) and implemented the control rules as envisaged by the tactical model, it can be considered as reasonably representative of the actual factory.

We see in the simulation results that even when the desired work release rate is partially met due to insufficient buffer inventory at the start of the segment, this only leads to a small delay in the release. This is because the upstream segment produces more parts within the period and replenishes the buffer, which in turn releases these parts into the downstream segment and fulfills the shortfall within the period. Thus, the disruptive effects on the system are very limited. Also, due to the difference in operations between the simulation and the tactical model, there is usually more inventory in the buffer than planned, which leads to less overtime needed. Hence, there is rarely a situation when the desired production rate cannot be met due to capacity constraints.

6.1 Managerial Insights

The tests provide insight about where decoupling buffers should be placed, and when should production smoothing be applied. They also highlight the trade-off between the different types of inventory and overtime costs when smoothing production, and help to establish the potential value from optimizing these tactics. We see three clear instances when decoupling buffers should be introduced:

- Before high value-add processes at which the inventory cost increases significantly. By decoupling the system at this point, we make an investment in a low cost inventory that allows one to optimize the upstream processes with substantially less inventory costs. This decoupling inventory also helps to minimize the more expensive inventory that is downstream.
- Before highly capacity-constrained or “bottleneck” processes. Decoupling prior to a bottleneck allows the upstream segment to operate with a different production policy than the downstream segment. This can result in less overtime costs for both segments.
- Between two sets of process that should operate at different levels of smoothing; for example, between two segments of processes, one with low-overtime, high-inventory holding costs and the other with high-overtime, low-inventory holding costs

We see significant value from reducing upstream variance propagation using the work release and production smoothing tactics. For manufacturing systems with high levels of utilization, there can be a substantial reduction in overtime costs and raw material holding costs that more than offsets the increase in finished goods inventory costs.

We see the trade-off between finished goods inventory and overtime costs in the choice of optimal production smoothing parameter, β. There is a high degree of smoothing (low β) when capacity is tight and overtime costs are high relative to the finished goods holding costs. Correspondingly, little or no smoothing (high β) is seen when utilization is low or when overtime costs are low relative to the finished goods holding costs.

The optimal choice of the work release smoothing parameter, α, is often quite close to that for the production smoothing parameter, β. This minimizes the variance of the WIP in the segment, implying that good solutions tend to have a constant level of WIP in the segment. This suggests that the optimal workflow policy is often a CONWIP type policy (Hopp and Spearman 2007).

6.2 Research Contribution and Limitations

We introduce a new framework to model forecast-driven manufacturing systems. We show that by modeling the forecast as an MMFE process, and having linear control rules for work release and production smoothing, the resulting work release, production and inventory processes are also MMFE processes. As the work release for a segment is effectively the demand forecast for its upstream segment, this allows us to use the same single-stage model for all production segments. This facilitates the analysis and provides for computational tractability.

We also develop a dynamic program to optimize the planning tactics: where to place decoupling buffers in a serial-line and what inventory and production control policies to use in each segment. We demonstrate the computational feasibility of the dynamic program and test it over a wide variety of cases, yielding good results in reasonable solve times.

As to limitations, this model was developed with a specific manufacturing system as the motivation. This model is designed for serial systems and cannot be used for assembly or job-shop type systems. It is also unsuitable for a high-volume demand environment or a short lead-time production environment. Chhaochhria (2011) develops an analogous model for a high-volume, multi-product assembly systems with changeovers. Some of the assumptions, particularly A3 and A4, are rather strong. Although these assumptions seem to “work” in the test environments, one might investigate how to build a more robust model that considers these supply and production capacity issues rather than assume that they do not occur.

References

1. Abhyankar H.S. and S.C. Graves. Creating an inventory hedge for Markov-modulated Poisson demand: An application and a model. Manufacturing & Service Operations Management, 3, 306-320, 2001.

2. Altug M.S. and A. Muharremoglu. Inventory management with advance supply information. Int. J. Production Economics, 129, 302-313, 2011.

3. Balakrishnan A., Geunes, J., Pangburn, M. Coordinating supply chains by controlling upstream variability propagation. Manufacturing & Service Operations Management, 6(2), 163-183, 2004.

4. Beckmann M. J. Production Smoothing and Inventory Control. Operations Research, 9, 456-467, 1961.

5. Bertrand J.W.M. Balancing production level variations and inventory variations in complex production systems. International Journal of Production Research, 24(5), 1059-1074, 1986.

6. Chhachhria, Pallav. Forecast-driven Tactical Planning Models for Manufacturing Systems. Ph. D. Thesis, Operations Research center, MIT, Cambridge MA, 2011.

7. Gallego G., and O. Ozer. Integrating replenishment decisions with advance order information. Management Science, 47, 1344-1360, 2001.

8. Gallego G., and O. Ozer. Optimal replenishment policies for multi-echelon inventory problems under advance demand information. Manufacturing and Service Operations Management, 5, 157-175, 2003.

9. Gaver D.P. Operating Characteristics of a Simple Production-Inventory Control Model. Operations Research, 9, 635-649, 1961.

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

11. Graves S. C., H. C. Meal, S. Dasu and Y. Qiu. "Two-Stage Production Planning in a Dynamic Environment" in Axsater S., C. Schneeweiss and E. Silver,(Eds.), Multi-Stage Production Planning and Inventory Control, Lecture Notes in Economics and Mathematical Systems, Vol. 266, Springer-Verlag, Berlin, pp 9-43, 1986.

12. Graves S., D.B. Kletter and W.B. Hetzel. A Dynamic Model for Requirements Planning with Application to Supply Chain Optimization. Operations Research, 46(3), S35-49, 1998.

13. Graves S.C. and S.P. Willems. Optimizing strategic safety stock placement in supply chains. Manufacturing & Service Operations Management, 2, 68-83, 2000.

14. Hausman W.H. and R. Peterson. Multiproduct Production Scheduling for Style Goods with Limited Capacity, Forecast Revisions and Terminal Delivery. Management Science, 18(7): 370-383, 1972.

15. Heath D.C. and P.L. Jackson, Modeling the Evolution of Demand Forecasts with Application to Safety Stock Analysis in Production/Distribution Systems, IIE Transactions, 26(3), 17-30, 1994.

16. Hoffman A.J. and W. Jacobs. Smooth Patterns of Production. Management Science, 1, 86-91, 1954.

17. Hollywood J. S. Performance Evaluation and Optimization Models for Processing Networks with Queue-Dependent Production Quantities. PhD Thesis, MIT, 2000.

18. Hopp W.J. and M.L. Spearman. Factory Physics (3rd edition). McGraw-Hill, 2007.

19. Huang B. A tactical planning model for a serial-flow manufacturing system. MS Thesis, MIT, 2010.

20. Iida T. and Zipkin P.H. Approximate Solutions of a Dynamic Forecast-Inventory Model. Manufacturing and Service Operations Management, 8(4), 407-425, 2006.

21. Johnson S. and G. Dantzig. A Production Smoothing Problem. Proceeding of 2nd Symposium in LP, 1, 151-176, 1955.

22. Kaminsky P. and J.M. Swaminathan. Utilizing Forecast Band Refinement for Capacitated Production Planning. Manufacturing and Service Operations Management, 3(1), 68-81, 2001.

23. Lu X., J. Song and A. Regan. Inventory Planning with Forecast Updates: Approximate Solutions and Cost Error Bounds. Operations Research, 54(6), 1079-1097, 2006.

24. Manne A.S. Note on the Modigliani-Hohn Production Smoothing Model. Management Science, 3, 371-379, 1957.

25. Modigliani, F. and F.Hohn. Production Planning Over Time. Econometrica, 23, 46-66, 1955.

26. Simon H. On the application of servomechanism theory in the study of production control. Econometrica, 20, 247-268, 1952.

27. Schoenmeyr T. and S.C. Graves. Strategic Safety Stocks in Supply Chains with Evolving Forecasts. Manufacturing and Service Operations Management, Articles in Advance, 1-17, 2009.

28. Simpson K.F. In-process inventories. Operations Research, 6(6), 863-873, 1958.

29. Symonds G.H. Stochastic Scheduling by the Horizon Method. Management Science, 8, 138-167, 1962.

30. Teo Chee Chong. A Tactical Planning Model for Make-To-Order Environment under Demand Uncertainty. PhD Thesis, NTU, 2006.

31. Toktay L.B. and L.M. Wein. Analysis of a Forecasting Production Inventory system with Stationary Demand. Management Science, 47(9), 1268-1281, 2001.