Setting Optimal Production Lot Sizes and Planned Lead Times in a Job Shop System

by

Rong Yuan

B.Eng. Industrial and Manufacturing Systems Engineering
University of Hong Kong, 2010

Submitted to the School of Engineering
in partial fulfillment of the requirements for the degree of

Master of Science in Computation for Design and Optimization

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2013

Author .....
School of Engineering
May 5, 2013

Certified by .....
Stephen C. Graves
Abraham J. Siegel Professor of Management Science & Engineering Systems
Thesis Supervisor

Accepted by .....
Nicolas Hadjiconstantinou
Professor of Mechanical Engineering
Co-director, Computation for Design and Optimization

Setting Optimal Production Lot Sizes and Planned Lead Times in a Job Shop System

by

Rong Yuan

Submitted to the School of Engineering

in partial fulfillment of the requirements for the degree of

Master of Science in Computation for Design and Optimization

Abstract

In this research we model a job shop that produces a set of discrete parts in a make-to-stock setting. The intent of the research is to develop a planning model to determine the optimal operating tactics that minimize the relevant manufacturing costs subject to workload variability and capacity limits. We model the interplay of three key components in the job shop, namely, the production frequency for each part, the variability of production at each work station, and the level of parts inventory. We consider two operating tactics (decision variables): the production lot size for each part and the planned lead time for each work station.

We model the relevant manufacturing costs, entailing production overtime costs and inventory-related costs (finished parts, work-in-process, and raw materials), as functions of these decision variables. We formulate a non-linear optimization model and implement it in the Excel Spreadsheet. We solve the model with the premium Excel Solver to determine the minimum-cost operating tactics. We test the model with both hypothetical and actual factory data from our research sponsor. The target factory processes 133 product parts on 59 work stations. The results are consistent with our intuition and demonstrate the potential value from optimizing over these tactics; these tests also provide some managerial insights on the application of these operating tactics.

Thesis Supervisor: Stephen C. Graves

Title: Abraham J. Siegel Professor of Management Science & Engineering Systems

Acknowledgement

I am most grateful for having Professor Stephen Graves as my academic advisor during the past three years at MIT. In Professor Graves's research group, I have been given the opportunities to work on real industry projects using analytical tools. I appreciated his patience and rigor in guiding me through every step in those exciting projects. I not only benefit from his analytical insights towards mathematical modeling, but also learn from his professional experience in managing projects with industrial partners. In particular, I would like to thank Professor Graves for financing me to various academic conferences and business trips and spending a great amount of time help me shape and polish my presentations and thesis. All of these are really instrumental for me to develop necessary skills for my future career.

I want to thank Pallav Chhaochhria, a PhD graduate from MIT Operations Research Center, for his guidance and encouragement in my first year at MIT. I thank Chen Song, a current CDO student who is also in Graves's group, for working with me side by side in my second year at MIT. I thank Ketan Nayak, a current CDO student who is also working with Professor Graves, for helpful discussions on my work. It was my honor to work with them at MIT.

This thesis cannot be completed without the support from our industrial partner Mitsubishi Heavy Industry. In particular, I would like to thank Manabu Kasano and Koki Tateishi for providing the access to the data and valuable suggestions on the project.

I also want to thank Laura Kollar and Barbara Lechner, the former and current CDO program coordinators, for offering a great amount of help on the academic procedures. They have been playing really supportive and caring throughout the entire period of my studies at the CDO program. I am really grateful to have both of you as my

program coordinators.

Finally, I dedicate this work to my parents back in China for their love and support.

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

Job Shop System

Our work is motivated by real-world planning challenges faced by manufacturers who run complex manufacturing systems such as job shops. A job shop contains multiple types of work stations that process multiple types of jobs or families of items. A great many manufacturing facilities can be described as job shop systems - machine tool shops, part fabrication shops, paint shops, commercial printing shops, to name a few.

Each product family (or job type) produced by the job shop can have a distinct processing route and thus the pattern of the work flow in a job shop is often quite complicated. This often leads to long product flow times due to system congestion and due to the complexity of flow planning. Typically a job spends a majority of its time waiting in the queues in front of work stations in job shops.

The planning and scheduling for a job shop has long been recognized as a hard problem. The performance of a job shop depends on the operating tactics applied, including production sequence scheduling, queueing rules, lot-sizing decisions, buffer-time planning, to name a few. One common objective is to meet the quoted delivery times for each job with the least manufacturing costs subject to fluctuating workload. There is an extensive literature dedicated to improving the operational efficiency in job shop systems and we review the most relevant ones in Chapter II.

Common Strategies to Demand Uncertainties

Manufacturers operating with job shop systems face new business challenges nowadays with an ever faster product life cycle and greater demand uncertainties. Manufacturers are not only required to provide quality products, but also better supply chain performance with short and reliable delivery lead times. To remain competitive in the market, manufacturers seek better operating tactics to improve their efficiency in production and inventory management without increasing operating costs.

Common strategies to deal with demand uncertainties include building safety stock inventory and investing in production flexibility. A higher inventory level is associated with higher inventory management costs, higher capital losses, and risks of product obsolesce. Investment in production flexibility might entail the acquisition of more production equipment and tooling to raise the production capacity above the average demand or the development of an expediting capability such as working overtime or sub-contracting. There exists a trade-off between these two strategies, namely one can choose to increase inventory levels to handle anticipated fluctuations in demand while keeping a steady production level, or to invest in production flexibility that would allow the production level to vary with demand, resulting in lower inventory requirements.

An alternative strategy for building more final inventories (or downstream inventory) is to build more work-in-process inventories at the upstream stages of production. Many manufactures find this alternative useful since it mitigates the risk of holding large amount of products at the final stage. Moreover, inventory holding cost at upstream stages is often less expensive. And in some contexts there might be risk pooling benefits from holding an upstream inventory if a raw part or intermediate part is common to several final products. A challenging question associated with this strategy, however, is how to design the buffer sizes at each stage of production line.

In some environments, the manufacturer may be able to vary the quoted delivery times to customers; in this way the manufacturer might be able to accommodate varying demand by adjusting the delivery times while keeping its production level smooth. Possibly the manufacturer can associate a delay cost with each type of customer and choose to satisfy the most important customers first when production capacity is tight. However, this may not always be a good strategy since varying the delivery time might hurt the business in the long-run.

Our Business Case

Our work is motivated by the real world planning challenges faced by a large factory. The factory produces a variety of piston pumps that are used as sub-assemblies for high value industrial applications such as iron manufacturing machinery, deck cranes, and concrete pump cylinders. The factory operates in a business environment characterized by short lead times, and with a moderate number of customers who view the products (i.e., the pumps) as important components in the assembly of their final products.

The factory manufactures 14 product models and each product model is an assembly of about a 10 to 20 parts. Since product models can share common parts, this translates into 133 parts in total to be produced in the factory. The factory consists of a job shop that manufactures each part, plus assembly lines that assemble sub-assemblies and then the final assembly for the product models (see Figure 1).

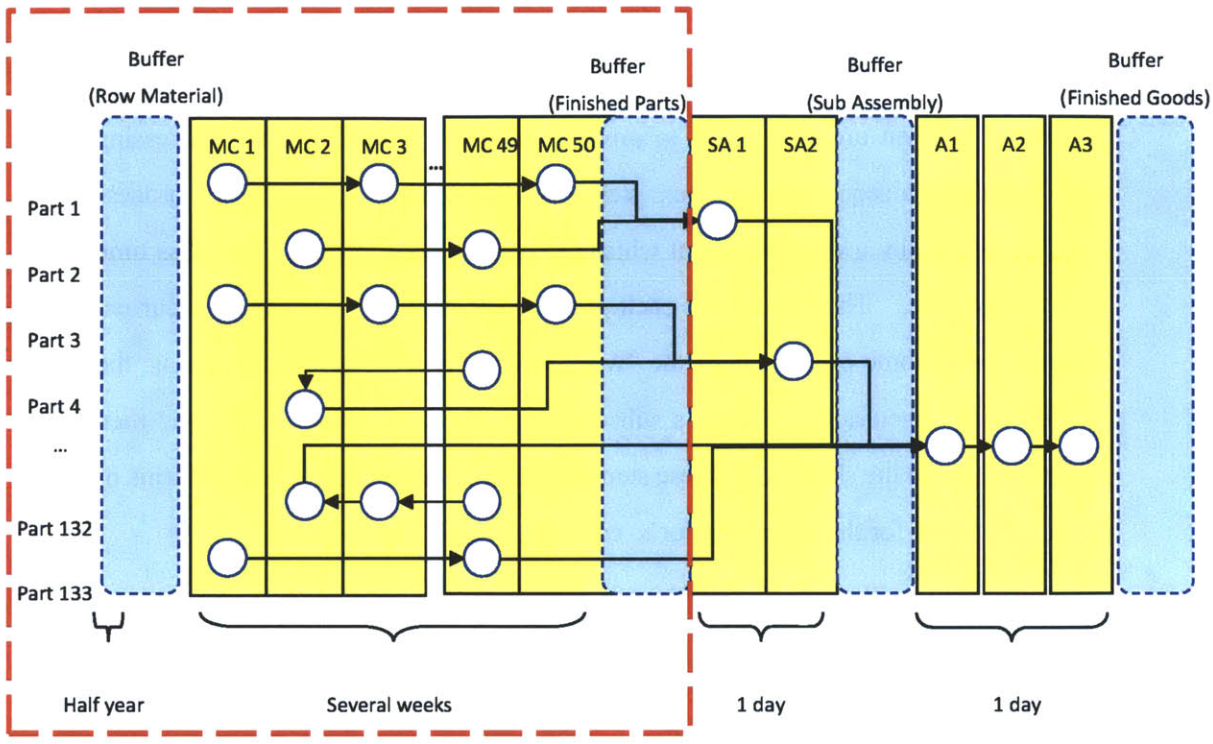

Figure 1: The Target Factory: Job Shop, Sub-assembly Line and Assembly Line

Our focus is on the job shop, which consists of 59 work stations and processes 133 parts. The assembly lines which have relatively short lead times will not be considered in this research.

The operating policy for the factory is to hold an inventory of each part in a finished parts inventory, and then follow an assemble–to-order policy for meeting demand. That is, when a demand order is received, the factory will pull the necessary parts from the finished parts inventory and then build the product assemblies to meet the order. Our understanding is that this is a feasible policy, as once an order is received there is sufficient time to perform the assembly, provided the parts are available. The assembly lines are quite flexible and the assembly times are relatively short.

The job shop produces each part to stock, namely to the finished parts inventory. A typical part starts with a piece of raw material, and then gets processed through a series of work stations in the job shop. Parts are produced in lots in the job shop

with production flow times of a few weeks. The production release for a lot is triggered by the finished parts inventory; that is, when the parts inventory drops below a reorder level, an order is issued to produce another lot for the part. Processing each lot entails a sequence of process steps, usually around a dozen. Each process step is specified by a work station at which the work is done, a per-unit process time and a setup time. The lot size for each product part is set by experience in current practice. For some process steps the "work station" is a subcontractor; that is, the job leaves the shop and is sent to a subcontractor who performs the work and then returns the job to the shop. For these steps, the specification of the step is in terms of the time allowed for the subcontractor to complete the job (e.g., 5 days).

The congestion level is currently high in the job shop. Due to high demand variability and the strict requirement for on-time delivery, the manufacturer currently spends a lot of money on expediting activities such as working overtime. The manufacturer is reluctant to build up a large finished parts inventory to handle the fluctuating demand due to physical constraints and the holding costs. The manufacturer is thus seeking tactics to reduce overall production overtime without compromising its service level or adding more inventories in the system.

1.2 Two Operating Tactics

We consider two operating tactics, namely the production lot sizes and planned lead times, which are used to manage the work flow in the factory in order to reduce aggregate workload of the system, smooth the production workload at each station, and reduce the inventory levels.

Production Lot Sizes

A lot consists of multiple units of the same type of part. Raw materials are first pulled from the raw materials inventory to form a lot according to the pre-determined lot size. A lot then enters the job shop and arrives at the queue of a work station, waiting to be processed. The processing of a lot typically involves a single setup followed by a processing time for each unit in the lot. The lot becomes available to be moved to the next stage only when the entire lot is complete.

Increasing the lot size will reduce the number of setups required and consequently reduce workload on the work stations. If setup time is small, the time that a lot spends at a work station increases almost linearly with the lot size. When the lot size is large, a lot needs longer time to be processed and more work will arrive during the time the lot being processed; an arriving lot also sees more work waiting ahead of it and thus will have longer waiting time.

The mechanics described above lead to a complex trade-off presented by lot-sizing decisions. On the one hand, large lot sizes cause higher work-in-process inventories since they tend to induce longer waiting times. Large lot sizes also imply a less fluid work flow which results in more lumpy arrivals to each work station. An increase in arrival variability translates into higher production variability which leads to more queuing. On the other hand, small lot sizes can reduce production variability, but imply more frequent setups and changeovers which will increase the workload utilization on the work stations. The higher workload induces greater congestion which eventually drives up both inventory and overtime costs.

According to the above reasoning, lot sizes are usually set larger for items requiring larger setup times and for items that heavily use the bottleneck work stations. Nevertheless, analyzing the effect of lot-sizing in a job shop is difficult since changing one item's lot size potentially has an impact on all other items; similarly, the flow time for an item can be potentially affected by the lot-sizing decisions made for

other items.

Planned Lead Times

In the job shop environment, the heterogeneity and complexity of the work flow creates great difficulties for production planning and scheduling. Actually, it is common to see that we have long queues in front of some work stations, while others are idle.

The planned lead times have been used as key inputs for Material Requirement Planning to project how long each job should spend in each production step or work station. The idea is to associate a reasonable amount of time for each work station and each product item based on the analysis of their workload and routes.

The planned lead time is always longer than the required processing time; it includes an additional buffer time in order to allow for queuing and to provide planning flexibility. Prescribing the planned lead time on each work station is an effective approach to regulate work flow within the system in Material Requirement Planning.

The planned lead times are usually determined according to a management policy, taking into account the trade-off between the level of inventory and production variability. Long planned lead times imply higher level of work-in-process inventory (by Little's law) and higher level of safety stocks of finished goods. However, long planned lead times can help to dampen the workload requirements at each work station, and thus can effectively smooth out production and reduce production overtime.

1.3 Model Overview and Thesis Outline

We model a job shop that produces a set of discrete parts in a make-to-stock environment. The intent is to develop a planning tool to determine the optimal operating tactics that minimize the relevant manufacturing costs subject to workload variability and capacity limits.

Our model considers the two sets of operating tactics (decision variables): the production lot size for each product part and the planned lead time for each work station. Our model provides tractability for a general job shop system with any type of flows. Our emphasis is on capturing the key interaction of demand and production variability, production capacity and the level of inventory requirement.

We model the relevant manufacturing costs, including the production overtime costs and the inventory-related costs (finished parts, work-in-process, and raw materials), as functions of the production lot size (or production frequency) for each part and the planned lead time for each work station. We pose a non-linear optimization model in Microsoft Excel spreadsheet and solve the model with the premium Excel build-in Solver.

Our model can be used to devise optimal lot-sizing policies or evaluate the performance of a factory for given lot-sizing policies. Our model can also provide a guideline for setting the planned lead times for each work station; these planned lead times can then be used to guide overtime decisions, the appropriate release schedule for the job shop, and the replenishment lead times for a make-to-stock job shop. In a make-to-order job shop the planned lead times will inform the quotation of the delivery times for customer jobs.

The model has been tested with data from the job shop from our research sponsor. The analysis shows that the model captures the key uncertainties and trade-offs in the manufacturing system, and determines the minimal-cost operating tactics. The

chosen tactics closely match the manufacturer's existing operations and our model provides new managerial insights on the application of these operating tactics.

The remainder of the thesis is as follows:

In Chapter 2 we review the related literature and point out how our work complements the previous contributions to tactical planning in job shop systems. Chapter 3 provides an extensive review on the Tactical Planning Model that is in the center of our model. Chapter 4 describes the assumptions and the formulation of the model. Chapter 5 discusses the implementation of the model in the Microsoft Excel environment and presents the numerical test results on three test cases of different scales. We finally conclude our research in Chapter 6.

2 Literature Review

There has been a continuous effort in understanding and estimating the performance of complex manufacturing systems like job shops. The well-known Jackson's Network (Jackson 1957, 1963) is one of the first models that describes the queueing dynamics in a job shop and provides insights into the relationship between capacity planning and job flow times. There is now an abundant literature on queueing networks that extends the Jackson's model. We refer the reader to Bitran and Dasu (1992) for an extensive review on queueing network models of the job-shop like manufacturing systems.

There are two streams of literatures that are in particular relevant to our work. The relationship of the literatures with our work is shown in Figure 2.

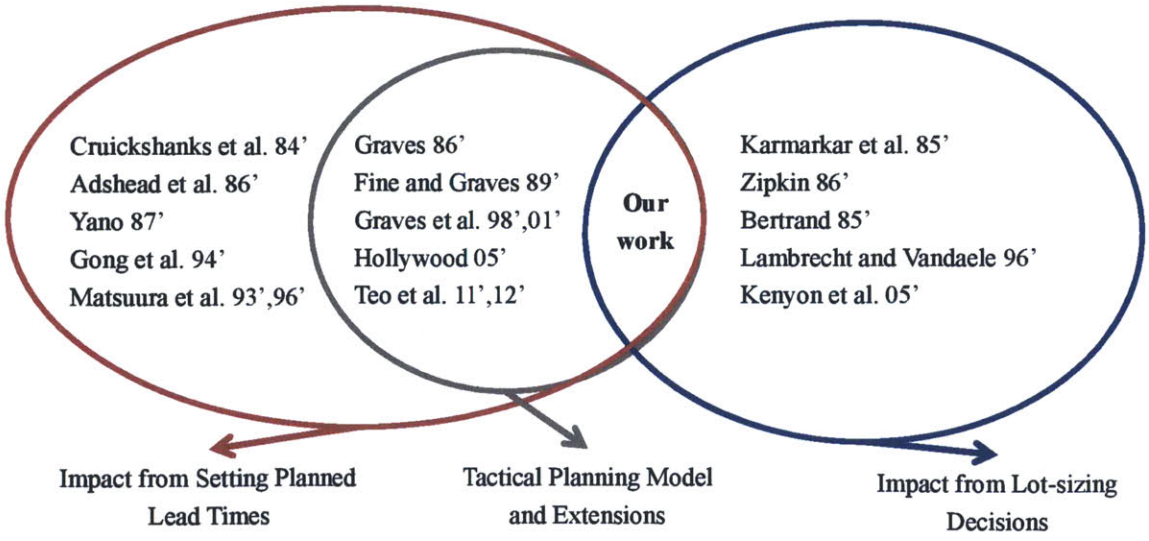

Figure 2: Relationship of the Literatures with Our Work

One stream of research works focus on setting *Planned Lead Times*. The planned lead time of a job is defined as the time span one plans for a job to spend at a step or stage in the system. The major question is how to set planned lead times and how to quantify the economic benefits when using planned lead times as control variables. Cruickshanks et al. (1984) first introduced the concept of a *Planning Window* in the environment of a job shop. Their model showed that small changes in the planning window can provide substantial benefits for production smoothing, in exchange for an increase in the flow time allowed for each order. Adshead et al. (1986) experimented with different rules in a simulation model in order to understand how to set planned lead times for each job under different circumstances. They found when the lead time is tight with respect to the due delivery time, an assignment rule based on the workload and number of operations in the job usually results in better performance. Efforts have also been paid on developing analytical models to determine the optimal planned lead times in terms of inventory holding and job tardiness costs. Yano (1987) and Gong et al. (1994) showed that there exists optimal planned lead times for serial production system in terms of inventory holding and tardiness costs. They also described computational procedures to find the optimal solutions. Matsuura and Tsubone (1993) developed a single-stage model to minimize workload variability under a particular rule of setting the planned lead times. Matsuura et al. (1996) extended the single-stage model to a multiple-stage one with the objective of minimizing the workload variability for the bottleneck machine in the system.

Our work is based on the Tactical Planning Model (TPM) developed by Graves (1986). The TPM describes a network of work stations in a job shop and characterizes the operational behavior within the system. In the TPM model, the production rates are determined according to the workload in the queue. In particular, it applies the linear production rule which sets the production rate as a fixed proportion of the queue level at the beginning of the time period for each work station. The TPM then associates a planned lead time with each work station and determines the first two

moments of the production levels at each work station, as well as the distribution of queue lengths. (The linear production rule has later been generalized as clearing functions that specify the fraction of current work-in-process to be "cleared" by a resource in a given period of time. We refer the reader to Karmarkar (1989), Karmarkar (1993), Asmundsson et al. (2006), and Selcuk et al. (2008) for more literature along this line of research.)

The TPM provides a tractable way to characterize the planned lead times in complex manufacturing systems such as job shops and to model the relationship to production smoothing. The model can be used to model the variability in the shop, balance the WIP inventory levels vis-à-vis the capacity levels for equipment and labor, examine the consequences when the workload release rule is changed, and evaluate the network/layout design of the job shop.

There are several works that extend the TPM to different contexts. Fine and Graves (1989) applied the model to a manufacturing line of components for mainframe computers. Graves and Hollywood (2001) and Hollywood (2005) adapted the model to allow for a CONWIP policy for work release. Graves et al. (1998) applied the similar linear control framework to a requirements planning context with an evolving demand forecast process. Teo et al. (2011, 2012) extended the model to production contexts that are planned by means of a Master Scheduling Planning framework. Teo et al. (2011) also relaxed the discrete-time assumption in the TPM and introduced a finer control to the system with a continuous-time model. Since the TPM and the continuous-time version of it serve as important building blocks of our work, we will review the major developments in Chapter 3.

The other stream of the research focuses on the impact of lot-sizing policies on the performance of job shops. Similar to the classic Economic Lot-Size Model (Wagner and Whitin 1958), conventional research has focused on the trade-off between the setup costs by producing too many small lots and the additional inventory holding

cost by tying up individual products in a larger lot. Karmarkar et al. (1985a), however, pointed out that in production contexts there can be another important tradeoff, namely the impact of lot sizes on manufacturing lead times could be substantial. They explained that the performance of the job shops can depend on the lot-sizing policies since queuing delays are directly related to lot sizes. In particular, they showed that the average waiting time in queue is a convex function of the lot size and consequently lot sizes can affect work-in-process costs, safety stock requirements, and scheduling performance through queueing behavior. In their model, they assume that lot arrivals can be modeled as a Poisson process, so standard queueing models such as M/M/1 and M/G/1 can be applied. Karmarkar et al. (1985b) extended the analysis to multiple-item and multiple-machine environment where each machine is modeled as an M/G/1 system. To obtain the optimal lot size for each product item, they solved an optimization problem with the total inventory and work-in-process costs as the objective function.

The model developed independently by Zipkin (1986) explored the performance of multiple-item production system with lot sizes in a similar way except it does not explicitly consider the impact of lot sizes on queueing delays. Zipkin's model also uses M/M/1 and M/G/1 queueing models as building blocks and it includes backlogging costs in the optimization problem. Bertrand (1985) also analyzes the impact of lot sizes on inventory, work-in-process and fixed ordering costs in a multi-item multi-station setting. Bertrand models each work station as a queueing system but he uses a subroutine to calculate queue length and does not consider the setup time on the work stations in the model. Finally, we note that Lambrecht and Vandaele (1996) analyzed the impact of lot sizes on queueing delay in a more general setting using approximation techniques. Kenyon et al. (2005) studied the same problem using statistical and simulation techniques which provided similar insights.

From the literatures mentioned above, we can observe that there have been two streams of research on the tactical planning level of the job shop systems. One of

them focuses on setting planned lead times in accordance with production planning activities. The other stream focuses on finding the optimal lot-sizing policies in order to achieve the balance between queueing delays and number of setups. Essentially, both streams aim at the same goal, namely to minimize the manufacturing costs in the system by choosing the optimal operating tactics. This motivates us to apply the two sets of tactics in the same shot in order to further move towards the global optimum and have a more comprehensive understanding of the system.

Our model explicitly sets the planned lead time for each work station and the lot size for each product part. Our focus is to investigate the interaction of demand and production variability, production capacity and the level of inventory requirement according to the changes in planned lead time and lot-sizing decisions.

3 Review of the Tactical Model

3.1 Introduction of Tactical Planning Model

Job shops are manufacturing systems that contain multiple work stations and process a wide variety of product parts. In a job shop system, each product part typically has a distinct process route through the work stations, along which an ordered set of tasks are performed. Due to the large variety of product parts (each with their own routing) in the system, there is usually no dominant work flow path and there is often significant uncertainty in process routing and workloads.

While job shops are designed to provide highly customized and flexible services, their production control has been known as a difficult question. It is not uncommon for a job to spend most of its time waiting for the work stations to become available. Furthermore, a control rule for job shops should also be relatively robust considering the heterogeneity of the routings and complexity of the system.

Graves (1986) formalized the Tactical Planning Model (TPM) to explore the interrelationship of production capacity, uncertainty of production requirements, and the level of work-in-process inventory in a complex discrete-part manufacturing operation, as typified by a job shop.

The TPM is a discrete-time, continuous flow framework. It assumes an underlying time period (e.g. an hour or a day) and assumes that all job movements can only occur at the start (or end) of each time period. The TPM models the job flow in the system in terms of the workloads imposed on the work stations, rather than in terms of

individual jobs. The TPM assumes that each work station operates with a linear production control rule, based on the planned lead time at the work station.

The TPM views the job shop as a queueing system with planned lead times. At each work station, arriving jobs wait in front of it in a queue. Each work station is assigned a planned lead time which represents the time, both waiting and in-process, that we *plan* for a job to spend at that work station. While the actual time spent at the work station will vary about the planned lead time, the model assumes that each work station will vary its production rate so that each job's time at the work station is close to the planned lead time.

The TPM is not concerned with the detailed scheduling of the system. Instead of suggesting how to prioritize the jobs on each work station, the model focuses on planning the operations at a tactical level, i.e., how to regulate the aggregate work flow in the system so as to achieve an efficient work flow in terms of flow time and resource utilization. Further theoretical extensions and applications of the TPM can be found in Fine and Graves (1989), Graves (1988), and Graves and Hollywood (2001).

A well-known model that describes the queueing behavior of a job shop is the Jackson's network. A key difference between the TPM and Jackson's network is in terms of the intent of these models. The primary intent of the Jackson's model is performance evaluation of the system. The intent of the TPM is to provide guidance in setting operational tactics for a production system; in particular, it can help to decide the planned lead times for a job shop. Jackson's network assumes that a work station will process its queue according to a fixed production rate whenever the queue is not empty. The TPM on the other hand assumes that a work station can adjust its production rate according to the queue length in front of it. In other words, the TPM assumes the shop has some degree of flexibility to control the speed of its production. This assumption taken by the TPM is consistent with the observations that most job shops can adjust their production rates to some extent by shifting the work force,

working overtime, outsourcing, etc. Furthermore, the TPM tracks the aggregate workload of jobs instead of tracking individual jobs as Jackson's network does.

Our model considers a make-to-stock job shop system that involves a large amount of overtime costs due to expediting activities. One of our major concerns is how to achieve the right balance of production overtime and inventory holding cost. The TPM provides an approach to model this tradeoff. In particular, the TPM that we develop will characterize the production at each work station as a random variable, whose first two moments depend on the tactical decision variables, namely the lot sizes and the planned lead times. With a few necessary assumptions, this provides a handy way for us to express the cost components as functions of these decision variables. For these reasons, we apply the TPM as the building block to our model.

3.2 Key Assumptions of the TPM

The TPM assumes there is no explicit capacity limit and that the production rate can vary according to the queue size. More specifically, the TPM assumes that whenever the queue grows at a work station, it can always work at a higher production rate to ensure the queue gets processed (on average) at the planned lead time. The assumption is only appropriate if adjusting the production rate is possible at the factory. However, we note that a capacity limit can be implicitly modeled by setting planned lead time reasonably.

The TPM assumes that each work station produces a stable mix of jobs and that the work flow has Markovian property, i.e., the work flow from one work station i to another work station j only depends on station i and does not depend on how the work got to station i.

The TPM assumes that work moves at discrete-time periods in the system, i.e., all movement of jobs from one work station to another, as well as new job arrivals can

only occur at the start of a time period. The length of the time period is pre-determined based on the context of the manufacturing system. On the one hand, the time period should be short enough so that a job is unlikely to move through two consecutive work stations in one time period. On the other hand, the time period should be long enough so that a work station is always able to complete a certain number of jobs within one period of time. In real practice, the time period is usually decided to be consistent with the time periods for the planning system, e.g. a shift or a day. The validity of the discrete-time assumption depends on the context of the manufacturing system. For example, if we consider the system where jobs usually travel through several work stations within one time period, we would prefer to set a smaller time period. We introduce later an extension of the TPM to a continuous-time model to overcome this limitation.

3.3 Discrete-time Model

In the TPM, Each work station operates with a linear production control rule, i.e. the amount of work processed at each time period is a fixed portion of the queue in front of that work station*:

$$P_t = \alpha Q_t \quad (3.1)$$

Where P_t is the amount of work processed in time period t, Q_t is the queue level at the start of time t, and $\alpha(\in (0,1])$ is a smoothing parameter which specifies the proportion of work in the queue that should be completed within one unit of time period. This linear relationship between production rate and queue size describes the dynamics when the queue grows, the work station will speed up; when the queue is small, the work station will slow down. Note that both variables are expressed in terms of the amount of work load (in time units) at the work station.

* The original TPM models a network of stations but we only present the results for the single station in this chapter since in this research we only apply the results for the single station.

The linear production control rule is directly connected with the planned lead time. One can interpret the inverse of the smoothing parameter α as the planned lead time:

$$\tau = \frac{1}{\alpha} \quad (3.2)$$

Where τ is the planned lead time for the work station. If the planned lead time is τ time periods, then the work station on average is expected to process $1/\tau$ amount of the queue within each time period. This aligns exactly with the definition of α.

The system satisfies the inventory balance equation:

$$Q_t = Q_{t-1} - P_{t-1} + A_t \quad (3.3)$$

Where A_t is the amount of work that arrives at the work station at the start of time period t.

Substituting (1) into (3) we can express the production in a recursion equation:

$$P_t = (1-\alpha)P_{t-1} + \alpha A_t \quad (3.4)$$

Equation (4) is similar to an exponential smoothing equation. It limits the variance of the production rate by only taking α portion of the new arrivals into account at each time period.

By repeated substitution and assuming an infinite history of arrivals, we can express the production as an infinite sum:

$$P_t = \sum_{s=0}^{\infty} \alpha(1-\alpha)^s A_{t-s} \quad (3.5)$$

If we assume the arrival streams A_t is i.i.d. with mean $E[A]$ and $Var(A)$, one can determine the first two moments of production using (3.5):

$$E[P_t] = E[A] \quad (3.6)$$

$$Var(P_t) = \frac{\alpha Var(A)}{2-\alpha} \quad (3.7)$$

Although the arrival time series are usually correlated since some parts will serve for the same product model, as long as the number of parts is large (and no product model predominates the work flow), the dependence on an individual model will be slight. Thus assuming independent arrivals will provide fairly accurate approximation. This assumption has also been observed in previous literatures (Zipkin 1986).

To investigate the case where arrival time series are not i.i.d., one needs to consider the amount of work generated from one work station to the other work stations and the variability in the workload transition. The analysis is carried out in the original paper, when modeling a network of stations.

3.4 Continuous-time model

One limitation of the discrete-time TPM is the contradictory objectives when choosing the time period. The time period should be set short enough to ensure that a single job is unlikely to move across multiple work stations in a single time period. But on the other hand, the time period should be set long enough to maintain a fluid work flow in the system, or in other words, to allow a work station to process certain amount of jobs in one time period. Furthermore, the Markovian property of the work flow is based on the assumption of a stable work flow moving among work stations, which requires a longer time period.

A continuous-time extension of the original TPM has been developed (C.C. Teo et al. 2011) to overcome the limitation of the TPM discussed above. The continuous-time model permits work to flow through more than one work station within a time period and determines the first two moments of the production and queue level.

The continuous-time model essentially divides each time period t into m equal sub-periods s, where $s = 1,2,\ldots,m$. The length of each sub-period is defined as $\Delta = 1/m$.

Assuming that work arrives at work station at the start of each sub-period, we can restate the linear production rule in (3.1):

$$Y(\Delta, s) = \alpha \Delta X(\Delta, s) \quad s = 1,2, \ldots m \quad (3.8)$$

Where $Y(\Delta, s)$ is the production level in sub-period s of length Δ, $X(\Delta, s)$ is the queue length at the start of sub-period s. Here, $1/\alpha$ is still interpreted as the planned lead time (in time periods) associated with the work station; however, α can take any positive value, i.e., the planned lead time can be less than one time period.

The inventory balance equation becomes:

$$\begin{cases} X(\Delta, s) = Q_t + \dfrac{A_t}{m} & s = 1 \\ X(\Delta, s) = X(\Delta, s-1) - Y(\Delta, s-1) + \dfrac{A_t}{m} & s > 1 \end{cases} \quad (3.9)$$

Equation (3.9) assumes that the work arrivals A_t do not arrive at the start of time period t, but rather arrive uniformly over the period t. Note that for $s > 1$ we can substitute equation (3.8) into equation (3.9) and get rid of term $Y(\Delta, s-1)$.

The production at period t can be expressed as:

$$P_t = \sum_{s=1}^{m} Y(\Delta, s) = \alpha \Delta \sum_{s=1}^{m} X(\Delta, s) \quad (3.10)$$

We sum up the expressions for $X(\Delta, s)$ from (3.9), using (3.8), to find

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

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

Substituting (3.10) into the above expression, we obtain for the production:

$$P_t = Q_t + A_t - (1 - \alpha\Delta) X(\Delta, m) \quad (3.11)$$

The remaining work is to find $X(\Delta, m)$. From equation (3.9) we know

$X(\Delta, s) = (1 - \alpha\Delta)X(\Delta, s-1) + \frac{A_t}{m}$ for $s > 1$. We can then obtain $X(\Delta, m)$ by repeated substitution using equation (3.9):

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

Substitute the above expression of $X(\Delta, m)$ into (3.11), we can re-write the production as

$$P_t = Q_t + A_t - (1 - \alpha\Delta)^m Q_t - (1 - \alpha\Delta)\left(\frac{1 - (1 - \alpha\Delta)^m}{\alpha\Delta}\right)\frac{A_t}{m}$$
$$= (1 - (1 - \alpha\Delta)^m)Q_t + \left(1 - \left(\frac{1 - \alpha\Delta}{\alpha}\right)(1 - (1 - \alpha\Delta)^m)\right)A_t$$
$$= \beta Q_t + \gamma A_t \quad (3.12)$$

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

Using the standard inventory balance equation

$$Q_t = Q_{t-1} - P_{t-1} + A_{t-1} \quad (3.13)$$

Assuming an infinite history for the arrivals, the queue length at the start of period t, Q_t, can be expressed by repeated substituting (3.13) into (3.12):

$$Q_t = (1 - \gamma)\sum_{s=1}^{\infty}(1 - \beta)^{s-1}A_{t-s} \quad (3.14)$$

Production level at time t can then be expressed in terms of arrival streams by substituting (3.14) into (3.12):

$$P_t = \beta(1 - \gamma)\sum_{s=1}^{\infty}(1 - \beta)^{s-1}A_{t-s} + \gamma A_t \quad (3.15)$$

If we further assume the arrivals are i.i.d., then the model determines the first two moments of production from (3.15):

$$E[P_t] = E[A_t] \quad (3.16)$$

$$Var(P_t) = \left(\frac{\beta}{2-\beta}(1-\gamma)^2 + \gamma^2\right) Var(A_t) \quad (3.17)$$

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

If further letting $m \to \infty$, we obtain $\beta = 1 - e^{-\alpha}$ and $\gamma = 1 - \beta/\alpha$.

Comparing the results of the continuous-time model with the original TPM, we observe that the mean production level is the same as in the TPM but the continuous-time model has a smaller variance given the same smoothing parameter. The reason of this lies in the assumption made by the continuous-time model. The continuous-time model assumes work arrives uniformly within each time period which also implies the production level changes gradually and uniformly within the time period. However, in the original TPM, the production level can only be changed at the start of each time period and remains fixed for the rest of the time period. Thus, the original TPM is more sensitive to the arrival variability than the continuous-time model. This difference grows as the planned lead time of the work station becomes smaller. This effect is illustrated in Figure 2 for the single work station case where the continuous-time curve is drawn for m at infinity.

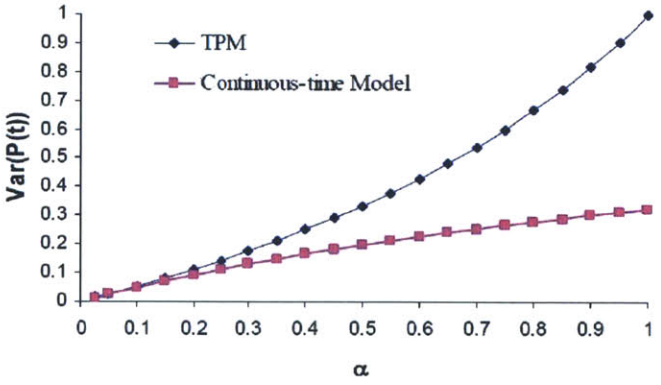

Figure 3: Comparison: Discrete and Continuous-time Tactical Planning Model[†]

[†] The continuous-time curve is drawn for m at infinity.

Here we provide an intuitive interpretation for the parameter m. In the model, a larger m corresponds to a more 'continuous' work flow arrival to the work station. In the factory practice, the value of m depends on the capability to adjust production rate. A larger m corresponds to a more frequent production adjustment rate. However, it is usually not practical to have a very large m since the production level cannot change too frequently in the factory. It makes more practical sense for the manager to set an appropriate adjustment frequency m based on the context of the manufacturing system and use formula (3.16) and (3.17) for the first two moments of production level accordingly. We apply those formulas in our model and please refer to section 4.2 for more details on setting m in practice.

4 Model Formulation

4.1 Model Assumptions

The basic idea of our model is to represent the inventory-related costs (raw materials, work-in-process and finished parts inventory) and production overtime costs in terms of the decision variables, namely the production lot size for each part and the planned lead time for each work station. The model needs to capture the essential trade-offs and interactions while maintaining tractability. We explain the assumptions for our model in this section.

Demand Process

A1. We assume that the daily demand for each product part follows an i.i.d. distribution with known first two moment parameters. This is a simplification as the demand for two parts will be highly correlated if the parts go into the same final assemblies. For instance, if part A and part B are only used for product model 123, then the two parts will effectively see the same demand process, as derived from the demand for product model 123. However, we think the independence assumption is reasonable given the large number of parts (133 in total) in the pilot factory. Moreover, most parts in the job shop go into a single product model and hence their demand is highly correlated with only a small subset of the entire part population as each assembly consists of about a dozen parts.

Inventory Review Policy

A2. We assume that a continuous review, reorder-quantity, reorder-point (QR) policy is used for the control of the inventory for each finished part. We assume for raw materials that we use a periodic review policy. We note that these policies are similar to that used in the target factory where value of parts in the finished parts inventory is much higher than that in the raw materials inventory.

A3. We assume that the raw material inventory does not stock out; hence, when a part triggers replenishment, there is no delay in the release of the order to the job shop. We also assume that when the finished parts inventory stocks out, demand is backordered. For the model, we assume that we are given service level targets for both the finished parts inventory and for the raw material inventory. These assumptions are consistent with the factory practice where both the raw material and finished parts inventories are managed with quite high service levels, and shortages will trigger managerial attention so as to avert any delays in the assembly of the final products (pumps).

Production Planning

A4. We assume that the replenishment of each part is done in a lot (equal to the reorder quantity), that lots are not split or combined within the job shop, and that the transfers from one work station to another are done in whole lots. We assume that the lot sizes for each product part are to be chosen from a given discrete set of candidate lot sizes. And each part has assigned to it a single lot size.

A5. We model the lot arrival process for each part at each work station as a Poisson process, independent of the arrival processes of the other parts. This implies that the processes are memoryless (thus, the number of lots that arrive

in one day is independent of how many arrived in the prior day). This assumption is justified by the large number of product parts in the system; thus the superposition of the arrival processes for the individual parts is approximately Poisson and independent. We tested this assumption with a simulation. The results show that the assumption seems OK, provided that the arrival rate for a part is less than 3 lots per day. We found that Poisson processes tend to overestimate the variability of lot arrivals if the arrival rates are large (e.g. more than 4 lots per day). Moreover, it will be quite rare for a part to generate multiple replenishment orders in a given day, resulting in 3 or more lots per day; if this were to happen, the shop would insist on increasing the lot size so as to reduce the frequency of lot arrivals and eliminate unnecessary setups.

A6. We assume the processing time for each unit at each work station is deterministic and no uncertainty is involved. We assume all units within a lot are processed in sequence; hence, the processing time of a lot is the sum of the processing times of each individual job or unit in the lot plus the setup time for the lot. We also assume the setup time is station dependent, but does not depend on the setup of the prior lot, i.e., setup times are independent of sequence.

We assume that we can adjust the production rate at each work station, at some frequency (e.g., twice per day), based on the quantity of work in queue at the work station. In other words, when the amount of work in the queue grows, the manager can expand the production by scheduling overtime, shifting workers, or outsourcing production.

4.2 Model Formulation

We first list the notation we use in the model. Note that index i is always associated with product parts and index j is always reserved for work stations. The definition of a "job" may raise some confusion. When we say job i, we mean a production lot for part i. We note that q_i and τ_j are the decision variables in our model.

μ_i – demand mean of product part (units/day)

σ_i – demand standard deviation of product part i (units/day)

P_{ij} – processing time of product part i on work station j (days per unit)

s_j – setup time for work station j (days per setup)

q_i – lot size of product part i (units)

τ_j – planned lead time on work station j (days)

T_i – total planned lead time of product part i in the system (days)

λ_i – lot arrival rate (lots/day)

w_{ij} – processing time of one lot of product part i on work station j (days)

$N(j)$ – index set of product parts that need to be processed at work station j

$M(i)$‡ – index set of work stations that are on the process route of product part i

h_i^R – holding cost for raw material of product part i (dollars per unit per day)

h_i^G – holding cost for finished goods of product part i (dollars per unit per day)

g_j – overime cost at work station j (dollars per worker per hour)

C_j – nominal production capacity of work station j (days/day)

L^r – review period of the raw materials inventory (days)

L^d – delivery lead time of the raw materials (days)

---

‡ A work station can be counted multiple times if it appears multiple times on the process route.

z^R – safety factor of the raw materials inventory

z^G – safety factor of the finished goods inventory

N – number of product parts in the system

M – number of work stations in the sytem

Given lot sizes (q_i), processing time (P_{ij}) and setup time (s_j), we can find w_{ij}, the processing time of each lot at each work station by

$$w_{ij} = P_{ij}q_i + s_j \quad (4.1)$$

We only define w_{ij} for pairs (i,j) such that part i requires processing at work station j. The above formula assumes that for each lot arrival, we need to perform one setup on the corresponding work station before processing the job.

We define the planned lead time for one lot of product part i at work station j (T_{ij}) as the sum of planned lead time of work station j (τ_j) and the processing time w_{ij}. We note that this definition differs from how the planned lead time is defined in the earlier literature where the planned lead time for a lot of product part i at work station j is just τ_j. In our context the processing times vary quite a lot for different product parts on the same work station, therefore we specify the planned lead time for a part to depend on both the work station and the part processing time.

We can then calculate the total planned lead time for product part i by adding up T_{ij} for all work stations j that part i visits:

$$T_i = \sum_{j \in M(i)} T_{ij} = \sum_{j \in M(i)} (\tau_j + w_{ij}) \quad (4.2)$$

The lot arrival rate for part i is given by $\lambda_i = \mu_i/q_i$. According to the Poisson assumption of the lot arrival, the variance of the daily lot arrival of product part i is also λ_i. The number of lots of part i that arrive to work station j is a Poisson random variable with mean λ_i; each of these lots brings a workload of w_{ij}. Hence,

the expected workload arrival for part i to work station j has expectation given by $\lambda_i w_{ij}$ and has a variance given by $\lambda_i w_{ij}^2$. By summing over all parts that are processed at work station j, we can then characterize the first two moments of the workload arrival A_j for each work station j:

$$E[A_j] = \sum_{i \in N(j)} \lambda_i w_{ij} = \sum_{i \in N(j)} \left(\mu_i P_{ij} + \frac{s_j \mu_i}{q_i} \right) \quad (4.3)$$

$$Var(A_j) = \sum_{i \in N(j)} \lambda_i w_{ij}^2 = \sum_{i \in N(j)} \left(\mu_i P_{ij}^2 q_i + \frac{\mu_i s_j^2}{q_i} + 2\mu_i P_{ij} s_j \right) \quad (4.4)$$

(4.3) describes that the workload arrival rate on work station j is the aggregate workload generated per unit time by all product parts that need to be processed at work station j. (4.4) describes the similar fact for the workload variance. From (4.3) and (4.4) we see that $E[A_j]$ is strictly deceasing and convex in each q_i if the setup time on work station j is non-zero and $Var(A_j)$ is a convex function in terms of each q_i. In particular, increasing lot size q_i will reduce workload arrival variability associated with setup time but increase the variability associated with the lot size effect.

The expected workload is what we expect to arrive to the work station j per unit of time (days), which is equivalent to the utilization of the work station j. The workload arrival variance in (4.4) denotes the variability of the incoming workload and will be seen to be directly related to the calculation of overtime. Notice that for the above calculation, we assume that the arrival process for each part is independent of that for all other parts.

We are now ready to proceed to describe the formula used to calculate the cost components, namely, *Raw Materials Inventory Cost, Finished Goods Inventory Cost, Work-in-process Cost* and *Production Overtime Cost* as functions of lot sizes and planned lead times.

Raw Materials Inventory Cost

We use a periodic review policy for raw materials inventory. That is, we review the raw materials inventory according to a fixed review period (L^r) and order up to a baseline at each review period. We will then wait a constant delivery lead time (L^d) for the order to arrive and the inventory to be replenished. We assume both the review period and the delivery lead time are given as managerial inputs.

In addition, we assume the release of raw materials from inventories to the corresponding work stations can be done in real time. Thus, the lot release process of product part i to the first work station can be seen equivalently as the demand process for the raw material of part i. Since the lot release process for part i is assumed to be Poisson with rate λ_i, the mean demand rate (in units) is thus $\lambda_i q_i$ and the demand variance is $\lambda_i {q_i}^2$. We model the daily raw materials inventory cost for part i using the following standard formula, consisting of approximations for the cycle stock and safety stock cost:

$$Cost_i^{RM} = h_i^R \left(\frac{\lambda_i q_i L^r}{2} + z^R \sqrt{\lambda_i {q_i}^2} \sqrt{L^d + L^r} \right)$$

$$= h_i^R \left(\frac{\mu_i L^r}{2} + z^R \sqrt{\mu_i q_i} \sqrt{L^d + L^r} \right) \quad (4.5)$$

where μ_i represents the daily demand mean of part i. (4.5) directly shows that the raw material cost of product part i is strictly increasing and concave in its lot size q_i.

Finished Goods Inventory Cost

We assume a QR continuous review policy for the Finished Goods Inventory. That is we will order a fixed quantity, in our case the lot size, every time the inventory position falls below a reorder point. The lead time of replenishment is simply the planned lead time for the part, T_i. Instead of the periodic review policy, we believe the continuous review policy is a more reasonable approximation to the factory

practice since the value of good in the FGI is higher. Following the standard formula, we can model the daily cycle stock and safety stock of the FGI of product part i by the approximation:

$$Cost_i^G = h_i^G \left(\frac{q_i}{2} + z^G \sigma_i \sqrt{T_i}\right) \quad (4.6)$$

where σ_i indicates the daily demand standard deviation. Substituting (4.1) and (4.2) into (4.6), we have

$$Cost_i^{FGI} = h_i^G \left(\frac{q_i}{2} + z^G \sigma_i \sqrt{\sum_{j \in M(i)} (\tau_j + P_{ij} q_i + s_j)}\right) \quad (4.7)$$

(4.7) shows that the FGI cost of product part i is strictly increasing with its lot size and the planned lead times of the work stations on its process route.

Work-in-process (WIP) Cost

By Little's Law, the WIP for product part i is directly proportional to the total planned lead time. We value the holding cost of WIP approximately at the average of the raw material and finished goods costs; thus we use the following approximation:

$$Cost_i^{WIP} = \frac{h_i^R + h_i^G}{2} T_i \mu_i \quad (4.8)$$

Substituting (4.1) and (4.2) into (4.8), we have

$$Cost_i^{WIP} = \frac{h_i^R + h_i^G}{2} \sum_{j \in M(i)} (\tau_j + P_{ij} q_i + s_j) \mu_i \quad (4.9)$$

We observe from (4.9) that the WIP cost of product part i is also increasing with its lot size and the planned lead times of the work stations on its process route.

Production Overtime Cost

If we model each work station using the continuous version of the Tactical Model, then we can characterize the production at each work station as a random variable with moments;

$$E[P_j] = E[A_j] \quad (4.10)$$

$$Var(P_j) = \left(\frac{\beta_j}{2-\beta_j}(1-\gamma_j)^2 + {\gamma_j}^2\right) Var(A_j) \quad (4.11)$$

where

$\beta_j = 1 - (1 - \alpha_j/m)^m$;

$\gamma_j = 1 - \frac{(1-\alpha_j/m)}{\alpha_j}\beta_j$;

$\alpha_j = 1/\tau_j$

$m - production\ adjustment\ frequency$

From previous discussion on $E[A_j]$, we know that $E[P_j]$ is a strictly decreasing and convex function in each q_i. This result matches our intuition since larger lot sizes imply fewer setups on the work stations, which further reduce the workload on the work station. We also see that $Var(P_j)$ is convex in each q_i. This is consistent with our previous discussion that increasing q_i can have a two-sided effect on the production variability.

We note that the planed lead time τ_j does not appear in the calculation of the expected workload $E[A_j]$. From (4.11), we can prove that $Var(P_j)$ is strictly decreasing with τ_j (see Appendix 2). This again is consistent with our intuition since longer planned lead times allow for more production smoothing.

We further assume that the production at each work station is normally distributed; we can then approximate the daily production overtime cost using the normal loss function:

$$Cost_j^{OT} = g_j \int_{C_j}^{\infty} (x - C_j) f_P(x) dx \quad (4.12)$$

where we assume $f_P(x)$ is a normal probability density function with mean $E[P_j]$ and variance $Var(P_j)$; C_j is the normal production capacity of the work station j. We can then express the production overtime cost by solving the integration (derivations in Appendix 1):

$$Cost_j^{OT} = g_j \int_{C_j}^{\infty} (x - C_j) \frac{1}{\sqrt{2\pi Var(P_j)}} e^{-\frac{(x-E[P_j])^2}{2Var(P_j)}} dx$$

$$= g_j \left(\sqrt{\frac{Var(P_j)}{2\pi}} e^{-\frac{\rho_j^2}{2}} + (E[P_j] - C_j)\Phi(-\rho_j) \right) \quad (4.13)$$

where $\rho_j = \frac{C_j - E[P_j]}{\sqrt{Var(P_j)}}$ and $\Phi(\cdot)$ is the CDF of standard normal distribution N(0,1).

We can prove that the production overtime cost expressed in (4.13) is a convex and increasing function in both $E[P_j]$ and $Var(P_j)$ (see proofs in Appendix 1) by taking the first and second derivative of (4.13). This is not surprising since we expect that overtime increases more rapidly when the production capacity is tighter or the production variance is larger.

The Optimization Problem

We have expressed each cost component as a function of the lot size for each product

part (q_i) and the planned lead time for each work station (τ_j). We pose the optimization problem as follows:

$$\min_{q_i,\tau_j} \sum_{i=1}^{N} \left(Cost_i^{RM} + Cost_i^{FGI} + Cost_i^{WIP}\right) + \sum_{j=1}^{M} Cost_j^{OT} \qquad (4.14)$$

$$s.t. \quad \underline{q_i} \le q_i \le \overline{q_i} \quad \forall\, i$$

$$\frac{1}{m} \le \tau_j \le \overline{\tau_j} \;\forall\, j$$

The objective function is simply to minimize the total relevant manufacturing costs. The two constraints set the lower and upper bounds for the decision variables subject to management concerns and physical limitations in the factory. In particular, m is a managerial parameter specifying the maximum production adjustment rate on the work stations. For example, if the factory works with a morning shift and an evening shift for every work station, the production adjustment rate can be naturally set to two for all work stations. As one can imagine, m indicates the ability to impose inter-period production control. A higher value of m indicates that the factory can adjust production rate according to job arrivals more frequently per time period.

The optimization problem, as stated here, allows for fractional lot sizes, which is unrealistic. Nevertheless, we can view the formulation as the linear relaxation problem of the "real" problem, and then use the optimal solution for this problem as a starting point to find integral solutions. As will be shown in the next Chapter, our numerical tests confirm that the rounding algorithm often works very well, partly due to the fact that the objective function is relatively smooth around the optimal solution point.

4.3 Analysis of the Cost Components

First, it is easy to see from (4.5), (4.7) and (4.9) that the inventory-related costs $Cost^{RM}, Cost^{WIP}, and\ Cost^{FGI}$ strictly increase in each q_i and each τ_j. A more careful look shows that $Cost^{RM}\ and\ Cost^{FGI}$ increase concavely in each q_i and each τ_j while $Cost^{WIP}$ increases linearly in each q_i and each τ_j.

We now discuss the impact from the lot size q_i and the planned lead time τ_j on the production overtime $Cost^{OT}$. For each lot size q_i, we show $Cost_j^{OT}$ is convex in both $E[P_j]$ and $Var(P_j)$ (see Appendix 1). Since $E[P_j]$ and $Var(P_j)$ are convex in each q_i (by observing (4.10) and (4.11)), by composition of convexity, we know $Cost_j^{OT}$ is also convex in each q_i. For each planned lead time τ_j, since $Cost_j^{OT}$ is increasing in $Var(P_j)$ (see Appendix 1) and $Var(P_j)$ is decreasing in each τ_j (see Appendix 2), $Cost_j^{OT}$ is also decreasing in each τ_j. Since the total overtime costs are the sum of the overtime cost of each individual work station, we conclude that the total overtime cost is convex in each lot size q_i and is strictly decreasing in each planned lead time τ_j.

Now, from the convexity of the total overtime cost function with respect to q_i, we can see that each lot size q_i has a two-sided impact. On one hand, increasing q_i leads to lower expected workloads on the set of work stations $M(i)$ that will further reduce the overtime costs on them. On the other hand, increasing q_i will eventually raise the production variability and increase the overtime cost. Notice that the only benefit from increasing lot sizes is to reduce the number of setups. Hence, we expect that larger lot sizes are applicable only for the product parts that heavily use the bottleneck work stations.

We finally observe that the planned lead time τ_j captures the trade-off between inventory-related costs and the production overtime cost. We can reduce overtime

on work station j by increasing τ_j; but this requires holding more safety stock and WIP. We can reduce the inventory level by decreasing the planned lead times for some work stations.

The relationships between the cost components and the decision variables are summarized in Table 1. The positive (negative) sign in the table indicates the cost component is positively (negatively) correlated with the decision variable.

	τ_j	q_i
$Cost^{RM}$	Not related	(+)
$Cost^{FGI}$	Not related	(+)
$Cost^{WIP}$	(+)	(+)
$Cost^{OT}$	(-)	(+) and (-)

Table 1: The Relationship between Cost Components and Decision Variables

5 Implementation and Numerical Tests

In this chapter, we discuss in section 5.1 how to implement our model in the Excel Spreadsheet and solve it with the premium Excel Build-in Solver. We then report the numerical tests of the model with three different test cases in sections 5.2, 5.3, and 5.4. We first demonstrate our model with a single-product single work station example. We then use a medium case example to illustrate the trade-offs in a job shop system. We finally report on a more extensive numerical example using real factory data. We perform sensitivity analysis to discover how the changes in parameters and input data affect the optimal cost and the operating tactics. We use a day as the time unit in the rest of the discussion and the numerical tests. The primary purpose of this is to be consistent with the actual planning and scheduling in the target factory.

5.1 Model Implementation

We implemented the mathematical model in the Microsoft Excel Spreadsheet and solve it using the premium Excel built-in Solver. We solve all examples with the large-scale GRG Nonlinear solver engine (developed by Frontline Systems).

Input Data Requirement

In order to apply the model, one needs to collect of the following input information -- the data regarding the product parts and the work stations, and the managerial parameters. We list the input data requirements in Table 2.

For the input data regarding product parts and work stations, we expect to get most of them from the factory database. For example, the demand mean and variance for each product part can be generated from forecasts on future demand (See Appendix 3 for more details); raw material and finished part cost for each product part are estimated from historical data. Other data like raw material delivery lead times, processing times, the setup time per lot on each work station, and the normal capacity of the work station are often readily available in the factory. We note that by normal capacity we mean the planned working schedule for the factory, by which the operating costs are incurred regardless of whether we work overtime or not.

Product Parts and Work Stations Data	***Managerial Parameters***
Part Demand Mean	*Number of Days in a Month*
Part Demand Std	*Inventory Holding Cost*
Raw Material Cost	*Safety Stock Factor (Service Level)*
Finished Part Cost	*Raw Materials Inventory Review Period*
Raw Material Lead Time	*Production Adjustment Frequency*
Processing Time	*Maximum Number of Lot Arrivals Per Day*
Setup Time per Lot on Work Station	*Overtime Cost per Hour*
Normal Capacity of Work Station	

Table 2: Input Data for the Model Implementation

Most of the managerial parameters are readily available from the factory manager. We would like to highlight two parameters that are particular to our model. We discussed the *Production Adjustment Frequency* in section 4.2 and observed that the inverse of it serves as a lower bound for the planned lead times. The manager should set this parameter according to how flexible the factory is in terms of adjusting the production rate on its work stations. Another parameter is the *Maximum Number of Lot Arrivals Per Day*. We assume that the lot arrival process to each work station can be modeled as a Poisson process, which allows us to simplify the mathematical analysis. Our simulation results show that as long as we have less than three lot arrivals in a day, the lot arrival distribution looks similar to Poisson process. Thus we recommend the manager to set this parameter to be no larger than 3 to maintain the accuracy of the solution. We expect this is done without any loss of optimality as

it is likely impractical to have lot arrivals of more than 3 per day for any part.

Lightly Loaded and Outsourced Work Stations

We exclude from the optimization lightly loaded and outsourced work stations. For the outsourced work stations, the associated planned lead time is given, and is not a decision variable. For lightly-loaded work stations, we pre-set the planned lead times for these stations to be at their minimum bound, as there is no need for smoothing given that there is ample capacity at these stations. We determine a work station to be a lightly loaded work station if the following condition is satisfied:

$$E[A_j] + v \times \sqrt{Var(A_j)} < C_j$$

where v is a preset constant subject to management decision. If this condition does not hold, we term the work station to be "not-lightly" loaded. One can view the parameter v as a threshold value being analogous to setting the probability that the arriving daily workload is less than the nominal daily capacity at a lightly loaded work station. For instance if $v = 3$, then a lightly loaded machine is any work station for which its daily capacity will exceed the daily workload arrivals with probability of at least .997 (under assumption that workload arrivals is normally distributed). If the probability that arrivals exceed capacity is .003 or more, then we would categorize the workstation as not-lightly loaded for $v = 3$.

Find Optimal Solution

We set lower and upper bounds for the decision variables τ_j and q_i. The lower bound for the lot sizes accounts for both the smallest possible lot size option and the maximum number of lot arrivals allowed per day:

$$LB\ for\ lot\ size\ of\ part\ i = max\{lowest\ lot\ size\ option, \mu_i/\lambda_{max}\}$$

Where λ_{max} is the managerial parameter *Maximum Number of Lot Arrival Per Day.*

We set the lower bounds for the planned lead times to the inverse of the production rate adjustment frequency, as described in section 4.2. The upper bounds for the lot sizes and planned lead times are both managerial inputs which depend on the physical constraints and the policies adopted in the factory.

In executing the Solver we do not impose integer constraints on the decision variables for computational efficiency reasons. Hence, the lot sizes directly found by the Excel Solver are continuous numbers, i.e., they are not necessarily integers or lot sizes that can be processed on the work stations. We then conduct a simple search based on the continuous solution to find the best nearest integer (or lot size) solution around the continuous solution. For instance, if the Solver determines a lot size to be 7.63, we then re-compute the costs assuming a lot size of 7 and a lot size of 8; the lot size that yields the lower cost is the best unrestricted integer lot size solution. We next do a similar comparison, but for a given set of the restricted lot sizes; for instance, if lot sizes are restricted to be 4, 8, 12,... then we would compare a lot size of 4 with a lot size of 8, and in this manner find the best restricted integer lot size solution. We then fix the lot sizes and re-optimize to find the best planned lead times for work stations for both cases, i.e., for the unrestricted and restricted integer lot size solutions.

To sum up, we provide 3 solutions in our software: continuous solution, best nearest integer solution (unrestricted) and best nearest lot size solution (restricted). The best nearest lot size (restricted) solution can be directly applied to the job shop since it is restricted to lot size values previously chosen by the manager.

For reasonable size problems we can reliably solve the optimization problem with the premium Excel built-in Solver. For instance, we show in section 5.4 that we can optimize the operating tactics in the target factory that processes 133 parts on 59 work stations, among which 23 work stations are not lightly loaded.

Due to the non-convexity in the problem, the performance of the Excel Solver

sometimes depends on the starting point, i.e., the initial values for the decision variables. These initial values represent the starting point for the search procedure that underlies the Solver's algorithm. That being said, we tested the algorithm over a set of test problems, and we observed that we obtained the same solution from many different starting points. As we understand from the analysis in section 4.3, although convergence to a global optimum cannot be guaranteed, the objective cost function appears to be reasonably smooth.

When trying to implement the model, we usually will be able to identify a few good starting points, given the cost parameters and the knowledge about how each component of the objective cost function behaves according to the decision variables. We confirmed that the search is very reliable as long as we start the search from one of these reasonable starting points. By performing some extensive tests, we found that a good starting point for our problem is to set the decision variables to their lower bounds. Our computational experience has been that with this starting point, the algorithm always converges to a stable set of values for the lot sizes and planned lead times.

5.2 Numerical Tests - A Simple Example

We will provide a single product part, single work station example in this section to illustrate our model. The simple setting allows us to show the relationship between costs and decision variables on 2-D graphs.

In this simple example, we assume the work station has a nominal working capacity of 10 hours per day and it takes 1 hour for a part to be processed by the work station. We assume the demand distribution is normal with mean equal to 8 units of product parts per day and standard deviation equal to 4 units per day. The setup time for one lot of parts is 15 minutes (0.25 hour).

For this case, it is not hard to see that some overtime has to be incurred in order to keep the balance of the workload inflow and outflow. For example, if the lot size is 1, the average total workload per day will be 8 + 0.25*8 = 10 hours; therefore, if there is variability in the arriving workload, there is a good chance that the work station needs to work overtime if it is to clear the workload in a day.

On the cost side, we assume the overtime cost is $1.5 per hour, the raw materials inventory holding cost is $0.1 per day, and the finished part inventory holding cost is $0.2 per day. We also assume the raw material delivery lead time is 3 days and raw materials inventory review period is 1 day. The safety stock factor is chosen to be 2 (97.7% service level).

The first result that we show is the inventory and overtime costs as functions of the lot size for a fixed planned lead time of one day. As can be seen from Figure 3, the inventory cost (including raw material, finished parts, and work-in-process) increases in the lot size while the overtime cost is convex in the lot size and attains its minimum when lot size equals 2. Under the one day planned lead time, we conclude that choosing a lot size of two is optimal, although it is only slightly better than a lot size of one.

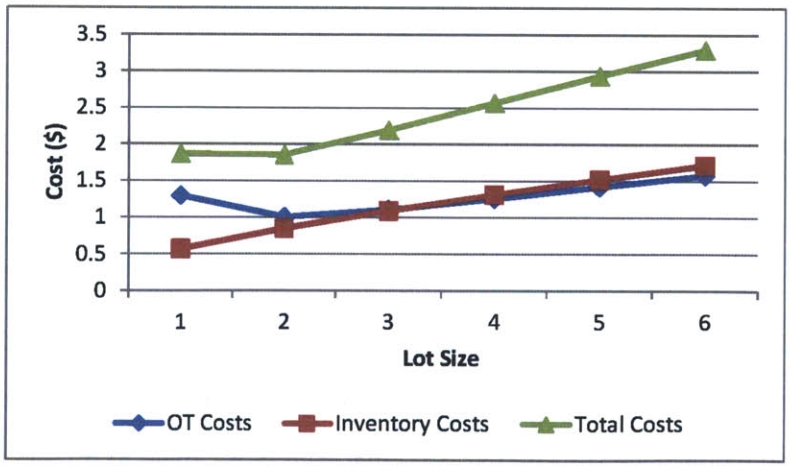

Figure 4: Inventory and Overtime Costs for Fixed Planned Lead Time (1 day)

We then fix the lot size and change the planned lead time. Figure 4 shows how the

inventory and overtime costs change along with the planned lead time when the lot size equals two. We observe that the inventory cost goes up gradually in the planned lead time. We also observe that the overtime cost drops quickly as the planned lead time increases. However, as the planned lead time gets larger, the marginal benefit from overtime cost reduction gets smaller. As a result, when we increase the planned lead time, the change of the total cost is dominated by the change of overtime cost up to a point, and then moves in accordance with the inventory cost. In this simple example, the total cost obtains its minimum when planned lead time equals 3 days and the cost function is flat around its minimum.

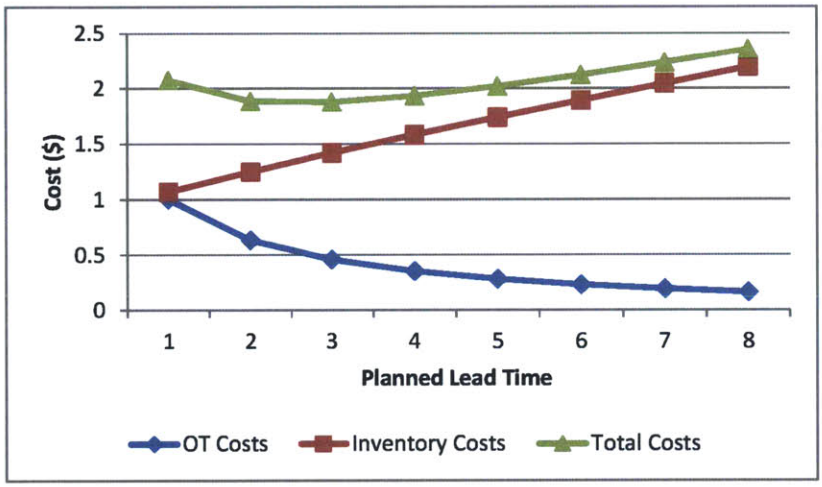

Figure 5: Inventory and Overtime Costs for Fixed Lot Size (2 units per lot)

With this simple example, we are able to show how the manufacturing costs behave as functions of the decision variables. The trade-off between the inventory cost and the production overtime cost is clear. For this simple case, the optimal solution $(q = 2;\ \tau = 3)$ matches the solution we found with the greedy scheme we described. However, the greedy scheme may not always work for multiple work station cases. The optimal non-integer solution for this simple case is $q = 1.47$ and $\tau = 2.44$. In this case the optimal solution from the linear relaxation gives a good start point for the optimal integer solution.

5.3 Numerical Tests - An Illustrative Case

We will illustrate our model in a more comprehensive way with a small scale job shop example. The job shop processes 4 product models that require 8 product parts that are processed on 5 work stations; each product part has its own processing route.

In this job shop, the need to replenish the parts inventory results in an order being placed on the job shop. Since each order is associated with a lot size, the raw material of the order will be pulled from the raw materials inventory to form a lot of the ordered parts. The lot will then enter the shop and follows its processing route until it reaches the finished parts inventory. For example, a lot of product part 8 (the last route shown in Figure 5) will enter the shop and visit work station 4 first. It will then proceed to work station 2 after finishing at work station 4. The lot will finally enter work station 5 and upon completion, end up at final parts inventory. Each process step adds value to the product part and we assume the value of a unit of the final part inventory is that for the unit of the raw material.

In this example, product models 1 and 2 have higher demand volume, higher demand variability, less production cost, and shorter raw material delivery lead times. Product models 3 and 4 have higher cost parts and have relatively lower demand mean and variance, higher production cost and longer delivery lead time. We list the parameters for the four models in Table 3.

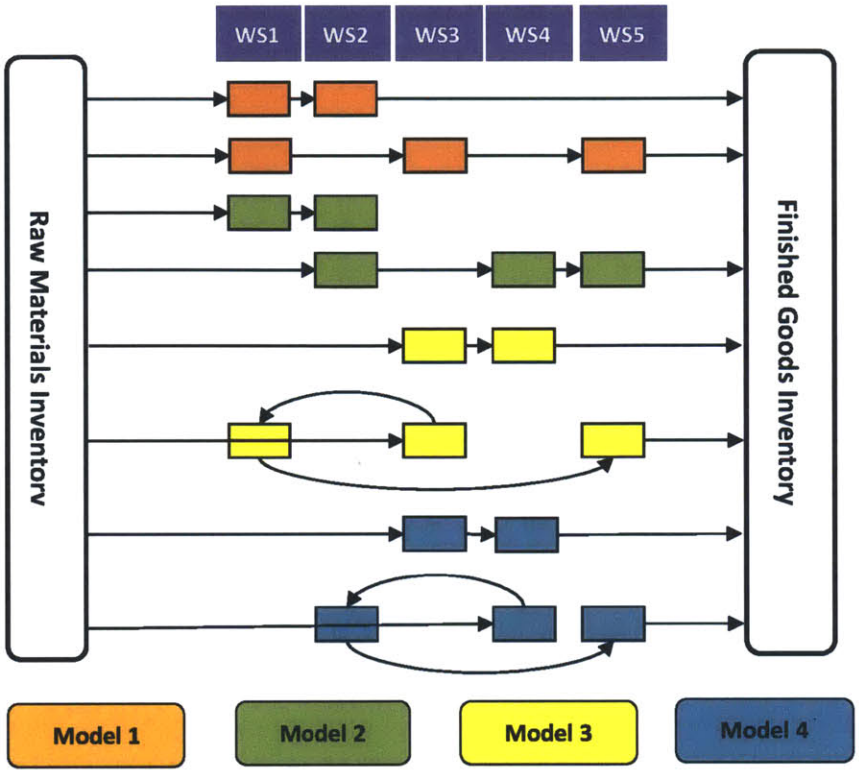

Figure 6: Overview of a Small Scale Job Shop System

	Model 1	Model 2	Model 3	Model 4
Demand mean	250	200	150	100
Demand std/demand mean	0.5	0.5	0.2	0.2
Raw material cost	$500	$500	$2000	$2000
Finished part cost	$1000	$1000	$4000	$4000
Raw material delivery lead time	1 month	1 month	2 months	2 months

Table 3: Input Data for the Illustrative Example

For simplicity, we assume all work stations have a capacity of 8 hours a day and the setup time for any lot of parts is 30 minutes on each work station. We also assume the processing time is 5 minutes per unit for each part on each work station. Other parameters we use for this example are listed in the following:

- Overtime cost for all work stations: $1000/hour
- Holding cost for all parts: 15%/unit per year

- Number of days in a month: 20 days
- Safety stock factor: 2.6
- Raw materials inventory review period: 20 days
- Upper bound for number of lot arrivals per day for each part: 3 units/lot
- Production rate adjustment frequency: 4 times a day

The Base Case

We first consider a base case where we fix the lot sizes for all parts as 5 units per lot and the planned lead times for all work stations as 0.25 day. Under this setting, we can calculate the mean and variance of the daily workload arrival, daily production variance, and expected overtime on each work station (Table 4), as well as the cost breakdown (Table 5) associated with the base case.

As shown in Table 4, when we set the planned lead times to the lower bounds (0.25 day, the inverse of the production frequency), we essentially do not have any production smoothing at the work stations. Also note that the expected workload arrivals (or equivalently the utilization rate) on the work stations are high, especially on work station 1. Work stations with large and variable workloads need to incur substantial production overtime costs, given their planned lead times of 0.25 days. In the next two alternative cases we examine how to set the tactical decisions to reduce the expected overtime by reducing the production variance.

Base Case	**WS1**	**WS2**	**WS3**	**WS4**	**WS5**
$E[A_j]$ (day/day)	0.97	0.86	0.74	0.63	0.80
$STD(A_j)$ (day/day)	0.33	0.31	0.29	0.27	0.30
$STD(P_j)$ (day/day)	0.33	0.31	0.29	0.27	0.30
Expected OT (hours/day)	**0.965**	**0.538**	**0.246**	**0.083**	**0.375**

Table 4: Work Station Statistics for the Base Case in the Illustrative Example

Daily Cost Breakdown	***Base Case***
Raw Materials Inventory Cost	***1167***
Finished Parts Inventory Cost	***356***
Work-in-process Cost	***62***
Overtime Cost	***2208***
Total Cost	***$ 3,793***

Table 5: Cost Breakdown for the Base Case in the Illustrative Example

Case 1: Increase Planned Lead Time for WS1

In Case 1, we increase the planned lead time of work station 1 from 0.25 day to 1 day while keeping the lot sizes the same as in the base case. We expect to see less overtime on work station 1. The results are shown in Table 6 and Table 7.

As we can see, by increasing the planned lead time of work station 1 to 1 day, we obtain 18.7% reduction in the overtime cost and 8.8% reduction in total cost. At the same time we observe that the safety stock and WIP increase slightly due to longer planned lead time at work station 1.

	WS1	
	Base Case	**Case 1**
$E[A_j]$ (day/day)	0.97	0.97
$STD(A_j)$ (day/day)	0.33	0.33
$STD(P_j)$ (day/day)	0.33	0.20
Expected OT (hours/day)	**0.965**	**0.553**

Table 6: Work Station Statistics for the Case 1 in the Illustrative Example

Daily Cost Breakdown	**Case 1**
Raw Materials Inventory Cost	***1167***
Finished Parts Inventory Cost	***413***
Work-in-process Cost	***85***
Overtime Cost	***1795***
Total Cost	***$ 3,461***

Table 7: Cost Breakdown for Case 1 in the Illustrative Example

Case 2: Increase Lot Sizes for Model 1 and 2

In Case 2, we increase the lot sizes of parts 1 to 4 from 5 units to 10 units while keeping the planned lead times the same as in the base case. The intention is to have fewer setups on work stations that process these high volume parts, and thus reduce the utilizations at these work stations. The results are shown in Table 8 and Table 9.

Case 2	WS1	WS2	WS3	WS4	WS5
$E[A_j]$ (day/day)	0.76	0.66	0.67	0.57	0.66
$STD(A_j)$ (day/day)	0.34	0.32	0.30	0.27	0.31
$STD(P_j)$ (day/day)	0.34	0.32	0.30	0.27	0.31
Expected OT (hours/day)	**0.380**	**0.188**	**0.153**	**0.051**	**0.171**

Table 8: Work Station Statistics for Case 2 in the Illustrative Example

As we see, by increasing the lot sizes of part 1 to 4 from 5 units to 10 units, we reduce the average workload by a substantial amount which leads to 57.3% reduction on overtime cost and 31.0% reduction on total cost. We also observe that the inventory costs increase but only by a small amount.

Daily Cost Breakdown	Case 2
Raw Materials Inventory Cost	**1231**
Finished Parts Inventory Cost	**379**
Work-in-process Cost	**65**
Overtime Cost	**943**
Total Cost	**$ 2,618**

Table 9: Cost Breakdown for Case 2 in the Illustrative Example

Case 3: Optimal Case

We finally determine the optimal lot sizes (the best nearest integer solution) and planned lead times (continuous solution) for this illustrative example. Please refer to section 5.1 for the rounding algorithm). The comparison results are shown in Table 10.

Part No.	*Base Case Lot Size*	*Optimal Lot Size*
1	5	**12**
2	5	**13**
3	5	**11**
4	5	**11**
5	5	**4**
6	5	**6**
7	5	**4**
8	5	**4**
Machine	*Base Case Planned Lead Time*	*Optimal Planned Lead Time*
1	0.25	**1.10**
2	0.25	**0.88**
3	0.25	**0.74**
4	0.25	**0.56**
5	0.25	**0.77**

Table 10: Optimal Case Solution in the Illustrative Example

We observe that the optimal solution is consistent with our intuition from case 1 and case 2 where we have larger lot sizes for part 1 to 4 and longer planned lead times on all work stations. These results are driven by the requirement to lower the production overtime cost. However, we also observe a slight decrease in the lot sizes for part 5, 7 and 8 from 5 units to 4 units. Although this change may possibly generate more setups on the corresponding work stations, it helps reduce the demand variability on the raw materials inventory. Since product parts 5 to 8 have relatively high values and small demand volume, the gain from reducing the raw material inventory cost dominates the increased cost due to more production overtime.

Under the optimal tactics, we will be able to further reduce the average workload and variability which leads to 91.8% reduction in overtime cost and 44.3% reduction in total cost. Compared with the base case results, the optimal cost structure suggests that we should allow more inventory in the system in order to provide more production smoothing. Since the unit overtime cost is high, there is almost no overtime work in the optimal solution. As a final note, we observe that with the chosen operating tactics, the expected workload and workload variance on the work

stations are close to each other. This indicates that our model sets the tactics to balance the workload to different work stations for production smoothness.

Optimal Case	WS1	WS2	WS3	WS4	WS5
$E[A_j]$ (day/day)	0.70	0.65	0.67	0.62	0.64
$STD(A_j)$ (day/day)	0.35	0.33	0.30	0.28	0.32
$STD(P_j)$ (day/day)	0.21	0.21	0.21	0.21	0.21
Expected OT (hours/day)	**0.055**	**0.032**	**0.040**	**0.023**	**0.031**

Table 11: Work Station Statistics for the Optimal Case in the Illustrative Example

Daily Cost Breakdown	Optimal Case
Raw Materials Inventory Cost	***1221***
Finished Parts Inventory Cost	***552***
Work-in-process Cost	***157***
Overtime Cost	***182***
Total Cost	***$ 2,112***

Table 12: Cost Breakdown for the Optimal Case in the Illustrative Example

5.4 Numerical Tests – Large Scale Application

In this large scale application, we use the data from the target factory from our industry partner. The job shop consists of 133 major product parts and 59 work stations, among which 23 work stations are characterized as not lightly loaded. Demand is high for many parts and the ratio of the standard deviation of daily demand to the mean is around 30%. Around one-third of the work stations are not-lightly loaded. Setup times on a few work stations are long. The per-unit processing times of a part on a work station varies from 20 to 90 minutes. We will not disclose other managerial parameters used in this application due to confidentiality, but we note that most of them are very similar to the parameters we used in the illustrative example in section 5.3. We first introduce a base case scenario and determine its optimal solution from our model. We then present sensitivity analysis using several test cases by varying input parameters. We note that all results are rescaled to disguise

the company data.

In the base case, we set the initial lot sizes and the planned lead times to their lower bounds and we will refer to this as the initial tactics. We believe that this set of initial values is a reasonable representative of the current operating policy applied by our industry partner. We observe and record the performance of the system under the initial tactics. We then solve the optimization model and determine the optimal tactics for the base case problem. The performance improvement from the initial tactics to the optimal tactics is shown in Table 13 and Figure 6. In particular, Table 13 shows the daily cost breakdown of both the initial and optimal tactics, the optimal lot sizes for only the product parts that have unit arrivals larger than 150 pieces per month, and the optimal planned lead times for only the not-lightly loaded work stations. Figure 6 presents the average workload (utilization) and expected overtime on the 23 not-lightly loaded work stations.

In this base case, the raw material inventory cost accounts for a large portion of the total manufacturing cost, mainly because of the long replenishment lead time (e.g. 5 months) of the product parts. The overtime cost for the initial tactics accounts for half of the total costs. However, under the optimal tactics, the production overtime cost is very low; the optimal tactics increase the planned lead times and increase the lot sizes for key product parts. This results in an increased production smoothness in the system, resulting in moderate increases in the inventory-related costs.

Daily Cost Breakdown	Initial Tactics	Optimal Tactics
Raw Materials Inventory Cost ($ per day)	**846.8**	**894.1**
Finished Parts Inventory Cost ($ per day)	**195.8**	**225.9**
Work-in-process Cost ($ per day)	**79.9**	**107.3**
Overtime Cost ($ per day)	**1159.0**	**92.8**
Total Cost ($ per day)	**$ 2281.5**	**$ 1320.0**

Part No.	Initial Tactics for Lot Sizes	Optimal Tactics for Lot Sizes
17	**13**	**14**
19	**13**	**14**
58	**4**	**4**
61	**4**	**11**
68	**4**	**4**
72	**4**	**4**
75	**4**	**9**
112	**6**	**7**
116	**6**	**7**
119	**6**	**13**
153	**5**	**5**
158	**12**	**14**
230	**7**	**7**

Work Station No.	Initial Tactics for Planned Lead Times	Optimal Tactics for Planned Lead Times
1	**0.25**	**0.25**
6	**0.25**	**1.79**
9	**0.25**	**1.55**
11	**0.25**	**0.71**
12	**0.25**	**3.00**
21	**0.25**	**0.81**
22	**0.25**	**0.38**
23	**0.25**	**0.31**
26	**0.25**	**1.51**
27	**0.25**	**0.51**
28	**0.25**	**3.00**
29	**0.25**	**2.74**
30	**0.25**	**2.18**
31	**0.25**	**0.89**
32	**0.25**	**2.28**
34	**0.25**	**0.26**
35	**0.25**	**0.25**
37	**0.25**	**0.35**
40	**0.25**	**1.75**
44	**0.25**	**3.00**
45	**0.25**	**3.00**
46	**0.25**	**1.90**
50	**0.25**	**3.00**

Table 13: Optimal Costs and Some Optimal Solutions of the Base Case

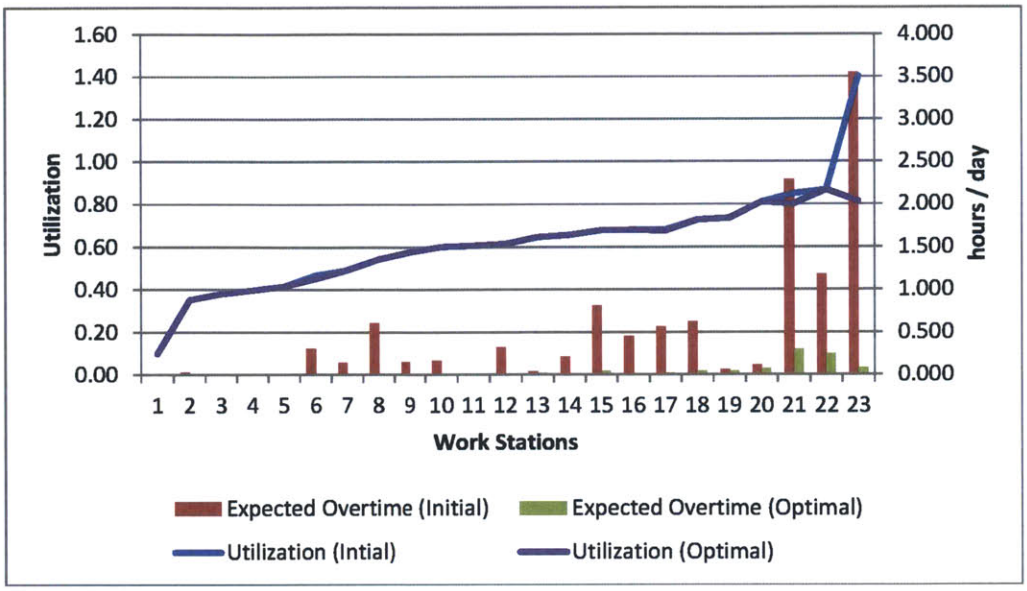

Figure 7: Utilization and Expected Overtime on the 23 Not-lightly Loaded Stations

A closer look at the loads and expected overtime on the work stations shows that we have greatly reduced the utilization and production overtime on several bottleneck work stations which leads to huge cost reduction. In the optimal tactics, the optimal planned lead times on the not-lightly loaded work stations range from 0.25 day to 3 days, which are the lower and upper bounds we impose on these work stations. Among the 23 not-lightly loaded work stations, 13 of them have planned lead times of more than one day. This indicates that the planned lead times are substantially extended on the set of most heavily loaded work stations in exchange for production smoothness. We also observe for most of the parts the optimal lot sizes are close to their lower bounds. This is because very few work stations have large setup time in the system. Exceptions are part 61, 75, and 119 which heavily use the work stations that have large setup times.

We will explain the key findings of the sensitivity analysis with test cases in the rest of this section. We first introduce the seven experiments we consider in the sensitivity analysis and the corresponding parameters changes in Table 14.

Experiments	Descriptions	Base Case	Test Cases
1	Production rate adjustment frequency (times/day)	4	1 2 6 8
2	Overtime cost per hour ($)	100	50 150
3	Threshold for deciding lightly loaded work station	3	2
4	Setup time for work station 6 and 28(min)	60 & 110	120 & 110 60 & 220
5	Capacity for work station 21,27 and 40 (hours/day)	16	10 12 14
6	Demand Standard Deviation	Base case	+50% -50%
7	Demand Mean	Base case	+10% -10%

Table 14: Key Parameters and Input Data Used for Base Case and Testing Cases

Experiment 1: Production Rate Adjustment Frequency

We describe the production rate adjustment frequency (m) in section 4.2 as the ability to impose inter-period production control. A higher value of m indicates that we can adjust production rate according to job arrivals more frequently. In this analysis, we allow m to be 1, 2, 4, 6 or 8.

The results show that increasing m can reduce overall manufacturing cost, but at a diminishing rate (Figure 7). In Figure 7, we show the percentage change of the inventory-related and production overtime costs relative to the base case where $m = 4$. The cost reduction comes from lower production variance and shorter total planned lead times at the work stations. Actually, we observe that the planned lead times reduce continuously on the majority of the heavily loaded work stations when m increases (Figure 8). We note that the optimal lot sizes remain nearly the same. We also omit the raw materials inventory cost in the comparison since it only changes by negligible amount.

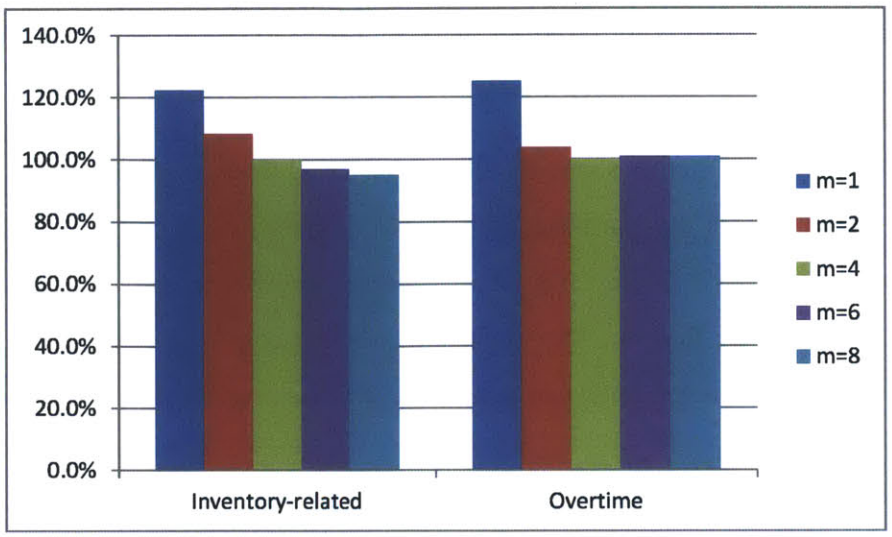

Figure 8: Cost Comparison under Different Production Rate Adjustment Frequency

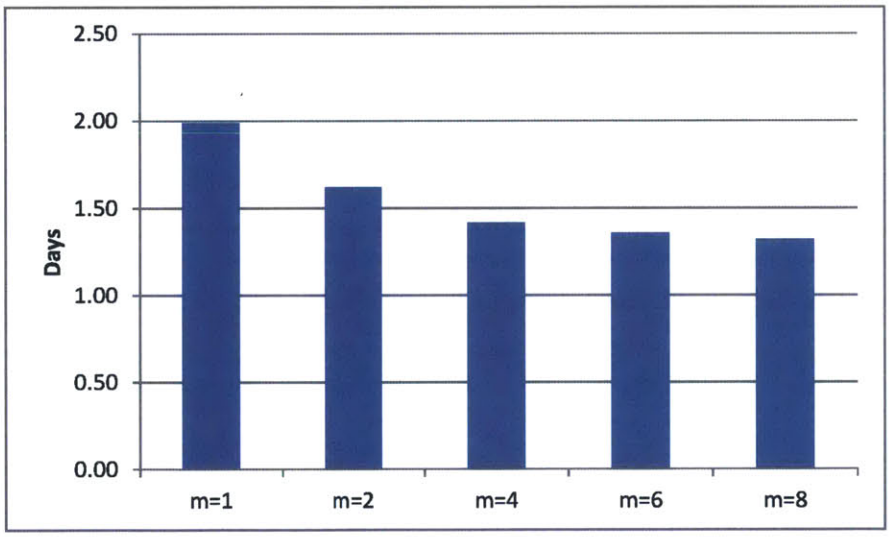

Figure 9: Average Planned Lead Time on the 23 Not-lightly Loaded Work Stations

Experiment 2: Unit Overtime Cost

We increase and decrease the unit overtime cost by 50% in this experiment. We observe when we increase the unit overtime cost by 50 dollars per hour (+50%), we perform 0.18 hours less overtime in total every day and when we decrease the unit overtime cost by 50 dollars per hour (-50%), we perform 0.38 hours more overtime in

total every day. This is not surprising since we expect to take advantage of a lower unit overtime cost by performing more overtime.

At the same time, Figure 9 shows that the planned lead times increase (decrease) by about 1/5 (1/4) of a day on average on 16 work stations when unit overtime cost increases (decreases) The optimal lot sizes remain almost unchanged for all product parts. Considering that we change the unit overtime cost by a large amount, we think the optimal solution is not very sensitive to the unit overtime cost in this case.

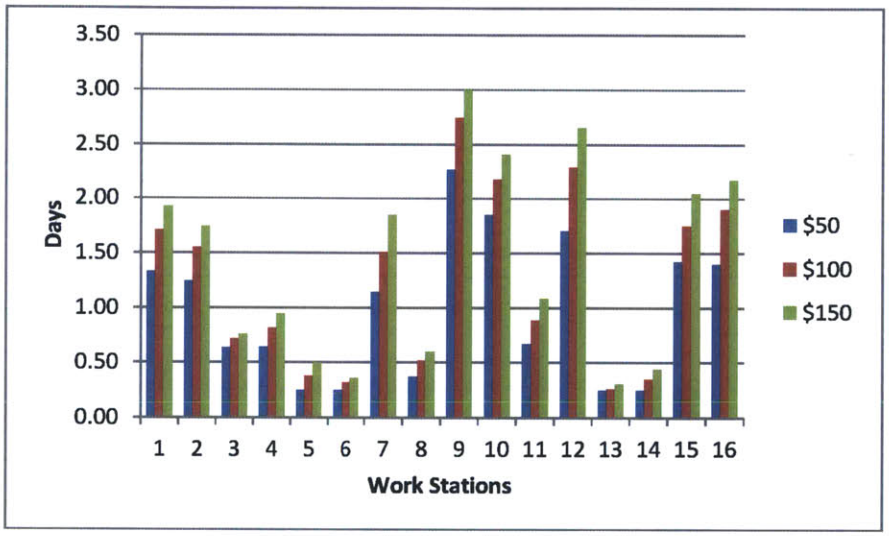

Figure 10: Planned Lead Times under Different Unit Overtime Cost

Experiment 3: Threshold for Deciding Lightly Loaded Work Stations

Recall that the threshold value is used to decide which work stations are lightly loaded work stations (see section 5.1). By reducing this parameter from 3 to 2, we essentially loosen our criteria and more work stations may be characterized as lightly loaded.

The analysis shows that neither the costs nor the solutions are sensitive to this change, but 5 work stations (M1, M11, M23, M34 and M35), originally marked as not-lightly loaded, are characterized as lightly loaded after the change.

Experiment 4: Setup Time for Work Station 6 and 28

Recall that we benefit from large lot sizes when the setup times are substantial on the corresponding resources. In this analysis, we double the setup time of work station 6 and then 28. The setup time on work station 6 (28) in the base case is 60 (11) minutes which is fairly large comparing to the processing times of a lot.

We observe that the optimization chooses larger lot sizes for all parts that visit work station 6 (28) (Figure 10 and Figure 11). In particular, the average lot sizes for parts visiting work station 6 (28) increases by 90% (80%) when setup time is doubled on the work station.

The optimal planned lead times on some work stations are also affected by the change in setup time. When doubling the setup time of work station 6, the optimal planned lead time of work station 6 increases from 1.79 days to 3.00 days. When doubling the setup time of work station 28, the optimal planned lead time of work station 30 increases by 0.7 day. In the second case, since the planned lead time of work station 28 has already reached the upper bound (3 days), the request for longer queues is thus shifted to work station 30 since most of the parts that visit work station 28 also visit work station 30.

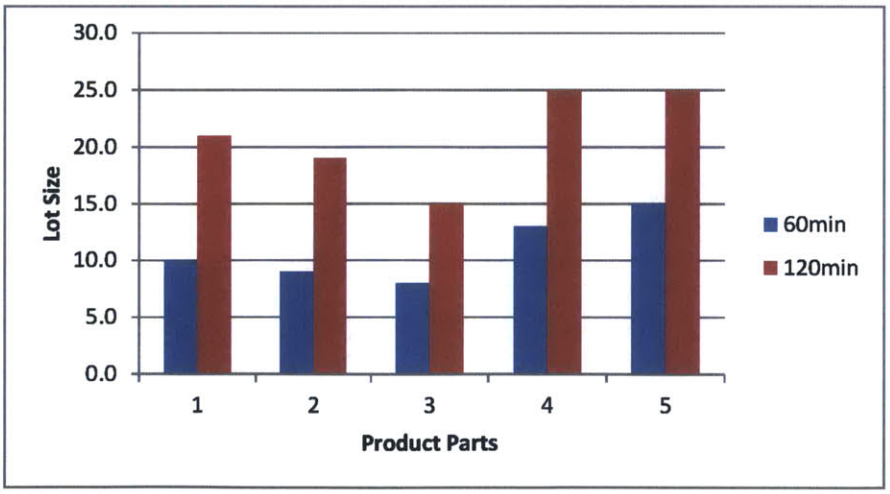

Figure 11: Optimal Lot Sizes of Parts that Visit WS6 under Different Setup Times

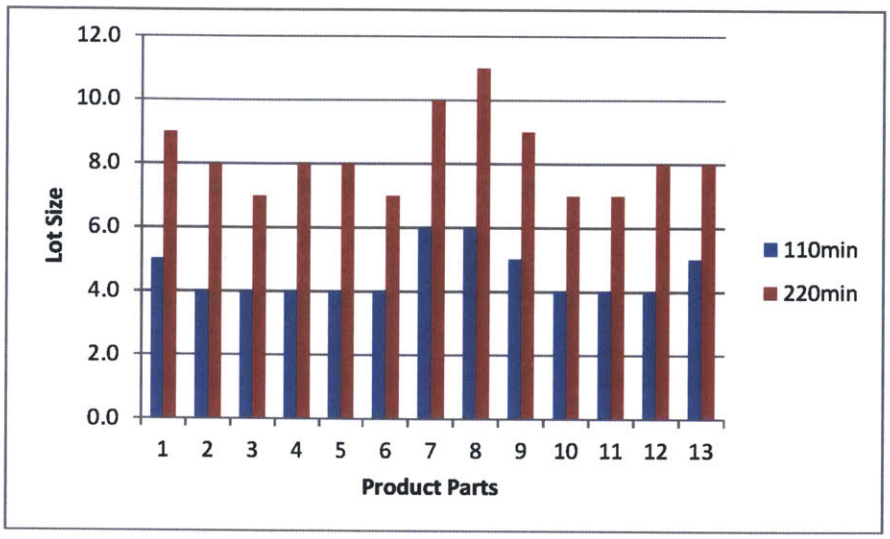

Figure 12: Optimal Lot Sizes of Parts that Visit WS28 under Different Setup Times

Experiment 5: Production Capacity for Work Station 21, 27 and 40

In this analysis, we reduce the normal capacity of work station 21, 27 and 40 from 16 hours per day to 14 hours, 12 hours, and 10 hours. We choose these three work stations since we observe they are the potential bottlenecks of the system.

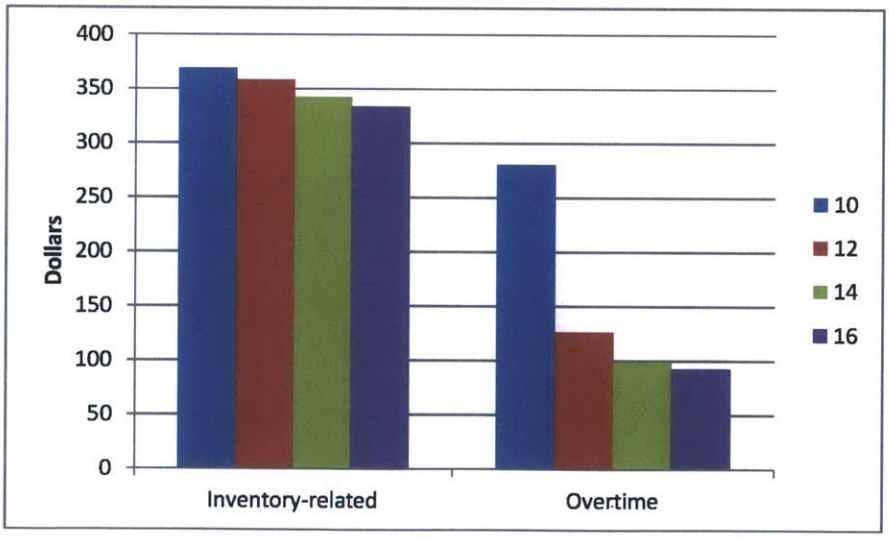

Figure 13: Cost Comparison under Different Capacity on Bottleneck Work Stations

As shown in Figure 13, the overtime cost dominates the increase in total cost as capacity decreases (raw materials inventory cost not included). As shown in Figure

14, the optimal planned lead times of the work stations increase as we decrease the capacities, until the planned lead time reaches the upper bound. The optimal lot size solution is in general insensitive to the changes.

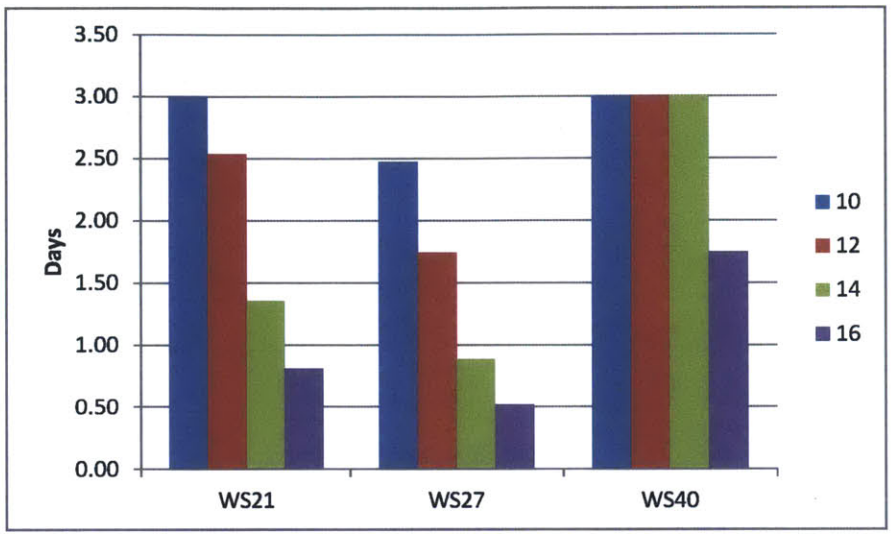

Figure 14: Planned Lead Time under Different Capacity on Bottleneck Work Stations

Experiment 6: Demand Standard Deviation

We adjust the standard deviation of daily demand by 50% around the base case in this analysis. In Figure 15, we observe that finished goods inventory cost, which increases linearly in the demand standard deviation (equation 4.7), dominates the change in total cost. A larger standard deviation also induces a greater overtime cost since more demand variability leads to more overtime. The WIP inventory cost decreases slightly when the demand standard deviation increases, because the planned lead times actually decrease on average (discussed below). The change in demand standard deviation has very little impact on the optimal lot sizes and thus changes in raw material are neglected.

As mentioned above, the planned lead times decrease on average as demand standard deviation increases. This is because a larger demand standard deviation implies more inventory holding cost for finished goods; to mitigate this, the solution

decreases the planned lead times which determine the replenishment times for the finished goods inventory. We observe in the optimal solution where the planned lead time for each product part increases by 0.18 days on average when the daily demand standard deviation is cut by half; the optimal planned lead time for each product part decreases 0.14 days on average when we increase the demand standard deviation by 50%.

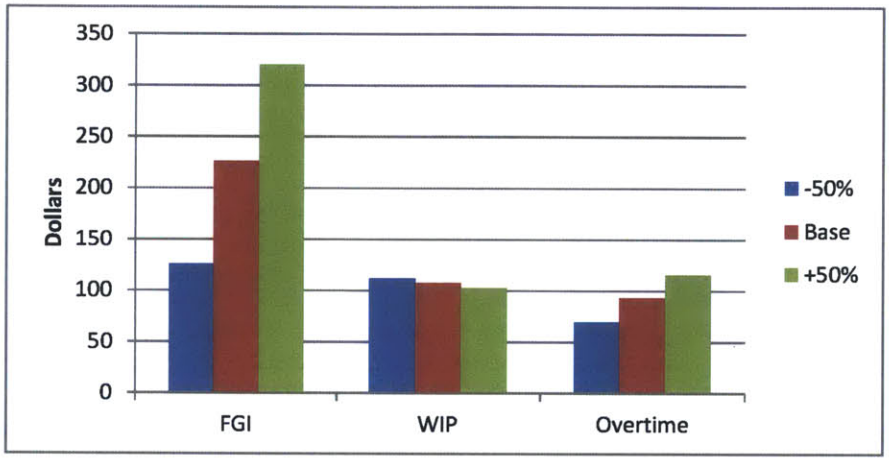

Figure 15: Cost Comparison under Different Demand Variability

Experiment 7: Demand Mean

We adjust the average demand by 10% around the base case in this analysis. The adjustment has a relatively large impact on both the cost and the optimal solution. We observe that as demand mean increases, all cost components increase by a considerable amount, especially raw material cost and production overtime cost (Figure 15).

When the demand mean increases (decreases) by 10%, we observe that the lot sizes increase (decreases) on 25 (14) product parts. This observation confirms our intuition that we can leverage the lot sizes to reduce the aggregate workload on work stations. On the other hand, three work stations (work station 1, 34 and 35) become lightly loaded as average demand decreases by 10%. The planned lead times of the

not-lightly loaded work stations increase (decrease) by 0.4 (0.3) day on average as the demand mean increases (decreases) by 10%. The planned lead times of the product parts increase (decrease) by 0.6 (0.5) days on average as the demand mean increases (decreases) by 10%.

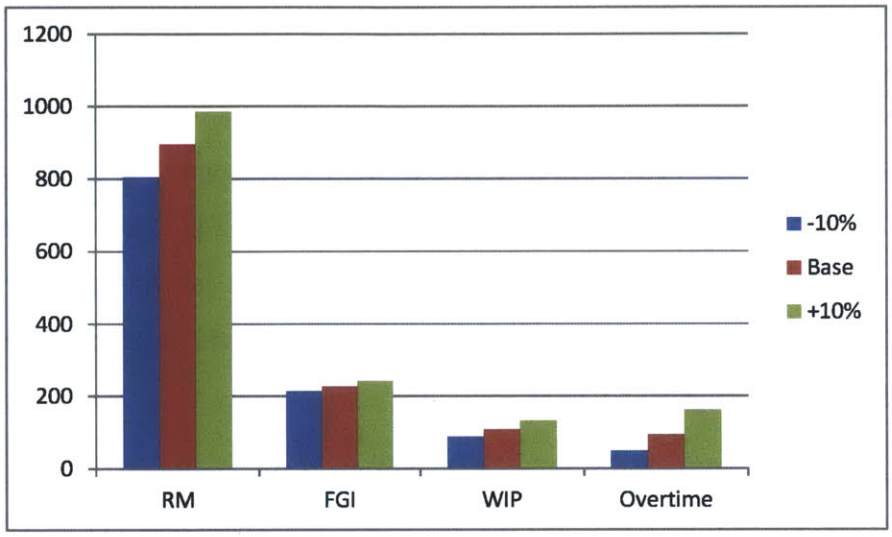

Figure 16: Cost Comparison under Different Demand Mean

6 Conclusions

In this research, we build a tactical model to analyze the operating tactics in discrete-part manufacturing systems, as typified by job shops. The model finds the minimum inventory-related costs and production overtime cost and determines the optimal operating tactics for the production lot size for each product part and the planned lead time for each work station.

Our model assumes that the product part lot arrival is a Poisson process. We then characterize the production at each work station as a function of the planned lead time through applying the linear production control rule described in Graves (1986). Our model sets as decision variables the production lot sizes for each part and the planned lead times for each work station, and calculates the inventory-related costs and production overtime cost. We implement the model in Excel spreadsheet and solve the non-linear optimization problem using the premium Excel built-in Solver to determine the optimal solution.

We have tested our model on both the hypothetical and the large-scale actual factory data which consists of 133 parts and 59 work stations; the results highlight and confirm managerial intuition. We also performed sensitivity analysis on the managerial parameters and other factory data to evaluate the influence from different inputs. Our computational experience shows that the problem in this scale can be solved very efficiently on a personal computer and the Solver is reasonably reliable if we use the lower bounds of the decision variables as the initial searching point. Furthermore, the model has been set up in a flexible way so that with minimum adjustment it can be applied to more general manufacturing systems.

Our model provides a convenient tool for the manager to establish intuition and test different operating policies. Our model essentially captures the several tradeoffs commonly seen in a complex manufacturing system. Our model confirms that extending the planned lead times for work stations is effective for production smoothing, however longer planned lead times require higher inventory stocks. We also found that the lot sizes for product parts can have a two-sided impact on the system: on the one hand, larger lot sizes reduce the number of setups which cuts the aggregate workload to be processed on the work stations; on the other hand, larger lot sizes induce higher production variability which can increase both production overtime and inventory-related costs. By solving the optimization problem, our model determines the optimal operating tactics in order to maintain the optimal balance between the inventory-related costs and the production overtime cost.

We finally discuss a few possible extensions of our model. First, in many manufacturing systems, the factory is able to manage the service level or delivery due time for each product part to the next stage in the supply chain. For example, important customers will be assigned higher priorities. Incorporating the backlog cost or delay penalty into the model can help capture the effect from service level management of the supply chain system. A second extension of the model is to relax the Poisson process assumption for the product part lot arrivals. Assuming Poisson lot arrivals provides some convenience for the mathematical modeling, in particular in expressing the workload arrivals on each work stations. We recommend setting the number of lots to be less than three per time period in order to make the Poisson assumption more realistic. However, any other rules that capture the impact from the lot sizes on the workload arrival variance can fit into our model. While we may lose some tractability, a more complicated rule can potentially improve the accuracy of our approximation. Lastly, one can extend the model to include shortages or machine breakdown in the system. For example, it is actually fairly easy to incorporate backlog for the raw material and finished parts inventory. As a final note, the model has been set up in a flexible way so that with minimum adjustment it

can be applied to more general cases, which may depend on the actual situation of the manufacturing system.

Reference

Adshead, N. S., & Price, D. H. R. (1986). Experiments with stock control policies and leadtime setting rules, using an aggregate planning evaluation model of a make-for-stock shop. *International journal of production research,24*(5), 1139-1157.

Asmundsson, J., Rardin, R. L., & Uzsoy, R. (2006). Tractable nonlinear production planning models for semiconductor wafer fabrication facilities.*Semiconductor Manufacturing, IEEE Transactions on, 19*(1), 95-111.

Bertrand, J. W. M. (1985). Multiproduct optimal batch sizes with in-process inventories and multi work centers. *IIE Transactions, 17*(2), 157-163.

Bitran, G. R., & Dasu, S. (1992). A review of open queueing network models of manufacturing systems. *Queueing systems, 12*(1), 95-133.

Cruickshanks, A. B., Drescher, R. D., & Graves, S. C. (1984). A study of production smoothing in a job shop environment. *Management Science, 30*(3), 368-380.

Fine, C. H., & Graves, S. C. (1989). A tactical planning model for manufacturing subcomponents of mainframe computers. *Journal of Manufacturing and Operations Management, 2*(1), 4-34.

Graves, S. C. (1986). A tactical planning model for a job shop. *Operations Research, 34*(4), 522-533.

Graves, S. C. (1988, December). Extensions to a Tactical Planning Model for a job shop. In Decision and Control, 1988., Proceedings of the 27th IEEE Conference on (pp. 1850-1855). IEEE.

Graves, S. C., & Hollywood, J. S. (2001). *A constant-inventory Tactical Planning Model for a job shop*. Working paper, January 2001, revised March 2004.

Graves, S. C., Kletter, D. B., & Hetzel, W. B. (1998). A dynamic model for requirements planning with application to supply chain optimization. *Operations Research, 46*(3-Supplement-3), S35-S49.

Gong, L., de Kok, T., & Ding, J. (1994). Optimal leadtimes planning in a serial production system. *Management Science, 40*(5), 629-632.

Hollywood, J. S. (2005). An approximate planning model for distributed computing

networks. *Naval Research Logistics (NRL)*, *52*(6), 590-605.

Jackson, J. R. (1957). Networks of waiting lines. *Operations Research*, *5*(4), 518-521.

Jackson, J. R. (2004). Jobshop-like queueing systems. *Management science*,*50*(12 supplement), 1796-1802.

Karmarkar, U. S., Kekre, S., Kekre, S., & Freeman, S. (1985a). Lot-sizing and lead-time performance in a manufacturing cell. *Interfaces*, *15*(2), 1-9.

Karmarkar, U. S., Kekre, S., & Kekre, S. (1985b). Lotsizing in multi-item multi-machine job shops. *IIE transactions*, *17*(3), 290-298.

Karmarkar, U. S. (1987). Lot sizes, lead times and in-process inventories.*Management science*, *33*(3), 409-418.

Karmarkar, U. S. (1989). Capacity loading and release planning with work-in-progress (WIP) and leadtimes. *Journal of Manufacturing and Operations Management*, *2*(105-123).

Karmarkar, U. S. (1993). Manufacturing lead times, order release and capacity loading. *Handbooks in operations research and management science*, *4*, 287-329.

Kenyon, G., Canel, C., & Neureuther, B. D. (2005). The impact of lot-sizing on net profits and cycle times in the *n*-job, *m*-machine job shop with both discrete and batch processing. *International Journal of Production Economics*, *97*(3), 263-278.

Lambrecht, M. R., & Vandaele, N. J. (1996). A general approximation for the single product lot sizing model with queueing delays. *European Journal of Operational Research*, *95*(1), 73-88.

Matsuura, H., Tsubone, H., & Kanezashi, M. (1996). Setting planned lead times for multi-operation jobs. *European journal of operational research*, *88*(2), 287-303.

Matsuura, H., & Tsubone, H. (1993). Setting planned leadtimes in capacity requirements planning. *Journal of the Operational Research Society*, 809-816.

Selcuk, B., Fransoo, J. C., & De Kok, A. G. (2008). Work-in-process clearing in supply chain operations planning. *IIE Transactions*, *40*(3), 206-220.

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

Teo, C. C., Bhatnagar, R., & Graves, S. C. (2012). An Application of Master Schedule Smoothing and Planned Lead Time Control. *Production and Operations Management*.

Wagner, H. M., & Whitin, T. M. (1958). Dynamic Version of the Economic Lot Size

Model. *Management Science*, *5*(1), 89-96.

Yano, C. A. (1987). Setting planned leadtimes in serial production systems with tardiness costs. *Management science*, *33*(1), 95-106.

Zipkin, P. H. (1986). Models for design and control of stochastic, multi-item batch production systems. *Operations Research*, *34*(1), 91-104.

Appendix

Appendix 1: Production Overtime Function and Its Properties

From formula (4.12), we know the production overtime cost in work station j can be expressed as the partial normal function where the production is assumed to be normally distributed with mean $E[P_j]$ and variance $Var(P_j)$:

$$Cost_j^{OT} = g_j \int_{C_j}^{\infty} (x - C_j) f_P(x) dx \quad (4.12)$$

Where g_j is the unit production overtime cost, C_j is the nominal workload capacity of work station j, and $f(\cdot)$ is the normal probability density function with mean $E[P_j]$ and variance $Var(P_j)$. W.l.o.g., we omit the subscript j for each work station for simplicity in the following. Let $\mu = E[P], \sigma^2 = Var(P)$, we can rewrite the partial normal function as follows:

$$Cost^{OT} = g \int_{C}^{\infty} (x - C) \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x-\mu)^2}{2\sigma^2}} dx$$

Change variable in the above expression with $y = \frac{x-\mu}{\sigma}$, we have

$$Cost^{OT} = g \int_{\frac{C-\mu}{\sigma}}^{\infty} (\sigma y + \mu - C) \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy$$

Let $\rho = \frac{C-\mu}{\sigma}$, we have

$$Cost^{OT} = g \int_{\rho}^{\infty} \sigma(y - \rho) \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy = g\sigma \int_{\rho}^{\infty} y \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy - g\sigma\rho \int_{\rho}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy$$

$$= g\sigma \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \Big|_{\rho}^{\infty} - g\sigma\rho(1 - \Phi(\rho)) = g\left(\frac{\sigma}{\sqrt{2\pi}} e^{-\frac{\rho^2}{2}} + (C - \mu)\Phi(-\rho)\right)$$

where $\Phi(\cdot)$ is the CDF of standard normal distribution $N(0,1)$. The above formula is the same as formula (4.13) in section 4.2.

We then show two properties of the production overtime cost function.

Proposition 1: The production overtime on the work station is an increasing and convex function in the expected workload (μ).

Proof:

Let us explicit express the production overtime cost $Cost^{OT}$ as a function μ and σ:

$$f(\mu,\sigma)=\int_C^{\infty}(x-C)\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx$$

Take derivative in terms of μ, we have

$$\frac{\partial f(\mu,\sigma)}{\partial\mu}=\int_C^{\infty}(x-C)\frac{1}{\sqrt{2\pi\sigma^2}}\frac{(x-\mu)}{\sigma^2}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx$$

Substitute $y=\frac{x-\mu}{\sigma}$, $\rho=\frac{C-\mu}{\sigma}$, we have

$$\frac{\partial f(\mu,\sigma)}{\partial\mu}=\int_\rho^{\infty}(y-\rho)\frac{y}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}dy=\int_\rho^{\infty}\frac{y^2}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}dy-\rho\int_\rho^{\infty}\frac{y}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}dy$$

$$=-\frac{1}{\sqrt{2\pi}}ye^{-\frac{y^2}{2}}\Big|_\rho^{\infty}+\int_\rho^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}dy+\frac{\rho}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}\Big|_\rho^{\infty}=\int_\rho^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}dy$$

$$=1-\Phi(\rho)>0$$

Take the second derivative in terms of μ, we have

$$\frac{\partial^2 f(\mu,\sigma)}{\partial\mu^2}=\frac{\partial\left(\int_{\rho(\mu)}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}dy\right)}{\partial\mu}=-\rho'(\mu)\frac{1}{\sqrt{2\pi}}e^{-\frac{\rho^2}{2}}=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{\rho^2}{2}}>0$$

Since both the first and the second derivatives of $f(\mu,\sigma)$ are positive in terms of the expected workload μ, we proved the production overtime cost function is increasing and convex in the expected workload.

Proposition 2: The production overtime on the work station is an increasing and convex function in the standard deviation of the workload (σ).

Proof:

Similarly, we take the first and the second derivatives of $f(\mu,\sigma)$in terms of σ.

$$f(\mu,\sigma)=\int_C^{\infty}(x-C)\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx$$

$$\frac{\partial f(\mu,\sigma)}{\partial\sigma}=-\int_C^{\infty}\frac{1}{\sigma^2}(x-C)\frac{1}{\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dp+\int_C^{\infty}\frac{(x-\mu)^2}{\sigma^3}(x-C)\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx$$

Substitute $y=\frac{x-\mu}{\sigma}$, $\rho=\frac{C-\mu}{\sigma}$, we have

$$\frac{\partial f(\mu,\sigma)}{\partial\sigma}=-\int_\rho^{\infty}(y-\rho)\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}dy+\int_\rho^{\infty}y^2(y-\rho)\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}dy$$

$$=\int_\rho^{\infty}y^3\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}dy-\rho\int_\rho^{\infty}y^2\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}dy-\int_\rho^{\infty}y\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}dy+\rho\int_\rho^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}dy$$

$$= \left(\frac{1}{\sqrt{2\pi}}\rho^2 e^{-\frac{\rho^2}{2}} + 2\frac{1}{\sqrt{2\pi}}e^{-\frac{\rho^2}{2}}\right) - \rho\left(\frac{1}{\sqrt{2\pi}}\rho e^{-\frac{\rho^2}{2}} + \int_\rho^\infty \frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}dy\right) - \frac{1}{\sqrt{2\pi}}e^{-\frac{\rho^2}{2}}$$
$$+ \rho\int_\rho^\infty \frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}dy = \frac{1}{\sqrt{2\pi}}e^{-\frac{\rho^2}{2}} > 0$$

Take the second derivative in terms of σ, we have

$$\frac{\partial^2 f(\mu,\sigma)}{\partial\sigma^2} = \frac{\partial\left(\frac{1}{\sqrt{2\pi}}e^{-\frac{\rho(\mu)^2}{2}}\right)}{\partial\sigma} = \frac{\rho^2}{\sigma}\frac{1}{\sqrt{2\pi}}e^{-\frac{\rho^2}{2}} > 0$$

As both the first and the second derivatives of $f(\mu,\sigma)$ are positive in terms of the standard deviation of the workload σ, we proved the production overtime cost function is increasing and convex in the standard deviation of the workload.

Appendix 2: Property of the Production Variance Function

We found in formula (4.11) that the production variance on work station j can be calculated as a function of the decision variables, the production lot size q_i and the planned lead time τ_j. We show in this section that the production variance is strictly decreasing in τ_j by observing the first order derivative of the variance is negative. The monotonicity of the production variance in terms of the planned lead time makes intuitive sense since longer planned lead time is intended to provide more production smoothness. We omit the subscript for work station for simplicity and rewrite (4.11) as follows:

$$Var(P) = \left(\frac{\beta}{2-\beta}(1-\gamma)^2 + \gamma^2\right) Var(A) \quad (4.11)$$

where

$\beta = 1 - (1 - \alpha/m)^m$;

$\gamma = 1 - \frac{(1-\alpha/m)}{\alpha}\beta$;

$\alpha = 1/\tau$

Note that $\tau \geq \frac{1}{m}$ is the boundary condition for the decision variable τ.

Since τ does not appear in $Var(A)$, we can focus on the term $\left(\frac{\beta}{2-\beta}(1-\gamma)^2 + \gamma^2\right)$. Our objective is to show the derivative of this term is non-positive in τ, i.e.

$$\left(\frac{\beta}{2-\beta}(1-\gamma)^2 + \gamma^2\right)' = \left(\frac{\beta}{2-\beta}\right)'(1-\gamma)^2 + \left(\frac{\beta}{2-\beta}\right)((1-\gamma)^2)' + (\gamma^2)'$$

$$= \frac{2\beta'}{(2-\beta)^2}(1-\gamma)^2 - 2\left(\frac{\beta}{2-\beta}\right)(1-\gamma)\gamma' + 2\gamma\gamma' \quad (A2.1)$$

is always a non-positive number in τ when $\tau \geq \frac{1}{m}$.

We first write down the expression for the each term appeared in (A2.1) in terms of τ:

$$\beta = 1 - \left(1 - \frac{1}{\tau m}\right)^m$$

$$\gamma = 1 - \tau(1 - \frac{1}{\tau m})\beta$$

Taking derivatives w.r.t. τ, we have

$$\beta' = -\frac{1}{\tau^2}\left(1 - \frac{1}{\tau m}\right)^{m-1}$$

$$\gamma' = -\beta - \tau(1 - \frac{1}{\tau m})\beta'$$

$$\left(\frac{\beta}{2-\beta}\right)' = \frac{2\beta'}{(2-\beta)^2}$$
$$((1-\gamma)^2)' = -2(1-\gamma)\gamma'$$
$$(\gamma^2)' = 2\gamma\gamma'$$

We observe that the first term $\frac{2\beta'}{(2-\beta)^2}(1-\gamma)^2$ in (A2.1) is always non-positive since $\beta' = -\frac{1}{\tau^2}\left(1-\frac{1}{\tau m}\right)^{m-1} \leq 0$ (note that $\tau \geq \frac{1}{m}$ is the given condition). We can then focus on the second and the third term.

Rewrite the sum of the second and the third term, we have

$$-2\left(\frac{\beta}{2-\beta}\right)(1-\gamma)\gamma' + 2\gamma\gamma' = -\frac{2\beta\gamma'}{2-\beta} + \frac{2\beta\gamma\gamma'}{2-\beta} + 2\gamma\gamma' = -\frac{2\beta\gamma'}{2-\beta} + \frac{4\gamma\gamma'}{2-\beta}$$
$$= \frac{2\gamma'}{2-\beta}(2\gamma-\beta) \quad (A2.2)$$

To show this term is also non-positive, we consider γ' and $2\gamma-\beta$ respectively.

Consider γ'', we have

$$\gamma'' = \left(-\beta - \tau\left(1-\frac{1}{\tau m}\right)\beta'\right)' = \left(-1+\left(1-\frac{1}{\tau m}\right)^m + \frac{1}{\tau}\left(1-\frac{1}{\tau m}\right)^m\right)'$$
$$= \frac{1}{\tau^2}\left(1-\frac{1}{\tau m}\right)^{m-1} - \frac{1}{\tau^2}\left(1-\frac{1}{\tau m}\right)^m + \frac{1}{\tau^3}\left(1-\frac{1}{\tau m}\right)^{m-1}$$
$$= \frac{1}{\tau^3 m}\left(1-\frac{1}{\tau m}\right)^{m-1} + \frac{1}{\tau^3}\left(1-\frac{1}{\tau m}\right)^{m-1} \geq 0$$

Furthermore, since

$$\gamma'|_{\frac{1}{m}} = -1+\left(1-\frac{1}{\tau m}\right)^m + \frac{1}{\tau}\left(1-\frac{1}{\tau m}\right)^m\Big|_{\frac{1}{m}} = -1$$

$$\gamma'|_{\infty} = -1+\left(1-\frac{1}{\tau m}\right)^m + \frac{1}{\tau}\left(1-\frac{1}{\tau m}\right)^m\Big|_{\infty} = 0$$

we conclude that γ' is an non-decreasing function on [-1,0] when $\tau \in [\frac{1}{m}, \infty)$ and thus

$$\boldsymbol{\gamma' \leq 0}$$

Consider $(2\gamma-\beta)''$, we have

$$(2\gamma-\beta)'' = (2\gamma'-\beta')' = \left(-2\beta-2\tau\left(1-\frac{1}{\tau m}\right)\beta'-\beta'\right)'$$

$$= \left(-2+2\left(1-\frac{1}{\tau m}\right)^m+\frac{2}{\tau}\left(1-\frac{1}{\tau m}\right)^m+\frac{1}{\tau^2}\left(1-\frac{1}{\tau m}\right)^{m-1}\right)'$$

$$= \frac{2}{\tau^2}\left(1-\frac{1}{\tau m}\right)^{m-1}-\frac{2}{\tau^2}\left(1-\frac{1}{\tau m}\right)^m+\frac{2}{\tau^3}\left(1-\frac{1}{\tau m}\right)^{m-1}-\frac{2}{\tau^3}\left(1-\frac{1}{\tau m}\right)^{m-1}$$

$$+\frac{1}{\tau^4}\left(\frac{m-1}{m}\right)\left(1-\frac{1}{\tau m}\right)^{m-2} = \frac{2}{\tau^3 m}\left(1-\frac{1}{\tau m}\right)^m+\frac{1}{\tau^4}\left(\frac{m-1}{m}\right)\left(1-\frac{1}{\tau m}\right)^{m-2}$$

$$\geq 0$$

Observe also

$$(2\gamma-\beta)'|_{\frac{1}{m}} = -2+2\left(1-\frac{1}{\tau m}\right)^m+\frac{2}{\tau}\left(1-\frac{1}{\tau m}\right)^m+\frac{1}{\tau^2}\left(1-\frac{1}{\tau m}\right)^{m-1}\Big|_{\frac{1}{m}} = -2$$

$$(2\gamma-\beta)'|_{\infty} = -2+2\left(1-\frac{1}{\tau m}\right)^m+\frac{2}{\tau}\left(1-\frac{1}{\tau m}\right)^m+\frac{1}{\tau^2}\left(1-\frac{1}{\tau m}\right)^{m-1}\Big|_{\infty} = 0$$

So $(2\gamma-\beta)'$ is an non-decreasing function on [-2,0] when $\tau \in [\frac{1}{m},\infty)$ and thus $(2\gamma-\beta)' \leq 0$ when $\tau \in [\frac{1}{m},\infty)$. Furthermore, since

$$(2\gamma-\beta)|_{\frac{1}{m}} = 2-\left(2\tau\left(1-\frac{1}{\tau m}\right)+1\right)\left(1-\left(1-\frac{1}{\tau m}\right)^m\right)\Big|_{\frac{1}{m}} = 1$$

$$(2\gamma-\beta)|_{\infty} = 2-\left(2\tau\left(1-\frac{1}{\tau m}\right)+1\right)\left(1-\left(1-\frac{1}{\tau m}\right)^m\right)\Big|_{\infty} = 0$$

The last equality applies the identity $\lim_{\tau\to\infty}\tau\left(1-\left(1-\frac{1}{\tau m}\right)^m\right)=1$ which can be shown by L'Hopital's rule. This shows that $(2\gamma-\beta)$ is non-increasing on [1,0] when $\tau \in [\frac{1}{m},\infty)$ and thus

$$\boldsymbol{(2\gamma-\beta)\geq 0}$$

Finally, since $\gamma' \leq 0$ and $(2\gamma-\beta)\geq 0$ when $\tau \in [\frac{1}{m},\infty)$, we know the expression (A2.2) is always non-positive and consequently the derivative (A2.1) of the production overtime cost function is non-positive, which implies that the production cost is non-increasing in the planned lead time τ.

Appendix 3: Generating Demand Mean and Variance Based on Given Demand Forecast

We explain how we generate demand mean and variance from the forecast data. The method accounts for order cancellation and variability in the due dates for the orders. We do not consider the movement of product orders (advanced and delayed orders).

Mean and Variance of Number of Orders per Month

N_j : number of orders need to deliver at month j

$X_{i,j}$: number of orders generated at month i and to be due at month j

M_i: number of orders generated at month i

The number of orders to deliver at month j is the sum of orders due at month j:

$$N_j = \sum_{i=1}^{t} X_{j-i,j}$$

We can characterize $X_{i,i+1}, X_{i,i+2}, \dots, X_{i,i+t}$ as a multinomial distribution:

$$\boldsymbol{Mult(M_i, \alpha_1, \alpha_2 \dots \alpha_t)}$$

where $\sum_{k=1}^{t} \alpha_k = 1$; $\sum_{k=1}^{t} X_{i,i+k} = M_i$; $M_i \sim normal(\mu, \sigma^2)$

We then can derive the conditional expectation and variance of $X_{i,j}$:

$$E[X_{i,i+k} | M_i = m] = \alpha_k m$$

$$Var(X_{i,i+k} | M_i = m) = \alpha_k(1 - \alpha_k)m$$

Based on the conditional expectation and variance, we can proceed to obtain the expectation and variance of N_j:

$$E[N_j] = \sum_{i=1}^{t} E[X_{j-i,j}] = \sum_{i=1}^{t} \alpha_i E[M_{j-i}] = \mu$$

$$Var(N_j) = \sum_{i=1}^{t} Var(X_{j-i,j}) = \sum_{i=1}^{t} \{E[Var(X_{j-i,j} | M_{j-i})] + Var(E[X_{j-i,j} | M_{j-i}])\}$$

$$= \sum_{i=1}^{t} \{\alpha_i(1 - \alpha_i)E[M_{j-i}] + Var(\alpha_i M_{j-i})\} = \sum_{i=1}^{t} \{\alpha_i(1 - \alpha_i)\mu + \alpha_i^2 \sigma^2\}$$

$$= \mu + (\sigma^2 - \mu) \sum_{i=1}^{12} \alpha_i^2$$

Mean and Variance of Number of Units per Month

Y: number of units to be delivered

D_i: number of units in order i

N: number of orders to be delivered

π: Expected preservation rate

$$Y = \sum_{i=1}^{N} \widetilde{D}_i$$

$$\widetilde{D}_i = \begin{cases} D_i & \pi \\ 0 & 1-\pi \end{cases}$$

$$D_i \text{ are } i.i.d \text{ with } N(u, v^2)$$

The expectation of Y is:

$$E[Y] = E\left[E\left[\sum_{i=1}^{N} \widetilde{D}_i \middle| N = n\right]\right] = E[Nu\pi] = u\pi E[N]$$

The variance of $\widetilde{D}_i$ is:

$$Var(\widetilde{D}_i) = E[\widetilde{D}_i^2] - E^2[\widetilde{D}_i] = \pi E[D_i^2] - \pi^2 E^2[D_i] = \pi Var(D) + \pi(1-\pi)E^2[D]$$
$$= \pi v^2 + \pi(1-\pi)u^2$$

We derive the variance of Y by conditioning on N:

$$Var(Y) = E\left[Var\left(\sum_{i=1}^{N} \widetilde{D}_i \middle| N\right)\right] + Var\left(E\left[\sum_{i=1}^{N} \widetilde{D}_i \middle| N\right]\right) = E[NVar(\widetilde{D}_i)] + Var(Nu\pi)$$
$$= E[N]Var(\widetilde{D}_i) + (u\pi)^2 Var(N)$$

By substituting$Var(\widetilde{D}_i) = \pi v^2 + \pi(1-\pi)u^2$, we have

$$Var(Y) = E[N](\pi v^2 + \pi(1-\pi)u^2) + (u\pi)^2 Var(N)$$

Numerical Comparison

	Simulation		Analytical Computation	
Model No.	**Ave**	**StdDev**	**Ave**	**StdDev**
2	107.4	29.6	105.7	26.7
5	24.3	6.2	24.1	6.1
6	7.2	3.3	7.2	3.4
7	7.2	3.4	7.2	3.4
8	9.5	3.8	9.7	3.9
10	31.1	7.0	30.6	5.9

13	10.3	4.0	10.2	3.7
15	10.1	3.9	10.2	3.7
16	39.1	8.5	39.8	8.6
20	5.4	4.1	5.2	3.9
23	8.0	5.3	8.1	5.7
25	51.2	14.8	52.6	13.6