To appear in the *International Journal of Production Research*
Vol. 00, No. 00, Month 2015, 1–14

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

Rong Yuan[a*] and Stephen C. Graves[b]

[a] *Operations Research Center, Massachusetts Institute of Technology,Cambridge, Massachusetts 02139-4307;* [b] *Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139-4307*

(*version released May 2015*)

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 tactical policies that minimize the relevant manufacturing costs subject to workload variability and capacity limits. We consider two tactical decisions, namely 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, as functions of these tactical decisions. We formulate a non-linear optimization model and implement it in the Excel spreadsheet. We test the model with actual factory data from our research sponsor. The results are consistent with our intuition and demonstrate the potential value from jointly optimizing over these tactical policies.

Keywords: production planning, planned lead time, lot size, job shop

1. Introduction

Our work is motivated by real-world planning challenges faced by manufacturers who run complex manufacturing systems such as job shops. Manufacturers operating with job shops face business challenges with shorter product life cycles and increased demand uncertainty. In the job shop environment, multiple types of work stations can process multiple types of jobs. Each job type can have a distinct processing route and thus the work flow in a job shop is often quite complicated. The heterogeneity and complexity of the work flow creates great difficulties for production planning and scheduling.

We consider two tactical decisions in a make-to-stock job shop context, namely the setting of the production lot sizes and the planned lead times. A lot consists of multiple units of the same type of part. The processing of a lot at each work station 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 station only when the entire lot is complete. Increasing the lot size will reduce the number of setups and thus reduce loads on the work stations. However, large lot sizes also imply longer processing times and a less fluid work flow, which results in lumpier arrivals to each work station. An increase in arrival variability then translates into increased production variability.

The planned lead times are planning parameters that correspond to how much time is allowed for each job at each production step or work station. The idea is to provide a reasonable amount of time for each job at each work station to accommodate the amount and variability of the station's workload. The planned lead time is always longer than the required processing time; it includes an additional buffer to permit queuing and to provide planning flexibility. Longer planned lead

*Corresponding author. Email: rongyuan@mit.edu

times imply a higher level of work-in-process inventory. However, longer planned lead times can help to dampen the variability in workload arrivals to each work station, resulting in production smoothing and less overtime.

Our intent is to develop a planning tool to determine the optimal tactical policies that minimize the relevant manufacturing costs subject to demand variability and capacity limits. Our model can determine the optimal lot sizes or evaluate the factory performance for a given lot-sizing policy. It can also determine the planned lead time for each work station; these planned lead times can then be used to inform decisions for overtime, for work release and for the replenishment policy for a make-to-stock job shop.

The remainder of the paper is organized as follows. In Section 2 we briefly review the related literature and address how our work complements the previous contributions to tactical planning in job shops. Section 3 describes the assumptions and the formulation of our model. Section 4 discusses the implementation of the model and presents our numerical testing with real factory data. We conclude and discuss future research opportunities in Section 5.

2. Literature Review

There are two streams of literature that are of particular relevance to our work. One stream focuses on setting planned lead times and the other on the impact of lot-sizing policies on the performance of job shops.

The planned lead time of a job is defined as the planned amount of time for a job to spend at a step or stage in the system. Cruickshanks et al. (1984) introduced the concept of a planning window in the environment of a job shop. Their model shows 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 job. Adshead et al. (1986) experimented with different rules to set planned lead times in a simulation model. Analytical results have been developed for determining the optimal planned lead time under different assumptions by Yano (1987), Gong et al. (1994), Matsuura and Tsubone (1993), and Matsuura et al. (1996).

Our work is closely related to the Tactical Planning Model (TPM) developed by Graves (1986). The TPM considers a network of work stations in a job shop in which the production rate is modeled as a fixed proportion of the queue level at the beginning of each time period for each work station. This "fixed proportion" translates into a planned lead time for each work station. From the TPM, one can determine the first two moments of the production levels at each work station, as well as the distribution of queue lengths. The TPM provides a tractable way to characterize the production and inventory requirements in complex job shops, as functions of the planned lead times.

There are several papers that extend the TPM to different contexts. See Fine and Graves (1989), Graves and Hollywood (2001), Hollywood (2005), Graves et al. (1998), Teo et al. (2011) and Teo et al. (2012). In particular, Teo et al. (2011) relax the discrete-time assumption in the TPM to allow additional modeling flexibility. We utilize this in our work.

The other stream of research focuses on the impact of lot-sizing policies on the performance and operational costs of job shops. Karmarkar (1987) and Zipkin (1986) were among the first to model the queueing impact from lot-sizing decisions in a single work station setting. Both papers model the work station as an M/M/1 queue with Poisson lot arrivals and exponential lot processing times; Karmarkar (1987) considers the impact of lot-sizing decisions on queueing delays and finds the optimal lot-sizing policy that minimizes the work-in-process and finished goods inventory. Zipkin (1986) solves for optimal lot-sizing policy with additional consideration of backlog costs and develops an extension to a simple network of work stations. The line of research has been further developed with more general arrival stream assumptions by Lambrecht and Vandaele (1996) and Tielemans and Kuik (1996), with a re-order point policy considering safety stock and inventory backlogs by Vaughan (2006), and in the context of a multi-period setting by Kang et al. (2014).

Substantial progress has been made in extending the single station model to a multi-item and multi-machine environment. Karmarkar et al. (1985b) extends the model in Karmarkar (1987) and assumes an M/G/1 queue at each work station. Söhner and Schneeweiss (1995) study the lot-sizing decisions in a hierarchical planning and scheduling context. Their model assumes a D/G/1 queue at each work station and solves a non-convex optimization problem to determine the lot-sizing policy.

Both streams of literature share the same goal, namely to minimize the manufacturing costs in the system by choosing the optimal operating policies. This motivates us to apply the two sets of policies concurrently to achieve better planning. Similar to the literature, our model captures the impact of lot sizes on work station utilization and on both flow variability and flow times. In addition, our model allows for a certain degree of production flexibility, through the choice of the planned lead times, that introduces an additional trade-off between production and inventory-related costs. We discuss the difference of our model and the conventional queueing-based model in section 3.5 of this paper.

3. Model

In our model, we express the inventory-related costs (raw materials, work-in-process and finished goods inventory) and production overtime costs as functions 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.

The job shop produces each part to a finished goods inventory. Each 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. Processing each lot entails a sequence of process steps. Each process step is specified by a work station at which the work is done, a per-unit processing time and a setup time.

The production release for a lot is triggered by the finished goods inventory; when the inventory position for a part drops below a reorder point, an order is issued to produce another lot for the part. To determine the reorder point, we need to estimate the replenishment lead time for each part, i.e., how long will it take to make the part. This is complicated as each part needs to visit a sequence of work stations for processing and there can be substantial waiting at each station, due to the fact that a station might be part of the replenishment process for several different parts. The wait time depends on the workload at each work station, as well as its variability.

We associate a planned lead time with each work station as a way to simplify the problem, and as a way to articulate and quantify the key trade-offs. One can use the planned lead times to estimate how long it will take to replenish the inventory for a part and hence can determine what its reorder point should be. But there is a cost associated with the planned lead time for each work station, as the variability of production varies inversely with the planned lead time. In our context, a shorter planned lead time results in more variable production, which incurs overtime costs.

3.1 *Assumptions*

We list the main assumptions for the model in the following:

A1. Demand Process The daily demand for each part follows an i.i.d. distribution with known first two moments. We think this assumption is reasonable as the parts are components for a family of final assemblies that have regular and stable demand.

A2. Inventory Policy We use a continuous review, order quantity, reorder point (QR) policy for the control of the inventory for each finished part. We use a periodic review policy for control of each raw material.

A3. Inventory Shortage When a part triggers replenishment, there is no delay in the release of the order to the job shop. In effect, we assume that there is a very high service level for the raw material inventory and that shortages are of short duration. We also assume that

when the finished goods inventory stocks out, demand is back-ordered. For the model, we assume that we are given service level targets for both the finished goods inventory and the raw material inventory.

A4. Lot Processing 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.

A5. Arrivals at Work Stations 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. We tested this assumption with a discrete-event simulation. For this simulation we first fix the lot size for each part q_i. At time period t, we generate demand d_{it} (units) for each part i using a normal random variable generator with known demand mean and standard deviation. We then calculate the number of lots to be released at time t by $\lfloor d_{it}/q_i \rfloor$. The remaining units $d_{it} - \lfloor d_{it}/q_i \rfloor q_i$ are carried over and added to the demand for the next period. We simulate this for 1000 time periods and record the number of lots released each period for each part. The results show that the assumption is reasonable, provided that the arrival rate for a part is less than or equal to 3 lots per day. We found that the Poisson assumption overestimates the variability of lot arrivals if the arrival rate is more than 3 lots per day. However, it will be quite rare for a part to generate multiple replenishment orders in a given 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. Processing Time We assume the processing time for each unit at each work station is deterministic. As lots are processed without interruption, the processing time for a lot is its setup time, plus the time to process all of the units in the lot.

A7. Setup Time We 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.

A8. Production Rate We assume that we can adjust the production rate at each work station, at some frequency (e.g., twice per day), based on the amount 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.

3.2 *Notation*

We associate the index i with parts and index j with work stations. When we say job i, we mean a production lot for part i.

Input Parameters

μ_i - demand mean of part i (units/day)
σ_i - demand standard deviation of part i (units/day)
p_{ij} - unit processing time of part i on work station j (hours per unit)
s_j - setup time for work station j (hours per setup)
b - inverse of the number of working hours in a day
$N(j)$ - index set of parts that need to be processed at work station j
$M(i)$ - index set of work stations that are on the process route of part i
h_i^R - holding cost for raw material of part i (dollars per unit per day)
h_i^G - holding cost for finished goods of part i (dollars per unit per day)
g_j - overtime cost at work station j (dollars per hour)
C_j - nominal production capacity of work station j (hours/day)
L^r - review period of the raw material inventory (days)
L^d - delivery lead time of the raw material (days)
z^R - safety factor for the raw material inventory
z^G - safety factor for the finished goods inventory

N - number of part types in the system
M - number of work stations in the system
m - production adjustment frequency

Decision Variables

q_i - lot size of part i (units)
τ_j - planned lead time of work station j (days)

Derived Variables

λ_i - lot arrival rate (lots/day)
T_i - total planned lead time of one lot of part i in the system (days)
T_{ij} - planned lead time for one lot of part i at work station j (days)
w_{ij} - processing time of one lot of part i on work station j (hours/lot)

3.3 *Formulation*

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

$$w_{ij} = p_{ij}q_i + s_j \tag{1}$$

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

We define the planned lead time for one lot of part i at work station j (T_{ij}) as the sum of the planned lead time of work station j (decision variable τ_j) and the lot processing time bw_{ij} (expressed in days). The planned lead time of work station j, τ_j, is the time that one plans for a job to wait at work station j. We note that this definition differs from how the planned lead time is defined in the earlier literature where the planned lead time for any part (or job) i at work station j is just τ_j. In our context the processing times vary quite a bit for different 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 part i by summing T_{ij} over the work stations j that part i visits:

$$T_i = \sum_{j \in M(i)} T_{ij} = \sum_{j \in M(i)} (\tau_j + bw_{ij}) \tag{2}$$

The lot arrival rate for part i is given by $\lambda_i = \mu_i/q_i$. Given assumption A5, we model the number of lots of part i that arrive each day to work station j as a Poisson random variable with mean and variance both equal to λ_i. Each lot brings a workload of w_{ij}. Hence, the daily workload for part i at work station j has expectation and variance $\lambda_i w_{ij}$ and $\lambda_i w_{ij}^2$, respectively. 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 (expressed in hours per day):

$$E[A_j] = \sum_{i \in N(j)} \lambda_i w_{ij} = \sum_{i \in N(j)} (\mu_i p_{ij} + \frac{s_j \mu_i}{q_i}) \tag{3}$$

$$Var(A_j) = \sum_{i \in N(j)} \lambda_i w_{ij}^2 = \sum_{i \in N(j)} (\mu_i p_{ij}^2 q_i + \frac{\mu_i s_j^2}{q_i} + 2\mu_i p_{ij} s_j) \tag{4}$$

Equation (3) specifies the workload arrival rate on work station j as the aggregate workload generated per unit time by all parts that need to be processed at work station j. Similarly, equation (4) describes the workload variance. Notice that for the above calculation, we assume that the arrival process for each part is independent of that for all other parts. We observe that $E[A_j]$ is strictly deceasing and convex in each q_i if the setup time on work station j is non-zero. We also see that $Var(A_j)$ is a convex function in each q_i, and that increasing a lot size q_i will reduce workload arrival variability due to setup time (second term in bracket) but increase the variability associated with the lot size effect (first term in bracket).

We are now ready to describe the formula used to calculate the cost components as functions of the lot sizes and planned lead times.

3.3.1 Raw Materials Inventory Cost

We use a periodic review policy for the raw materials inventory with a review period (L^r) and an order-up-to level at each review period. We then wait a constant delivery lead time (L^d) for the order to arrive and the inventory to be replenished. In addition, the lot release process of part i to the first work station is effectively 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 material (RM) inventory cost for part i using the following standard formula for periodic review policy (Silver et al. 1998, Formula 7.37), consisting of approximations for the cycle stock and safety stock levels.

$$\begin{aligned} Cost_i^{RM} &= h_i^R\left(\frac{\lambda_i q_i L^r}{2} + z^R\sqrt{\lambda_i q_i^2}\sqrt{L^d + L^r}\right) \\ &= h_i^R\left(\frac{\mu_i L^r}{2} + z^R\sqrt{\mu_i q_i}\sqrt{L^d + L^r}\right) \end{aligned} \tag{5}$$

Equation (5) directly shows that the RM cost of part i is strictly increasing and concave in its lot size q_i.

3.3.2 Finished Goods Inventory Cost

We assume an order quantity, reorder point (QR) continuous review policy for the finished goods inventory. That is, we will order a fixed quantity, the lot size, every time the inventory position falls below a reorder point. The replenishment lead time is simply the planned lead time for the part T_i. Following the standard formula for QR continuous review policy (Silver et al. 1998, Formula 7.14), we can model the daily cycle stock and safety stock of the finished goods inventory (FGI) of part i by the approximation:

$$\begin{aligned} Cost_i^{FGI} &= h_i^G\left(\frac{q_i}{2} + z^G\sigma_i\sqrt{T_i}\right) \\ &= h_i^G\left(\frac{q_i}{2} + z^G\sigma_i\sqrt{\sum_{j\in M(i)}(\tau_j + bp_{ij}q_i + bs_j)}\right) \end{aligned} \tag{6}$$

where the second equality is obtained by substituting equation (1) and (2). For this approximation we treat the replenishment time as being deterministic. Equation (6) shows that the FGI cost of part i is strictly increasing with its lot size and the planned lead times of the work stations on its process route.

3.3.3 Work-in-process Cost

By Littles Law, the work-in-process (WIP) for part i is directly proportional to the total planned lead time. We approximate the holding cost of WIP at the average of the RM and FGI costs; thus we use the following approximation:

$$\begin{aligned} Cost_i^{WIP} &= \frac{h_i^R + h_i^G}{2} T_i \mu_i \\ &= \frac{h_i^R + h_i^G}{2} \sum_{j \in M(i)} (\tau_j + bp_{ij} q_i + bs_j) \mu_i \end{aligned} \tag{7}$$

where the second equality is obtained by substituting equation (1) and (2). We observe from (7) that the WIP cost of part i is also increasing with its lot size and with the planned lead times of the work stations on its process route.

3.3.4 Review of the Single-station Tactical Planning Model (TPM)

We use the single-station TPM to model the overtime costs. The TPM is a discrete-time, continuous flow framework. The TPM views the job shop as a stochastic flow system with planned lead times. The TPM assumes that each work station operates with a linear production control rule, i.e. the amount of work processed at each time period is a fixed fraction of the workload queue in front of that work station: $P_t = \alpha Q_t$, where P_t is the amount of work processed in time period t, Q_t is the workload 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 time period. We emphasize that both P_t and Q_t are measured in workload (e.g. hours of work per day).

The TPM assumes some amount of production flexibility is possible and that there is not a hard capacity constraint. More specifically, the TPM assumes that whenever the queue grows at a work station, it can always increase its production rate, at some cost, to ensure the queue gets processed (on average) at the planned lead time. In this paper, we assume there is an option to work overtime whenever the required production exceeds the nominal production capacity. By adjusting the overtime cost, one can impose various degrees of hardness to the nominal capacity limit. For example, one can model the overtime cost as a piece-wise function where the overtime cost is an increasing function of the workload.

The linear production control rule enables the planned lead time. One can interpret the inverse of the smoothing parameter α as the planned lead time $\tau = 1/\alpha$ for the work station. If the planned lead time is τ time periods, then the work station is expected to process $1/\tau = \alpha$ amount of the queue within each time period.

The TPM is a discrete-time model that assumes the production rate to be constant over the planning period. For our motivating example the planning period is a day, as this is the periodicity of the demand arrivals. However, the production rate can be readily adjusted at least twice a day as there are normally two working shifts in a day, each with an option for overtime. We thus applied the extension of the TPM made by Teo et al. (2011) to allow for this production flexibility.

The extension of the TPM model essentially divides each planning period t into m equal-length sub-periods where m reflects the frequency with which the production rate could be adjusted. The model then assumes the workload arrival A_t is uniformly spread over the period, i.e., at the beginning of each sub-period the workload arrival is A_t/m. This assumption tends to be more reasonable if the workload arrivals have short inter-arrival times relative to the time period (Teo et al. 2011, pp 404-405). We think this assumption is acceptable given the time period is a day and the workload arrivals are the superposition of many Poisson processes, one for each part produced at a station.

Assuming the arrival process A_t at a work station is i.i.d with mean $E[A]$ and variance $Var(A)$, we can then determine the first two moments of production (refer to Teo et al. 2011 for more

details):

$$E[P_t] = E[A] \tag{8}$$

$$Var(P_t) = (\frac{\beta}{2-\beta}(1-\gamma)^2 + \gamma^2)Var(A) \tag{9}$$

where $\beta = 1-(1-\alpha/m)^m$; $\gamma = 1-\beta(\frac{1-\alpha/m}{\alpha})$; $\alpha = 1/\tau$. Here, $\tau = 1/\alpha$ is the planned lead time (in time periods) associated with the work station; α can take any positive value, which allows the planned lead time to be less than one time period.

We provide an intuitive interpretation for the parameter m. A larger m corresponds to a more continuous work flow arrival to the work station. A larger m also 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.

3.3.5 Production Overtime Cost

We now proceed to model each work station using the single-station TPM. We apply formula (8) and (9) to characterize the production at work station j with moments:

$$E[P_{jt}] = E[A_j] \tag{10}$$

$$Var(P_{jt}) = (\frac{\beta_j}{2-\beta_j}(1-\gamma_j)^2 + \gamma_j^2)Var(A_j) \tag{11}$$

where $\beta_j = 1-(1-\alpha_j/m)^m$; $\gamma_j = 1-\frac{1-\alpha_j/m}{\alpha_j}\beta_j$; $\alpha_j = 1/\tau_j$. τ_j is the planned lead time associated with work station j defined earlier, m is the production adjustment frequency, and $Var(A_j)$ is calculated by formula (4). In order to calculate the overtime cost, we now assume that the production at each work station is normally distributed. As justification, we found by simulation that a typical heavily-loaded work station (note that we only care about heavily-loaded work stations as those are the work stations that will incur overtime costs) usually processes 5 to 20 parts every day, each part has 1 to 3 lot arrivals and the lot processing time of those parts range from 30 minutes to 5 hours. We verified that the workload arrivals are very close to a normal distribution for the given range of the parameters as for each heavily-loaded work station the number of lot 10 to 20 lots per day. Normality is preserved for production workload as we apply a linear production rule to process the queue at the work stations. We can then approximate the daily production overtime cost using the normal loss function:

$$Cost_j^{OT} = g_j \int_{C_j}^{\infty} (x - C_j) f_P(x) dx \tag{12}$$

where we assume $f_P(x)$ is a normal probability density function with mean $E[P_{jt}]$ and variance $Var(P_{jt})$ as specified in (10) and (11); C_j is the nominal production capacity of the work station j, measured in hours of work per day. We can then determine the production overtime cost by solving the normal loss function as:

$$\begin{aligned} Cost_j^{OT} &= g_j \int_{C_j}^{\infty} (x - C_j) \frac{1}{\sqrt{2\pi Var(P_{jt})}} e^{-\frac{(x-E[P_{jt}])^2}{2Var(P_{jt})}} dx \\ &= g_j \left(\sqrt{\frac{Var(P_{jt})}{2\pi}} e^{-\rho_j^2/2} + (E[P_{jt}] - C_j)\Phi(-\rho_j) \right) \end{aligned} \tag{13}$$

where $\rho_j = \frac{C_j - E[P_{jt}]}{\sqrt{Var(P_{jt})}}$ and $\Phi(\cdot)$ is the CDF of standard normal distribution $N(0,1)$. We can prove that the production overtime cost expressed in (13) is a convex and increasing function in both $E[P_{jt}]$ and $Var(P_{jt})$ (see online companion Appendix 1 for proofs) by taking the first and second derivatives. This is not surprising since we expect that overtime increases more rapidly when the production capacity is tighter or the production variance is larger.

3.3.6 The Optimization Problem

We have expressed each cost component as a function of the lot size for each part (q_i) and the planned lead time for each work station (τ_j). We pose the optimization problem as follows:

$$\begin{aligned} \min_{q_i,\tau_j} \quad & \sum_{i=1}^{N}(Cost_i^{RM} + Cost_i^{FGI} + Cost_i^{WIP}) + \sum_{j=1}^{M} Cost_j^{OT} \\ s.t. \quad & \underline{q_i} \le q_i \le \bar{q_i} \qquad \forall i \\ & 1/m \le \tau_j \le \bar{\tau_j} \qquad \forall j \end{aligned} \tag{14}$$

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 and may represent management considerations and/or physical limitations in the factory. 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 might be set to two per day 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. Thus we use the optimal solution for this problem as a starting point to find integral solutions. Our numerical tests confirm that rounding often works very well, partly due to the fact that the objective function is relatively smooth around the optimal solution point.

3.4 *Analysis of the Cost Components*

First, we observe from (5), (6) and (7) that the inventory-related cost $Cost^{RM}$ increases concavely in each q_j; $Cost^{FGI}$ increases linearly in each q_i and concavely in each τ_j; $Cost^{WIP}$ 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 $Cost^{OT}$. For each lot size q_i, we show that $Cost_j^{OT}$ is convex in both $E[P_j]$ and $Var(P_j)$ in the online companion Appendix 1. Furthermore, since $E[P_j]$ and $Var(P_j)$ are convex in each q_i, by composition of convexity, we know $Cost_j^{OT}$ is also convex in each q_i. For each planned lead time τ_j, we find that $Cost_j^{OT}$ is increasing in $Var(P_j)$ and $Var(P_j)$ is decreasing in each τ_j (see online companion Appendix 1 and 2 for proofs). Thus, $Cost_j^{OT}$ 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 the one hand, increasing q_i leads to lower expected workloads on the set of work stations $M(i)$ that will reduce the overtime costs on them. On the other hand, increasing q_i increases the production variability, which increases overtime cost. We thus expect that larger lot sizes are applicable mainly for the parts that heavily use the bottleneck work stations.

Table 1. The Relationship between Cost Components and Decision Variables

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

We finally observe that in setting the planned lead time τ_j, there is a trade-off between inventory-related costs and the production overtime cost. We can reduce overtime on work station j by increasing τ_j; but increasing τ_j requires holding more finished goods safety stock and more WIP.

The relationships between the cost components and the decision variables are summarized in Table 1. The sign in the table indicates the sign of the correlation with the decision variable.

3.5 *Comparison with Queueing-based Models*

As we discussed in the Literature Review, there are several papers (e.g. Karmarkar 1987, Zipkin 1986, Tielemans and Kuik 1996) that use queuing models to examine lot-sizing decisions. The classical properties of a queueing system are then invoked for characterizing the impacts due to lot size decisions: the expected waiting time increases non-linearly with the utilization level, and increases with the variability in arrival and service processes. In particular, when considering the choice of lot size, the key insights are that there is an interesting trade-off between the queueing effects due to utilization and the effects due to variability. Reducing the lot size increases the time spent on setups, which increases the utilization level resulting in more waiting. But reducing the lot size can result in a less variable arrival stream and/or a shorter service time, both of which reduce waiting.

Our model captures the same phenomena: reducing the lot sizes increases the workload at each station but reduces both the lot processing time and the variability of the arrivals. In addition, we assume that there is some degree of production flexibility but at a cost, as is often the case in practice. A shop can extend its work-hours through (say) overtime, as one additional way to handle variability in the workload across the shop. To capture this flexibility, we use the planned lead time as a planning parameter for each work station, in addition to the lot size decisions for each part. Unlike the queueing-based models, we permit another control lever, namely the choice of the planned lead time at each work station. Given the workload and the arrival process to a work station, the planned lead time then determines the level of production variability, which determines the production cost; in our context, this was expressed as an overtime cost. The planned lead times at the work stations then combine to give the planned lead time for each part, which dictates the WIP and FGI in a make-to-stock setting. In Yuan (2013), these trade-offs are illustrated by some simple hypothetical cases.

In contrast, the queueing-based models generally take the lead times for the parts as outputs from the queueing model, as they do not permit adjustment or control of the production rate. Thus they do not explicitly capture this trade-off between production and inventory-related costs.

To compare the different approaches, one would either need to remove production flexibility from the model in this paper, or add it to the queueing-based models. For the former, this could be done by increasing the overtime cost so that it is prohibitively expensive. We expect the current model would not fare well in comparison to the queueing-based models, as the model's solutions would become overly conservative so as to minimize the probability of any overtime; in effect, the solution would set the planned lead times to be very long. On the other hand, it might be possible to add production flexibility to the queueing-based models; for instance, Agnew, C. E. (1976) gives a single-station model that captures both production flexibility and queuing effects. Nevertheless, we expect additional research is needed to determine how to extend this or other approaches to model a network of stations. We leave this for future research.

4. Implementation and Numerical Tests

We implemented the optimization model (14) in the Microsoft Excel spreadsheet and solved it using the Premium Excel built-in Solver. This provides great convenience for practical use. In this section, we report a numerical example with representative factory data that has been modified to protect proprietary information.

We note that due to the non-convexity of the problem, the performance of the Excel Solver will depend on the starting point, i.e., the initial values for the decision variables. That being said, we tested the solver from twenty feasible random starting points in the solution space and we always obtained the same solution. We conjecture that this is because the objective function is actually well behaved around the optimal point given the scale of our problem. Also, as we understand from the analysis, although convergence to a global optimum cannot be guaranteed, the objective cost function appears to be reasonably smooth. The similar behavior of the output is also observed by Kang et al. (2014) where they found non-convexity of the cost function abates considerably as the WIP level and lot size increase. Anli et al. (2007) and Albey et al. (2014) provide additional evidence supporting this observation.

The job shop we considered consists of 133 major 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%. Setup times on a few work stations are long. The per-unit processing times of a part on a work station varies from 5 to 110 minutes.

We show an initial scenario and determine its optimal solution from our model. We set the initial lot sizes to 4 units per lot for all parts except a few. Given that the assumption of Poisson lot arrivals tends to overestimate the variability of lot arrivals if the arrival rate is more than 3 lots per day (A5), we constrain the lot size choice so that there can be at most 3 lot arrivals per day for each part. So for the parts that have more than 3 lot arrivals per day with 4 units per lot, we increase the lot size (effectively combine the lots) until the number of lots is less than or equal to 3 lots in a day. We also set the planned lead times to 1 day for all work stations. We refer to these settings as the initial policy. We believe that this set of initial values is a reasonable representation of the current operating policy applied by our industry partner. We use the model to compute and record the performance of the system under the initial policy. We then solve the optimization model and determine the optimal policy for the problem. The performance improvement relative to the initial policy is shown in Figures 1 to 5. In particular, we show the daily cost breakdown of both the initial and optimal policies in Figure 1, the lot sizes for the 13 parts with the highest demand, namely with demand rates of more than 150 units per month, in Figure 2 (parts are sorted by initial lot sizes), the planned lead times for only the 23 not-lightly loaded work stations in Figure 3 (work stations are sorted by optimal planned lead times), the average workload (utilization) for these work stations in Figure 4 and the expected overtime on these work stations in Figure 5.

In the initial policy, the raw material inventory cost accounts for a large portion of the total manufacturing cost, mainly because of the long procurement lead time (e.g. 5 months) of the parts. The overtime cost for the initial policy is also a significant portion of the total costs. However, under the optimal policy, the production overtime cost is very low.

A closer look at the loads and expected overtime on the work stations (Figure 4 and 5) shows that we have greatly reduced the utilization and production overtime on several bottleneck work stations which leads to a large cost reduction. In the optimal policy, the 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. We find that it is beneficial to extend the planned lead times on the most heavily-loaded work stations to achieve less variable production. We also observe that 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 times. There do exist a few exceptions (part #2, #5 and #9) which warrant larger lot sizes due to having larger setup times. More numerical

results and discussion are reported in Yuan (2013).

5. Conclusions

In this research, we build a tactical model to minimize the inventory-related costs and production overtime costs and to determine the production lot size for each part and the planned lead time for each work station. Our model provides a manager with a simple-to-use tool for testing different operating policies and for gaining intuition on how shop performance depends on both lot sizing and planned lead time decisions.

We discuss a few possible research opportunities. First, one might try to find useful structural properties that might provide managerial guidance. For example, Karmarkar (1987) found conditions for which the optimal lot sizes and the total cost depend linearly on the setup time when the ratio of WIP and FGI holding costs is half. We expect that it may be possible with our model to find similar relationships across the setup times, the station utilization levels, and the ratio between different cost coefficients.

Secondly, in our model we assume that we are given a service level for the finished parts inventory. In many manufacturing systems, one might also consider how to set this service level for each part to the next stage in the supply chain. The backlog cost is then an important component of the cost function and will also depend on the production variance in the system. Papers by Zipkin (1986), Vaughan (2006), and Kang et al. (2014) are good references that incorporate the backlog cost in the optimization.

Thirdly, we assume Poisson lot arrivals at each work station. It would be important to examine how this assumption might be relaxed and with what consequences, as this assumption will not be reasonable for many settings. We refer the readers to Tielemans and Kuik (1996) and Lambrecht and Vandaele (1996) for more discussions about generalizing this assumption.

Lastly, in our model we do not explicitly consider the dependency across work stations, due to the output (production) from one station being the input (arrivals) to another station. Furthermore, the production process at a work station is auto-correlated, and thus so will be each downstream arrival process. In Graves (1986) this interdependency among work stations is explicitly modeled as a network of queues. It could be useful to see how to incorporate this interdependency into the optimization problem developed in the current paper, as a way to improve the accuracy and applicability of the model.

References

Adshead, N. S., and 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.

Agnew, C. E. 1976. "Dynamic modeling and control of congestion-prone systems." *Operations Research* 24(3): 400–419.

Albey, E., Bilge, Ü., and Uzsoy, R. (2014). "An exploratory study of disaggregated clearing functions for production systems with multiple products." *International Journal of Production Research* 52(18): 5301–5322.

Anli, O. M., Caramanis, M. C., and Paschalidis, I. C. 2007. "Tractable supply chain production planning, modeling nonlinear lead time and quality of service constraints." *Journal of Manufacturing Systems* 26(2): 116–134.

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

Fine, C. H., and 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., and 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., and 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., and 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* 52 (6): 590–605.
Kang, Y., Albey, E., Hwang, S., and Uzsoy, R. 2014. "The impact of lot-sizing in multiple product environments with congestion." *Journal of Manufacturing Systems*.
Karmarkar, U. S., Kekre, S., Kekre, S., and Freeman, S. 1985a. "Lot-sizing and lead-time performance in a manufacturing cell." *Interfaces* 15 (2): 1–9.
Karmarkar, U. S., Kekre, S., and 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.
Lambrecht, M. R., and 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., and Kanezashi, M. 1996. "Setting planned lead times for multi-operation jobs." *European journal of operational research* 88 (2): 287–303.
Matsuura, H., and Tsubone, H. 1993. "Setting planned leadtimes in capacity requirements planning." *Journal of the Operational Research Society* 809–816.
Selcuk, B., Fransoo, J. C., and De Kok, A. G. 2008. "Work-in-process clearing in supply chain operations planning." *IIE Transactions* 40(3): 206-220.
Silver, E. A., Pyke, D. F.,and Peterson, R. 1998. *Inventory Management and Production Planning and Scheduling*. New York: Wiley.
Söhner, V., and Schneeweiss, C. 1995. "Hierarchically integrated lot size optimization." *European Journal of Operational Research* 86(1): 73-90.
Teo, C. C., Bhatnagar, R., and 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., and Graves, S. C. 2012. "An Application of Master Schedule Smoothing and Planned Lead Time Control." *Production and Operations Management.* 21(2): 211–223.
Tielemans, P. F., and Kuik, R. 1996. "An exploration of models that minimize leadtime through batching of arrived orders." *European Journal of Operational Research* 95(2): 374-389.
Vaughan, T. S. 2006. "Lot size effects on process lead time, lead time demand, and safety stock." *International Journal of Production Economics* 100(1): 1-9.
Yano, C. A. 1987. "Setting planned leadtimes in serial production systems with tardiness costs." *Management Science* 33 (1): 95–106.
Yuan, Rong 2013. "Setting Optimal Production Lot Sizes and Planned Lead Times in a Job Shop System." Master Thesis, Computation for Design and Optimization, Massachusetts Institute of Technology.
Zipkin, P. H. 1986. "Models for design and control of stochastic, multi-item batch production systems." *Operations Research* 34 (1): 91–104.

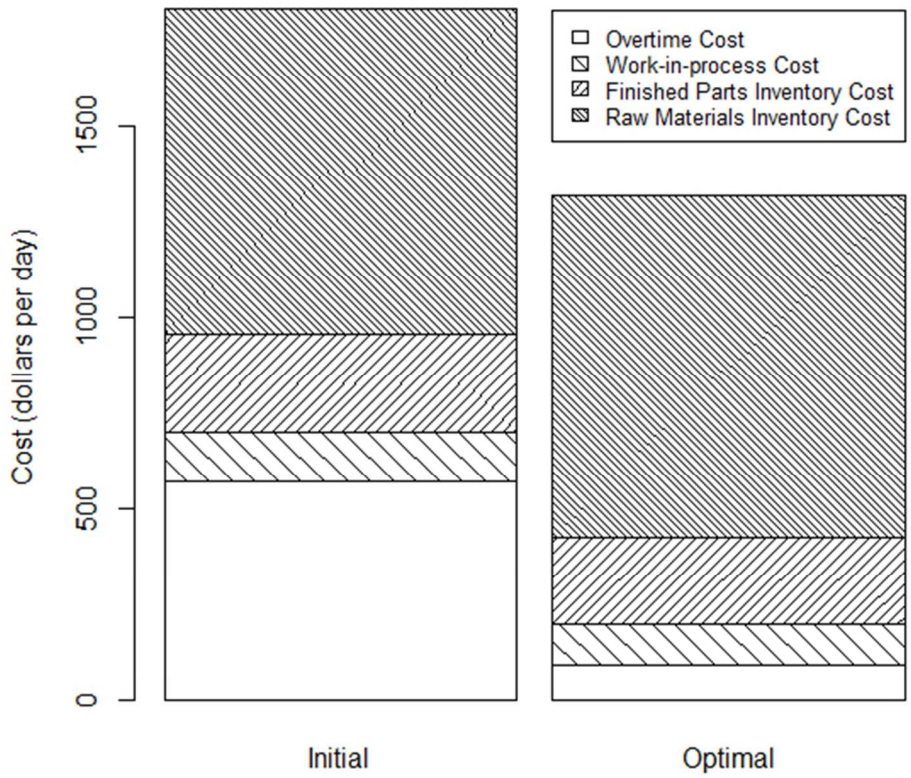

Figure 1. Initial and Optimal Cost Breakdown
208x207mm (72 x 72 DPI)

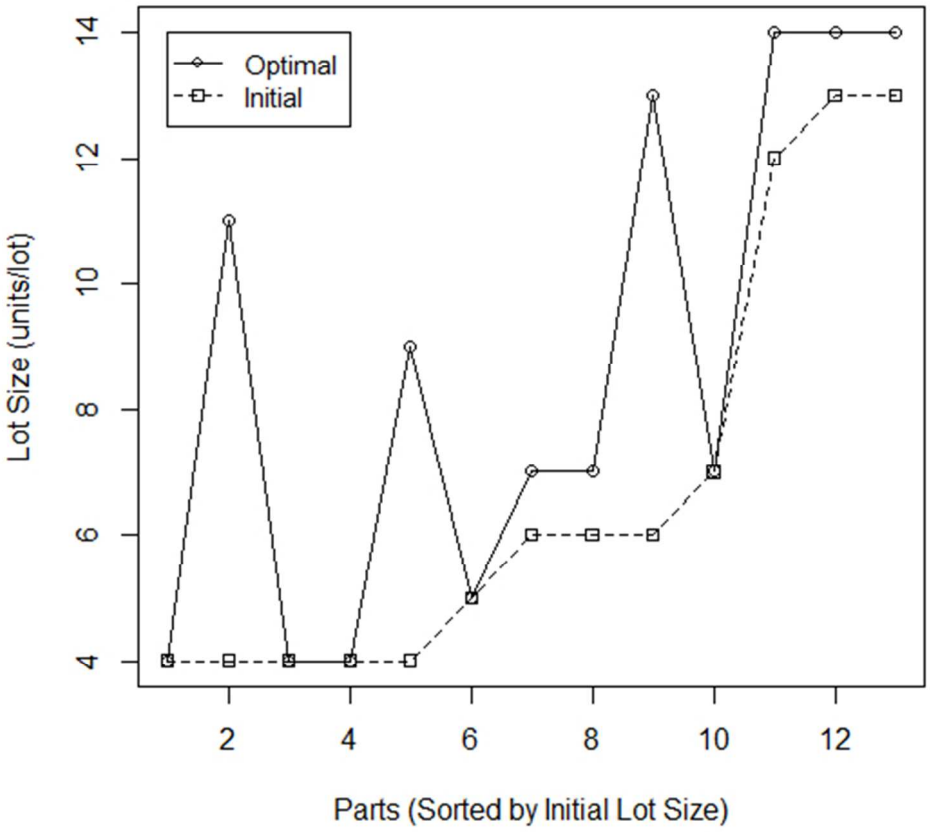

Figure 2. Initial and Optimal Lot-size Decisions for the High Demand Parts
195x194mm (72 x 72 DPI)

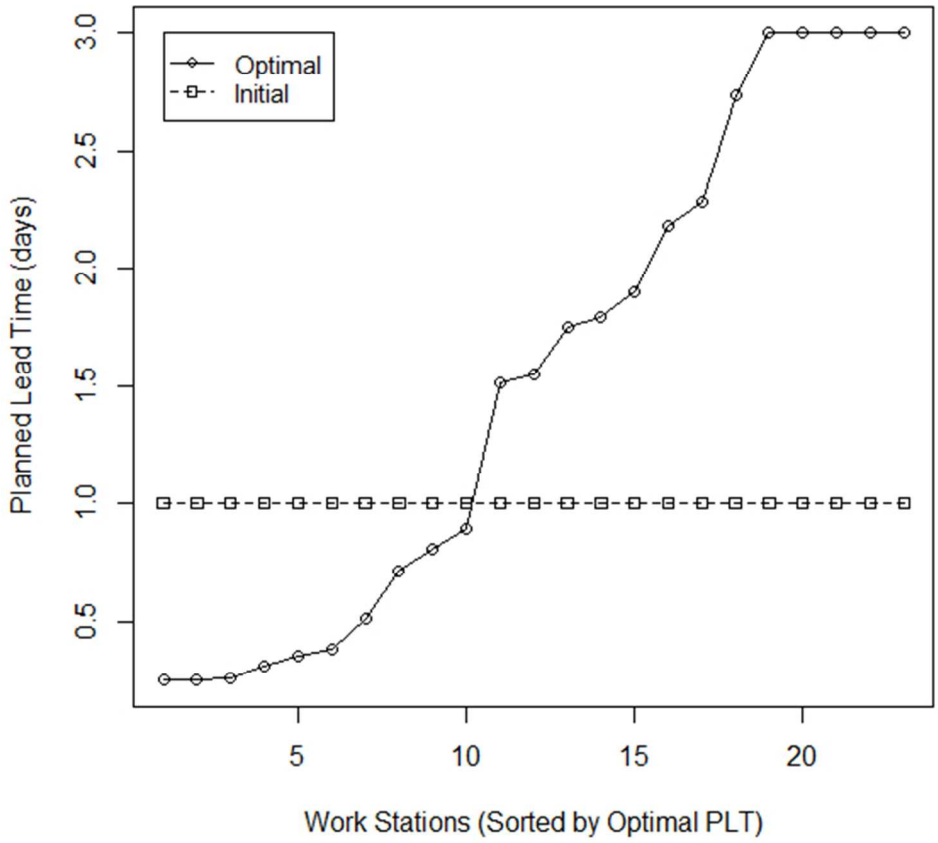

Figure 3. Initial and Optimal Planned Lead Times on the 23 Not-lightly Loaded Work Stations
208x207mm (72 x 72 DPI)

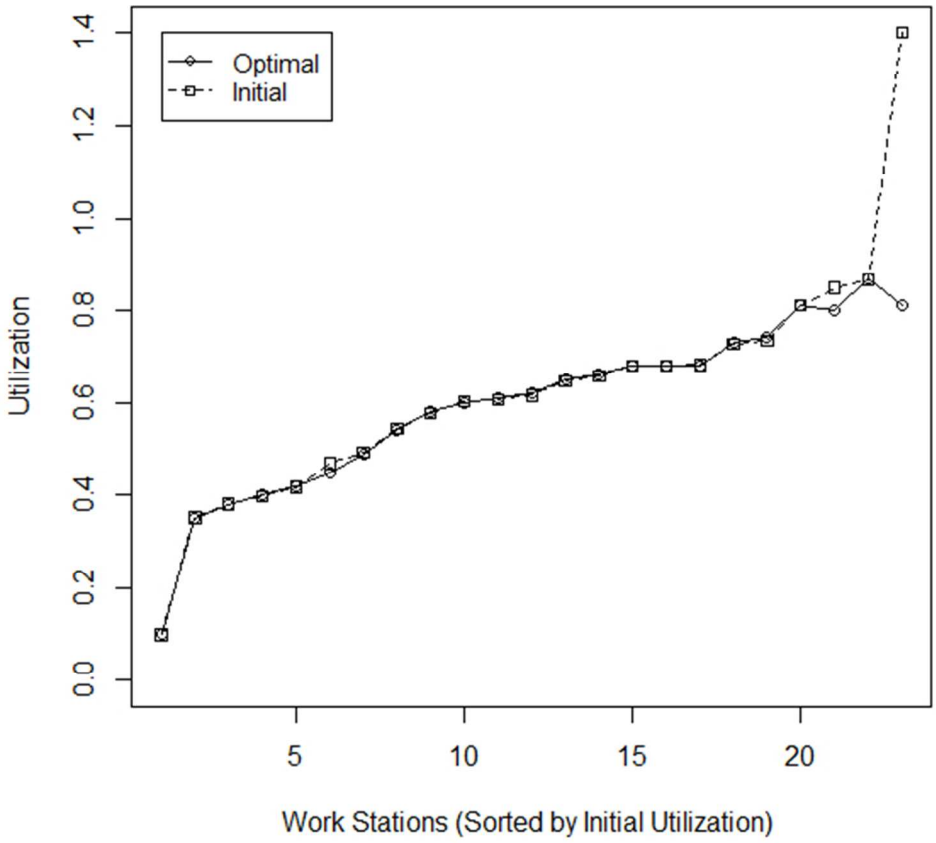

Figure 4. Initial and Optimal Average Workloads on the 23 Not-lightly Loaded Work Stations
208x207mm (72 x 72 DPI)

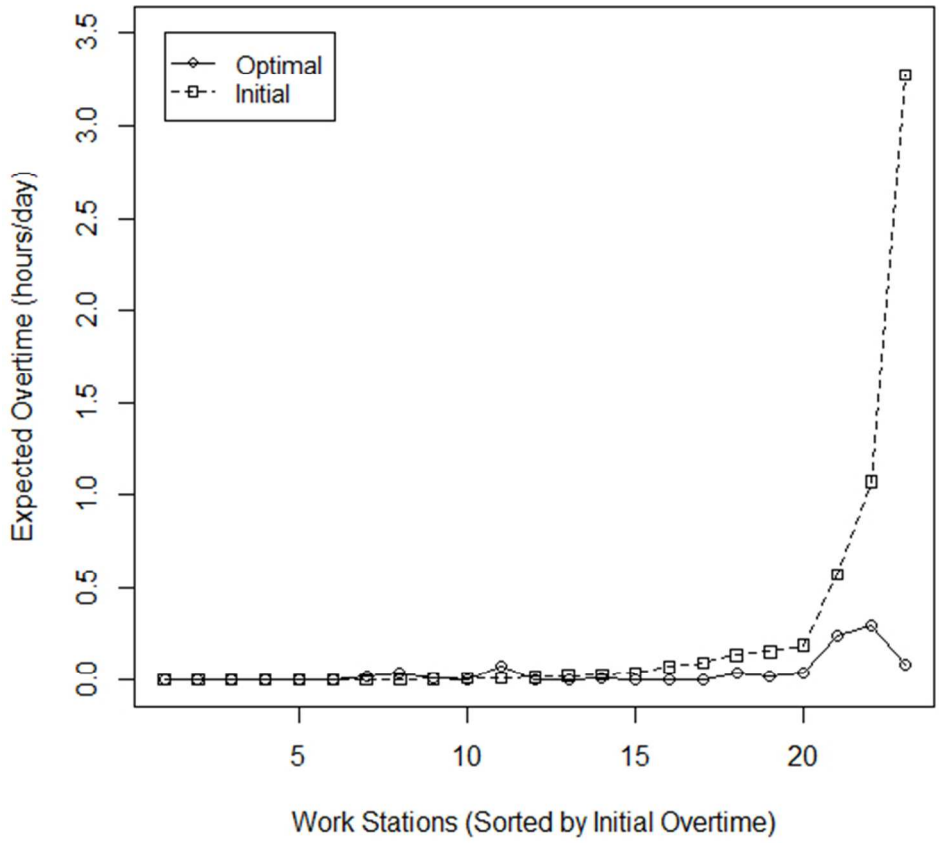

Figure 5. Initial and Optimal Expected Overtime on the 23 Not-lightly Loaded Work Stations
208x207mm (72 x 72 DPI)

Appendix 1: Production Overtime Function and Its Properties

From equation (13), we know the production overtime cost in work station j can be expressed as the partial normal function where workload is assumed to be normally distributed with mean $E[P_{tj}]$ and variance $Var(P_{tj})$. In this appendix, we 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.

Proof: To simplify notation, given the nominal working capacity C, we express the production overtime cost $Cost^{OT}$ as a function of the expected production workload ν and the standard deviation θ:

$$Cost^{OT}(\nu,\theta) = \int_C^{\infty} (x-C)\frac{1}{\sqrt{2\pi\theta^2}} e^{-\frac{(x-\nu)^2}{2\theta^2}} dx$$

Take derivative in terms of ν, we have

$$\frac{\partial Cost^{OT}(\nu,\theta)}{\partial \nu} = \int_C^{\infty} (x-C)\frac{1}{\sqrt{2\pi\theta^2}} \frac{(x-\nu)}{\theta^2} e^{-\frac{(x-\nu)^2}{2\theta^2}} dx$$

By substituting $y = \frac{x-\nu}{\theta}, \rho = \frac{C-\nu}{\theta}$, we have

$$\begin{aligned}
\frac{\partial Cost^{OT}(\nu,\theta)}{\partial \nu} &= \int_{\rho}^{\infty} (y-\rho)\frac{y}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy \\
&= \int_{\rho}^{\infty} \frac{y^2}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy - \rho \int_{\rho}^{\infty} \frac{y}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy \\
&= \int_{\rho}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy \\
&= 1 - \phi(\rho) > 0
\end{aligned}$$

Take the second derivative in terms of ν, we have

$$\begin{aligned}
\frac{\partial^2 Cost^{OT}(\nu,\theta)}{\partial \nu^2} &= \frac{\partial(\int_{\rho(\nu)}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy)}{\partial \nu} \\
&= -\rho'(\nu)\frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \\
&= \frac{1}{\sqrt{2\pi}\theta} e^{-\frac{y^2}{2}} > 0
\end{aligned}$$

Since both the first and the second derivatives of $Cost^{OT}(\nu,\theta)$ are positive in terms of the expected workload ν, we proved the production overtime cost function is increasing and convex in the expected workload ν. ■

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

Proof: We take the first and the second derivatives of $Cost^{OT}(\nu,\theta)$ in terms of θ:

$$\frac{\partial Cost^{OT}(\nu,\theta)}{\partial\theta} = -\int_C^{\infty}\frac{1}{\theta^2}(x-C)\frac{1}{\sqrt{2\pi}}e^{-\frac{(x-\nu)^2}{2\theta^2}}dp + \int_C^{\infty}\frac{(x-\nu)^2}{\theta^3}(x-C)\frac{1}{\sqrt{2\pi\theta^2}}e^{-\frac{(x-\nu)^2}{2\theta^2}}dx$$

By substituting $y=\frac{x-\nu}{\theta}, \rho=\frac{C-\nu}{\theta}$, we have

$$\begin{aligned}\frac{\partial Cost^{OT}(\nu,\theta)}{\partial\theta} &= -\int_{\rho}^{\infty}(y-\rho)\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}dy + \int_{\rho}^{\infty}y^2(y-\rho)\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}dy \\ &= \int_{\rho}^{\infty}y^3\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}dy - \rho\int_{\rho}^{\infty}y^2\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}dy - \int_{\rho}^{\infty}y\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}dy + \rho\int_{\rho}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}dy \\ &= \frac{1}{\sqrt{2\pi}}e^{-\frac{\rho^2}{2}} > 0\end{aligned}$$

Take the second derivative in terms of θ, we have

$$\begin{aligned}\frac{\partial^2 Cost^{OT}(\nu,\theta)}{\partial\theta^2} &= \frac{\partial(\frac{1}{\sqrt{2\pi}}e^{-\frac{\rho(\nu)^2}{2}})}{\partial\theta} \\ &= \frac{\rho^2}{\sigma\sqrt{2\pi}}e^{-\frac{\rho^2}{2}} > 0\end{aligned}$$

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

Appendix 2: Property of the Production Variance Function

We showed that the production variance on work station j can be expressed 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 makes intuitive sense since longer planned lead time is intended to provide more production smoothness.

Proposition 3 : Production variance is strictly decreasing in planned waiting time τ_j

We omit the subscript for work station for simplicity and rewrite the variance of the production rate as follows:

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

where $\beta = 1-(1-\frac{\alpha}{m})^m$; $\gamma = 1-\frac{1-\frac{\alpha}{m}}{\alpha}\beta$; $\alpha = 1/\tau$. 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 $(\frac{\beta}{2-\beta}(1-\gamma)^2+\gamma^2)$.

Our objective is to show the derivative of this term is non-positive in τ, i.e.

$$\begin{aligned}(\frac{\beta}{2-\beta}(1-\gamma)^2+\gamma^2)' &= (\frac{\beta}{2-\beta})'(1-\gamma)^2+(\frac{\beta}{2-\beta})((1-\gamma)^2)'+(\gamma^2)' \\ &= \frac{2\beta'}{(2-\beta)^2}(1-\gamma)^2-2(\frac{\beta}{(2-\beta)})(1-\gamma)\gamma'+2\gamma\gamma' \end{aligned} \tag{A1}$$

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 $(A1)$ in terms of τ:

$$\begin{aligned}\beta &= 1-(1-\frac{1}{\tau m})^m \\ \gamma &= 1-\tau\beta(1-\frac{1}{\tau m}) \\ \beta' &= -\frac{1}{\tau^2}(1-\frac{1}{\tau m})^{m-1} \\ \gamma' &= -\beta-\tau\beta'(1-\frac{1}{\tau m}) \\ (\frac{\beta}{2-\beta})' &= \frac{2\beta'}{(2-\beta)^2} \\ ((1-\gamma)^2)' &= -2(1-\gamma)\gamma' \\ (\gamma^2)' &= 2\gamma\gamma'\end{aligned}$$

We further observe the first term $\frac{2\beta'}{(2-\beta)^2}(1-\gamma)^2$ in equation (A1) is always non-positive since $\beta' \leq 0$. We can then focus on the second and the third term:

$$(\frac{\beta}{2-\beta})((1-\gamma)^2)'+(\gamma^2)' = \frac{2\gamma'}{2-\beta}(2\gamma-\beta) \tag{A2}$$

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

Consider γ'', we have

$$\begin{aligned}\gamma'' &= (-\beta-\tau\beta'(1-\frac{1}{\tau m}))' \\ &= (-1+(1-\frac{1}{\tau m})^m+\frac{1}{\tau}(1-\frac{1}{\tau m})^m)' \\ &= \frac{1}{\tau^2}(1-\frac{1}{\tau m})^{m-1}-\frac{1}{\tau^2}(1-\frac{1}{\tau m})^m+\frac{1}{\tau^3}(1-\frac{1}{\tau m})^{m-1} \\ &= \frac{1}{\tau^3 m}(1-\frac{1}{\tau m})^{m-1}+\frac{1}{\tau^3}(1-\frac{1}{\tau m})^{m-1} \geq 0\end{aligned}$$

Furthermore, since $\gamma|_{\tau=\frac{1}{m}} = 1$ and $\gamma|_{\tau=\infty} = 0$, we conclude that γ' is an non-decreasing function on $[-1,0]$ when $\tau \in [\frac{1}{m},\infty)$ and thus $\gamma' \leq 0$.

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

$$\begin{aligned}(2\gamma-\beta)'' = (2\gamma'-\beta')' &= (-2\beta-2\tau\beta'(1-\frac{1}{\tau m})-\beta')' \\ &= \frac{2}{\tau^2}(1-\frac{1}{\tau m})^{m-1}-\frac{2}{\tau^2}(1-\frac{1}{\tau m})^{m}+\frac{2}{\tau^3}(1-\frac{1}{\tau m})^{m-1} \\ &-\frac{2}{\tau^3}(1-\frac{1}{\tau m})^{m-1}+\frac{1}{\tau^4}(1-\frac{1}{m})(1-\frac{1}{\tau m})^{m-2} \\ &= \frac{2}{\tau^3 m}(1-\frac{1}{\tau m})^{m}+\frac{1}{\tau^4}(1-\frac{1}{m})(1-\frac{1}{\tau m})^{m-2} \\ &\geq 0\end{aligned}$$

Observe that $(2\gamma-\beta)'|_{\tau=\frac{1}{m}}=-2$ and $(2\gamma-\beta)'|_{\tau=\infty}=0$, so $(2\gamma-\beta)'$ is an non-decreasing function on $[-2,0]$ when $\tau\in[\frac{1}{m},\infty)$ and thus $(2\gamma-\beta)'\leq 0$. Furthermore, since $(2\gamma-\beta)|_{\tau=\frac{1}{m}}=1$ and $(2\gamma-\beta)|_{\tau=\infty}=0$ (this equality applies the identity $\lim_{\tau\to\infty}\tau(1-(1-\frac{1}{\tau m})^m)=1$ which can be shown by LHopitals rule.), $(2\gamma-\beta)$ is non-increasing on $[1,0]$ when $\tau\in[\frac{1}{m},\infty)$ and $(2\gamma-\beta)\geq 0$.

As we show $\gamma'\leq 0$ and $(2\gamma-\beta)\geq 0$ when $\tau\in[\frac{1}{m},\infty)$, we know expression (A2) is always non-positive and consequently the derivative of the production overtime cost function (A1) is non-positive, which implies that the production cost is non-increasing in the planned lead time τ. ■