Proceedings of the 2015 Winter Simulation Conference
L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, eds.

LEAD TIME MODELING IN PRODUCTION PLANNING

Erinc Albey
Reha Uzsoy

Edward P. Fitts Department of Industrial and Systems Engineering
College of Engineering
North Carolina State University
Raleigh, NC 27695-7906, USA

ABSTRACT

We use two mathematical models to represent the dependency between workload releases and lead times: a linear programming model with fractional lead times (FLT) and a clearing function (CF) based nonlinear model. In an attempt to obtain a reference solution, a gradient based simulation optimization procedure (SOP) is used to determine the lead times that, when used in the FLT model, yield the best performance. Results indicate that both FLT and CF models perform well, with CF approach performing slightly better at very high workload scenarios. The SOP is able to improve upon the performance of both models across all experimental conditions, suggesting that FLT and CF models are limited in representing the lead time dynamics. All three models yield quite different lead time patterns at critical machines, suggesting the need for further study of the behavior of these models.

1 INTRODUCTION

In most optimization models for production planning (Johnson and Montgomery 1974; Voss and Woodruff 2003) the capability of the shop floor is represented as the "capacity" of the production system, expressed as the total available time of resources. However, capacity is, in fact, is a very complex concept (Elmaghraby 2010) which is manifested in lead times, the delay between work being released into the system and its emergence as finished product. Accurate modeling of lead times is essential to optimally coordinate the release of new work into the plant. However, lead times are, in fact, strongly dependent on the release decisions that determine the evolution of the system workload over time (Missbauer 2002; Pahl et al. 2005; Pahl et al. 2007). As shown by queuing models, system performance measures, especially lead times, start deteriorating well before resource utilization reaches 100%, and are influenced by both the mean and variance of the service and arrival processes (Hopp and Spearman 2008). This mutual dependency between lead times and releases is known as the planning circularity.

The planning circularity has been addressed in the literature using different techniques. The most common approaches, which will be discussed in Section 2, are the use of nonlinear clearing functions (CF) and iterative multi-model approaches combining linear programming (LP) and simulation. In this paper we use a fractional lead time based LP model (the FLT model); and a clearing function based model (the CF model) to obtain production/release plans taking the planning circularity into account. The FLT model is adapted from the work of Kacar et al. (2014) and the CF model from Asmundsson et al. (2009). We also use a simulation optimization procedure (SOP), similar to that of Kacar and Uzsoy (2015), that considers a more general fractional lead time model and searches for the set of planned lead times that yield the highest profit. Although our SOP does not guarantee global optimality, its solution is used as a

benchmark to investigate the solution quality of the two models and the lead time behaviour observed in their respective optimal solutions.

The remainder of the paper is organized as follows. Section 2 briefly reviews the literature. Section 3 presents the production planning models and SOP. Section 4 presents the numerical experiments, while conclusions and directions for future research are presented in Section 5.

2 LITERATURE REVIEW

There have been two basic approaches to the planning circularity in the literature to date. Clearing functions (CF) explicitly represent the nonlinear dependency between lead times and workload as a nonlinear function that can be incorporated into optimization models for production planning (Missbauer and Uzsoy 2010). Some authors (Srinivasan et al. 1988; Karmarkar 1989; Selçuk et al. 2007) derive CFs analytically for simple systems, while others such as Asmundsson et al. (2009), consider a complex production system and use off-line simulation to estimate the parameters of the CF. Albey et al. (2014) propose a family of product-based multi-dimensional CFs that use the disaggregated WIP levels for individual products as state variables. CFs that consider lot sizing decisions are introduced by Kang et al. (2014).

In iterative approaches, production/release planning is achieved by iterations between a LP model and a simulation model. The estimates of the capacity related parameters used in the LP model are updated at each iteration until convergence (Hung and Leachman 1996; Byrne and Bakir 1999; Kim and Kim 2001; Byrne and Hossain 2005; Bang and Kim 2010). However, the convergence behavior of these techniques is often inconsistent and not well understood (Irdem et al. 2010; Albey et al. 2014). In addition, computational experiments have shown that appropriately fitted CF models yield superior production plans to those obtained by these iterative approaches (Kacar et al. 2012),

Tekin and Sabuncuoglu (2004), Fu (2002) and Zapata et al. (2010) present comprehensive surveys of simulation optimization techniques. They classify the existing techniques according to problem characteristics (global vs. local optimization), objective functions (single or multiple objectives) and parameter spaces (discrete or continuous parameters). Continuous parameter space methods, which are relevant to this study, are further categorized into response surface methodology, gradient-based methods and stochastic approximation methods. In this paper, we use a gradient based approach where the gradient is estimated using perturbation analysis (Ho et al. 1979). Our reasons for choosing perturbation analysis are: i) when applied properly, it is able to estimate all gradients from a *single* simulation run (Tekin and Sabuncuoglu 2004), hence is more efficient than other techniques; ii) it performs well for discrete-event dynamic systems that can be modeled as queueing networks, hence has been used for manufacturing systems (Donohue and Spearman 1993; Liberopoulos and Caramanis 1994; Yan et al. 1994).

3 LEAD TIME MODELING: PLANNING FRAMEWORK

The production system under investigation is assumed to process a set of products, each requiring a specified set of operations. Products that complete processing at a particular machine immediately become available to the next operation on their process routing. The production/release plan considers a predetermined number of periods and seeks to maximize the total net present value of the profit obtained in each planning period, which is given by the difference of the period revenue and the sum of backorder, finished good inventory (FGI) holding, work in process (WIP) holding and material (i.e. release) costs incurred in the period. The central decision variables are the amount of each product to be released into the production system in each period.

In our experiments, the two planning models are executed to obtain a release plan for the entire horizon, yielding the quantity of each product to be released to the shop floor in each period. These aggregate release quantities are disaggregated using the heuristic of Askin and Standridge (1993) and released uniformly over the period. The First Come First Served (FCFS) dispatching rule is used to select parts from machine queues. The disaggregated release plan is then simulated using the software

environment of Albey and Bilge (2011). The realized profit at the end of the simulation as well as the observed lead times are recorded for performance comparison.

To provide a benchmark, the FLT model is incorporated into a SOP that searches for the lead time values that allow this model to obtain the maximum expected profit. In this mode, the framework is operated as follows: Starting from an initial estimate of lead times for each period, the fractional lead time model presented in Section 3.2 is solved to generate a release plan for the initial set of parameters. The release plan is simulated and the resulting objective is recorded. The parameter vector is randomly perturbed, following the Simultaneous Perturbation Stochastic Approximation approach of Spall (1998) and a new parameter vector is obtained (the random perturbation approach is also used successfully in Kacar and Uzsoy (2015) for a similar setting to ours). A new iteration is triggered and the production planning framework is executed for the new set of parameters. The execution terminates once the predefined iteration limit is reached. Details of the SOP are given in Section 3.3.
The following sub-sections present the mathematical models and the SOP.

3.1 Fractional Lead Time (FLT) Model

The FLT model used in this study extends that of Kacar et al. (2014) by considering fractional lead times varying from period to period and maximizes discounted total profit. The notation is presented in Table 1.

Table 1: Notation.

Indices:	Sets:
t: $Period\ index, t = 1,2, \dots T$	$AllO(i)$: $Set\ of\ all\ operations\ of\ product\ i$
i: $Product\ index$	$AltM(o)$: $Set\ of\ alternative\ machines\ for\ operation\ o$
o, n : $Operation\ index$	$Opr(m)$: $Set\ of\ operations\ that\ machine\ m\ can\ process$
m, g : $Machine\ index$	$ImP(o)$: $Immediate\ predecessor\ of\ operation\ o$
L_i: $Terminal\ operation\ of\ product\ i$	
F_i: $First\ operation\ of\ product\ i$	
Decision Variables:	
R_{it}: $Release\ quantity\ of\ product\ i\ at\ the\ beginning\ of\ period\ t$	
I_{it}: $Finished\ goods\ inventory\ (FGI) of\ product\ i\ at\ the\ end\ of\ period\ t$	
B_{it}: $Backorder\ for\ product\ i\ at\ the\ end\ of\ period\ t$	
X_{omt}: $Number\ of\ operation\ o\ completed\ on\ machine\ m\ in\ period\ t$	
W_{ot}: $WIP\ amount\ of\ operation\ o\ at\ the\ end\ of\ period\ t$	
Parameters:	
φ_i: $Unit\ selling\ price\ of\ product\ i$	
ρ_i: $Unit\ material\ cost\ of\ product\ i$	
π_i: $Unit\ inventory\ holding\ cost\ of\ product\ i$	
β_i: $Unit\ backorder\ cost\ of\ product\ i$	
d_{it}: $Demand\ of\ product\ i\ at\ the\ end\ of\ period\ t$	
ε_{om}: $Unit\ processing\ time\ of\ operation\ o\ at\ machine\ m$	
C_t: $Machine\ capacity\ in\ period\ t\ (planning\ period\ length)$	
L_{ot}: $Estimated\ time\ elapsing\ from\ the\ release\ of\ the\ raw\ material\ of\ product\ i\ \ at\ period\ t\ to\ the\ completion\ of\ the\ o'th\ operation\ of\ product\ i$	
ϑ_{ot}: The fractional portion of the lead time L_{ot} from the beginning of the process to the start of operation o, $i.e. \vartheta_{ot} = L_{ot} - \lfloor L_{ot} \rfloor$	
FF_i: $estimated\ flow\ factor\ of\ product\ i, defined\ as\ the\ ratio\ of\ the\ average\ time\ required\ for\ material\ started\ into\ the\ process\ to\ become\ available\ as\ FGI$	
f: $Discount\ factor$	

FLT Model:

$$Max\ z = \sum_{\mathrm{t}} \frac{1}{(1+f)^t} \sum_{\mathrm{i}} \left\{ \varphi_i(d_{it} - B_{it} + B_{it-1}) - \left(\pi_i I_{it} + \rho_i R_{it} + \beta_i B_{it} + \sum_{o \in AllO(i)} \omega_o W_{ot} \right) \right\} \quad (1)$$

s.t.

$$I_{it-1} + \sum_{m \in AltM(L_i)} X_{L_i mt} + B_{it} - B_{it-1} - I_{it} = d_{it} \qquad \forall i,t \quad (2)$$

$$W_{ot} = \sum_{p=1}^{t} R_{ip} - \sum_{m \in AltM(o)} \sum_{p=1}^{t} X_{omp} \qquad \forall i,t,o \in F_i \quad (3)$$

$$W_{ot} = \sum_{n \in ImP(o)} \sum_{g \in AltM(n)} \sum_{p=1}^{t} X_{ngp} - \sum_{m \in AltM(o)} \sum_{p=1}^{t} X_{omp} \qquad \forall i,t,o \in AllO(i) \backslash F_i \quad (4)$$

$$X_{omt} = \vartheta_{ot} R_{i,t-\lfloor L_{ot} \rfloor} + (1 - \vartheta_{ot}) R_{i,t-\lfloor L_{ot} \rfloor} \qquad \forall i,t,o \in AllO(i),\ m \in AltM(o) \quad (5)$$

$$\sum_{o \in Opr(m)} \varepsilon_{om} X_{omt} \leq C_t \qquad \forall m,t \quad (6)$$

$$I_{it}, X_{omt}, B_{it}, R_{it}, W_{ot} \geq 0 \qquad \forall i,o,m,t$$

The FLT model maximizes the net present value of cash flows composed of revenue minus the sum of holding, material, backorder and work-in-process costs. Constraints (2) ensure finished inventory balance, where $\sum_{m \in AltM(L_i)} X_{L_i mt}$ represents the total completed amount of product *i* as the sum of the amounts completed at the final operation for each product over all machines that can process this operation. Constraints (3) are the WIP balance constraints for the first operation of product *i*. Similarly, constraints (4) ensure WIP balance for the remaining operations of product *i*. Constraints (5) estimate the output of each operation *o* considering the noninteger lead times. In the FLT model, we assume constant lead time estimates independent of planning period *t* throughout the planning horizon, i.e. $L_{ot} = L_o\ \forall t$. This assumption is based on the observation that in practice there is no clear way to predict future lead times unless a release plan of some form is developed first. We compute the L_o values for each operation as $L_o = \varepsilon_{om} FF_i + L_{ImP(o)}\ \forall i, o \in AllO(i) \backslash F_i$ and $L_{F_i} = \varepsilon_{om} FF_i\ \forall i$. Constraints (6) represent the machine capacity.

3.2 Clearing Function Model: CF

Our CF model is based on the nonlinear allocated CF model of Asmundsson et al. (2009) where the decision variables Z_{omt} are the fraction of the output of machine *m* allocated to operation *o* in period *t*:

CF Model:

$Max\ (1)$

s.t.

(2)-(4)

$$\varepsilon_o X_{omt} \leq Z_{omt} \left[a_m (1 - e^{-b_m (WIP_{omt}^{avg} / Z_{omt})}) \right] \qquad \forall i,t,o \in AllO(i), m \in AltM(o) \quad (7)$$

$$\sum_{o \in Opr(m)} Z_{omt} = 1 \qquad \forall\ m,t \quad (8)$$

$$Z_{omt} \in \{0,1\} \qquad \forall i,t,o \in AllO(i)$$

However, unlike the final model of Asmundsson et al. (2009), which is reduced to an LP by outer linearization of the CF, we use the CF in nonlinear form, which results in a nonconvex optimization model. We take this approach to eliminate the possible impact of alternative piecewise linearization approaches on the model.

Our CF model is analogous to FLT in terms of its objective function and flow balance equations. Parameters a_m and b_m are found using a fitting procedure similar to that of Albey et al. (2014). Constraint (8) ensures the total allocated capacity cannot exceed the capacity of the machine.

We solve the resulting nonlinear model with the KNITRO NLP solver Version 7.0.0, which is a local solver for general purpose nonlinear models (Byrd et al. 2006). KNITRO is able to converge in all cases, requiring CPU times that are similar in magnitude to those of the LP solver (CPLEX 11.2, with default options) used to solve FLT model.

3.3 Simulation Optimization Procedure: SOP

Our SOP utilizes the FLT model described in Section 3.2. However, the SOP does not assume constant lead time estimates independent of planning period t throughout the planning horizon. Instead, SOP seeks a set of values for the parameters L_{ot} for all *(o, t)* that maximize the objective function value of the FLT model in which they are inserted. Denoting the parameter vector of lead time estimates L_{ot} by μ and an individual simulation replication for a given μ by *r*, the SOP seeks to solve the maximization problem:

$$\max_{\mu} E_r[z(\mu, r)]. \tag{9}$$

In (9), $E_r[z(\mu, r)]$ denotes the expected value of the FLT objective over all replications *r*, for a given set of lead time parameters, μ.

The SOP searches the parameter set μ, which is a continuous set bounded by the interval [0, Δ], where Δ is an upper bound on the lead time whose choice affects the SOP run time. In other words, for every period pair *(o, t)*, the parameter L_{ot} is in the interval [0,Δ]. Since our SOP is searching a continuous parameter space, a gradient based approach with perturbation analysis is selected as mentioned in Section 2. The notation and the pseudocode for the SOP are given below.

Notation:
k: Iteration index
r: Replication index
ItLim: The iteration limit
Δ: Lead time upper bound.
a_k and c_k: The gain sequences at iteration *k*. They are updated at each iteration using the following relations: $a_k = a/(A + k)^{\alpha}$ and $c_k = c/(k)^{\gamma}$. The parameters in these recursive relations need to be selected carefully before SOP execution. In this work we set $\alpha = 0.602$ and $\gamma = 0.101$ following Spall (1998). The values of A, a and c are set to 1, 10^{-5}, and 1 respectively based on a set of preliminary runs.
σ_k: Random perturbation vector at iteration *k*. In generating perturbations a Bernoulli distribution with probability of 0.5 is used as recommended by Spall (1998). The outcome values for each component, σ_k^i, are selected as ±0.1.
∇z_k: Estimated gradient vector at iteration *k*. Each component ∇z_k^i of the gradient is updated at each iteration as $\nabla z_k^i = \frac{z(\mu_k + c_k\sigma_k) - z(\mu_k - c_k\sigma_k)}{2c_k\sigma_k^i}$.
μ_k: The parameter vector at iteration *k* consisting of lead time estimates, L_{ot}^k for all periods *t* in the planning horizon. μ_k is updated at each iteration following the recursive relation $\mu_{k+1} = \mu_k + a_k\nabla z_k$.
H: User defined tolerance parameter. As in Kacar and Uzsoy (2015), this parameter *H* allows the objective function value to decrease by some fraction of the previous iteration's objective value. If a

larger decrease is observed a new μ vector is obtained by applying a new independent random perturbation to the current μ vector.

The SOP can now be stated as follows:

Algorithm SOP:

Step 1: Initialization: Set k=1 and initialize lead time parameters.

Step 2: Construct μ_k, run FLT model and obtain release plan.
Simulate the plan for each replication r and obtain $z(\mu_k)$.

Step 3: Generate a perturbation vector σ_k.
Estimate the gradient as $\nabla z_k = \frac{z(\mu_k + c_k\sigma_k) - z(\mu_k - c_k\sigma_k)}{2c_k\sigma_k}$.

Step 4: Compute the new parameter vector: $\mu_{k+1} = \mu_k + a_k \nabla z_k$.
Compute the objective value, $z(\mu_{k+1})$.

Step 5: Check the tolerance:
If $\frac{z(\mu_{k+1})}{z(\mu_k)} \leq H$, then repeat the above steps starting from Step 3.
Otherwise go to step 6.

Step 6: If $k > ItLim$, then STOP and report the $z^* = \max_k z(\mu_k)$.
Otherwise set $k = k + 1$ and go to Step 7.

Step 7: Update gain sequences using $a_k = a/(A + k)^\alpha$ and $c_k = c/(k)^\gamma$, then go to Step 3.

The set of values for the parameters used in the SOP are summarized in the next section, which presents the experiments conducted to test the performance of SOP.

4 EXPERIMENTS

Experiments are conducted on a scaled down multi–stage multi-product wafer fabrication system previously studied by several authors (Kayton et al. 1997; Irdem et al. 2010; Kacar et al. 2012). As shown in Figure 1, the system used in these experiments is composed of 11 machines producing three products. The system reflects the major characteristics of semiconductor wafer fabrication, including a re-entrant bottleneck process, batching machines, and multiple products. Each row in the figure represents the routing of each product. The material flows from left to right, i.e. all products start the process at Station 1 and complete their processing at Station 10. Product 1 has 22 operations including 6 visits to machine 4, which is the bottleneck station for a system producing only products 1 and 2. Products 2 and 3 have 14 operations. Processing time parameters and batch sizes are listed in Table 2. All processing times follow a lognormal distribution and are given in minutes. The processing times for all operations on a given machine are the same, and we assume instantaneous material transfer between consecutive operations on a routing. We use backorder cost of 50, WIP holding cost of 35, release cost of 3, FGI holding cost of 15 for all products. Unit selling prices are set to 60, 90 and 120 for products 1, 2 and 3 respectively.

The base demand scenario, S1, assumes a 3:1:1 product mix. The mean and standard deviation of the demand are $\mu_i = \{60, 20, 20\}$ and $\sigma_i = \{6, 2, 2\}$ for Products 1, 2 and 3 respectively. For each product, a demand series for a 26 week (six months) period is used. The workload scenarios S2 and S3 are generated based on S1 by increasing the demand (10% and 20% for S2 and S3 respectively) of all products throughout the planning horizon while maintaining the 3:1:1 mix. The resulting overall expected system utilization are 89%, 98% and 107% for S1, S2 and S3. For each workload scenario, 5 different demand realizations are generated, hence in total 15 different workload scenarios are used in the experiments.

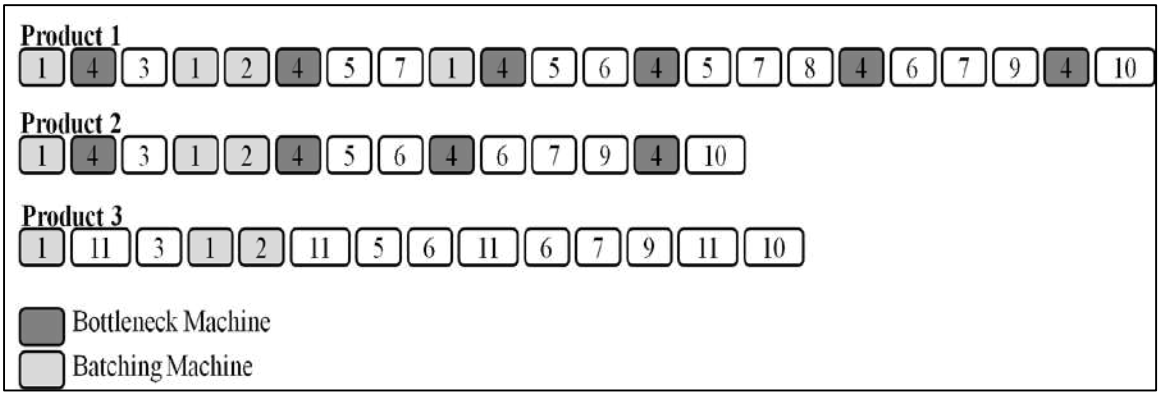

Figure 1: Product flow routes over the machines for the studied semiconductor fab.

Table 2: Processing time distributions and batch sizes.

Workcenter #	Mean	Std. Dev.	Batch Size
1	80	7	4
2	220	16	4
3	45	4	1
4	40	4	1
5	25	2	1
6	22	2.4	1
7	20	2	1
8	100	12	1
9	50	4	1
10	50	5	1
11	70	2.5	1

The execution of the FLT and CF models consist of a single run whereas the number of iterations for SOP is set to 250. In deciding the number of iterations, scenario S1 is used and the number of iterations is increased starting from 50 by increments of 50 until the performances of two consecutive executions are approximately equal. For the FLT and CF models, 10 simulation replications are performed. At each SOP iteration we use two perturbations and simulate each of these for ten replications. The average of 10 replications is used to estimate the objective function value for gradients. The CPU time for a single execution of the simulation varies between 5 and 10 seconds on a computer with 1.60 GHz Intel Core i5-2567M processor and 4GB RAM. The initial values of the backorder, FGI and WIP are taken as zero. The parameter values used in SOP are due to Kacar and Uzsoy (2015) and summarized in Table 3.

Table 3: SOP parameter values used in the experiments.

Δ	γ	α	A	a	c	H	$ItLim$
3	0.101	0.602	1	10^{-5}	1	0.99	250

Table 4 compares the realized profits from SOP, FLT and CF. The mean and the standard deviation of the profits obtained by 10 independent replications of the final release plans are presented for all models. The performances of CF and FLT are very close, with the average profit obtained by CF slightly higher than that of FLT. Similarly, for all cases, SOP has higher average profit values than the CF and FLT models. This is intuitive as SOP spends considerable time in searching the solution space to improve the initial solution found by simply using FLT model with time independent lead time estimates. It is

interesting to note the magnitude of the improvement between SOP and FLT. the result of SOP can be interpreted as representing the performance of FLT with perfect a priori knowledge of the realized lead times. The results in Table 4 indicate that both the CF and the FLT models approach quite close to the performance possible with perfect information. As pointed out by Kacar et al. (2014), the use of fractional lead times allows FLT to close the performance gap with the CF model. These results suggest that the ability of the CF model to capture the effects of workload upon lead times does not result in a major gain in expected profit, which requires further investigation to determine the causes.

Table 4: Comparison of realized average profits and standard deviations (x 10^3).

	SOP		CF		FLT	
	Mean	Stdev	Mean	Stdev	Mean	Stdev
S1	120.3	1.1	115.4	1.1	114.6	0.9
S2	114.5	0.8	109.5	0.5	107.5	0.8
S3	93.4	1.5	85.4	0.7	81.4	0.8

To determine whether there are statistically significant differences between the realized profit values of the methods, the Wilcoxon rank sum test (Wilcoxon 1945) which is a non-parametric alternative to the paired *t*-test when the populations cannot be assumed normally distributed or the data is on the ordinal scale, is used. The test results are presented in Table 5. The cases where the performances of the compared models are not statistically different indicated by "0"; and the cases where one of the models outperforms the other is shown using "+", where "+" indicates that the model shown in the row outperforms the model in the column. It is seen that CF and FLT are alike in performance, except in the very high workload scenario, S3. SOP always outperforms FLT, but for scenario S1, SOP and CF are indistinguishable.

Table 5: Wilcoxon rank sum test results.

Scenario		CF	FLT
S1	SOP	0	+
	CF		0
S2	SOP	+	+
	CF		0
S3	SOP	+	+
	CF		+

The small differences in performance between the different production planning models and the SOP suggest a closer investigation of the specific lead times resulting from each of the models. For this purpose, we analyzed the average lead times at machines 1, 4 and 11 in scenario S1 as these machines constitute a representative sample of batching, bottleneck and near-bottleneck stations. Figure 2a-c present the realized machine lead times in periods 1 through 23 as a time series for these machines for all three models. To test whether there is significant difference in realized lead times a set of paired t-tests are executed. Table 6 shows the average and standard deviation values of the paired differences, as well as the resulting p-value for each possible 2-way combination. At a significance level of $\alpha = 0.05$, the lead time estimates of SOP for machine 1 are significantly lower than those of CF and FLT (Figure 2a) and both CF and SOP have lower lead time estimates than FLT for machine 11 (Figure 2c). All three methods estimate the lead times for designed bottleneck machine 4 quite accurately. The realized lead times at machine 1 vary widely across the three approaches, and have quite different dynamics. It is striking that even though the FLT model assumes a constant lead time estimate across all periods, the realized lead times vary quite substantially as seen in Figure 2. To illustrate, the sample standard

deviations of mean lead time for FLT are 0.83, 0.20 and 0.19 for machines 1, 4 and 11 respectively; whereas the same values are for SOP are 0.88, 0.20 and 0.26.

Table 6: Paired t-test results.

	Machine 1			Machine 4			Machine 11		
	SOP-CF	SOP-FLT	CF-FLT	SOP-CF	SOP-FLT	CF-FLT	SOP-CF	SOP-FLT	CF-FLT
Avg.	0.95	1.22	0.99	0.22	0.21	0.25	0.29	0.29	0.33
St.dev.	0.83	0.88	0.64	0.15	0.18	0.19	0.21	0.23	0.21
p-value	**0.042**	**0.002**	0.129	0.760	0.483	0.719	0.632	**0.003**	**0.022**

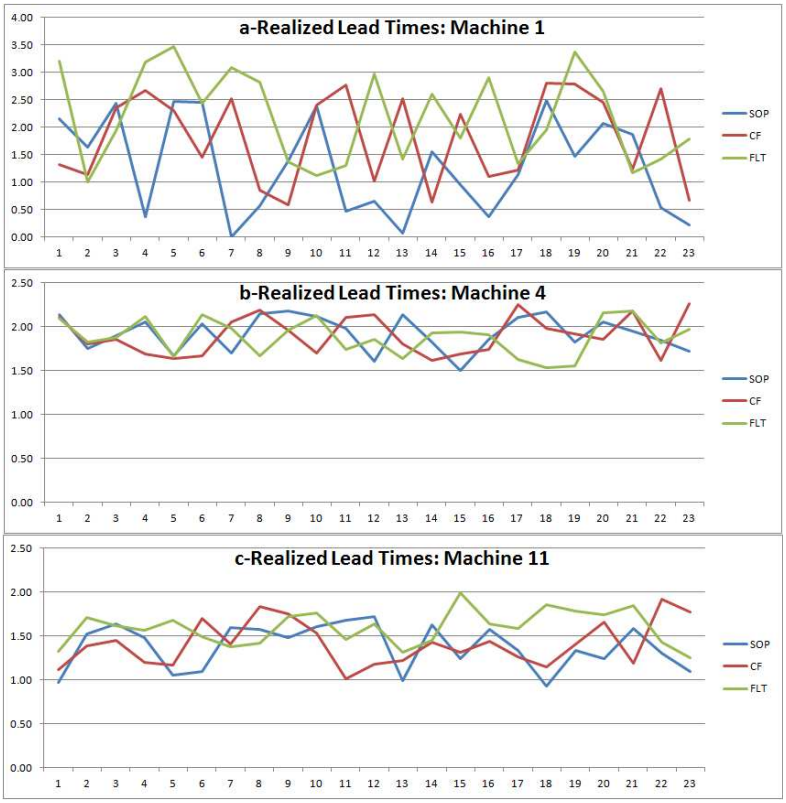

Figure 2: Realized lead times.

5 CONCLUSIONS AND FUTURE RESEARCH

We have developed a gradient based SOP for estimating the set of lead times that maximize expected profit when used in a LP model with dynamic fractional lead times. Although the developed approach requires high computational time, it allows us to estimate the performance of models with perfect a priori estimates of lead times. The results indicate that the use of fractional lead times, even when the same lead time is assumed across the entire horizon, results in almost equal performance to the CF model, and is quite close to the performance achievable with perfect lead time estimates. The causes of this behavior are not clear at present, and require further analysis of the lead time estimates obtained by each model in order to clarify the situation. A number of possible explanations arise. There may be many different release plans that lead to very similar realized lead times; it is well known that many production planning formulations have many alternative optima. The use of an aggregate CF for the entire system which is then allocated to products may reduce the effectiveness of the CF models. Taken as a whole, the very

limited results in this paper suggest the possibility that the quality of the solutions produced by LP models may be rather robust to lead time estimates.

Recent innovations in computation increase the applicability of simulation optimization methods. Despite its promising performance in our test environment, many issues still need to be investigated before we can assess the actual value and applicability of SOP in production planning. Some future research directions are:

- Systematically analyzing the sensitivity of SOP to the initial parameter selection, threshold parameters and range of the perturbations.
- Comparing the performance of perturbation analysis to other gradient estimation techniques.
- Testing the developed SOP in larger, more realistic production systems.

Developing other simulation optimization procedures, which aim to optimize other decisions (such as releases) in the production planning model. Approaches of this type have already been proposed by Liu et al.(2011) and Homem de Mello et al. (Homem de Mello et al. 1999). Also some performance improvement can be achieved by changing the search space of SOP. For example, instead searching over operation lead times, lead times over machines can be used, which would give a much smaller search space for SOP.

ACKNOWLEDGMENTS

The research of Reha Uzsoy was supported by the National Science Foundation under Grant No. CMMI-1029706. The opinions on this paper reflect those of the authors and not those of the National Science Foundation.

REFERENCES

Albey, E., and U. Bilge. 2011. "A Hierarchical Approach to Fms Planning and Control with Capacity Anticipation." *International Journal of Production Research* 49(11): 3319-3342.

Albey, E., U. Bilge, and R. Uzsoy. 2014. "An Exploratory Study of Disaggregated Clearing Functions for Multiple Product Single Machine Production Environments." *International Journal of Production Research* 52(18): 5301-5322.

Askin, R. G., and C. R. Standridge. 1993. *Modeling and Analysis of Manufacturing Systems*. New York, John Wiley.

Asmundsson, J. M., R. L. Rardin, C. H. Turkseven, and R. Uzsoy. 2009. "Production Planning Models with Resources Subject to Congestion." *Naval Research Logistics* 56: 142-157.

Bang, J. Y., and Y. D. Kim. 2010. "Hierarchical Production Planning for Semiconductor Wafer Fabrication Based on Linear Programming and Discrete-Event Simulation." *IEEE Transactions on Automation Science and Engineering* 7(2): 326-336.

Byrd, R. H., J. Nocedal, and R. A. Waltz. 2006. "Knitro: An Integrated Package for Nonlinear Optimization". In *Large-Scale Nonlinear Optimization*, edited by G. Di Pillo and M. Roma. Heidelberg, 35-59. New York: Springer.

Byrne, M. D., and M. A. Bakir. 1999. "Production Planning Using a Hybrid Simulation-Analytical Approach." *International Journal of Production Economics* 59: 305-311.

Byrne, M. D., and M. M. Hossain. 2005. "Production Planning: An Improved Hybrid Approach." *International Journal of Production Economics* 93-94: 225-229.

Donohue, K. L., and M. L. Spearman. 1993. "Improving the Design of Stochastic Production Lines—an Approach Using Perturbation Analysis." *International Journal of Production Research* 21: 2789-2806.

Elmaghraby, S. E. 2010. Production Capacity: Its Bases, Functions and Measurement. *Planning Production and Inventories in the Extended Enterprise: A State of the Art Handbook*. K. G. Kempf, P. Keskinocak and R. Uzsoy. New York, Springer. **1:** 119-166.

Fu, M. C. 2002. "Optimization for Simulation: Theory Vs. Practice." *INFORMS Journal on Computing* 14(3): 192-215.

Ho, Y. C., M. A. Eyler, and T. T. Chien. 1979. "A Gradient Technique for General Buffer-Storage Design in a Serial Production Line." *International Journal of Production Research* 17: 557-580.

Homem de Mello, T., A. Shapiro, and M. L. Spearman. 1999. "Finding Optimal Material Release Times Using Simulation-Based Optimization." *Management Science* 45(1): 86-102.

Hopp, W. J., and M. L. Spearman. 2008. *Factory Physics: Foundations of Manufacturing Management*. Boston, Irwin/McGraw-Hill.

Hung, Y. F., and R. C. Leachman. 1996. "A Production Planning Methodology for Semiconductor Manufacturing Based on Iterative Simulation and Linear Programming Calculations." *IEEE Transactions on Semiconductor Manufacturing* 9(2): 257-269.

Irdem, D. F., N. B. Kacar, and R. Uzsoy. 2010. "An Exploratory Analysis of Two Iterative Linear Programming-Simulation Approaches for Production Planning." *IEEE Transactions on Semiconductor Manufacturing* 23(3): 442-455.

Irdem, D. F., N. B. Kacar, and R. Uzsoy. 2010. "An Exploratory Analysis of Two Iterative Linear Programming-Simulation Approaches for Production Planning." *IEEE Transactions on Semiconductor Manufacturing* 23: 442-455.

Johnson, L. A., and D. C. Montgomery. 1974. *Operations Research in Production Planning, Scheduling and Inventory Control*. New York, John Wiley.

Kacar, N. B., D. F. Irdem, and R. Uzsoy. 2012. "An Experimental Comparison of Production Planning Using Clearing Functions and Iterative Linear Programming-Simulation Algorithms." *IEEE Transactions on Semiconductor Manufacturing* 25(1): 104-117.

Kacar, N. B., L. Moench, and R. Uzsoy. 2014. Modelling Cycle Times in Production Planning Models for Wafer Fabrication. Technical Report, *Edward P. Fitts Department of Industrial and Systems Engineering*. Raleigh, NC, North Carolina State University.

Kacar, N. B., and R. Uzsoy. 2015. "Estimating Clearing Functions for Production Resources Using Simulation Optimization." *IEEE Transactions on Automation Science and Engineering* 12(2): 539-552.

Kang, Y. H., E. Albey, S. Hwang, and R. Uzsoy. 2014. "The Impact of Lot Sizing in Multiple Product Environments with Congestion." *Journal of Manufacturing Systems* 33: 436-444.

Karmarkar, U. S. 1989. "Capacity Loading and Release Planning with Work-in-Progress (Wip) and Lead-Times." *Journal of Manufacturing and Operations Management* 2(1): 105-123.

Kayton, D., T. Teyner, C. Schwartz, and R. Uzsoy. 1997. "Focusing Maintenance Improvement Efforts in a Wafer Fabrication Facility Operating under Theory of Constraints." *Production and Inventory Management*(Fourth Quarter): 51-57.

Kim, B., and S. Kim. 2001. "Extended Model for a Hybrid Production Planning Approach." *International Journal of Production Economics* 73: 165-173.

Liberopoulos, G., and M. Caramanis. 1994. "Infinitesimal Perturbation Analysis for 2nd Derivative Estimation and Design of Manufacturing Flow Controllers." *Journal of Optimization Theory and Applications* 81: 297-327.

Liu, J., C. Lii, F. Yang, H. Wan, and R. Uzsoy. 2011. "Production Planning for Semiconductor Manufacturing Via Simulation Optimization". In *Proceedings of the 2011 Winter Simulation Conference*, edited by S. Jain, R. R. Creasey, J. Himmelspach, K. P. White and R. Fu. 3617-3627. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Missbauer, H. 2002. "Aggregate Order Release Planning for Time-Varying Demand." *International Journal of Production Research* 40: 688-718.

Missbauer, H., and R. Uzsoy. 2010. Optimization Models for Production Planning. *Planning Production and Inventories in the Extended Enterprise: A State of the Art Handbook*. K. G. Kempf, P. Keskinocak and R. Uzsoy. New York, Springer**:** 437-508.

Pahl, J., S. Voss, and D. L. Woodruff. 2005. "Production Planning with Load Dependent Lead Times." *4OR: A Quarterly Journal of Operations Research* 3: 257-302.

Pahl, J., S. Voss, and D. L. Woodruff. 2007. "Production Planning with Load Dependent Lead Times: An Update of Research." *Annals of Operations Research* 153: 297-345.

Selçuk , B., J. C. Fransoo, and A. G. de Kok. 2007. "Work in Process Clearing in Supply Chain Operations Planning." *IIE Transactions* 40: 206-220.

Spall, J. C. 1998. "Implementation of the Simultaneous Perturbation Algorithm for Stochastic Optimization." *IEEE Transactions on Aerospace and Electronic Systems* 34(3): 817-823.

Srinivasan, A., M. Carey, and T. E. Morton. 1988. Resource Pricing and Aggregate Scheduling in Manufacturing Systems. *Graduate School of Industrial Administration, Carnegie-Mellon University*. Pittsburgh, PA.

Tekin, E., and I. Sabuncuoglu. 2004. "Simulation Optimization: A Comprehensive Review on Theory and Applications." *IIE Transactions* 36: 1067-1081.

Voss, S., and D. L. Woodruff. 2003. *Introduction to Computational Optimization Models for Production Planning in a Supply Chain*. Berlin ; New York, Springer.

Wilcoxon, F. 1945. "Individual Comparisons by Ranking Methods." *Biometrics Bulletin* 1(6): 80-83.

Yan, H., G. Yin, and S. X. C. Lou. 1994. "Using Stochastic Optimization to Determine Threshold Values for the Control of Unreliable Manufacturing Systems." *Journal of Optimization Theory and Applications* 83: 511-539.

Zapata, J. C., J. Pekny, and G. V. Reklaitis. 2010. Simulation-Optimization in Support of Tactical and Strategic Enterprise Decisions. *Planning Production and Inventories in the Extemnded Enterprise: A State of the Art Handbook*. K. G. Kempf, P. Keskinocak and R. Uzsoy, 1:593-628. New York, Springer.

AUTHOR BIOGRAPHIES

ERINC ALBEY received his B.Sc., M.Sc. and Ph.D. degrees in Industrial Engineering from Bogazici University, Istanbul, Turkey. He was a researcher at Bogazici University Flexible Automation and Intelligent Manufacturing Laboratory during his graduate education. He is currently a postdoctoral research associate in the Edward P. Fitts Department of Industrial and Systems Engineering, North Carolina State University. His research interests are planning under uncertainty, forecasting and capacity modeling in congested systems, predictive modeling and decision making in the presence of alternative and flexible resources, optimization and simulation. His email address is ealbey@ncsu.edu.

REHA UZSOY is Clifton A. Anderson Distinguished Professor in the Edward P. Fitts Department of Industrial and Systems Engineering at North Carolina State University. He holds BS degrees in Industrial Engineering and Mathematics and an MS in Industrial Engineering from Bogazici University, Istanbul, Turkey. He received his Ph.D. in Industrial and Systems Engineering in 1990 from the University of Florida. His teaching and research interests are in production planning, scheduling, and supply chain management. He was named a Fellow of the Institute of Industrial Engineers in 2005, Outstanding Young Industrial Engineer in Education in 1997, and has received awards for both undergraduate and graduate teaching. His email address is ruzsoy@ncsu.edu.