Proceedings of the 2008 Winter Simulation Conference
S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, J. W. Fowler eds.

AN EXPERIMENTAL STUDY OF AN ITERATIVE SIMULATION-OPTIMIZATION ALGORITHM FOR PRODUCTION PLANNING

Durmus Fatih Irdem
Necip Baris Kacar
Reha Uzsoy

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

ABSTRACT

It is well known from queueing and simulation models that cycle times in capacitated production systems increase nonlinearly with resource utilization, which poses considerable difficulty for the conventional linear programming (LP) models used for this purpose. Hung and Leachman (1996) propose a highly intuitive iterative approach where a detailed simulation model of the production facility is used to estimate flow time parameters used in an LP model We examine the convergence of this method under different experimental conditions, and conclude that it is hard to determine precisely when the method converges.

1 INTRODUCTION

The purpose of production planning is to match the output of production facilities to market demand in a manner that optimizes some performance measure for the firm. The basic actionable decision is the timing of material releases into the plant so that output meets customer demand in a timely fashion. This requires knowledge of the cycle time of the production facility, the time elapsing between the release of work into the plant and its emergence as finished product that can be used to meet demand. However, queueing models (e.g., Hopp and Spearman (2001); Buzacott and Shanthikumar (1993)) have shown that average cycle times will depend on the resource utilization, which is determined by the material release decisions made by the planning models. This creates a circularity that lies at the heart of the production planning field.

Semiconductor wafer fabrication facilities (fabs) are particularly exposed to this circularity for several reasons. Firstly, the capital intensive nature of the equipment requires the system to operate at high utilization levels to be financially viable, resulting in highly nonlinear relationships between resource utilization and cycle times. In addition, the production processes are complex, involving hundreds of unit operations and complex reentrant product flows, where lots of wafers visit the same equipment groups multiple times at different stages of their processing. Hence an effective production planning system must explicitly account for the cycle times, and the fact that small variations in resource utilization, due to small changes in work release patterns, may cause significant variation in cycle times. The estimates of cycle times used in production planning are referred to as lead times or flow times; we will use both terms interchangeably in this paper. A review of the semiconductor manufacturing process and its main complexities is given by Uzsoy et al. (1992).

Hung and Leachman (1996) have proposed an elegantly intuitive solution to this difficulty: an iterative algorithm that alternates between an LP model for production planning, which takes flow time estimates as inputs and determines a profit-maximizing release pattern over the planning horizon; and a detailed simulation model of the production facility, which takes as input the release pattern determined by the LP model and returns estimates of the flow times that would be realized by the facility under that release pattern. The new flow time estimates are then input into a new LP model, and the procedure iterates until some convergence criterion is satisfied. They apply their approach to an industrial data set, and report that the procedure converges according to their criteria. In a subsequent paper (Hung and Hou (2001)), they replace the simulation model with a queueing model and an empirical model for flow time estimation to reduce the computational burden due to the simulation model.

In this paper we report the results of an exploratory study of the convergence behavior of the HL procedure. We use a scaled-down model of a semiconductor wafer fabrication facility developed by the third author from industrial data and used extensively in previous research (Kayton et al. (1997); Barua et al. (2005)) as the testbed for the approach. Results indicate that while at low utilization the HL procedure produces very similar solutions at each iteration, at high utilization levels it is difficult to propose an

978-1-4244-2708-6/08/$25.00 ©2008 IEEE

objective, quantitative convergence criterion due to unpredictable oscillatory behavior between iterations.

In the following section we give a brief review of previous related work, focusing on attempts to model workload-dependent lead times in the production planning context. We then describe the HL procedure in detail, followed by a description of the testbed wafer fab used in our experiments. The design of our experiments is discussed in Section 4, followed by the results in Section 5. We conclude the paper with a summary and some directions for future research.

2 PREVIOUS RELATED WORK

Efforts to address the problem of workload-dependent lead times in production planning can be grouped under two main headings. The first of these are methods that assume lead times to be exogeneous parameters that are independent of resource utilization. These approaches, which in-clude the widely used Material Requirements Planning (MRP) procedure (Vollmann et al. (2005); Orlicky (1975)) and most LP models (e.g., Hackman and Leachman (1989); Johnson and Montgomery (1974); Voss and Woodruff (2003)), generally results in computationally tractable models whose accuracy at high utilization levels is questionable, especially under conditions of varying demands and workloads.

The second approach has been to use either a detailed scheduling algorithm (e.g., Dauzere-Peres and Lasserre (1994)) or a simulation model to verify that the release pattern proposed by the planning model does indeed result in a feasible production schedule. This approach generally captures the queueing behavior of the production resources quite accurately, but does not scale well to large systems due to the computational and data requirements of the detailed models of the shop floor it requires.

This latter approach has formed the basis for a number of approaches that combine LP and simulation models in an iterative manner. The procedure by Hung and Leachman (1996) which is the focus of this paper appears to be the best known of these. Hung and Hou (2001) examine a variant o the HL procedure where a queueing model is used to replace the simulation model in the interest of computational efficiency. However, other authors have proposed a number of variations that differ both in the specific LP formulation used and the information passed between the LP and simulation models. Byrne and Bakir (1999) propose an iterative LP-simulation approach where the results of the simulation model are used to update the right hand sides of the capacity constraints in a given period based on the realized resource utilization during that period; they extend the approach in Byrne and Hossain (2005). Kim and Kim (2001) propose a modification to the approach of Byrne and Bakir (1999). In all these papers, computational results are presented for a very small set of examples and a limited range of experimental conditions. Byrne and Bakir (1999), Byrne and Hossain (2005), and Kim and Kim (2001) use a small automated production system consisting of several machines linked by an automated conveyor, raising the question of how their procedures will perform in larger, more complex systems. Hung and Leachman (1996), on the other hand, use a single industrial data set for a large semi-conductor wafer fab. Although they conduct sensitivity analyses on the presence of randomness due to machine failures, a limited set of experiments is conducted, presumably due to the time-consuming nature of the experimentation (which we can vouch for from personal experience!)

In recent years a number of authors have proposed alternative models that capture the nonlinear relationship between resource utilization and workload explicitly. These approaches include the addition of a nonlinear term representing the cost of work in process inventories (WIP) to the objective function (e.g., Voss and Woodruff (2003)); constraints based on nonlinear clearing functions which relate the expected WIP level in a planning period to the expected output (e.g., Asmundsson et al. (2006); Karmarkar (1989); Missbauer (2002)); and enhanced LP formulations (e.g., Spitter et al. (2005); Lautenschlager and Stadtler (1998)). Extensive reviews of this work are given by Missbauer and Uzsoy (forthcoming) and Pahl et al. (2005).

This paper extends previous work by systematically examining the performance of the HL procedure under different experimental conditions, specifically the level of utilization at the bottleneck resource, the cost structure used in the LP model, and the implementation of the iterative algorithm. We examine the convergence behavior of the procedure with a view to developing simple, practical guidelines for when the procedure can be terminated with some confidence of having reached a good solution. We present a more detailed discussion of the HL procedure in the next section.

3 THE HUNG-LEACHMAN (HL) PROCEDURE

3.1 The Linear Programming Model

The HL procedure follows the conventional LP approach of dividing the planning horizon, the time interval over which decisions are to be made, into discrete planning periods. The production process for a product is represented as a series of operations; due to the reentrant routings in wafer fabs, multiple operations may use the same equipment. The model used is essentially the Step-Separated formulation of Leachman and Carmon (1992), which requires the estimated lead times F_{gl} required for a lot of product g to reach operation l after being released into the plant. However, instead of fixed lead times that remain constant over the entire planning horizon, the authors

associate values of the lead time parameters with the start of each planning period. In the following $p=0$ is the start of period 1, $p=1$ is the start of period 2, etc., that is, a time unit is the period length. The lead time parameters F_{gpl}, which may take fractional values, denote the expected time required for a lot of product g to reach operation l if the lot reaches operation l at the end of period p (i.e., at time p). In our experiments a planning period was defined to be seven days. Given the lead times, the loading of the production resource in period p is defined by releases occurring in the time interval $Q = [(p-1)- F_{g,p-1,l}, p- F_{gpl}]$, assuming planning period p starts at time $(p-1)$. The crux of the formulation is relating the resource loading Y_{gp} by product g in period p to the amount of product g released over time. We shall use the following notation:

$\tau_{g,p}$ *: number of working days for wafer type g from start of period 1 (time 0) until the end of period p, p=1,2,...,P.*
$[\tau]^+$ *: smallest index p such that* $\tau_{g,p} > \tau$.
$F_{g,p,l}$ *: the expected flow time from wafer release to operation l, occurring at epoch* $\tau_{g,p.}$
$F_{g,p}$ *: the expected flow time from wafer release to finish, occurring at epoch* $\tau_{g,p.}$
Y_{gpl} *: wafer quantity consuming machine hours at operation l for wafer type g in period p.*
Y_{gp} *: wafer output quantity for wafer type g in period p.*
X_{gp} *: wafer release quantity for wafer type g in period p.*

$$p^- = \left[\tau_{g,p-1} - F_{g,p-1,l}\right]^+ \quad (1)$$

$$p^+ = \left[\tau_{g,p} - F_{g,p,l}\right]^+ \quad (2)$$

There are two cases to consider here. In the first, simpler case, the time interval Q lies within a single planning period, and the amount Y_{gpl} of product g loading resources at operation l in period p is given by

$$Y_{gpl} = \frac{(\tau_{g,p} - F_{g,p,l}) - (\tau_{g,p-1} - F_{g,p-1,l})}{(\tau_{g,p^+} - \tau_{g,p^+-1})} X_{gp^+} \quad (3)$$

If, on the other hand, the time interval Q spans multiple planning periods, we allocate the load due to releases in that period in proportion to the fraction of that period's total duration included in the interval Q (again assuming uniform release rates within the planning periods). This yields

$$Y_{gpl} = \frac{\tau_{g,p^-} - (\tau_{g,p-1} - F_{g,p-1,l})}{(\tau_{g,p^-} - \tau_{g,p^--1})} X_{gp^-} + \sum_{p^-+1}^{p^+-1} X_{gp} + \frac{(\tau_{g,p} - F_{g,p,l}) - \tau_{g,p^+-1}}{(\tau_{g,p^+} - \tau_{g,p^+-1})} X_{gp^+} \quad (4)$$

The LP formulation meximizes profit subject to constraints on material flow and resource capacities. An artificial final period with length equal to the longest flow time over the horizon is added to ensure that an appropriate ending condition is achieved. We use the following notation:

Decision Variables:
X_{gp} *: wafer release quantity for wafer type g in period p.*
I_{gp}*: units of product g in finished goods inventory at the end of period p.*
B_{gp}*: units of product g backlogged at the end of period p.*

Parameters:
a_{glk} *:average machine hours of machine type k used in operation l of wafer type g.*
C_{kp} *: hours of mahine type k available in period p.*
v_{gp}*:Unit revenue from product g in period p*
c_{gp}*: Unit incremental production cost of product g in period p.*
h_{gp}*: Unit inventory holding cost for product g in period p.*
b_{gp}*: Unit backlogging cost for product g in period p.*
d_{gp} *:Demand for wafer type g in period p.*
fp_g *:First time period in which output of wafer type g is obtained.*
zp_g *:First frozen period of wafer type g. The production rates after this period will be set equal to the rate in this period in order to satisfy the steady-state horizon condition.*
sp_g *:Earliest nonpositive period number in which current WIP would have started considering the assumed flow times.*
X_{gp} *:Equivalent wafer releases generating the current WIP status of wafer type g, defined in periods before the start of the planning horizon, p=0,-1,-2,...,-* sp_g
$\overline{B_{gp}}$*: upper bound on backlogs for wafer type g in period p.*

The complete formulation is as follows:

$$\max \quad \sum_{g\in G}\sum_{p=1}^{P+1} v_{gp} Y_{gp} - \sum_{g\in G}\sum_{p=1}^{P+1} c_{gp} X_{gp} - \sum_{g\in G}\sum_{p=1}^{P+1} h_{gp} I_{gp} - \sum_{g\in G}\sum_{p=1}^{P+1} b_{gp} B_{gp}$$

Subject to:
1) Resource Capacity:

$$\sum_{g\in G}\sum_{l=1}^{l_g} a_{glk} Y_{gp}^l \le C_{kp} \qquad p=1,...,P+1 \qquad \text{for all } k \in K$$

2) Demand Equations:

$$Y_{gp} - I_{gp} + B_{gp} = \sum_{p=1}^{fp_g} d_{gp} \qquad g \in G, \quad p=fp_g$$

$$Y_{gp} + I_{g,p-1} - B_{g,p-1} - I_{gp} + B_{gp} = d_{gp}$$

$g \in G, \quad p=fp_g+1,...,P-1$

$$Y_{gp} - B_{g,p-1} + B_{gp} = d_{gp} \qquad g \in G, \quad p=P,\ldots,P+1$$

3) Variable Nonnegativity:

$$X_{gp} \geq 0 \qquad g \in G, \quad p=1,\ldots,zp_g$$

$$I_{gp} \geq 0 \qquad g \in G, \quad p=1,\ldots,P\text{-}1$$

$$I_{gp} = 0 \qquad g \in G, \quad p=P,\ldots,P+1$$

$$0 \leq B_{gp} \leq \overline{B_{gp}} \qquad g \in G, \quad p=1,\ldots,P+1$$

Details of the formulation are in Hung and Leachman (1996) and Leachman and Carmon (1992). The LP model was implemented in the OPL Version 5.5 language by ILOG <www.ilog.com/products/oplstudio> and run on a Intel PC with a Intel(R) Core(TM) 2 CPU 6700 2.66 GHz processor and 2GB of RAM, under MS Windows XP Professional.

3.2 The Simulation Model

The re-entrant bottleneck system was built with attributes of the real-world fab environment. The major characteristics of wafer fabrication, including a re-entrant bottleneck process, unreliable machines, batching machines, and multiple products with varying process routings are included in the model. The model was built by defining a distinct re-entrant bottleneck representing the photolithography process. The processing times for all other stations were scaled to the bottleneck processing time so that no non-bottleneck station would have a utilization approaching that of the bottleneck. The model has batching stations early in the process, representing the furnaces which perform the diffusion and oxidation processes. The minimum batch size required is two lots and the maximum batch size is four lots. The batching stations can be loaded with any product lot mix, that is, a batching station can run lots of one type of product or many product types at one time. The remaining stations process one lot at a time.

The simulation model is made up of 11 stations, each with one server except the bottleneck station (Station 4) that has two servers. The processing times for the stations are lognormally distributed with the standard deviation less than or equal to 10 percent of the mean. The low process variance is representative of automation and tight process specifications encountered in the semiconductor industry. There are three products produced in the system with different complexity. Product 1 has 22 process steps including 6 visits to the bottleneck station. Product 2 has 14 process steps with 4 visits to the bottleneck station. Product 3 has 14 process steps and does not visit the bottleneck, but instead visits Station 11. The system is required to produce a product mix in proportions of 3:1:1 of Product 1, 2, and 3 respectively.

In the model, there are two unreliable stations that create most of the starvation at the bottleneck. One station is visited only once by each product early in the process routings; for simplicity, we shall refer to this station as the "Single Entry Machine". The second unreliable station a station that is visited multiple times by the products and occurs later in the processing steps. This station is representative of a Chemical Vapor Deposition (CVD) process that is capable of producing a high output very quickly. This station will be referred to as the "Multiple Entry Machine". These two unreliable stations have the ability to produce a lot of product in a very short period of time but can starve the bottleneck due to poor availability. In our experiments, we just considered the failures at Machine 3 and Machine 7 and ignored the failures at the other machines. A detailed description of the model and the process flows can be obtained from the third author on request.

Lots are dispatched in First-in-First-Out order on all machines. The simulation model was implemented in Arena Version 10.0 <www.arenasimulation.com>, and integrated with the LP model using Excel and a number of Visual Basic scripts.

3.3 The Iterative Procedure

Given the LP formulation and the simulation model above, the HL procedure can be stated as follows:

Algorithm HL:

Step 1: Set $k = 1$; $MaxIT = 30$; obtain initial flow time estimates F^0_{gpl}.Set $\phi^k_{gpl} = F^0_{gpl}$. In our experiments the F^0_{gpl} were obtained from a steady state simulation run with releases set equal to period demand for each product.

Step 2: Solve the LP model using the flow time estimates F^k_{gpl} to obtain the material release schedule X^k_{gp}.

Step 3: Assuming the releases in each period are uniformly distributed over the period, use five independent replications of the simulation model to estimate the flow times F^k_{gpl}. The mean of the sample values obtained from the simulation replications is used as the estimator. The releases suggested by the LP model are rounded to integer quantities, and any additional lots thus generated (due to the difference between fractional and rounded values of the X^k_{gp}) are distributed evenly over the planning horizon to minimize their disruptive effects.

Step 4: If $k < MaxIT$, set $k = k+1$, $\phi^k_{gpl} = \alpha F^k_{gpl} + (1-\alpha)F^{k-1}_{gpl}$, where $0 \leq \alpha \leq 1$ is a user-defined smoothing constant, and go to Step 2. Otherwise, stop.

The number of simulation replications was selected based on a tradeoff between the need to obtain some statistical precision in our estimates of the flow times, while keeping the computational burden of the overall iterative procedure within reasonable limits. In the following section we describe the design of the computational experiments, and then proceed to present our results.

4 EXPERIMENTAL DESIGN

Our objective in this paper is to explore the behavior of the iterative HL procedure under a broader range of experimental conditions than those hitherto studied, with a particular interest in examining the convergence behavior of the procedure. It should be noted that there are several different potential ways of defining the convergence of this algorithm. In the original work, Hung and Leachman require the Mean Absolute Deviation of the average flow time across all products to be within 5% from one iteration to the next, but this leaves open the possibility of fluctuations in the flow times of each product that cancel out across products but can cause significant differences in the realized output. A more stringent criterion would be to require the Mean Absolute Deviation of flow times for all individual products to be below some tolerance. A less demanding approach would require the objective function values of the LP at successive iterations to converge to zero. Clearly a number of other criteria are possible.

In our description of the HL procedure above, we have based our convergence criterion of a maximum number of iterations (30 in our case). This decision is due to the fact that in preliminary experiments we were unable to obtain unambiguous convergence with any criteria except that of the LP objective function values. Our experiments were designed to examine the effects of three different factors on the performance of the HL procedure:

Smoothing Constant: The original HL procedure uses $\alpha=1$, which might result in significant oscillations if the simulation model returns significantly different flow time estimates at successive iterations. To this end we consider experiments with α values of 1, 0.5 and 0.2 to see whether a smoothing scheme is able to mitigate some of the oscillatory behavior observed between iterations.

Bottleneck Utilization: It is well known from queueing theory that the nonlinear relationship between resource utilization and flow times becomes more severe at high utilization levels. Hence one would expect an LP model using fixed, exogeneous flow time estimates to perform well at low utilization levels, but to degrade in performance at higher utilization. Hence we experiment with two bottleneck utilization values of 0.5 and 0.9. The utilization level is achieved by varying the demand of all products while maintaining the 3:1:1 product mix required by the testbed. In order to test the algorithm under favorable conditions, we maintain the demand constant across the planning horizon of 14 periods. Both the LP model and the simulation model are initialized with work in process levels obtained from a steady-state simulation made with the same demand levels and with releases equal to demand in each period.

Backorder and Inventory Costs: Examining the LP model above, it is apparent that when the revenue is much higher than the other costs, as is the case in practice, the greater part of the objective function value is determined by the demand level – the model will try to produce output that meets demand as far as possible, and inventory and backorder costs represent a relatively small fraction of the total value. Thus one might hypothesize that if inventory holding and backorder costs are of similar magnitude, there may exist many alternative LP solutions with very similar objective function values, yielding quite different release schedules which would affect factory performance but not necessarily the LP objective function value. To this end, we examine two cases, one where backorder cost is slightly higher than inventory holding cost, and another where backordering is more than twice as expensive as holding inventory.

Thus we have a total of twelve algorithm runs (3 x 2 x 2) in our experiment. For ease of reference, we shall denote each by a triple (u, α, b_{gp}), denoting the bottleneck utilization, the value of the smoothing parameter, and the value of the backorder cost, respectively.

Table 1: Experimental Design.

Factor	Levels
Smoothing Parameter α	1 ,0.5, 0.2
Bottleneck utilization	0.5, 0.9
Backorder costs (b_{gp})	(20, 40)

5 EXPERIMENTAL RESULTS

We shall focus our discussion on the convergence behavior of the HL procedure. We must emphasize that these results are exploratory in nature, and will require further experiments and analysis before firm conclusions can be reached. However, we believe they provide some useful insights, highlighting the difficulty of determining an objective, quantitative convergence criterion for this type of procedure.

Since the results obtained at low bottleneck utilization did not exhibit major changes from iteration to iteration, we focus on those for the high utilization case.

Given the highly nonlinear relationship between flow times and utilization suggested by queueing theory (e.g., Hopp and Spearman 2001), we would expect serious difficulties under this situation; the LP model's ability to represent the actual behavior of the capacitated production resources is likely to be quite poor, especially at the bottle-

neck resources. Our base case for this discussion is the (0.9, 1.0, 20) case, as shown in Figures 1 and 2 below.

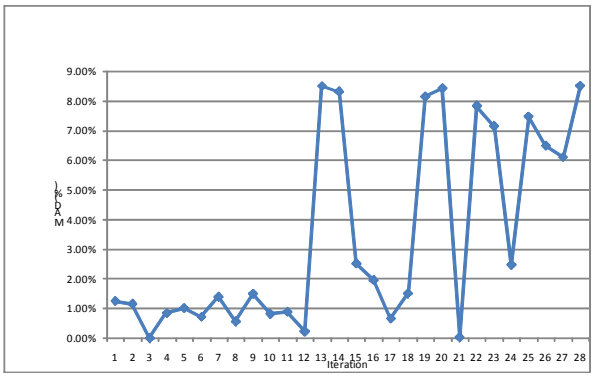

Figure 1: MAD of LP Objective for (0.9, 1.0, 20) Case

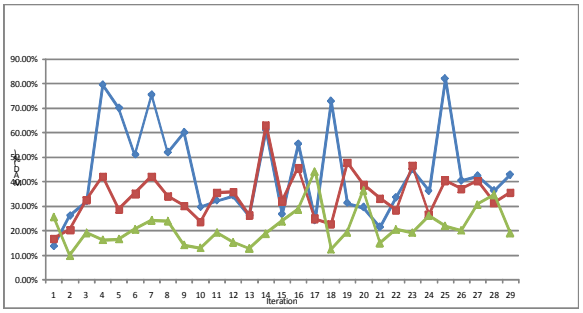

Figure 2: MAD in Product Flow Times for (0.9, 1.0, 20) Case

These results are again discouraging in terms of convergence. The MAD of the LP objective has quite low variability for the first 12 iterations, but then begins to oscillate strongly. Figure 2 shows that in the same set of iterations, the flow time estimates the product with large number of bottleneck visits varies quite significantly; MADs are now up to 80% from one iteration to another, despite the MAD of the LP objective being less than 10%.

Figures 3 and 4 show the effects of smoothing on this case (the (0.9, 0.2, 20) scenario), while Figures 5 and 6 present those for the smoothed case with high backlog cost (the (0.9, 0.2, 40) scenario). The smoothing approach does seem to reduce the MAD in LP objective function value significantly, but severe variability remains in the individual product flow time estimates, especially for Product 1, which uses the bottleneck heavily and returns to it for many visits. Once again, examining the MADs for the individual products, they do not appear to converge to zero; at best they maintain a constant average, and in some cases suggest they might actually diverge if more iterations were performed.

A number of interesting observations can be made from these results. For example, in Figure 4, the MAD in the product flow time estimates between iterations 4 and 5 is actually zero; the corresponding MAD in objective functions is positive, though small. However, the next iteration yields a substantial deviation which never returns to the previous level. We might conjecture this is due to

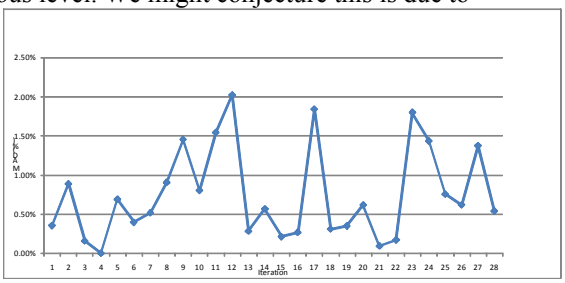

Figure 3: MAD in LP Objective Value for (0.9, 0.2, 20) Case

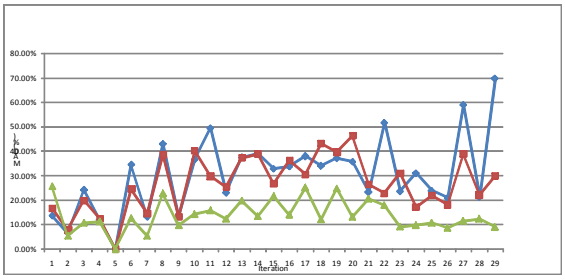

Figure 4: MAD in Product Flow Times for (0.9, 0.2, 20) Case

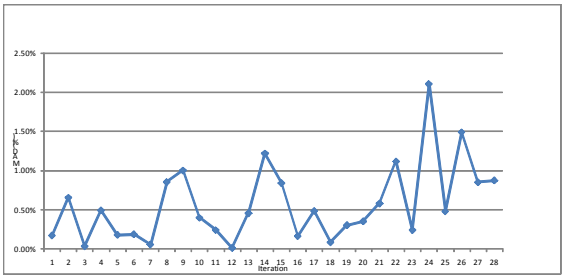

Figure 5: MAD in LP Objective Value for (0.9, 0.2, 40) Case

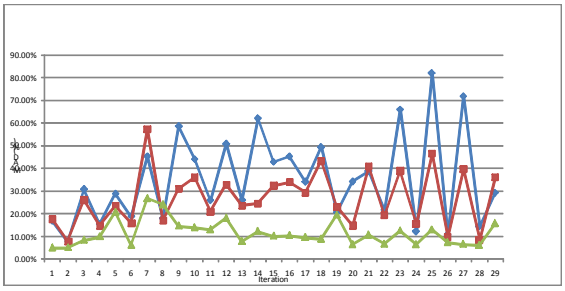

Figure 6: MAD in Product Flow Times for (0.9, 0.2, 40) Case

sampling error in the simulations; due to random variation, the simulations return a different flow time estimate, which, due to the high utilization level, leads to a sufficiently different release pattern that the next iteration produces rather different flow time estimates from the simula-

tion. However, this is also difficult to defend as a root cause. Figure 7 presents the results of the (0.9, 1.0, 40) case with random processing times and machine failures per the simulation model above; Figure 8 shows the results of the same case when all machine failures and processing time variability are removed, yielding a fully deterministic simulation. Substantial variations are still seen in Figure 8, although not as persistent as in Figure 7. Hence the sampling error from the simulation cannot be the sole source of the algorithm's erratic behavior. The deterministic simulation does lead to a MAD of zero at iteration 23 in Figure 8, and the objective function value variation is less than 5% in both cases, but significant fluctuation in the flow times of individual products persist.

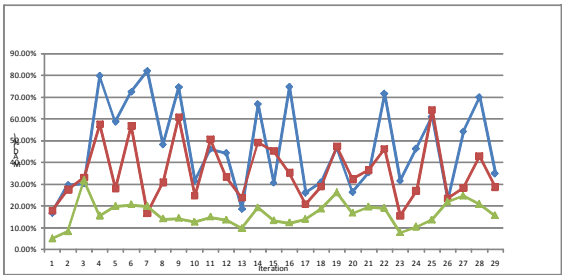

Figure 7: MAD for Product Flow Times for (0.9, 1.0, 40) Case

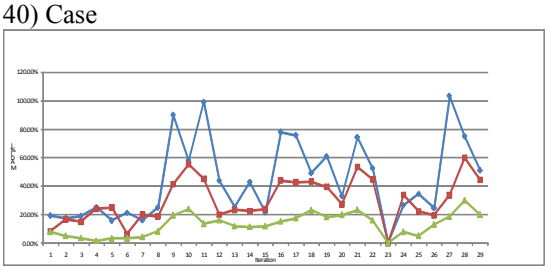

Figure 8: MAD for Product Flow Times for (0.9, 1.0, 40) Case with Deterministic Simulation

Another interesting observation is that the three products are affected differently. Products 1 and 2, which run through the principal bottleneck station, station 4, exhibit quite large variation from one iteration to the next. Product 3, which does not use Station 4, achieves the lowest deviations over the horizon – but still does not converge to zero.

5.1 Alternative Convergence Criteria

The observations above might suggest that terminating the algorithm either when the LP objective function value or the MAD of the individual product flow times is minimum might yield a suitably good solution. To examine this, we take the release pattern obtained at the most favorable iteration under either criterion and simulate its execution to compare the output predicted by the LP model and that realized in the simulation. Figure 9 shows the percent difference (as a percentage of the simulation output) between the output values predicted by the LP and those achieved in the simulation over the 14 period planning horizon, where a negative value indicates that the LP predicted higher output than the simulation achieved. For the (0.5, 0.2, 40) case, both the objective function and the MAD of product flow times gave their best values at the same iteration. The final period, period 14, shows the difference in total output over the planning horizon (13 periods) between the two models. Significant percentage differences between the output predicted by the LP and that achieved by the simulation are present in individual periods, although total output over the horizon is very close.

Figures 10 and 11 show the same results for the (0.9, 0.2, 40) case, where the two criteria suggested different iterations at which to terminate. In both cases, the differences between the output predicted by the LP model and that realized by the simulation are substantial, suggesting that when we terminate the iterations at what looks like a promising point, the release schedule suggested by the LP may not be feasible in terms of producing the desired output at the objective function value predicted by the LP model.

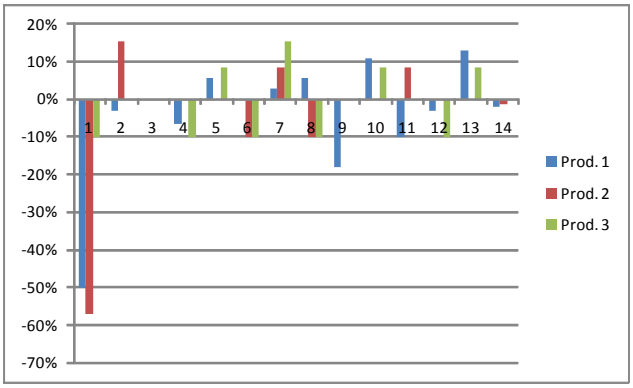

Figure 9: Differences in Realized Output between LP and Simulation for the (0.5, 0.2, 40) Case

6 SUMMARY AND FUTURE DIRECTIONS

Taken as a whole, these results appear to suggest that the convergence behavior of the HL procedure is quite complex, and that it is difficult to propose an unambiguous, quantitative termination criterion, especially at high utilization levels. While a smoothing constant helps to some degree, it does not bring an unambiguous convergence of the MAD values to zero. Random variation in the simulation estimates cannot be the sole culprit, since a completely deterministic simulation model does not result in a qualitative improvement in convergence behavior.

We believe the key to this dilemma is suggested in the last set of results in the previous section. These suggest that the LP model does not represent the behavior of the production system accurately, based on the deviations between the output predicted by the two models. Thus the release schedule produced by the LP solution at a given iteration results in flow times rather different than those that

were used to obtain the release schedule. Changes in the flow time estimates, in turn, change the LP formulation itself, altering the sets of variables that are connected to each other by the constraints. This, combined with the fact that the LP model will produce extreme point solutions, results in the LP models producing quite different release schedules from those at the previous iteration.

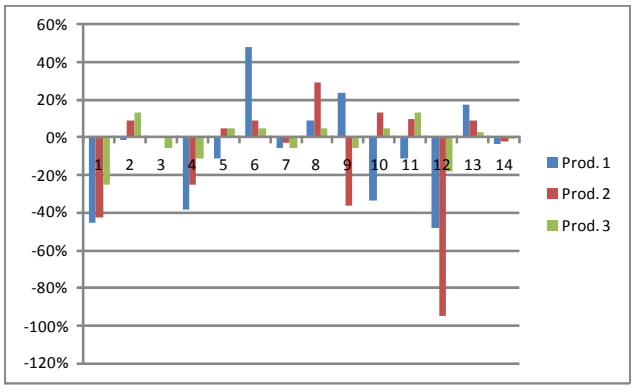

Figure 10: Differences in Realized Output between LP and Simulation for the (0.9, 0.2, 40) Case with Minimum Objective Function Value Criterion

Another potential difficulty lies in the fact that the flow time estimates are made at the beginning and the end of a planning period. This will induce significant variability in the estimates from individual simulation replications, which may cause problems with oscillations as discussed above.

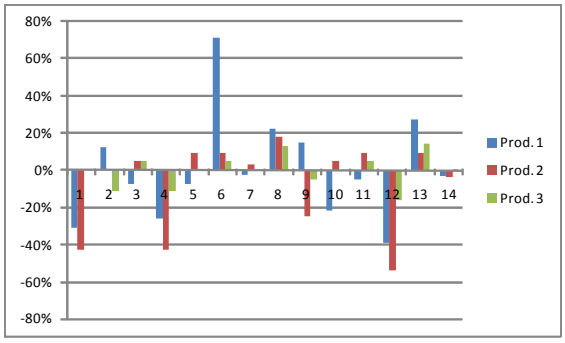

Figure 11: Differences in Realized Output between LP and Simulation for the (0.9, 0.2, 40) Case with Minimum Flow Time MAD Criterion

In addition, if the planning periods are long relative to the processing times at the operations, it is likely that the flow times during the period will vary substantially. This will limit the ability of the LP model to correctly capture the behavior of the queues correctly. This could be remedied by reducing the length of the planning period used in the LP model at the cost of a much larger LP formulation; also, once the planning period is reduced beyond a certain point, the usefulness of the LP model becomes questionable, as it becomes equivalent to shop-floor scheduling.

These results are clearly not conclusive, and several directions for future research remain. A detailed analysis of individual solutions, and individual runs of the procedure, is required to confirm or refute our conjectures made above. A satisfactory explanation of the behavior of the current procedure will naturally open the way for improved procedures with better performance. Finally, the large volumes of data produced by the approach requires some systematic thought as to how to present these results to best effect, providing the analyst with insight needed to determine whether convergence is present and if not, why.

ACKNOWLEDGMENTS

This research was supported by the E. P Fitts Department of Industrial and Systems Engineering, the National Science Foundation under Grant No. 080956, a research grant from the Intel Research Council, and an equipment grant from the Intel Corporation.

REFERENCES

Asmundsson, J. M., R. L. Rardin and R. Uzsoy. 2006. Tractable Nonlinear Production Planning Models for Semiconductor Wafer Fabrication Facilities. *IEEE Transactions on Semiconductor Manufacturing* 19: 95-111.

Barua, A., R. Narasimhan, R. Uzsoy and A. Upasani. 2005. Implementing Global Factory Schedules in the Face of Stochastic Disruptions. *International Journal of Production Research* 43: 793-818.

Buzacott, J. A. and J. G. Shanthikumar. 1993. *Stochastic Models of Manufacturing Systems.* Englewood Cliffs, NJ, Prentice-Hall.

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.

Dauzere-Peres, S. and J. B. Lasserre. 1994. *An Integrated Approach in Production Planning and Scheduling*. Berlin, Springer-Verlag.

Hackman, S. T. and R. C. Leachman. 1989. A General Framework for Modeling Production. *Management Science* 35: 478-495.

Hopp, W. J. and M. L. Spearman. 2001. *Factory Physics : Foundations of Manufacturing Management*. Boston, Irwin/McGraw-Hill.

Hung, Y. F. and M. C. Hou. 2001. A Production Planning Approach based on Iterations of Linear Programming

Optimization and Flow Time Prediction. *Journal of the Chinese Inst. of Industrial Engineers* 18(3): 55-67.

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.

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

Karmarkar, U. S. 1989. Capacity Loading and Release Planning with Work-in-Progress (WIP) and Lead-times. *Journal of Manufacturing and Operations Management* 2(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.

Lautenschlager, M. and H. Stadtler. 1998. *Modelling Lead Times Depending on Capacity Utilization*. Research Report, Technische Universitat Darmstadt.

Leachman, R. C. and T. F. Carmon. 1992. On Capacity Modeling for Production Planning with Alternative Machine Types. *IIE Transactions* 24(4): 62-72.

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

Missbauer, H. and R. Uzsoy. Forthcoming. Optimization Models for Production Planning. In *Planning Production and Invento-ries in the Extended Enterprise: A State of the Art Handbook*. K. G. Kempf, P. Keskinocak and R. Uzsoy. Norwell, MA, Springer.

Orlicky, J. 1975. *Material Requirements Planning: the New Way of Life in Production and Inventory Management*. New York, McGraw-Hill.

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.

Spitter, J. M., C. A. J. Hurkens, A. G. de Kok, J. K. Lenstra and E. G. Negenman. 2005. Linear Programming Models with Planned Lead Times for Supply Chain Operations Planning. *European Journal of Operational Research* 163: 706-720.

Uzsoy, R., C. Y. Lee and L. A. Martin-Vega. 1992. A Review of Production Planning and Scheduling Models in the Semiconductor Industry Part I: System Characteristics, Performance Evaluation and Production Planning. *IIE Transactions on Scheduling and Logistics* 24(47-61).

Vollmann, T. E., W. L. Berry, D. C. Whybark and F. R. Jacobs. 2005. *Manufacturing Planning and Control for Supply Chain Management*. New York, McGraw-Hill.

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

AUTHOR BIOGRAPHIES

DURMUS FATIH IRDEM is currently pursuing his Master's degree in Industrial Engineering and a Minor in Operations Research in the Edward P. Fitts Department of Industrial and Systems Engineering at North Carolina State University. He received his B.S. degree in Engineering Management from Istanbul Technical University, Turkey. His research interests are in production planning, application of simulation-based optimization to production planning models and supply chain management. He can be reached by e-mail at <dfirdem@ncsu.edu>.

NECIP BARIS KACAR is a Research Assistant and is pursuing an M.S degree with a concentration in Production Systems in Industrial Engineering in the Edward P. Fitts Department of Industrial and Systems Engineering at North Carolina State University. He received a BS degree in Mechanical Engineering from Bogazici University, Istanbul, Turkey. His research interests are in simulation based optimization, production planning and logistics. He can be reached via email at <nbkacar@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 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 is the author of one book, an edited book, and over eighty refereed journal publications.. He was named a Fellow of the Institute of Industrial Engineers in 2005, Outstanding Young Industrial Engineer in Education in 1997 and a University Faculty Fellow by Purdue University in 2001, and has received awards for both undergraduate and graduate teaching. He is currently serving on the Editorial Boards of *IIE Transactions on Scheduling and Logistics* and *International Journal of Computer-Integrated Manufacturing*. He can be reached by email at <ruzsoy@ncsu.edu>.