Proceedings of the 2019 Winter Simulation Conference
N. Mustafee, K.-H.G. Bae, S. Lazarova-Molnar, M. Rabe, C. Szabo, P. Haas, and Y.-J. Son, eds.

OPTIMZING ENGINEERING AND PRODUCTION LOTS DURING PRODUCT TRANSITIONS IN SEMICONDUCTOR MANUFACTURING

Atchyuta Bharadwaj Manda
Reha Uzsoy

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

ABSTRACT

The adverse impact of new product introductions on the performance of semiconductor wafer fabrication facilities (fabs) is widely acknowledged. In this paper, we develop a simulation model of a simplified production system that captures the impact of a new product introduction and the learning effects of production and engineering activities on the total system throughput. We use a simulation optimization procedure utilizing a Genetic Algorithm (GA) to obtain near optimal releases of production and engineering lots that maximize total contribution over the planning horizon. Numerical experiments provide insights into the structure of optimal release policies and illustrate the improvements that can be achieved through the strategic use of engineering lots.

1 INTRODUCTION

Introducing new products into the market frequently and effectively is important in high technology industries such as semiconductor manufacturing. Existing products are replaced in the market by new products that offer improved functionality at lower cost or higher price, especially during the early part of their life cycle when competition is still scarce. While these new product introductions are essential to competitiveness, they can have an adverse impact on the performance of manufacturing facilities, especially semiconductor wafer fabrication facilities (fabs) (Leachman and Hodges 1996). This paper examines the problem of managing such product transitions by building a mathematical model of a simplified production system.

There is considerable anecdotal evidence that when a new product is introduced into a fab that is also producing other products, it adversely affects the performance of the fab through several interrelated mechanisms. Examples of these are increased operator errors in processing, additional setup times and test wafer runs, and increased engineering holds on production equipment as problems with the new product are identified and remedied. This paper will focus on the last of these mechanisms, in which the new product creates disruptions in production as problems are uncovered and equipment put on hold for engineering work to diagnose and remedy the problem. These disruptions increase the average capacity consumption of the new product, reducing the available capacity for existing products. This in turn increases the variability of processing times, adversely impacting all products in the fab. However, as experience is gained in processing the new product, the frequency and duration of these disruptions decreases.

In earlier work (Manda and Uzsoy 2018b), we explored this problem of managing releases during a transition from an older product to a newer one using a simulation model of a simplified production system. The model incorporated the increased production disruptions induced by the new product and modelled the subsequent learning effects as a function of cumulative production experience.

This paper extends our earlier work by incorporating learning through experimentation, which occurs when engineers run experimental lots aimed at proactively identifying and fixing problems. These engineering lots do not result in products that can be sold to customers, but provide increased learning about how to better produce the new product. This is modelled as a reduced probability of a regular production lot of the new product inducing an engineering hold due to a previously unencountered problem. Thus in our previous model the only means for management to influence cycle times and throughput during the product transition is the releases of regular production lots of the old and new products. This paper introduces a new tradeoff between processing engineering lots aimed at debugging the process to boost learning and producing regular lots that satisfy demand but have a much slower learning effect. A simulation optimization procedure is applied to this enhanced simulation model to find approximately optimal releases for the production lots of both products and engineering lots of the new product. The objective is to maximize total expected contribution (revenue minus variable costs) over the duration of the product transition.

The rest of the paper is organized as follows: Section 2 provides a brief review of previous related work. Section 3 briefly reviews our previous simulation and mathematical models and describes the simulation model used in this work. The simulation optimization procedure is presented in Section 4. Section 5 presents the numerical experiments and results, while conclusions and future directions are presented in Section 6.

2 LITERATURE REVIEW

In this section we place this work in the context of the existing literature, beginning with the traditional learning curve literature that focuses on learning by doing. We then discuss models of learning through experimentation and how these are incorporated into decision support models. Finally we present some relevant simulation models and the contribution of our work.

2.1 Learning by Doing

Learning in manufacturing is the process of acquiring skills and improving productivity, usually through experimentation and experience. Learning leading to performance improvement through repetition of the specific task is called 'learning by doing', and the mathematical models that capture this learning process are referred to as "learning curves". The early learning literature focused on learning by doing in individual workers. The work of Wright (1936) is widely acknowledged as the first formal work on learning curves. The learning effect was subsequently identified and studied in several industries (Searle 1945; Hirsch 1952; Preston and Keachie 1964). Yelle (1979) and Dutton et al. (1984) present reviews of the early learning literature.

Much of the early literature in the semiconductor industry also focused on learning by doing. Webbink (1977) was one of the first systematic investigations of the learning curve in the semiconductor industry. Tirkel (2013) discusses the various yield learning models in the literature and proposes generalized compound learning curves based on power and exponential functions of factors like cumulative output, elapsed time and production rate.

Anzanello and Fogliatto (2011) review the extensive literature on learning curves and present various popularly used univariate learning curve models such as the log-linear, exponential and hyperbolic models. They also present multivariate learning curve models and forgetting models that have been used across different industries. Jaber (2006) provides a concise discussion of the various models that capture the learning and forgetting process and discusses why using cumulative production by itself might not be a good measure of learning.

A common theme among much of the early learning curve literature and some later works that incorporate this type of learning into planning models such as Liao (1979) and Reeves and Sweigart (1981), is their assumption that learning is driven directly by cumulative production. Our earlier work, Manda and Uzsoy (2018a) also used this type of learning model and ignored the possibility of learning through deliberate

experimentation. In that paper we examined how to manage the releases of production lots during a transition from an older product to a newer product. This work extends that model by considering learning through experimentation.

2.2 Learning Through Experimentation

Learning through experimentation occurs as a result of the direct managerial action of scheduling and performing controlled engineering trials, essentially using the production process as a laboratory. These engineering lots provide insights into manufacturing problems and permit the development and testing of proposed solutions. The potential learning from these lots is much higher than a regular production lot since the experiments can be designed carefully with a specific end in mind, thus providing managers with a way to increase the rate of learning. However, these engineering lots consume limited production capacity, presenting an interesting tradeoff between experimentation for faster learning versus regular production to generate revenue.

A distinction between learning by doing and by experimentation is made in later learning curve literature. Chand et al. (1996) consider learning as a process of systematic experimentation to improve process capability that consumes production capacity. Process improvement over time improves quality, which, in turn, helps to reduce production costs. Fine (1986) proposes an alternative formulation where learning depends on both cumulative production and the level of quality control activity. Adler and Clark (1991) build a learning model as a function of cumulative output and two managerial variables - engineering changes and workforce training. They distinguish between first order learning, which is learning by doing and driven by cumulative production and second order learning which results from explicit engineering or managerial action to increase capability by changing underlying technology, equipment, processes or human capital. They use data from two manufacturing departments in an electronic equipment company to propose a learning model capturing both first order and second order learning. Hatch and Mowery (1998) try to understand the processes underpinning learning by doing in a more generalized manner by using a learning model that incorporates managerial actions in the form of cumulative engineering activity. They perform statistical tests to assess the importance of cumulative engineering and conclude that learning does not directly result from higher cumulative output but is a result of systematic allocation of engineering resources to problem solving activities.

Terwiesch and Bohn (2001) distinguish learning through experimentation from learning by cumulative production. They formulate a dynamic programming model to decide when to experiment and when to produce but do not consider manufacturing lead times or the effects of congestion. Kim and Uzsoy (2008; 2013) propose integrated production planning models using clearing functions that incorporate learning effects by considering production and engineering lots separately. Production lots are sold to generate revenue, whereas engineering lots help increase capacity in future periods. They perform a marginal cost analysis to provide insights on managing the system. However, they do not explore the problem of ramp up and new product introduction. The idea of distinguishing production and engineering lots is applied to a full fab model by Ziarnetzky and Mönch (2016), who give three production planning formulations with fixed lead times that incorporate learning and different capacity allocation scenarios for the engineering and production lots.

A number of simulation models have studied the behavior of wafer fabs during product transitions under changing product mix (Nemoto et al. 2000; Dümmler 2000; Klein and Kalir 2006). Using a scaled down version of a wafer fabrication facility, Crist and Uzsoy (2011) use simulation to study the impact of several different policies for allocating resources to production and engineering work.

This work uniquely combines learning by doing and learning from experimentation in a product transition from an older product to a newer one. Apart from the question of how to handle the releases of production lots of the old and the new product, additional questions about how to manage experimentation and balancing the two learning sources is addressed in this work.

3 SIMULATION MODELLING AND ANALYSIS

As in our earlier work (Manda and Uzsoy 2018a), we begin by considering a simplified production system consisting of a single resource. This is clearly not representative of any actual wafer fab, but rather an attempt to examine the system dynamics induced by the combination of product transition and different modes of learning in an environment sufficiently simple to allow controlled experimentation and insight. At the beginning of the time horizon this resource is producing an older product with an initially stable demand. The processing time distribution of this product, henceforth referred to as $P1$, is assumed to have a mean of t_0 minutes and a standard deviation of σ_0 minutes. We assume that the production process of $P1$ has been thoroughly debugged, inducing disruptions requiring engineering action at a low, stable rate. A production lot of $P1$ induces an event requiring engineering activity on average once every Q_s lots. The duration of the engineering activity, and hence the production time lost, is a random variable with a mean of P minutes and standard deviation of σ_p minutes. At some point in the planning horizon, demand for the old product $P1$ starts to decrease and is replaced by demand for a new product, which we shall denote by $P2$. Without loss of generality the new product is assumed to have to the same processing time distribution (t_0, σ_0) and engineering activity duration distribution (P, σ_p) as the old product. However, this new product induces disruptions that require engineering activities much more frequently, on average every $Q_0 < Q_s$ lots. As experience is gained with the product, which involves finding and fixing problems that require engineering interventions, the frequency of the disruptions decreases, eventually reaching a steady state value Q_s. This represents reactive learning as learning occurs with production experience.

In this work, we extend the previous model with an additional mode of learning through deliberate experimentation. This is achieved by releasing engineering lots specifically designed to implement and test potential engineering improvements. When an engineering lot is released into the system and reaches the resource, its capacity consumption is significantly higher than that of a regular production lot, due to special processing requirements such as the presence of engineering personnel to run the lot, equipment configuration and preparation. It does not produce any revenue generating output like a regular production lot. However, completing an engineering lot induces more learning than a production lot, reducing the probability of disruption for subsequent production lots. We assume an engineering lot has a processing time distribution with a mean of E minutes and standard deviation of σ_E.

The learning effects of the new product production and engineering lots on the average number of lots between disruptions are then given by

$$Q_{2t} = Q_0 + (Q_s - Q_0)(1 - e^{-\alpha X_p(t) - \beta X_e(t)}) \tag{1}$$

where Q_{2t} denotes the average number of lots between disruptions for the new product $P2$ at time t, $X_{cumm}(t)$ the total number of production lots of $P2$ produced until time t, $X_e(t)$ the cumulative number of engineering lots of $P2$ produced till time t, and $X_p(t)$ the cumulative number of production lots of $P2$ produced up to time t.

The planning horizon over which the simulation optimization is performed consists of J discrete periods each of length T (equal to three months in our numerical experiments). The structure of the queuing system is presented in Figure 1. Three types of lots are released into the fab; production lots of $P1$ and $P2$, and engineering lots of the new product $P2$. The number of lots of a given type of lot i to be released into the fab in period t is a decision variable denoted by R_{it}. In the simulation the mean arrival rate of lots of type i in period t is given by

$$\lambda_{it} = \frac{R_{it}}{T} \ \ \forall i \in (1, 2, E) \tag{2}$$

When a production lot of either product is released and arrives at the resource it induces a requirement for engineering intervention with a probability of p_{it}. For the old, stable product, this value is constant and given by $p_{1t} = \frac{1}{Q_s}$. The probability that a production lot of the new product $P2$ will induce an engineering disruption is $p_{2t} = \frac{1}{Q_{2t}}$, where Q_{2t} is given by (1). Therefore, as more engineering and production lots

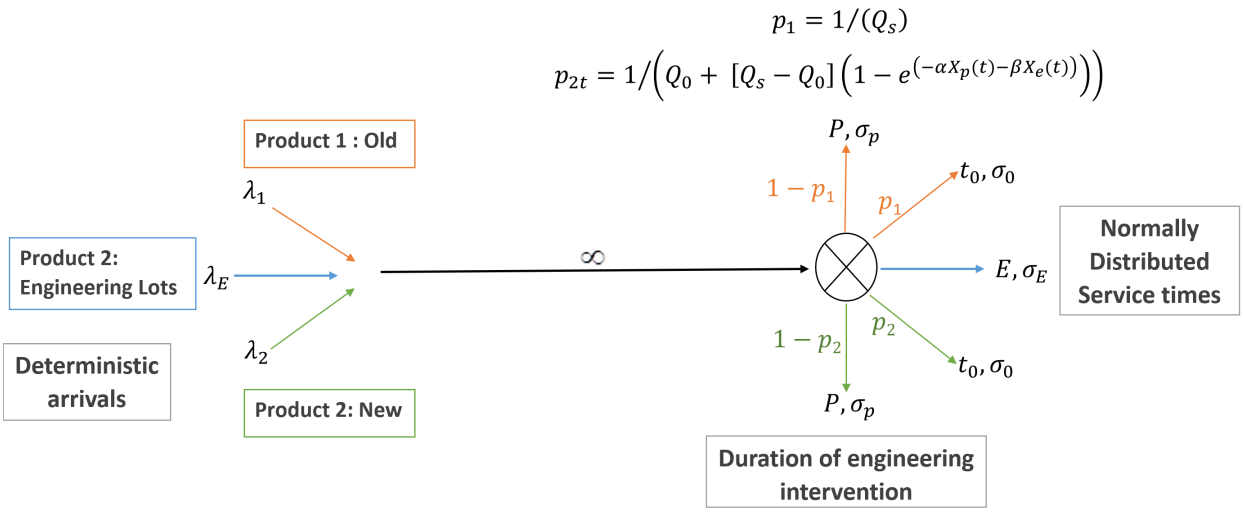

Figure 1: Schematic of queuing system.

are produced in the fab, the probability of an engineering disruption decreases. A production lot of either product undergoes regular production With a probability of $1 - p_{it}$ as depicted in Figure 1. At the end of the simulation, we compute the value of total contribution over the entire time horizon.

The simulation takes as inputs the values of the processing parameters (such the mean and variance of the different service and disruption distributions), the various costs (backorder and inventory holding costs) and a release vector specifying the number of production lots of each product and engineering lots of $P2$ released in each planning period. Each simulation call returns the expected total contribution for a given release vector. This simulation is used as the fitness function for a Genetic Algorithm when implementing the simulation optimization.

4 Simulation Optimization

We now present a simulation optimization procedure using the simulation model from the previous section. The model seeks a release vector that maximizes the expected total contribution over the planning horizon of J periods. The problem can be stated as:

$$\max \quad E\left[\sum_{i,t}[\pi_{it}(D_{it} + S_{it-1} - S_{it}) - (r_{it}R_{it} + i_{it}I_{it} + s_{it}S_{it})]\right]$$

s.t.

$$R_{it} \geq 0 \ \forall \ \ i \in (1,2,E); t \in J$$

where,

I_{it} Finished Goods Inventory (FGI) of product i at the end of period t, with unit cost of i_{it}.
R_{it} Total releases of product i into the system during period t, with unit cost of r_{it}.
S_{it} Total backorders of product i at the end of period t, with unit cost of s_{it}.
D_{it} Demand of product i in period t, with a unit revenue of π_{it}.

As in our previous work, we use a genetic algorithm (GA) as our simulation optimization engine. The fitness function for the GA is the simulation model which outputs the average total contribution for a given

Table 1: Demand data.

Demand	1	2	3	4	5	6	7	8	9	10
Product 1	1250	1250	1250	937.5	625	312.5	0	0	0	0
Product 2	0	0	0	312.5	625	937.5	1250	1250	1250	1250

Table 2: Parameter values.

Parameters	J	T	t_0	c_0	P	c_p	$Q_{1s} = Q_{2s}$	Q_{20}	α	β
Values	10	129600	80	0.25	800	0.5	50	10	0.0001	0.004/0.001

Table 3: Contribution and Cost Parameters.

Parameters	π_1	π_2	s_1	s_2	r_1	r_2	r_e	i_1	i_2
Values	5	7.5	8	8	1	1	2	3	3

release vector over multiple simulation replications. Due to the nature of the simulation, with planned and unplanned interruptions and variance in processing and disruption times, the total contribution value for a single simulation run is quite variable, resulting in wide confidence intervals. To obtain smaller confidence intervals, for each evaluation of the fitness function the mean value of 20 simulation runs is returned as an estimator of the expected total contribution value. To further reduce the size of confidence intervals an antithetic variates variance reduction approach was implemented for each replication of the simulation. The resulting final confidence interval of the expected total contribution are reported in the numerical examples performed in section 5.

A challenge in implementing the simulation optimization procedure using a GA is its slow convergence when using a completely random initial population. Each individual represents a release vector consisting of the release quantities of production lots for $P1$ and engineering and production lots of $P2$ in each planning period. Due to the nonconvex nature of the problem, random starting populations result in inconsistent solutions, with the GA reporting local minima instead of approaching a global minimum. To achieve quicker and more consistent convergence, the initial population of the GA is seeded with good starting vectors for production lots obtained by solving a relaxation of a deterministic optimization model Manda and Uzsoy (2018a) consistent with the simulation model discussed. We also seeded the initial population with release vectors obtained by solving the optimization model for random perturbations of the demand vector Manda and Uzsoy (2018a). However, since there is no explicit demand for engineering lots of $P2$ we randomly generate the number of engineering lots in each period in the initial population.

The `ga` function of the MATLAB Global Optimization Toolbox (2016) was used to implement the genetic algorithm. The total population size was set to 100. Of these, 10 good starting vectors were included in the initial population and the rest were generated randomly. The elite count was set to 10 and the crossover fraction was set to 0.6. This means that of the 90 remaining individuals, on average $(0.6)(90) = 54$ are from crossover children and the remaining 36 are generated through mutation. The mutation, crossover, migration, scale and shrink options were left at their default values. The default mutation function, `Gaussian`, adds a random number taken from a Gaussian distribution with mean 0 and standard deviation derived from the scale and shrink parameters to each entry of the parent vector. The default crossover function, `Scattered`, creates a random binary vector and selects the genes where the entry is a 1 from the first parent, and those whose entry is 0 from the second parent to form the child. We refer the reader to the Matlab global optimization toolbox documentation for a detailed explanation of the rest of the default functions and values of the Matlab `ga` parameters.

5 NUMERICAL EXPERIMENTS

To examine the performance of the model developed in this paper we report several numerical experiments. A product transition problem is considered with a planning horizon of 10 periods. As each planning period is 3 months long, the total planning horizon is two and half years in length. At the beginning of the planning horizon, in period 1, demand exists only for $P1$. Starting in period 4, the demand for $P1$ decreases and the demand for $P2$ increases until there is only demand for $P2$ starting in period 7. The values of the demand are given in Table 1.

Each production lot has a mean natural processing time of $t_0 = 80$ minutes with coefficient of variation $c_0 = 0.5$. The new product $P2$ causes a disruption on average every $Q_{20} = 10$ lots when it is first introduced into the production system. This number increases to $Q_{2s} = 50$ lots by a combination of engineering lots and production lots of the product. The increase in the average number of units between a disruption follows Equation 1. Unlike our previous work ((Manda and Uzsoy 2018b)) which does not model engineering lots, this model can reduce the disruption frequency of the new product $P2$ without producing unnecessary production lots if engineering lots provide an performance advantage.

The mature product $P1$ causes a disruption on average every $Q_{1s} = 50$ lots. During the initial periods, if production consists solely of $P1$, the average utilization will be approximately 89%. The utilization changes as engineering and production lots of $P2$ are introduced (which cause greater frequency of disruptions). Each disruption lasts an average of $P = 800$ minutes with coefficient of variation $c_p = 1$. Each engineering lot of $P2$ has a mean processing time E of 800 minutes with coefficient of variation $c_0 = 1$. Table 2 summarizes the values of the parameters.

The old product $P1$ generates a revenue of 5 per unit. The new product, on the other hand, generates 50% more revenue when it is first introduced into the system. This revenue of 7.5 per unit decreases each period by 6% until it reaches the same revenue of 5 per unit as $P1$. This revenue structure mimics the high prices commanded by a new product early on in its life cycle and the subsequent drop in prices with time. The unit backorder cost is assumed to be 8 for both products and the finished goods inventory holding cost to be 3. The cost of releasing a production lot (material cost) is set to 1 and that of an engineering lot to 2. An engineering lot does not generate any revenue or incur any backorder or holding cost. Table 3 summarizes the values of the parameters.

A problem we faced with the experimental design is the lack of clear benchmark policies with which to compare the solutions. Therefore, we decided to perform experiments with a high and low values of β relative to the α value. A high β value represents the case where engineering lots are attractive due to the high boost in learning they offer even though they do not provide any revenue and block production as they are processed. A low β value in comparison to α value makes engineering lots less attractive. A, extremely low β value reduces the value of engineering lots to a point where engineering lots yield no benefit over production lots and the problem becomes that of managing only the production lots as addressed in our previous work. For the first experiment we choose the β value to be 40 times greater than the α value. This means that a engineering lot of $P2$ produces 40 times more learning than a production lot of $P2$. For the second experiment we set β equal to 10 times the α value. Experiments have revealed to us that around this point, with the current setup, engineering lots are no longer attractive.

5.1 Experiment 1: High Engineering Learning

For this experiment we use $\beta = 0.004$ and a $\alpha = 0.0001$, making an engineering lot 40 times more effective than a production lot at learning. The average total contribution values for the three GA runs are given in Table 4. The results show that the different GA runs are in close agreement with each other, as indicated by the confidence interval on the expected total contribution.

Examining the engineering lot releases given in Figure 2, we observe that even though there is some variance in the number of engineering lots released in each GA run, a common theme is the lack of any engineering lot releases in the later planning periods (periods 6 to 10). This shows that the engineering

Table 4: Total Contribution: High β.

	GA Run1	GA Run2	GA Run3
Mean	52713	52682	51208
95% CI	[53398, 52028]	[52968, 52397]	[52362, 50055]

lots are being used early in the life cycle of the new product($P2$) to boost learning. At a certain point after production lots of $P2$ are introduced and enough learning has been accumulated, the marginal value of an engineering lot approaches zero and thus we do not see any engineering lot releases. The marginal value of an engineering lot remains constant before there is any demand (and thus production) of $P2$ and therefore from the perspective of the objective function, engineering lots can be produced in any of these periods without changing the final total contribution value. This explains the variance in the number of engineering lot releases between each GA run in Figure 2.

The narrow confidence intervals and the agreement among the total contribution values between the different GA runs is a result of the high amount of learning achieved early in the planning horizon from the engineering lots. The production of the engineering lots increases Q_2 value following (1). When the production lots of $P2$ are introduced into the system, they are less likely to induce disruptions and thus the total revenue generated is not subject to high variability.

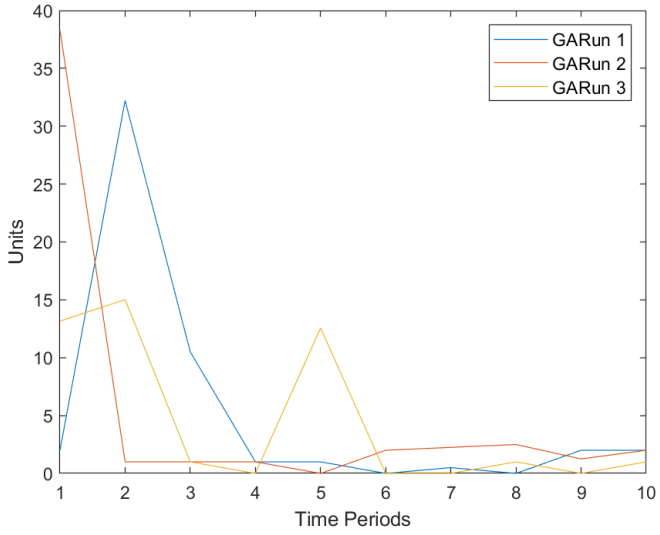

Figure 2: Engineering Lot Releases for High β.

5.2 Experiment 2: Low Engineering Learning

For this experiment we use a $\beta = 0.001$ and a α value of 0.0001, making an engineering lot 10 times more effective at learning than a production lot. The average total contribution values are given in Table 5. We can make two observations from the expected total contribution values and the confidence intervals. Firstly, the mean total contribution values are lower than under the previous case of high β. This is consistent with expectations, as the engineering lots are now less effective to the point where the disruption in processing and the loss in revenue they induce is greater than the benefit they provide. We can observe this clearly in Figure 4 where, in none of the time periods of the three GA runs, the engineering lots released exceed two lots. These very low engineering lot releases can be attributed to noise in the simulation results.

The other observation that can be made from Table 5 is the higher variance in the results of the GA runs and the larger confidence intervals when compared to the results of experiment 1. At first sight, the

Table 5: Total Contribution: Low β.

	GA Run1	Ga Run2	GA Run3
Mean	44642	50486	44987
95% CI	[47344 41939]	[51456 49516]	[47203 42771]

differences between Tables 4 and 5 seem counterintuitive; almost no engineering lots are processed in the low β case shown in Figure 3, so we would expect a lot less variability in this case than in the high β case in Figure 2, where more engineering lots are being processed. The answer is apparent after examining the solutions in the two cases. When β is high, the model releases engineering lots of the new product early in the horizon, before demand for the new product is substantial. Since most of the learning takes place early in the horizon, the new product induces far fewer engineering disruptions later in the horizon when demand for the new product is high, resulting in far more variability in the later periods. When β is low, in contrast, few engineering lots are released, since their benefit to learning is outweighed by the burden imposed on the system by their additional processing time. Hence the majority of learning must take place via the production lots, resulting in much slower learning and hence more engineering disruptions induced by the new product. This creates additional variability in production compared to the high β case, explaining the results in Figure 3.

Without the engineering lots to boost learning, the model uses releases to counter the effects of new product introduction and increased disruptions. These results are in line with our previous work. As can be seen from Figure 3, the optimal solutions in all three GA runs increase production of $P1$ and build up inventory in the initial periods. The model also tried to increase learning by introducing a small number of lots of $P2$ during these early periods. However, a lack of demand and high backorder cost prevent it from releasing more units of $P2$. During the transition period, the releases of $P2$ are increased and built up inventory of $P1$ is utilized to meet demand. This increase in releases of $P2$ helps in maintaining output of the product and in increasing learning. This is the exact pattern of behaviour observed in our previous work with no engineering lots and thus these results are consistent with our expectations.

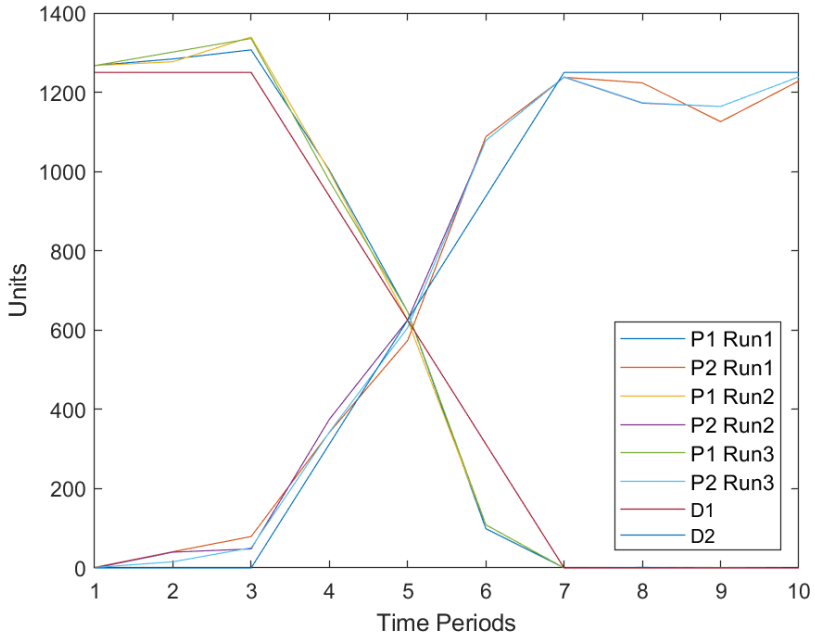

Figure 3: Production Lot Releases for Low β.

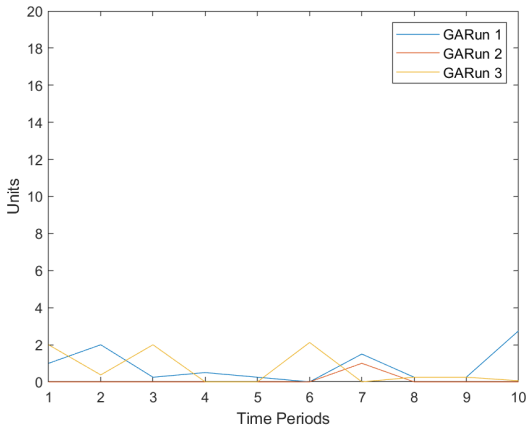

Figure 4: Engineering Lot Releases for Low β.

6 CONCLUSIONS AND FUTURE WORK

While the results presented in this paper are preliminary and exploratory in nature, a number of interesting conclusions can already be identified. Earlier work seeking to model the interaction of production and engineering lots while explicitly representing congestion (Kim and Uzsoy 2008; Kim and Uzsoy 2013) has focused on the impact of learning on the mean effective processing time alone, without explicit consideration of the impact of learning on its variance. Once the impact of learning on processing time variance is explicitly considered, making the learning rate endogenous to the model rather than an exogenous parameter, deterministic optimization models become highly non-convex (Manda et al. 2016), requiring the use of metaheuristics to obtain approximately optimal solutions. The inherently stochastic nature of the underlying problem, which is subject to many uncertainties in both shop floor events and the amount of learning realized from production or engineering lots will eventually require a solution to a complex stochastic optimization problem. The non-convex structure of the deterministic models suggest that approaches such a stochastic programming or robust optimization are unlikely to be computationally tractable, leaving simulation optimization as the principal alternative. This approach is explored in this paper, using a simple genetic algorithm as the simulation optimization engine.

Our results give a number of interesting insights. The results in Experiment 1 and 2 clearly illustrate the the ill effects of slow production learning and lack of effective engineering activity. They show how effectively designed engineering lots that can provide boosts in learning can be strategically used early in the life cycle of a product to decrease disruptions, increase total contribution and reduce variance in the long run, even at the expense of temporary revenue and capacity. They also show that an engineering lot requires to be able to boost learning by a certain level before it can be viable. These simple experiments at a high level illustrate the basic intuitive understating of how proactive engineering activity can be more beneficial than focusing engineering effort on reacting to problems that occur randomly.

The results we have presented are stylized and optimistic in several ways. Learning from engineering and production lots is always positive; there is no probability of zero learning, or of forgetting. Our exponential model of learning posits diminishing returns to additional learning effort over time, which is probably realistic at a high level, but calls into question the specific form of learning curve used. Most importantly, there are many different dimensions in which learning takes place on the shop floor, ranging from processing recipes and equipment adjustments to managerial practices and employee training which are all difficult to model in detail. Godinho Filho and Uzsoy (2013) and Wu (2013) present models of processing times showing that many different factors can contribute to improvements in effective processing times. Our model focuses on the adverse effects of new product introduction on the throughput of the

different products, but does not examine detailed models of the yield of good devices thus obtained. Finally, the choice of what engineering lots to run at what point in the life cycle, i.e., what specific processing steps to target and in what order, are important questions our model does not address. In particular, our model assumes all engineering lots are statistically the same in their effect on the process. In reality, the results of one engineering lot are likely to influence the design of the next one, limiting the ability of management to learn meaningfully from a large number of engineering lots released early in the life cycle.

In this paper we have aggregated all of these into one mechanism, the frequency of engineering holds induced by the new product relative to the older, more mature ones. Clearly at any point in time learning is continuing on older products; some improvements made for the new product will also improve the processing times of older products in different ways. Detailed modelling of all these issues is likely to be challenging, requiring major data analytics efforts to identify the mechanisms and estimate the parameters involved. The simple model we have presented appears to capture the high-level dynamics of the problem addressed, and opens the possibility for refining it in future work.

Several directions for future work are immediately apparent. The extension of this work to a multistage system, where engineering effort can be reallocated over time as learning takes place and bottlenecks shift is an important direction. While Theory of Constraints would suggest exclusive focus on bottlenecks, it is quite possible that once the bottlenecks have been improved to a high degree, improvement at non-bottlenecks will become desirable in order to improve material flow into the bottleneck, as suggested in Kefeli and Uzsoy (2016). Our results in this paper suggest that developing an effective simulation optimization approach for this problem is likely to be challenging. The simple genetic algorithm we have used is challenged by the high variability in system performance measures we encounter. A more efficient, gradient-based approach may provide a more effective approach.

REFERENCES

Adler, P. S., and K. B. Clark. 1991. "Behind the Learning Curve: A sketch of the Learning Process". *Management Science* 37(3):267–281.

Anzanello, M. J., and F. S. Fogliatto. 2011. "Learning Curve Models and Applications: Literature Review and Research Directions". *International Journal of Industrial Ergonomics* 41(5):573–583.

Chand, S., H. Moskowitz, A. Novak, I. Rekhi, and G. Sorger. 1996. "Capacity Allocation for Dynamic Process Improvement with Quality and Demand Considerations". *Operations Research* 44(6):964–975.

Crist, K., and R. Uzsoy. 2011. "Prioritising Production and Engineering Lots in Wafer Fabrication Facilities: A Simulation Study". *International Journal of Production Research* 49(11):3105–3125.

Dummler, M. A. 2000. "Analysis of the Nonstationary Behavior of a Wafer Fab During Product Mix Changes". In *Proceedings of the 2000 Winter Simulation Conference*, edited by J. A. Joines, R. R. Barton, K. Kang, and P. A. Fishwick, Volume 2, 1436–1442. Orlando, Florida: Institute of Electrical and Electronics Engineers, Inc.

Dutton, J. M., A. Thomas, and J. E. Butler. 1984. "The History of Progress Functions as a Managerial Technology". *Business History Review* 58(2):204–233.

Fine, C. H. 1986. "Quality Improvement and Learning in Productive Systems". *Management Science* 32(10):1301–1315.

Godinho Filho, M., and R. Uzsoy. 2013. "The Impact of Continuous Improvement in Setup and Repair Time on Manufacturing Cycle Time Under Uncertain Conditions". *International Journal of Production Research* 51(2):447 – 464.

Hatch, N. W., and D. C. Mowery. 1998. "Process Innovation and Learning by Doing in Semiconductor Manufacturing". *Management Science* 44(11-part-1):1461–1477.

Hirsch, W. Z. 1952. "Manufacturing Progress Functions". *The Review of Economics and Statistics*:143–155.

Jaber, M. Y. 2006. "Learning and Forgetting Models and Their Applications". *Handbook of Industrial and Systems Engineering* 30(1):30–127.

Kefeli, A., and R. Uzsoy. 2016. "Identifying Potential Bottlenecks in Production Systems Using Dual Prices from a Mathematical Programming Model". *International Journal of Production Research* 54(7):2000–2018.

Kim, S., and R. Uzsoy. 2013. "Modeling and analysis of integrated planning of production and engineering process improvement". *IEEE Transactions on Semiconductor Manufacturing* 26(3):414–422.

Kim, S., and R. M. Uzsoy. 2008. "Integrated Planning of Production and Engineering Process Improvement". *IEEE Transactions on Semiconductor Manufacturing* 21(3):390–398.

Klein, M., and A. Kalir. 2006. "A Full Factory Transient Simulation Model for the Analysis of Expected Performance in a Transition Period". In *Proceedings of the 2006 Winter Simulation Conference*, edited by

L. F. Perrone, F. P. Wieland, J. Liu, B. G. Lawson, D. M. Nicol, and R. M. Fujimoto, 1836–1839. Monterey, California: Institute of Electrical and Electronics Engineers, Inc.

Leachman, R. C., and D. A. Hodges. 1996. "Benchmarking semiconductor manufacturing". *IEEE transactions on semiconductor manufacturing* 9(2):158–169.

Liao, W. M. 1979. "Effects of Learning on Resource Allocation Decisions". *Decision Sciences* 10(1):116–125.

Manda, A. B., and R. Uzsoy. 2018a. "Optimizing Releases During New Product Introductions". Technical report, Edward P. Fitts Department of Industrial and Systems Engineering, NCSU, Raleigh, NC 27519.

Manda, A. B., and R. Uzsoy. 2018b. "Simulation Optimization for Planning Product Transitions in Semiconductor Manufacturing Facilities". In *2018 Winter Simulation Conference (WSC)*, edited by M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, 3470–3481. Gothenburg, Sweden: Institute of Electrical and Electronics Engineers, Inc.

Manda, A. B., R. Uzsoy, K. G. Kempf, and S. Kim. 2016. "Modeling the Impact of New Product Introduction on the Output of Semiconductor Wafer Fabrication Facilities". In *Proceedings of the 2016 Winter Simulation Conference*, edited by T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick, 2547–2558. Arlington, Virginia: Institute of Electrical and Electronics Engineers, Inc.

MATLAB 2016. *Global Optimization Toolbox*. R2016a ed. Natick, Massachusetts: MATLAB. https://www.mathworks.com/help/gads/.

Nemoto, K., E. Akcali, and R. M. Uzsoy. 2000. "Quantifying the Benefits of Cycle Time Reduction in Semiconductor Wafer Fabrication". *IEEE Transactions on Electronics Packaging Manufacturing* 23(1):39–47.

Preston, L. E., and E. C. Keachie. 1964. "Cost Functions and Progress Functions: An Integration". *The American Economic Review* 54(2):100–107.

Reeves, G. R., and J. R. Sweigart. 1981. "Product-Mix Models when Learning Effects are Present". *Management Science* 27(2):204–212.

Searle, A. D. 1945. "Productivity Changes in Selected Wartime Shipbuilding Programs". *Monthly Labor Review* 61:1132–1147.

Terwiesch, C., and R. E. Bohn. 2001. "Learning and Process Improvement During Production Ramp-up". *International Journal of Production Economics* 70(1):1–19.

Tirkel, I. 2013. "Yield Learning Curve Models in Semiconductor Manufacturing". *IEEE Transactions On Semiconductor Manufacturing* 26(4):564–571.

Webbink, D. W. 1977. "The Semiconductor Industry: A Survey of Structure, Conduct, and Performance". *NASA STI/Recon Technical Report N* 78.

Wright, T. P. 1936. "Factors Affecting the Cost of Airplanes". *Journal of the Aeronautical Sciences* 3(4):122–128.

Wu, K. 2013. "Classification of Queuing Models for a Workstation with Interruptions: A Review". *International Journal of Production Research* 52(3):902 – 917.

Yelle, L. E. 1979. "The Learning Curve: Historical Review and Comprehensive Survey". *Decision Sciences* 10(2):302–328.

Ziarnetzky, T., and L. Mönch. 2016. "Incorporating Engineering Process Improvement Activities into Production Planning Formulations Using a Large-Scale Wafer Fab Model". *International Journal of Production Research* 54(21):6416–6435.

AUTHOR BIOGRAPHIES

ATCHYUTA BHARADWAJ MANDA is a Doctoral student in the Edward P. Fitts Department of Industrial and Systems Engineering at North Carolina State University. He holds a Masters in Industrial and Systems Engineering from North Carolina State University, and a Bachelor of Technology degree in Mechanical Engineering from GITAM University. His research interests are in production systems, supply chain management and new product introduction, focusing on data oriented decision analysis using stochastic simulation. His e-mail address is amanda@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.