Proceedings of the 2016 Winter Simulation Conference
T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick, eds.

ROBUST SEMICONDUCTOR PRODUCTION PLANNING UNDER YIELD UNCERTAINTY

Jonathan J. Lowe
Amin Khademi
Scott J. Mason

Department of Industrial Engineering
Clemson University
Freeman Hall
Clemson, SC 29634, USA

ABSTRACT

Uncertainty throughout the semiconductor manufacturing process is both dependable yet inevitable. We present a robust optimization approach to production planning under uncertainty in semiconductor manufacturing. A sensitivity analysis was completed with a representative industrial dataset to verify those uncertain aspects that most affect our model: process yield out of wafer fabrication and test vendors. As the manufacturer can utilize multiple vendors for these steps, a case study was performed using actual industry data in which the number of vendors experiencing yield uncertainty was varied. Based on results, uncertainty in yield resulted in both higher first month costs and higher costs over the planning horizon; uncertainty at test results in greater cost increases than uncertainty at wafer fabrication, due to uncertainty occurring later in the supply chain. However, specific to our data, the cost of a robust solution is minimal when compared to demand lost under a deterministic solution.

1 INTRODUCTION

Production planning for semiconductor manufacturing, a mature industry, is still faced with challenges due to uncertainty. Bucy (1972) indirectly implies a mature semiconductor industry in a call for more growth and broader market bases. Peters (2005) states that the industry is becoming accustomed to its more mature compound growth rate. Although mature, uncertainty still exists; in pushing the leading edge of technology, the production process will continue to be affected by uncertainty. The costs of not considering uncertainty during the planning process could be considerable yet not fully known by manufacturers; for example, the costs may be masked through the use of inventory or not realized due to lost future sales. Barahona et al. (2005) state that by considering uncertainty via a stochastic programming approach, additional profits "on the order of tens of millions of dollars per year" could be possible from their specific real-life case.

Demand, capacity, cycle time, and yield are the most common uncertain facets of the planning process to be studied. The literature incorporating demand, capacity, and cycle time uncertainty is rich (see Section 1.2). However, few studies on semiconductor production planning consider the uncertainty of yield in the production process at the echelon level and if they do, it is assumed that the yield follows a known statistical distribution. Additionally, one manufacturer, from which our case study data came, employs sub-contractors (or "vendors") for the last two stages of production (assembly and final test). As these vendors can change frequently over time, the probability distributions are not available upfront and yields are estimated to be in a specific range. Therefore, we model production planning of semiconductor industry under yield uncertainty (at the echelon level) by a robust optimization approach. Our results from a real-life dataset 1) indicate that yield uncertainty can significantly change total cost, 2) show that under reduced

yields, customer satisfaction may be maintained through demand fulfillment, and 3) shed light on its effects in different phases of the production process.

1.1 Semiconductor Manufacturing Overview

The mass production of integrated circuits (ICs) can be segregated into four distinct phases: wafer fabrication (fab), sort (probe), assembly, and final test. Wafer fabrication is the most complex portion of the entire process. Wafer fabrication's complexity (for example, re-entrant flows, sequence- and part-specific setup times, unreliable equipment, and competition for capacity) contributes to uncertainty in process yield that is both dependable yet unpredictable.

After a wafer completes wafer fabrication, it proceeds to the sort (or probe) phase: the individual ICs on each wafer are tested for basic functionality. As in the wafer fabrication phase, there is uncertainty in the number of ICs that survive sort. After sort is complete, wafers go to die bank (i.e., the completed wafer work-in-process) where they may wait for the assembly and final test phases. Our definition of fab yield is based on the completion of sort: the product of line (or wafer) yield and sort yield (Cunningham 1990).

In the assembly phase, the wafer substrate is cut into individual ICs; failed ICs are discarded and functional ICs are then packaged and moved to the final test phase, at which point they are tested, rated, and ultimately date-stamped for inclusion in finished goods inventory. These last two phases are still complex, but not to the extent of the first two; assembly and final test deal with millions of devices versus the thousands of wafers handled in wafer fabrication and sort. Additionally, there is still uncertainty in IC yield; for example, some completed devices designed for a specific performance rating may fail testing at that rating, yet pass at a lower rating.

It has become standard to term the wafer fabrication and sort phases as the "front-end" with the assembly and final test phases being called the "back-end" (Uzsoy, Lee, and Martin-Vega 1992, Mönch et al. 2011). The front-end process is generally performed in-house; the back-end process is typically contracted to an outside vendor, and in general, located in lower-wage countries. The back-end vendors are engaged primarily on the basis of cost, then capacity; work they receive from fabs may vary from month to month based on prevailing market conditions and best available costs.

1.2 Semiconductor Planning Under Uncertainty Overview

Stochastic programming (SP) and dynamic programming have been the traditional methods for incorporating uncertainty in decision making. The underlying assumption in these methods is that probability distributions are available to the decision maker. The objective is then to find an optimal (or near-optimal) solution based on long-term effects of uncertainty. Unfortunately, this usually results in intractable models and answers that are appropriate for expected values, but not necessarily optimal under non-expected realizations.

In the SP arena, research specific to semiconductor planning has generally focused on demand uncertainty. Christie and Wu (2002) study strategic capacity planning with a multi-stage SP model in which demand and capacity estimates are uncertain. Barahona et al. (2005) use a two-stage SP model with demand uncertainty to identify the appropriate toolset to minimize unmet demand. At the operational level, Geng and Jiang (2007) also propose a two-stage SP model under demand uncertainty for capacity planning. Geng, Jiang, and Chen (2009) propose a SP model with demand and capacity uncertainty for capacity planning, but at a single fab. Finally, more recently, Fu et al. (2015) consider a stochastic mixed-integer linear program to minimize back-end machine qualifying costs and future back-order costs.

However, if it is important for a solution to remain feasible under all potential realizations or if data do not support a particular distribution, robust optimization (RO) may be a more appropriate tool. Robust optimization allows for data to not be specified exactly, but rather just requires that the data fall within an "uncertainty set." The resulting solution must satisfy any possible value of the data, provided data remain in that set. There are other advantages to RO other than the relaxed assumption of data fitting a distribution. For example, Ben-Tal and Nemirovski (1998) show that under specific uncertainty set forms,

the RO version of some generic convex optimization problems is still "either exactly, or approximately, a tractable problem" (Ben-Tal and Nemirovski, 1998, p. 769).

The work by Soyster (1973) on "inexact linear programming" is generally considered to be one of the earliest works to address data uncertainty. He showed that under a maximization objective function, a semi-infinite model results with constraints around any possible value (within a specified range) of the uncertain parameter. Soyster then proved that any solution will take the most conservative route in that the *supremum* of each uncertain parameter will be the binding value. Translated to a minimization problem, it is easy to see that the *infimum* of each uncertain parameter will be binding.

Ben-Tal and Nemirovski (1998) and Ben-Tal and Nemirovski (1999b) recognize that Soyster's approach addresses "column-wise" uncertainty. Under the assumption that it is a rare event for all uncertain parameters to take their worst value, they extend Soyster's work through the creation of uncertainty sets where only a specified quantity of uncertain parameters will be allowed to vary. Ben-Tal and Nemirovski (1999a) later show that as RO is applied to uncertain linear, conic quadratic, and semi-definite programs, computationally tractable robust versions can be realized.

Bertsimas and Sim (2004) propose an approach to allow some control over "the degree of conservatism for every constraint" (Bertsimas and Sim, 2004, p. 36). Their approach limits the number of coefficients within a constraint that may be allowed to vary. They introduce a parameter, Γ_i, that for every row i, specifies the number of coefficients on that row that will be allowed to vary within their allowed range. They further show that their method is at least as flexible as previous methods and their proposed robust versions are mixed-integer programs.

Semiconductor planning research utilizing RO may be less abundant, but does address uncertain aspects other than just demand. Ng and Fowler (2007) introduce a RO approach for the co-production planning problem (in which a single wafer could ultimately yield more than one SKU of finished product, either intentionally through direct down-configuration or unintentionally through binning). They use uncertainty in demand and bin fractions to schedule wafer starts; once uncertainty is resolved, production is allocated to satisfy demand (including product substitution). Later, Ng, Sun, and Fowler (2010) extend this work to lot allocation in order to meet demand. Variability in lot sizes are modeled with ellipsoidal uncertainty sets. However, both of these works are at the fab echelon and operational level.

Given the maturity of the industry, data should be available to reasonably fit statistical distributions to yield. But, as Kumar et al. (2006) have noted, the large amounts of data and complicated processes make yield difficult to quantify. With multiple assembly and test vendors from which to choose, it is highly unlikely that these distributions will be identical at the unit level across these locations; the lack of uniformity in yield distribution motivates the use of RO. However, we assume that if a unit's yield is allowed to vary at a specific location, it will be allowed to vary in all periods over the planning horizon at that location. While future research will address relaxing this assumption, it can be justified by considering that the vendor-reported yield either does not reflect a systematic (yet unknown) issue or an over-estimate of their capability. The impact of this assumption is that temporary reductions in yield (for example, affecting one period or one unit over the horizon) are not modeled; yet, these temporary reductions may be absorbed by the manufacturer's use of safety stock. Therefore, our contribution is to illustrate the use of robust optimization at a tactical level, across the entire supply chain, and to consider uncertainty other than demand.

In Lowe and Mason (2016), a deterministic production planning model is proposed to schedule fab, assembly, and test starts over a defined horizon (typically six months), driven by demand forecasts over that horizon. Following industry practices, inventory is maintained at each echelon, allowing for specification of minimum levels ("safety stock") at the unit level. Limited capacity at each echelon is also modeled, as are costs associated with qualifying back-end facilities. Forecasted demand will always be met, either through production or through the use of penalty quantities (to insure model feasibility and to represent demand not met through production—but with extremely high unit costs in order to encourage production).

2 ROBUST OPTIMIZATION MODEL

We have followed the approach detailed by Bertsimas and Sim (2004), but first need to determine if our solutions were significantly affected by yield uncertainty; we additionally explored the effects of variability in cycle time and capacity. A sensitivity analysis using the model developed in Lowe and Mason (2016) was performed to determine which parameters (yield, cycle time, and capacity) have the most effect on model results. Forecasts from two months were used to generate test data; with nine months of forecasts available from which to choose, the months with the greatest total demand, July, and least total demand, August, over the horizon were chosen; we chose these two datasets because they appropriately represent the "endpoints" (or "extremes") from the available forecasts. The 20, 50, 100, and 250 most demanded devices ("device-quantity") were selected from forecasts with over 3000 SKUs. Demand for these device quantities comprise a majority of each forecast's demand: the top 20 most demanded devices account for 63.5% of July's forecasted demand and 56.7% of August's forecasted demand. Demand coverage increases such that the top 250 most demanded devices account for almost 89% of July's total demand (and over 87% of August's total demand). Capacities at assembly and test were varied within $[75,125]\%$ in increments of 5%. Cycle times at assembly and test were varied ± 3 periods (in one period increments), with a floor of one period. Yields at both assembly and test were varied $[-25,-5]\%$ in 5% increments. Test yields were varied up by 5.0% and 10.0%, with a ceiling of 1.0; assembly yields, already defaulted to 0.99, were varied up to 0.999 and 1.0.

Including a baseline case (in which there are no variations on yield, cycle time, and capacity) for each device-quantity and month combination, a total of 376 test cases were run using AMPL and Gurobi, with runs limited to two hours or within a relative optimality gap of 0.0001%. First month costs were used for comparison as schedules are updated on a monthly basis (coinciding with forecast updates). For each month and device-quantity, a ratio of those costs to baseline costs was calculated. Ratios were then averaged across the months to determine the average change from baseline costs.

Total first month costs (production costs at fab, assembly, and test plus holding costs at Die Bank, Test WIP, and Finished Goods) displayed expected results. Costs were inversely affected by yield changes: as yields increased (decreased), costs decreased (increased). This is reasonable in that as more (less) units result in an echelon, fewer (more) additional units need to be produced. Costs were positively correlated to changes in cycle time: as cycle times increased (decreased), costs increased (decreased). With an increase to a unit's cycle time, production for that unit will have to occur earlier in the horizon, pushing that production into earlier periods; with a decrease in cycle time, production can be pushed later into the horizon. Finally, cost changes due to changes in capacity showed the least sensitivity, with average ratios within $\pm 2\%$ of the baseline. Consequently, the model is not as sensitive to capacity changes as to the other parameters.

The results show that the model is found to be sensitive to changes in yield. Therefore, yield uncertainty will be addressed, via robust optimization, using actual data provided by a manufacturer. Assembly yields are not reported by the manufacturer; considering the small, symmetrical range available around the default value of 0.99, assembly yields will not be considered as "uncertain." Yields at fab are also not reported, but the default value of 0.90 (as anecdotally reported by the manufacturer) allows a reasonable range such that fab yield uncertainty will be included. Finally, test yields are reported by the manufacturer and across devices, show enough variation and range such that test yield uncertainty will also be included. We have changed our model as follows; only those changes to our model will be detailed. See Lowe and Mason (2016) for a complete treatment of the deterministic model.

Three sets of changes were made to address: 1) yield uncertainty in test, 2) yield uncertainty in fab, and 3) yield uncertainty simultaneously in test and fab. For all three scenarios, parameter Γ_{ij} is introduced to control the number of locations, i representing the period and j in the appropriate set of fab or test locations, in which yield is allowed to vary. In this manner, we are emulating production changes at a location that has affected all production at that site; future research will address changes at the unit level (e.g. the introduction of improved processes or unexpected downtime on a specific machine).

2.1 Notations

First, we introduce the notation used in the model:

$D =$ Set of Die IDs $- d \in D$ | $F =$ Set of Front-end (Fab) locations $- l \in F$

$P =$ Set of Package IDs $- p \in P$ | $A =$ Set of Assembly locations $- l \in A$

$V =$ Set of Device IDs $- v \in V$ | $T =$ Set of Test locations $- l \in T$

$N =$ Periods of demand $- i \in N$

Additionally, indices j and k are used in the model as a working index and to indicate the step in the process. Further, $\boldsymbol{F}$, $\boldsymbol{A}$, and $\boldsymbol{T}$ are used to indicate production echelons: **F**ab, **A**ssembly, and **T**est. Inventory echelons are indicated by $\boldsymbol{B}$, $\boldsymbol{W}$, and $\boldsymbol{G}$ (Die **B**ank, Test **W**IP, and Finished **G**oods). Following are the inputs and parameters used in the model:

C^M Maximum overall cycle time, regardless of echelon (periods)

C^T Maximum Test cycle time (periods)

C^A Maximum Assembly cycle time (periods)

C^F Maximum Fab cycle time (periods)

δ Number of periods over which demand should be averaged (periods)

γ_{iv} Demand forecast for period i, $i = C^M, \ldots, (N + C^M + C^A + C^T + \delta + 1)$ of device $v \in V$
Only N periods are actually used; the remaining are necessary due to differences in cycle times for each echelon (devices)

u_{iv} Quantity in period i of device $v \in V$ expiring (devices)

$$\varepsilon_{dp} = \begin{cases} 1 & \text{if Die } d \in D \text{ can be used for Package } p \in P, \\ 0 & \text{otherwise} \end{cases}$$

$$\theta_{pl} = \begin{cases} 1 & \text{if location } l \in A \text{ is qualified to assemble Package } p \in P, \\ 0 & \text{otherwise} \end{cases}$$

$$\rho_{vl} = \begin{cases} 1 & \text{if location } l \in T \text{ is qualified to test Device } v \in V, \\ 0 & \text{otherwise} \end{cases}$$

$$\tau_{kl} = \begin{cases} 1 & \text{if Assembly } k \in A \text{ can send work to Test } l \in T, \\ 0 & \text{otherwise} \end{cases}$$

$$\psi_{ld} = \begin{cases} 1 & \text{if Fab } l \in F \text{ is qualified to produce Die } d \in D, \\ 0 & \text{otherwise} \end{cases}$$

e_{jlk} Yield of unit j ($j \in D, P, V$ appropriately) out of echelon k ($k =$ Fab, Assembly, or Test) at location l ($l \in F, A, T$ appropriately) (%)

c_{jlk} Integer cycle time in number of periods to completion of unit j in echelon k at location l (periods)

t_{kl} Number of periods required to transport (or transfer) units from location $k \in F$ or A to location $l \in A$ or T (periods)

g_d Maximum number of good die $d \in D$ physically capable per wafer (die/wafer)

Finally, pertinent decision variables used in the model are listed:

x_{idl} Quantity of wafers in period i, $i = (C^M - C^F), \ldots, (C^M - C^F + N)$ of die $d \in D$ to start in Fab location $l \in F$

z_{ivkl} Quantity of devices in period i, $i = (C^M - C^T), \ldots, (C^M + C^A + N + 1)$ of device $v \in V$ coming from Assembly $k \in A$ to start in Test location $l \in T$

I_{idlB} Inventory at beginning of period i, $i = (C^M - 1), \ldots, (C^M + N)$ of yielded die $d \in D$ in location $l \in F$ at completion of Sort (that is, Die Bank)

I_{ivG} Inventory at beginning of period i, $i = (C^M - 1), \ldots, (C^M + C^A + C^T + N)$ yielded device $v \in V$ at completion of Test (that is, Finished Goods)

2.2 Modeling Test Yield Uncertainty

The following parameters are used in addition to the original model in Lowe and Mason (2016):

e^U_{vlT} Maximum deviation allowed from the reported mean test yield, $v \in V, l \in T$

Γ_{iv} Parameter to control the number of test locations in which the mean test yield is allowed to deviate, $i = C^M, \ldots, C^M + C^A + C^T + N - 1, v \in V$

The following decision variables are considered:

α_{iv} Buffer variable associated with the quantity of test yields allowed to vary, $i = C^M, \ldots, C^M + C^A + C^T + N - 1, v \in V$

p_{ivkl} Buffer variable associated with the yield-deviated device production, $i = C^M - C^T, \ldots, C^M + C^A + N + 3, v \in V, k \in A, l \in T$

β_{iv} Buffer variable limiting specific device production to within the maximum deviation allowed, $i = C^M, \ldots, C^M + C^A + C^T + N - 1, v \in V$

Only one constraint set from the original model in Lowe and Mason (2016), the flow balance constraints on Finished Goods, must be changed to reflect the reduction in output as a result of reduced yield. The remaining constraint sets were added to the model:

$$I_{ivG} + \sum_{l \in T} \rho_{lv} \sum_{k \in A} \tau_{kl} e_{vkl} z_{(i - c_{vlT} - t_{kl})vkl} - \gamma_{iv} - u_{iv} - \alpha_{iv}\Gamma_{iv} - \sum_{j \in T} \rho_{jv} \sum_{h \in A} \tau_{hj} p_{(i - c_{vjT} - t_{hjT})vhj} = I_{(i+1)vG}$$

$$\forall i = C^M, \ldots, C^M + C^A + C^T + N - 1, v \in V \quad (1)$$

$$\alpha_{iv} + \sum_{h \in A} \tau_{hl} p_{(i - c_{vlT})vhl} \geq e^U_{vlT} \beta_{iv} \qquad \forall i = C^M, \ldots, C^M + C^A + N - 1, v \in V, l \in T \quad (2)$$

$$\sum_{l \in T} \rho_{lv} \sum_{k \in A} \tau_{kl} z_{(i - c_{vlT} - t_{klT})vkl} \leq \beta_{iv} \qquad \forall i = C^M, \ldots, C^M + C^A + N - 1, v \in V \quad (3)$$

$$\sum_{l \in T} \rho_{lv} \sum_{k \in A} \tau_{kl} z_{(i - c_{vlT} - t_{klT})vkl} \geq -\beta_{iv} \qquad \forall i = C^M, \ldots, C^M + C^A + N - 1, v \in V \quad (4)$$

$$p_{ivkl} = 0 \qquad \forall i = C^M - C^T, \ldots, C^M - 1, v \in V, k \in A, l \in T \quad (5)$$

Constraint sets (1) modify the yielded quantity of devices produced in a period by the quantity produced within the allowable deviation from the mean test yield, if yields in that test location are allowed to vary within their allowable range. Constraint sets (2) ensure that the *modifying* device production is at least the quantity of actual device production during that period under the maximum allowed deviation from the test yield. Constraint sets (3) and (4) combine to ensure that the quantity of devices produced within the yield deviation is limited to actual device production in that period. Finally, device production prior to the start of the planning horizon is not allowed via constraint set (5). Together, these constraints ensure that if yield is allowed to vary in this test location, actual production during this period will be modified by a yielded quantity of devices within the allowable range.

2.3 Modeling Fab Yield Uncertainty

The following parameters have been added to the model in Lowe and Mason (2016):

e^U_{dlF} Maximum deviation allowed from the reported mean fab yield, $d \in D, l \in F$
Γ_{id} Parameter to control the number of fab locations in which the mean fab yield is allowed to deviate, $i = C^M, \ldots, C^M + N - 1, d \in D$

The following decision variables have also been added to the model:

α_{id} Buffer variable associated with the quantity of fab yields allowed to vary, $i = C^M, \ldots, C^M + N - 1, d \in D$
p_{idl} Buffer variable associated with the yield-deviated die production, $i = C^M - C^F, \ldots, C^M - C^F + N, d \in D, l \in F$
β_{id} Buffer variable limiting specific die production to within the maximum deviation allowed, $i = C^M, \ldots, C^M + N - 1, d \in D$

Only one constraint set from the original model, flow balance constraints on Die Bank, must be changed; the following remaining constraint sets were added to the model:

$$\sum_{p \in P} \varepsilon_{dp} \sum_{k \in A} \theta_{kp} y_{iplk} \leq \psi_{ld}(e_{dlF} g_d x_{(i-c_{dlF})dl} - \alpha_{id}\Gamma_{id} - p_{(i-c_{dlF})dl} + I_{idlB} - I_{(i+1)dlB})$$

$$\forall i = C^M, \ldots, C^M + N - 1, d \in D, l \in F \quad (6)$$

$$\alpha_{id} + p_{(i-c_{dlF})dl} \geq e^U_{dlF} g_d \beta_{id} \qquad \forall i = C^M, \ldots, C^M - C^F + N - 1, d \in D, l \in F \quad (7)$$

$$\sum_{l \in F} x_{(i-c_{dlF})dl} \leq \beta_{id} \qquad \forall i = C^M, \ldots, C^M - C^F + N - 1, d \in D \quad (8)$$

$$\sum_{l \in F} x_{(i-c_{dlF})dl} \geq -\beta_{id} \qquad \forall i = C^M, \ldots, C^M - C^F + N - 1, d \in D \quad (9)$$

Constraint sets (6) modify the yielded quantity of die produced in a period by the quantity produced within the allowable deviation from the mean fab yield, if yields in that fab are allowed to vary within their allowable range. Constraint sets (7) ensure that the *modifying* die production is at least the quantity of actual wafer production during that period under the maximum allowed deviation from the fab yield. Constraint sets (8) and (9) combine to ensure that the quantity of wafers produced within the yield deviation is limited to actual wafer production in that period. Together, these constraints ensure that if yield is allowed to vary in this fab, actual production in this period will be modified by a yielded quantity of die within the allowable range.

2.4 Modeling Both Test and Fab Yield Uncertainty

This model is an aggregation of the previous changes and will not be detailed here; modifications to the previous changes are the addition of indices to appropriately separate the parameters and decision variables between the two echelons.

3 CASE STUDY

3.1 Problem Statement

A manufacturer whose portfolio consisted of Linear and Power Management products, LAN solutions, and Timing and Communications products that helped the company maintain a strong position in the Industrial, Automotive, and Communications markets provided actual data for the case study. They produce a large

variety of devices (over 3000 SKUs) at a high volume for low prices; forecasted monthly demand averages over 110 million devices. Their largest customers typically require delivery sooner than production can deliver; therefore, the manufacturer maintains safety stock (in terms of weeks of demand) at each echelon. Finally, the manufacturer performs the fabrication but then contracts the assembly and test phases to a primary vendor (if the primary vendor is not available, a secondary vendor may be chosen); assembly and test vendors are engaged on a month-to-month basis. The manufacturer's challenge is to meet demand at the lowest cost by scheduling fabrication starts and contracting appropriate assembly and test starts with the vendors.

Using this actual data, we perform a case study to investigate the effects of allowing yield to vary within a symmetrical range; for research purposes, symmetric yield ranges are used. As stated earlier, yields at fab are defaulted to 0.90, allowing a reasonable range for testing. Finally, using the past nine months of the manufacturer's data, test yields for each device in the test cases were queried. Unfortunately, the manufacturer does not update test yields as they change, feeling that the values are "close enough." However, yields across devices showed enough variation such that we felt comfortable in setting the range of each yield equal to the difference between 1.0 and the current value: each device's yield will be allowed to vary between $[\max\{2*yield-1,0\},1.0]$.

Test cases were then built and run utilizing the same data as the sensitivity analysis. The reasons for this are twofold: we are able to validate model changes by comparing results under no uncertainty to the sensitivity analysis results and we are able to draw some preliminary conclusions regarding the cost of robustness. Three sets of cases were built: cases under varied sizes of uncertainty sets for test yields, under varied sizes for fab yields, and combining both fab and test. Since there is only one fab available for production, robust parameters of $\Gamma_* = 0$ and $\Gamma_* = 1$ for the fab cases were tested (across all periods in the planning horizon). However, there are 12 test vendors available, so robust parameters of $\{0,1,2,3,6,9,12\}$ were used; 0 corresponds to no uncertainty, 1 allows uncertainty in one test vendor, $3,6,9$, and 12 represent 25% increments in the number of uncertain test vendors, and 2 was included to provide continuity. As previously noted, these parameters were uniformly set across all periods in the planning horizon. Finally, all combinations of both fab and test yield uncertainty were tested. With eight test instances (four each from July and August), this resulted in a total of 184 test cases. Runtimes were limited to one hour or optimality; validation tests showed that optimality gaps typically reach their lowest level within the first 30 minutes. All of the cases were run with AMPL and Gurobi.

3.2 Test Yield Uncertainty Results

As stated, test yield uncertainty was modeled with robust parameters of $\{0,1,2,3,6,9,12\}$. Ratios of total costs (not including penalty costs) over the planning horizon in the uncertain cases to total costs for the baseline were calculated and then averaged over all eight instances for each robust parameter setting. As expected, costs increased as more uncertainty was introduced. A simple linear regression shows that for each additional vendor in which yields may fall within their uncertainty range, total horizon costs increase by approximately 22.2% (with an R^2 value of 0.98). Furthermore, as more uncertainty is present, production is not able to meet early demand. Non-zero penalty quantities began to appear under the $\Gamma_* = 3$ case (25% of test vendors have uncertain yields): five of the eight instances with an average of 2.1% of devices had penalties. All eight instances were unable to meet early demand through production when 50% or more of test vendors have yield uncertainty, with increasing percentage of devices and penalty quantities. This is reasonable in that as test yields decrease, more production is necessary at all three echelons, yet capacities at all three echelons have not changed.

However, as forecasts are updated on a monthly basis, schedules are re-generated; a comparison of first month costs may be more appropriate. As in the horizon costs comparison, ratios of first month costs were calculated and averaged. These results are similar to horizon costs results: costs increase as more uncertainty is introduced, with a similar 21.3% cost increase ($R^2 = 0.95$) for each additional level of uncertainty. Table 1 summarizes the results of this suite of tests.

Table 1: Summary of Test Yield Uncertainty.

Uncertainty Level	Horizon Cost Ratio	Month 1 Cost Ratio	Cases with Penalties	Average Pen Qty (millions)	Average % of Devices w/ Pens
$\Gamma = 0$	1.00	1.01	0	0.0	0.0%
$\Gamma = 1$	1.14	1.13	1	0.0	1.0%
$\Gamma = 2$	1.39	1.38	5	1.3	2.1%
$\Gamma = 3$	1.70	1.75	7	3.1	4.0%
$\Gamma = 6$	2.66	2.70	8	25.0	12.5%
$\Gamma = 9$	2.87	3.16	8	37.7	19.8%
$\Gamma = 12$	3.62	3.31	8	50.7	24.0%

3.3 Fab Yield Uncertainty Results

As fab production may occur in only one location, robust parameter settings of 0 and 1 were tested. Similar to the *Test Uncertainty* analysis, ratios were calculated and averaged over all eight instances for each robust parameter setting. Under "certainty" ($\Gamma_* = 0$), costs showed no difference from the baseline. However, when yields fall within their allowable range ($\Gamma_* = 1$), horizon costs increased by 9% and first month costs increased by 11%. Table 2 summarizes these results. Notice that in all cases, production was able to meet all demand; given the quantity of die on a wafer, the model is able to "recover" from the reduced yields at that fab.

Table 2: Summary of Fab Yield Uncertainty.

Uncertainty Level	Horizon Cost Ratio	Month 1 Cost Ratio	Cases with Penalties
$\Gamma = 0$	1.00	1.01	0
$\Gamma = 1$	1.09	1.11	0

3.4 Combined Test and Fab Yield Uncertainty Results

We also looked at the effects of simultaneously allowing uncertainty in fab and test yields. For each test yield robust parameter setting, cases were run with both values of the fab yield robust parameter setting. We first reviewed those cases in which production was not able to meet early demand. With few exceptions, the number of devices and quantity of units penalized were un-affected by uncertainty in fab yields (six cases had a one-device difference in the *number* of devices penalized, but the *quantity* of units penalized in those cases were virtually identical). Then, ratios of costs under the various levels of uncertainty to baseline costs were calculated and averaged over all instances under each robust parameter setting; results show the ratios increasing under the various levels of uncertainty. Analyzing test yield uncertainty when no uncertainty was allowed for fab yields showed similar cost increases to the *Test Uncertainty* case: for each additional test vendor with yield uncertainty, total costs over the planning horizon increased by 22.1% ($R^2 = 0.97$). However, when fab yield uncertainty is allowed, for each additional test vendor with yield uncertainty, total costs over the horizon increased by 24.3% ($R^2 = 0.98$). Notice that the full increase due to fab yield uncertainty, 9%, is not realized. Curiously though, when analyzing the effect of fab yield uncertainty against each level of test yield uncertainty, the average increase in horizon costs was 9% (with a variance of $7.4x10^{-7}$).

First month costs were then analyzed. Under no fab yield uncertainty, for each additional test vendor with yield uncertainty, first month costs increased by 21.2% ($R^2 = 0.94$); this result is similar to the

Test Uncertainty result. When uncertainty is present in fab yields, each additional test vendor with yield uncertainty increased first month costs by 24.4% ($R^2 = 0.95$). Again notice that the full increase due to fab yield uncertainty is not realized. Finally, analysis of the effect of fab yield uncertainty at each level of test yield uncertainty showed an average increase in first month costs of 11% (with a larger variance of $2.6x10^{-4}$).

3.5 Robust Solution Justification

In order to address whether the additional cost of a RO solution is justified, we provide a simple example, specific to our case study, to calculate the benefit of robustness; a more comprehensive study is necessary to fully understand the cost effectiveness of robustness. We analyzed schedule performance by developing a system to simulate each schedule's effects on inventory and ultimately, the satisfaction of forecasted demand. We focused on test yields as there is little opportunity to recover from reduced yields in order to satisfy a customer's order. However, knowing that more test vendors experiencing yield variability results in more costs, we utilized the most conservative results from only one test vendor allowing yield uncertainty.

Each device's yield was gradually reduced to the lowest value that would still allow for demand satisfaction; any further reduction would result in the unfulfillment of at least one customer's order. The more conservative schedules, generated via RO (with larger start quantities), allowed at least equal (often more) reductions in yield than the non-RO schedules. In other words, the RO-based schedules were able to still satisfy all forecasted demand under test yields lower than those supported by the non-RO schedules. Then, in order to determine the quantity of lost demand, test yields were set to the RO levels for the non-RO schedules. For example, device ABCD's lowest test yield for the non-RO schedule was 0.74 but was 0.69 for the RO schedule; the non-RO schedule was simulated with device ABCD's test yield set to 0.69. Using the lower yield with the non-RO schedule results in unfilled demand for ABCD. In other words, all ABCD demand is satisfied under the 0.69 yield with the RO solution; some demand will not be satisfied under the 0.69 yield with the non-RO solution.

All demand not satisfied was then totaled and compared to total demand over the planning horizon. On average, 6.7% of total demand was not fulfilled with almost 60% of devices experiencing at least one period in which their demand could not be filled. The "best" performer, July's forecast with the top 100 demanded devices, had approximately 28 million devices not produced from a forecasted total demand of just over 573 million devices (with 35 of the 100 having at least one period in which demand was lost). The "worst" performer, August's forecast with the top 100 demanded devices, had roughly 38 million devices not produced from a forecasted total demand of almost 372 million devices (65 of the 100 devices had at least one period in which demand was not satisfied).

In order to provide for potential uncertainty, it is important to consider the increase in cost. If realized yields are equal to the reported yields, the additional costs associated with the RO solution are unnecessary. However, if yields are lower, the RO solution will satisfy more demand than the non-RO solution. Specific to these cases within our study (comparing *no test vendor yield uncertainty* to *one test vendor yield uncertainty*), the average cost increase was 13.1% (actual increases range from $\$484,000$ to $\$1,360,000$). Ratios of the cost increase to lost demand were calculated and averaged; the average cost per device not satisfied was \$0.03 per device. In other words, it costs roughly 3 cents per device to implement the RO solution such that all demand would be satisfied under reduced test yields.

4 CONCLUSIONS AND FUTURE RESEARCH

Uncertainty is inherently present throughout the system under study. Production at all levels is driven by monthly forecasts of demand. We know that forecast quantities will fluctuate with each update; by maintaining minimum levels of inventory (e.g. safety stock), we are able to mask (in a reactive manner) the effects of these fluctuations in an aspect of our system over which we have little control. However, we can take a proactive approach to uncertainty in yield: given ranges of variability in wafer fabrication and

test yields (and by extension, assembly), different choices can be made as to which vendors to utilize.

Robust optimization allows the model to adjust production levels to accommodate yields within an allowable range, specifically lower yields. Our results show that as the quantity of vendors allowing uncertainty in yields increases, costs also increase. As yields decrease, fewer units are output from the appropriate echelons; since demand does not change to reflect this reduced output, more production (and hence, higher costs) is necessary to meet that static demand. When only yield uncertainty in test vendors is considered, first month costs increased by 21.3% for each additional test vendor allowing yield uncertainty. Under only fab yield uncertainty, first month costs increased by 11% for each additional fab location allowing uncertainty. Finally, if both fab and test vendors allow yield uncertainty, first month costs increase 24.4% for every additional test vendor with yield uncertainty (only one fab location was available in the input data).

Furthermore, with reduced yields comes the potential that early demand in the horizon will not be met from initial inventories and early production. With yield uncertainty, production quantities have to increase, but with no corresponding increase in capacity, demand will not be satisfied (this is signalled in the form of "penalty quantities"). Of interest to semiconductor manufacturers, this unmet demand represents lost sales or having to purchase units from a competitor in order to meet that demand. Our results show that as more test vendors permit yield uncertainty, more devices will not have early demand satisfied; if at least half of test vendors experience yield uncertainty, non-zero penalty quantities result (with increasing quantities as the percentage of test vendors with uncertainty increases).

However, when we compare the performance of schedules generated under no uncertainty to schedules resulting from the first level of robust test uncertainty (only one test location allows reduced yields), the "certain" schedules were unable to meet demand using test yields under which the "uncertain" schedules would meet all forecasted demand. Almost 7% (tens of millions of devices) of total demand was not satisfied by the "certain" schedules under these conditions. Additionally, an average of 60% of devices had at least one period in which *some* demand was not filled. This illustrates the benefit of the schedules produced via RO in that customer orders will still be satisfied under lower test yields than those supported by the deterministic schedules.

Through our implementation, as stated earlier, we have assumed that if a device (or wafer) yield was allowed to vary, it would vary in all periods over the planning horizon at that location. Future research plans include incorporating an additional level of *temporary* uncertainty, emulating a specific problem affecting production in one period that is resolved later in the horizon.

REFERENCES

Barahona, F., S. Bermon, O. Gunlok, and S. Hood. 2005. "Robust capacity planning in semiconductor manufacturing". *Naval Research Logistics* 52:459–468.

Ben-Tal, A., and A. Nemirovski. 1998. "Robust convex optimization". *Mathematics of Operations Research* 23 (4): 769–805.

Ben-Tal, A., and A. Nemirovski. 1999a. "Robust optimization – methodology and applications". *Mathematical Programming, Series B* 92 (3): 453–480.

Ben-Tal, A., and A. Nemirovski. 1999b. "Robust solutions of uncertain linear programs". *Operations Research Letters* 25:1–13.

Bertsimas, D., and M. Sim. 2004. "The price of robustness". *Operations Research* 52 (1): 35–53.

Bucy, J. 1972. "For semiconductors. Growth, not maturity". *IEEE Spectrum* 9 (4): 55–58.

Christie, R., and S. Wu. 2002. "Semiconductor capacity planning: stochastic modeling and computational studies". *IIE Transactions* (4): 131–143.

Cunningham, J. 1990. "The use and evaluation of yield models in integrated circuit manufacturing". *IEEE Transactions on Semiconductor Manufacturing* 3 (2): 60–71.

Fu, M., R. Askin, J. Fowler, and M. Zhang. 2015. "Stochastic optimization of product–machine qualification in a semiconductor back–end facility". *IIE Transactions* 47 (7): 739–750.

Geng, N., and Z. Jiang. 2007. "Capacity planning for semiconductor wafer fabrication with uncertain demand and capacity". In *Proceedings of the 3rd Annual IEEE Conference on Automation Science and Engineering*.

Geng, N., Z. Jiang, and F. Chen. 2009. "Stochastic programming based capacity planning for semiconductor wafer fab with uncertain demand and capacity". *European Journal of Operational Research* 198:899–908.

Kumar, N., K. Kennedy, K. Gildersleeve, R. Abelson, C. Mastrangelo, and D. Montgomery. 2006. "A review of yield modeling techniques for semiconductor manufacturing". *International Journal of Production Research* 44 (23): 5019–5036.

Lowe, J., and S. Mason. 2016. "Integrated semiconductor supply chain production planning". *IEEE Transactions on Semiconductor Manufacturing* 29 (2): 116–126.

Mönch, L., J. Fowler, S. Dauzère-Pérès, S. Mason, and O. Rose. 2011. "A survey of problems, solution techniques, and future challenges in scheduling semiconductor manufacturing operations". *Journal of Scheduling* 14 (6): 583–599.

Ng, T., and J. Fowler. 2007. "Semiconductor production planning using robust optimization". In *2007 IEEE International Conference on Industrial Engineering and Engineering Management*.

Ng, T., Y. Sun, and J. Fowler. 2010. "Semiconductor lot allocation using robust optimization". *European Journal of Operational Research* 205:557–570.

Peters, L. 2005. "Semiconductor market steady as she goes?". *Semiconductor International* 28 (10): 19–21.

Soyster, A. 1973. "Technical note – convex programming with set-inclusive constraints and applications to inexact linear programming". *Operations Research* 12 (5): 1154–1157.

Uzsoy, R., C.-Y. Lee, and L. 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* 24 (4): 47–60.

AUTHOR BIOGRAPHIES

JONATHAN J. LOWE is a Lecturer of Industrial Engineering at Clemson University. He received the B.S. degree in Management Science from the Georgia Institute of Technology, Atlanta, the M.S. degree in Industrial Engineering from the University of Florida, Gainesville, and the Ph.D. in Industrial Engineering from Clemson University. Dr. Lowe's research is focused on deterministic applied optimization, primarily supply chain and scheduling problems. His email address is jjlowe@clemson.edu.

AMIN KHADEMI is an Assistant Professor of Industrial Engineering at Clemson University. He received his Ph.D. in Industrial Engineering from the University of Pittsburgh after earning B.Sc. and M.Sc. degrees from Sharif University of Technology. Dr. Khademi's research interest is on developing optimization algorithms for decision making problems under uncertainty. His email address is khademi@clemson.edu.

SCOTT J. MASON is the Fluor Endowed Chair in Supply Chain Optimization and Logistics and a Professor of Industrial Engineering at Clemson University. He received his Ph.D. in Industrial Engineering from Arizona State University after earning Bachelor's and Master's degrees from The University of Texas at Austin. Dr. Mason's research focuses on developing optimization-based solutions for planning, scheduling, and controlling large-scale supply chain systems. His email address is mason@clemson.edu.