*Proceedings of the 2009 Winter Simulation Conference*
*M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin and R. G. Ingalls, eds.*

# OPTIMIZING DEMAND FULFILLMENT FROM TEST BINS

Brittany M. Bogle
Scott J. Mason

Department of Industrial Engineering
4207 Bell Engineering Center
University of Arkansas
Fayetteville, AR 72701, USA

## ABSTRACT

A primary component of the wafer assembly and final testing phases of the semiconductor manufacturing process is the process of binning wherein integrated circuits are tested for speed, voltage, and other functionality requirements. Customer demand for products is satisfied using binned components. While higher functionality components can be used to satisfy lower-level demand at a profit loss, the reverse case is not an option. We investigate the important question of satisfy customer demand from available binned devices with maximum profit in terms of maximizing revenue and minimizing inventory holding costs using a mathematical programming-based solution approach. Initial results suggest our model is able to accurately produce cost-effective demand fulfilment strategies for semiconductor manufacturers in practice.

## 1 INTRODUCTION

Integrated circuits (ICs) are the heart of most electronic devices, toys, and appliances today. ICs are fabricated first as individual die on silicon wafers in wafer fabrication facilities. Next, once the die are electrically tested for functionality, the silicon wafers are sawed into individual circuits for subsequent assembly or packaging and final testing. A primary component of the wafer assembly and final testing phases is the idea of binning. During final test, completed ICs may be evaluated in terms of assessing their processing speed (i.e., 3.0 GHz) and/or voltage requirements (i.e., 1.45 volts). After each IC's functionality is assessed and recorded, it is placed (sorted) by the manufacturer (i.e., the supply side) based on capability testing results in a "bin" corresponding to different product qualities (Uzsoy *et al.*, 1992).

Consider the following binning scenario: bin 1 could be reserved for the "highest functionality" components; bin 2 could contain components with "medium functionality;" bin 3, in turn, could hold the "low functionality" components. In practice, devices are placed into bins when they fall into a certain functionality range. For example, one bin may contain devices that have a power rating of 89 Watts, a speed of 3.1 GHz, and a voltage level between 1.425 volts and 1.45 volts. Further, it is important to note that product binning is not deterministic, as wafer fab production variability (which is inevitable) leads to variable end product capabilities/functionality.

External customers (i.e., the demand side) may place orders with a semiconductor manufacturer for various quantities of products having a wide range of capabilities—these product capabilities typically map to specific bin designations (i.e., supplier bin designations often relate to customer ordering patterns/demands). If necessary, higher product functionality can be substituted for lower level product demand, but not vice versa. For example, if there is customer demand for Bin 3 products in excess of the available inventory, bin 1 and/or bin 2 products could be used by the supplier to meet the demand. However, this substitution often comes with a significant price in terms of lost potential profits due to the related selling prices of high vs. low functionality components.

As customer demands occur on a weekly basis, frequent high functionality product substitution for lower functionality product demand can result in significant lost profits. However, generating excess inventories of all product functionality types can lead to unnecessarily high inventory holding costs. As excess inventory creation is not desirable, this paper investigates the important question of satisfy customer demand from available binned devices with maximum profit in terms of maximizing revenue and minimizing inventory holding costs. Towards this goal, we present an mathematical programming model designed to achieve our research goals.

## 2 LITERATURE REVIEW

The tradeoff between maintaining appropriate inventory levels to satisfy customer demand while seeking to maximize profits has been addressed previously in the literature, especially as global planning has become more necessary in practice due to increasing product demands. Hsu and Bassok (1999) present a downgrading substitution model with random demand and random yield that is quite similar to our motivating case of using devices placed in higher functionality bins for lower functionality demand satisfaction. Bitran and Gilbert (1994) study the idea of co-producing different products in batches within the semiconductor manufacturing industry. They note that random production yield of wafer die can help to enable manufacturers to supply a range of product functionalities to a number of customers from one production batch. IMPReSS, an optimization production planning and scheduling tool developed at the University of California-Berkeley, was created in order to provide inventory planning assistance for Harris Corporation through more effective scheduling efforts (Leachman *et al.*, 1996).

Hung and Wang (1997) examine the problem of meeting customer demand requirements through effective inventory planning. Finally, a theoretical approach presented by Gallego *et al.* (2006) is the most closely-related prior effort to our research study. Gallego *et al.* (2006) explore the minimization of inventory through part downgrading for customers that desire low functionality products. In contrast, our approach is more applied in nature than Gallego *et al.* (2006) as (1) we develop a number of our model's constraints based on prior discussions with semiconductor industry personnel and (2) we focus on minimizing total inventory costs over a multi-time period planning horizon. Towards this end, we now present an initial, working deterministic optimization model for the demand fulfillment problem of interest that later can be enhanced with additional, realistic probabilistic customer demand requirements and stochastic device binning distributions/results.

## 3 MODEL DEVELOPMENT

### 3.1 Definitions

Throughout the remainder of this paper, we will use key terms such as devices, bins, and OPNs, as they relate to a semiconductor manufacturer with whom we have consulted on the problem area of interest. For reader clarification, these key terms are defined as follows:

- Devices: The products that are electrically tested to ascertain different levels of functionality during the final test stage of the semiconductor manufacturing process. Based on the results of these tests, devices are assigned to a single bin related to their functionality/capability.
- Bin: A repository or inventory location containing tested devices of some specified functionality or capability. Individual bins are used to satisfy customer demand requirements for one or more specific OPNs.
- OPN: Order Part Number. Customers place their demand orders for OPNs in varying quantities during each time period. This demand subsequently is satisfied by the supplier with tested devices from bins corresponding to required production functionality/capability specifications.

### 3.2 Model

We now present a mixed-integer programming model for optimizing customer demand fulfillment from bins such that company profits are maximized. First, we define relevant model notation in terms of the model's sets, parameters, and variables:

Sets:

| | |
|---|---|
| $B$ | Set of bins, indexed by $b$ |
| $D$ | Set of devices, indexed by $d$ |
| $O$ | Set of OPNs, indexed by $o$ |
| $T$ | Set of time periods, indexed by $t$ |

Parameters:

$a_{bd}$ percent of device *d* assigned/distributed to bin *b* during electrical testing; these values were obtained from historical data supplied by our partnering semiconductor manufacturer; $\geq 0$ , $\leq 1$
$m_{bo}$ =1 if demand for OPN *o* can be satisfied by devices in bin *b*; otherwise, =0
$v_d$ initial inventory for device *d*; $\geq 0$
$p_d$ cost to produce device *d*; $\geq 0$
$h_b$ holding cost per unit in bin *b* per time period; $\geq 0$
$n_{ot}$ customer demand for OPN *o* during time period *t*; $\geq 0$

Variables:

$x_{bot}$ integer variable representing the number of components selected from bin *b* to satisfy customer demand for OPN *o* in time period *t*; $\geq 0$
$y_{bt}$ integer variable representing the component inventory level in bin *b* during time period *t*; $\geq 0$
$z_{dt}$ integer variable representing the quantity of device *d* components produced in time period *t*; $\geq 0$

The objective function of our model minimizes the sum of inventory holding costs and production costs:

$$minimize \ \sum_{b\in B} \sum_{t\in T} \sum_{d\in D}((h_b y_{bt}) + (p_d z_{dt})) \tag{1}$$

A number of constraints are necessary in our model. First, constraint sets (2) and (3) maintain inventory balance across all time periods by reconciling inventory levels with production quantities and customer demand levels:

$$y_{b1} = \sum_{d\in D}((\ v_d + z_{d1}\, a_{bd}) - \sum_{o\in O} x_{bo1} \quad \forall b \in B \tag{2}$$

$$y_{bt} = y_{b(t-1)} + \sum_{d\in D}((\ v_d + z_{dt}\, a_{bd}) - \sum_{o\in O} x_{bot} \quad \forall b \in B, t \in T, t > 1 \tag{3}$$

Next, constraint set (4) ensures that demand for each OPN is only met by components contained in valid (i.e., functionally compatible) bins in each time period:

$$\sum_{b\in B} x_{bot} \geq n_{ot} \quad \forall o \in O, t \in T \tag{4}$$

Finally, constraint set (5) requires that only available binned components can be used to satisfy customer demand in each time period:

$$\sum_{b\in B} \sum_{o\in O} x_{bot} \leq \sum_{b\in B} \sum_{d\in D}(z_{dt}\, a_{bd} + y_{bt}) \quad \forall t \in T \tag{5}$$

Variable type constraints are not included here, as they are defined above in the model notation portion of the paper.

### 3.3 Model Validation

In order to assess the validity of our formulation, we now present an example toy data set to demonstrate the model's functionality. First, let the set of devices *D* = {A, B, C, D, E, F, G, H}. Assume that customers can place orders for OPNs in the set *O* = {I, II, III, IV, V, VI, VII, VIII}. Further, we defined the set *B* to contain bins 1 through 9 and initially examine a two-period time horizon (i.e, *T* = {1, 2}).

Table 1 contains an example matrix for parameter $a_{bd}$ describing the distribution for each device *d* into bin *b* resulting from the final test stage of semiconductor manufacturing. The values were obtained from our partnering semiconductor manufacturer. This distribution is expressed as the fraction of devices tested that are mapped/placed into each bin. Note that the column associated with each device sums to 1, as each device must be entirely mapped to the available bins.

Table 1: Example Data for Device-to-Bin Mapping Parameter $a_{bd}$

| | Device | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Bin** | **A** | **B** | **C** | **D** | **E** | **F** | **G** | **H** |
| **1** | 0 | 0 | 0 | 0 | 0 | 0 | .03 | 0 |
| **2** | 0 | 0 | 0 | 0 | 0 | 0 | .02 | 0 |
| **3** | 0 | 0 | 0 | 0 | 0 | 0 | .08 | 0 |
| **4** | .84 | .06 | .99 | 0 | .94 | .95 | .33 | 0 |
| **5** | .16 | .94 | .01 | 0 | .06 | .05 | .32 | 0 |
| **6** | 0 | 0 | 0 | 0 | 0 | 0 | .13 | .01 |
| **7** | 0 | 0 | 0 | 0 | 0 | 0 | .09 | .79 |
| **8** | 0 | 0 | 0 | .84 | 0 | 0 | 0 | .20 |
| **9** | 0 | 0 | 0 | .16 | 0 | 0 | 0 | 0 |

The distribution or mapping of devices into bins as given in Table 1 determines how much each bin will/can be utilized when fulfilling customer demand orders. For example, consider the bin mappings for devices D and G in Table 1. Between these two device types, all bins can be covered. However, the cost of making each device type comes into play when a semiconductor manufacturer is deciding on which device to produce in order to meet customer demand. For example, if it is more expensive to produce device type G, a semiconductor manufacturer may elect to produce device H products in order to satisfy inventory requirements for bin 7.

Next, Table 2 displays example production cost data for each device type. Even though device types D and G can be used solely to produce inventories for all bin types (see Table 1), they are also the most costly to produce, according to Table 2. This tradeoff is especially evident when one considers the inventory holding costs incurred when extra devices are not sold to customers in response to OPN orders. For example, observe bin 5 and the low yield of device C into this bin (1%) and it's cost in Table 2 (1 unit). These values are in stark contrast to the same bin 5 values for device B (94% and a cost of 5 units). It is possible that the 500% cost increase may be warranted if significant amounts of demand exist for bin 5 components, given the difference between the two device's mapping percentage into bin 5.

Table 2: Example Device Production Cost Data

| **Device** | **A** | **B** | **C** | **D** | **E** | **F** | **G** | **H** |
|---|---|---|---|---|---|---|---|---|
| $p_d$ | 10 | 5 | 1 | 20 | 4 | 8 | 20 | 16 |

Table 3 displays example inventory holding cost data for the components in each bin. Each bin has a different carrying cost per piece held, as the value of the components contained in each bin vary according to the specifications of each bin. As a reminder, cost is the main objective of our proposed optimization model. Next, Table 4 displays the initial inventory (in number of items) of each device available for binning. From these initial inventories, subsequent binning and production decisions follow.

Table 3: Example Bin Holding Cost Data

| **Bin** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
|---|---|---|---|---|---|---|---|---|---|
| $h_b$ | .01 | .02 | .01 | .02 | .05 | .07 | 1 | .05 | .04 |

Table 4: Example Initial Device Inventory Levels

| **Device** | **A** | **B** | **C** | **D** | **E** | **F** | **G** | **H** |
|---|---|---|---|---|---|---|---|---|
| $v_d$ | 10439 | 2052 | 636 | 64560 | 12648 | 5700 | 373142 | 81659 |

In Table 5, an example bin-to-OPN mapping is shown. This is the map of bins from which components can be used to fulfill customer OPN demands. For example, the only way that any customer's demand for OPN I can be satisfied is by having an appropriate amount of inventory in bin 1. This mapping data, viewed in concert with Table 1's specification that only one device (G) bins out to bin 1, suggest that sufficient inventories and/or production levels of device G are critical. Lastly, Table 6 shows an example set of customer OPN demand quantities for use in our model verification efforts.

Table 5: Example Bin-to-OPN Mapping Data ($m_{bo}$)

| | OPN | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Bin** | **I** | **II** | **III** | **IV** | **V** | **VI** | **VII** | **VIII** |
| **1** | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **2** | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| **3** | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| **4** | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| **5** | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| **6** | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 |
| **7** | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| **8** | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| **9** | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |

Table 6: Example Customer Demand Orders by OPN in Each Time Period

| | OPN | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Time** | **I** | **II** | **III** | **IV** | **V** | **VI** | **VII** | **VIII** |
| **1** | 19806 | 320 | 946 | 71831 | 115822 | 6839 | 155606 | 13208 |
| **2** | 18636 | 33323 | 7317 | 23533 | 148777 | 30000 | 270479 | 7317 |

The model is implemented in AMPL v10.1 for solution and analysis using CPLEX v10.1 on a standard desktop PC. As it is in the class of optimization problems known as assignment problems, solution time is negligible. After analyzing this example problem, we achieve an objective function value of $343,843,106.90 for total production and inventory holding costs. The model outputs for primary decision variable $x_{bot}$ resulting from this solution are presented in Tables 7 and 8 for time periods 1 and 2, respectively.

Table 7: Number of Units Taken from Each Bin to Fulfill Demand in Time Period 1

| | OPN | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Bin** | **I** | **II** | **III** | **IV** | **V** | **VI** | **VII** | **VIII** |
| **1** | 27294 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **2** | 0 | 320 | 0 | 0 | 0 | 0 | 0 | 0 |
| **3** | 0 | 0 | 953 | 71831 | 0 | 0 | 0 | 0 |
| **4** | 0 | 0 | 0 | 0 | 300234 | 0 | 0 | 0 |
| **5** | 0 | 0 | 0 | 0 | 291136 | 0 | 0 | 0 |
| **6** | 0 | 0 | 0 | 0 | 0 | 0 | 118274 | 0 |
| **7** | 0 | 0 | 0 | 0 | 0 | 81882 | 0 | 0 |
| **8** | 0 | 0 | 0 | 0 | 0 | 0 | 37348 | 0 |
| **9** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 13211 |

Table 8: Number of Units Taken from Each Bin to Fulfill Demand in Time Period 2

| | OPN | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Bin** | **I** | **II** | **III** | **IV** | **V** | **VI** | **VII** | **VIII** |
| **1** | 23172 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **2** | 0 | 33324 | 0 | 0 | 0 | 0 | 0 | 0 |
| **3** | 0 | 0 | 32715 | 23533 | 0 | 0 | 0 | 0 |
| **4** | 0 | 0 | 0 | 0 | 254892 | 0 | 0 | 0 |
| **5** | 0 | 0 | 0 | 0 | 247168 | 0 | 0 | 0 |
| **6** | 0 | 0 | 0 | 0 | 0 | 0 | 100412 | 0 |
| **7** | 0 | 0 | 0 | 0 | 0 | 69516 | 0 | 0 |
| **8** | 0 | 0 | 0 | 0 | 0 | 0 | 154130 | 0 |
| **9** | 0 | 0 | 0 | 0 | 0 | 0 | 15943 | 7317 |

Table 9 displays the model's output for the amount of device production required each time period to satisfy customer demand. Careful observation confirms that this indeed is the least costly option, based on the multiple competing factors—production costs are balanced with inventory holding costs in order to produce the most cost-effective solution. The primary decision is to only produce devices D and G, and then allow the bin mapping percentages result in appropriate amounts of all customer-demanded OPNs. In fact, the majority of all demanded OPNs are satisfied from device G production.

Table 9: Model Outputs for Weekly Device Production

| Device | Week 1 | Week 2 |
|---|---|---|
| A | 0 | 0 |
| B | 0 | 0 |
| C | 0 | 0 |
| D | 82575 | 145375 |
| E | 0 | 0 |
| F | 0 | 0 |
| G | 909800 | 772400 |
| H | 0 | 0 |

Building on the production results in Table 9, Table 10 displays the inventory levels in each bin during each time period. As production efforts build device inventory levels, final testing distributes these devices to their appropriate bins. Then, customer demand for OPNs consumes binned ICs in the most cost-effective manner, resulting in the bin inventory levels described in Table 10.

Table 10: Bin Inventory Levels Per Week

| Bin | Week 1 | Week 2 |
|---|---|---|
| 1 | 0 | 0 |
| 2 | 17876 | 0 |
| 3 | 0 | 5544 |
| 4 | 0 | 0 |
| 5 | 0 | 0 |
| 6 | 0 | 0 |
| 7 | 0 | 0 |
| 8 | 32015 | 0 |
| 9 | 1 | 1 |

### 3.4 Model Sensitivity Analysis

In order to test the model's sensitivity to various input parameter values, the OPN-to-Bin map parameter $m_{bo}$ was varied to assess the model's ability to select different bins for customer OPN demand. However, it is clear from Table 1 that device G (as described by our partnering semiconductor manufacturer) often is the sole source for several bins' inventories. Therefore, regardless of a variety of sensitivity analysis changes, device G must always be chosen for production in order to satisfy customer demand, regardless of device G's associated costs.

It is important to realize that care must be taken to insure proper coverage of all bins is possible. Preliminary experiments reveal that when device G's mapping parameters were changed to not produce any bin 1 components ($a_{1G} = 0$), demand for OPN I could not be satisfied at all, as bin 1 is the only bin allowed to fulfill OPN I demand. This highlight the importance of proper bin coverage and confirms the model's ability to seek lowest cost, feasible solutions for customer OPN demand fulfillment.

## 4 CURRENT RESEARCH—SCALING UP TO REALITY

Now that our initial model is verified and its functionality has been validated, we are working with our partnering semiconductor manufacturer to generate a larger, more realistic data set. For this effort, the model scale of interest contains 300 OPNs available for customer demand and a product mix of 350 different device types. Some initial lessons learned from this

effort confirmed the importance of ensuring appropriate bin coverage for all bins are available for device-to-OPN translation mapping.

Upon generating our first realistically-sized dataset, the optimization model described above was run again in CPLEX. Unfortunately, our initial solution was determined to be infeasible, which at first we thought was the result of infeasible/uncovered bin mappings. However, subsequent investigations revealed that the integer restrictions on our model's primary decision variables were the reason for this infeasibility.

While customer demand requirements are for integer quantities of OPNs, the device-to-bin mapping parameters cause a fractional number of items to be present in bin inventories. We continue to investigate the appropriate method for relaxing the integrality restrictions in practice. We have verified that individually relaxing the integrality restriction on each of the model's three primary decision variables, $x_{bot}$, $y_{bt}$, and $z_{dt}$, separately results in optimal solutions that are within .000001% of each other, regardless of which variable was chosen for integrality relaxation. Therefore, our future efforts will focus on this area of model modification to produce cost-effective demand fulfillment solutions for our partnering semiconductor manufacturer.

## ACKNOWLEDGMENTS

This research was partially supported by an Arkansas State Undergraduate Research Fellowship (SURF) grant.

## REFERENCES

Bitran, G., and Gilbert, S. 1994. Co-Production Processes with Random Yields in the Semiconductor Industry. *Operations Research*, 42 (3): 476–491.

Gallego, G., Katircioglu, K., and Ramachandran, B. 2006. Semiconductor Inventory Management with Multiple Grade Parts and Downgrading. *Production Planning and Control*, 17 (7): 689–700.

Hsu, A., and Bassok, Y. 1999. Random Yield and Random Demand in a Production System with Downward Substitution. *Operations Research*, 47 (2): 277–290.

Hung, Y., and Wang, Q. 1997. A New Formulation Technique for Alternative Material Planning. *Computers and Engineering*, 32 (2): 281–297.

Leachman, R., Benson, R., Liu, C., and Raar, D. 1996. IMPReSS: An Automated Production-Planning and Delivery-Quotation System at Harris Corporation-Semiconductor Sector. *Interfaces*, 26 (1): 6–37.

Uzsoy, R., Lee, C.-Y., and Martin-Vega, L. A. 1992. A Review of the 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

**BRITTANY M. BOGLE** is a senior industrial engineering undergraduate research assistant in the Honors College at the University of Arkansas. She has won a number of scholarships and undergraduate research grants during her academic career and has studied abroad in China and Singapore. In addition, Brittany has completed two National Science Foundation Research Experiences for Undergraduates (REUs). Her future plans are to pursue industrial engineering graduate studies in order to become a faculty member. Brittany can be reached via e-mail at <bmbogle@uark.edu>.

**SCOTT J. MASON** is an Associate Professor and the Associate Department Head of Industrial Engineering at the University of Arkansas (UA). He received his BS in mechanical engineering and MS in engineering (emphasis in operations research) from The University of Texas at Austin and his PhD in industrial engineering from Arizona State University. Scott has published journal articles, book chapters, and conference papers in the fields of production planning and scheduling in semiconductor manufacturing facilities; supply chain optimization and analysis; and transportation logistics. Currently, he is an Associate Editor for IEEE Transactions on Electronics Packaging Manufacturing. Scott is a member of INFORMS, a senior member of the Institute for Industrial Engineers (IIE), and can be reached via e-mail at <mason@uark.edu>.