WORKING PAPER
ALFRED P. SLOAN SCHOOL OF MANAGEMENT

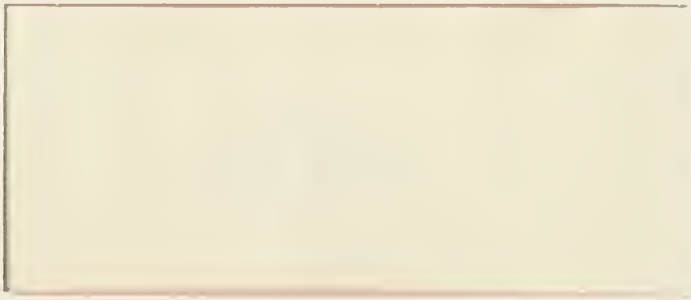
"A Product Design Problem
in Semiconductor Manufacturing"

Florin Avram
and
Lawrence M. Wein

MIT Sloan School Working Paper #3075-89-MS

August 1989

MASSACHUSETTS
INSTITUTE OF TECHNOLOGY
50 MEMORIAL DRIVE
CAMBRIDGE, MASSACHUSETTS 02139



"A Product Design Problem
in Semiconductor Manufacturing"

Florin Avram
and
Lawrence M. Wein

MIT Sloan School Working Paper #3075-89-MS

August 1989

# A PRODUCT DESIGN PROBLEM IN SEMICONDUCTOR MANUFACTURING

**Florin Avram**

*Department of Mathematics, Northeastern University*

and

**Lawrence M. Wein**

*Sloan School of Management, M.I.T.*

## Abstract

We consider the product design problem of allocating the chip sites on a semiconductor wafer to various types of chips. The manufacturing facility sells chips to its customers in sets (a specified number of several different types of chips), and the objective of the facility is to maximize the average production rate of sets. Variability in the wafer fabrication process, in particular random yield, poses a major obstacle in producing sets in a reliable fashion. A stochastic analysis is employed to develop an effective wafer design, and to measure the improvement in performance of the multi-type wafer over the traditional single-type wafer. The analysis reveals that multi-type wafers regularize the production flow of non-defective chips of each type and cause these flows to be positively correlated, both of which help to improve performance. A numerical example is provided that illustrates the analysis and demonstrates the design's effectiveness.

**August 1989**

# A PRODUCT DESIGN PROBLEM IN SEMICONDUCTOR MANUFACTURING

**Florin Avram**

*Department of Mathematics, Northeastern University*

and

**Lawrence M. Wein**

*Sloan School of Management, M.I.T.*

It is well-known that reducing variability in production systems will increase performance. Some of the conventional ways to reduce variability include the reduction of set-up times, greater use of preventive maintenance in lieu of lengthy machine breakdowns, the regulation of job releases, the cross-training of personnel, and the use of statistical process control. This paper provides an example from the semiconductor industry where the design of the product can also reduce variability, and hence improve performance.

The key step in semiconductor manufacturing is wafer fabrication, which consists of the production of disc-like wafers about four to six inches in diameter. The surface of the wafer is partitioned into chip sites in a grid-like fashion, and each wafer typically has between 20 and several hundred chips. Traditionally, all chips on a given wafer are of the same type. Such wafers will be referred to as *single-type wafers*. However, recent technological developments now permit different types of chips to be produced on the same wafer. This development is the catalyst for the *wafer design problem*: how does one allocate the chip sites on a wafer to the various types of chips. The resulting wafer will be referred to as a *multi-type wafer*.

This paper addresses two issues concerning the wafer design problem. The first issue is to find an effective design for a multi-type wafer, and the second is to assess how much improvement in performance a multi-type wafer offers over a single-type wafer. In order to

define the performance measure used in this paper, we will first describe the manufacturing environment of the particular facility that motivated this study; however, we believe that this environment is typical of large semiconductor manufacturers. Moreover, although the analysis is stimulated by a specific industry, some of our results apply to more genreal production settings.

Because of the complexity of the technology and the fast-changing nature of the semiconductor industry, the yield of non-defective chips in wafer fabrication can be low and very erratic. Average yield rates can vary from several percent up to eighty or ninety percent, depending upon the maturity and complexity of the product. Chips are made on wafers, and wafers are produced in lots, where the lot size is typically between five and fifty. In this paper, three different possibilities for producing defective chips will be explicitly considered: an entire lot can be defective, an entire wafer in a non-defective lot can be defective, and finally an individual chip on a non-defective wafer can be defective. As will be described in Section 1, the yield of a particular chip type in a particular chip site depends upon both the location of the chip site on the wafer and on the type of chip.

In the manufacturing environment under study, customers order chips in *sets*, where a set of chips consists of a specified number of chips for each of several chip types. In our particular application, a set of chips are required to build a circuit board, and thus the customers are unable to assemble a circuit board until they receive a set of chips. The variability in yield described in the previous paragraph makes it difficult for the chip manufacturer to provide reliable delivery performance to its customers.

The wafer fabrication facility, or *fab*, is modeled here as a network of quasi-reversible queues (see Kelly [4]), and the objective of the design problem is to maximize the long-run expected average number of non-defective sets produced by the wafer fab. This objective will get the sets of chips to the customers faster and will hence increase customer service, which is an overriding objective in the semiconductor industry. The particular manufacturing facility under consideration makes chips for several types of boards; however, in

order to gain insight into the nature of the problem, we restrict our analysis to designing one multi-type wafer for use in assembling one type of board.

In order to produce a wafer of a particular design using traditional photolithographic equipment, a *mask set* must first be produced that dictates the circuitry design that will appear on the wafer. A mask set costs on the order of $20,000, and thus once a mask set for a particular design is produced, it is desirable to produce many wafers with that mask set. Therefore, the chip site allocation problem described earlier is truly a static design problem, as opposed to a dynamic control problem. If electron beam machines are used in the lithographic process, then mask sets are not necessary for wafer production. However, there is still the need in this case for a simple and effective mechanism to design the wafers.

The design decision has been described as one of allocating the chip sites on a wafer to the various chip types. Currently, the problem being considered by the designers is to allocate the rows of a wafer to a particular chip site. The main constraint that restricts the design in this way is not technological in nature, but informational. To keep track of which chip type is in each of several hundred sites on a wafer (and then to use this information in a constructive way, such as for process control) is considered too large an informational burden. Instead, it is more desirable to just keep track of which chip type is in each of ten or twenty rows on a wafer. Our analysis remains essentially unchanged whether one allocates chip sites or rows to the chip types, and we will continue to refer to the design problem as one of allocating chip sites. To summarize, the problem considered in this paper is to allocate the chip sites (to various chip types) on a wafer that is to be mass-produced in a highly stochastic environment (a network of queues with three manners of defects) with the objective of maximizing the long-run expected average number of non-defective sets produced.

This appears to be the first study that has related the design of a product to the process flow of finished goods. The only paper related to the wafer design problem to our knowledge is Singh et al. [5]. They allocate the chip sites of a single wafer among

chip types in order to maximize the probability that a non-defective set can be produced from one wafer. Unlike our study, they assume that the yield of a chip type in a chip site is a Bernoulli random variable that depends only on the chip type and not on the site location. They also determine how the probability of completing a set increases as the number of wafers produced is increased. Although the design derived there may maximize the probability of obtaining a single set from a single (or several) wafers, the resulting wafer, if mass-produced, may not be effective in the long-run because each chip type may not be produced (after accounting for defectives) in the same proportions as the demand. Considering the cost involving in making a mask set, and the fact that the long-run demand for the computer boards appears to be steady, the long-term view is more appropriate for our setting.

A deterministic analysis that considers only long-run averages is employed in Section 2 to address the wafer design problem. There are two cases to analyze, depending upon the assumption made with regard to yield of wafers on a chip. Let $\mu_{sk}$ equal the probability that a type $k$ chip allocated to site $s$ on a non-defective wafer will be non-defective. If there exists positive numbers $c_s$ and $\mu_k$ such that $\mu_{sk} = c_s\mu_k$ for all sites $s$ and types $k$, then the yield will be called *multiplicative*; otherwise, the yield will be referred to as *non-multiplicative*. Although the yield was thought to be multiplicative in our setting, we know of no published data analysis that addresses this issue.

The wafer design problem is formulated as a linear program in Section 2; the decision variable $x_{sk}$ is the fraction of site $s$ devoted to type $k$ chips; thus, the integrality constraint that each site consists of a single chip that is of a single type is ignored in our formulation. The zero-one constrained version of the proposed LP will lead to an effective and feasible wafer design; however, ignoring this integrality constraint allows for a tractable stochastic analysis. Since there are typically many more sites than chip types, very few of the optimal $x_{sk}$ variables will have values other than zero or one. Thus, the performance of the integer and non-integer solutions will be very close, and the LP solution could be implemented

directly for the row allocation problem currently under consideration, without the fear of causing an information "explosion".

The results of the deterministic analysis are the following: under multiplicative yield (non-multiplicative, respectively) the multi-type wafer will perform the same as (better than, respectively) the traditional single-type case (under the appropriate mix of wafers entering the fab). Furthermore, under multiplicative yield, there are many possible optimal designs. Thus, under the multiplicative yield assumption that is believed to hold in practice, the naive deterministic analysis cannot distinguish among many possible wafer designs, or between the best multi-type wafer and the single-type wafer.

A stochastic analysis is employed in Section 3 that takes into account the variability inherent in the queueing network and in the random yield. We derive the point process that counts the number of non-defective sets completed up to time $t$ for an arbitrarily-designed multi-type wafer and for the single-type wafer. In each case, the central limit theorem for renewal processes is used to reduce the asymptotic (as $t \to \infty$) analysis of the point process to the analysis of the minimum of a multivariate normal random variable. A Slepian inequality (see Slepian [6]) is then employed to show that a particular multi-type wafer will perform better than the single-type wafer, and to give sufficient conditions under which one multi-type wafer design will dominate another design. This dominance condition is used to develop a surrogate objective function for the LP in Section 2 that will identify an effective wafer design. We also show that the difference in cumulative set production up to time $t$ between a multi-type policy and the single type policy is directly proportional to $\sqrt{t}$.

The stochastic analysis shows that if one wants to maximize the number of sets produced, and hence to maximize the minimum of a multivariate normal, it is desirable to have the covariance matrix of the multivariate normal contain large off-diagonal elements (that is, positive correlation among the elements) and small diagonal elements. Under the traditional single-type wafer policy, the resulting departure streams of non-defective chips

of each type are independent and thus the covariance matrix has zeroes on the off-diagonal, whereas the multi-type wafer leads to a covariance matrix that has positive off-diagonal elements. The reason for the positive correlation is that entire wafers and entire lots of wafers can be defective. Furthermore, the single-type wafer policy leads to much lumpier departure streams of non-defective chips of each type (departures of a particular type occur whenever a non-defective lot of wafers of that type exits the fab) than under the multi-type wafer policy (where each non-defective wafer of each non-defective lot may contain some non-defective chips of each type), and hence the corresponding covariance matrix has larger diagonal elements.

In addition to offering an effective multi-type wafer design, the stochastic analysis developed in Section 3 can also be used to plan production. In particular, one can use the algorithm of Clark [2] to approximate the mean of the minimum of a set of (possibly correlated) normal random variables. This will allow the production planner to predict the cumulative production of sets as a function of the rate at which lots of wafers are released into the fab. Thus, a simple one-dimensional search can be used to find the appropriate rate of lot releases that will attain a specified average output rate of sets. A naive deterministic analysis of the production rate, which uses only average yields, ignores the time that completed chips wait to form sets, and hence *overestimates* the production rate of sets for a given input rate, thereby leading to poor customer service. This production planning tool is not specific to the problem at hand. For any stochastic production system with variable yield that can be modeled as a quasi-reversible network of queues, we can estimate the cumulative production of sets as a function of the start rate of individual units that make up the set.

In Section 4, a numerical study is undertaken to assess the effectiveness of the proposed multi-type wafer design and measure its relative advantage over the single-type wafer. The numerical example assumes that yield is multiplicative, and a simulation model is used to measure the production rate of sets for the single-type policy and several multi-

type policies. The numerical results show that, after roughly 70,000 sets were completed, the single-type policy had only produced 88.7% of the naive deterministic estimate of the production rate, whereas the proposed multi-type design had produced 98.6% of the deterministic production rate. Other multi-type policies produced at over 96% of the deterministic production rate, and thus choosing among the solutions to the LP via the Slepian inequality has only a marginal effect on performance. However, the stochastic analysis did provide an accurate estimate (typically within $2-3\%$ of the simulated values) of the observed cumulative production in the single-type and multi-type cases. Thus the stochastic analysis is a useful tool to assess the relative impact of a multi-type wafer design and to plan production under any wafer design.

It is interesting to contrast the results of the deterministic and stochastic analysis. Although the naive deterministic analysis offers an effective design in the non-multiplicative yield case, it is not detailed enough to distinguish among many competing designs in the multiplicative yield case and, more importantly, is unable to detect the large difference in performance that exists between the single-type and multi-type wafers. Thus, although the deterministic analysis of yields can be useful, it is our hope that the stochastic analysis performed here will urge researchers and practitioners to delve deeper into the problems of random yields in manufacturing, as opposed to only analyzing yields in terms of long-run averages.

## 1. The Model

A wafer consists of $S$ chip sites, indexed by $s = 1, ..., S$, and a set of chips is composed of $n_1$ type 1 chips, $n_2$ type 2 chips,..., and $n_K$ type $K$ chips. As mentioned earlier, the wafer fab is being modeled as an open network of quasi-reversible queues. We assume that the network possesses a stationary distribution (which amounts to assuming that the network is not loaded up to or beyond its capacity), and that it is initialized with its stationary distribution. The individual customers in the queueing system represent a lot

of wafers, and the nodes of the queueing network represent the workcenters in the fab. In particular, this assumes that, under a multi-type policy, lots of multi-type wafers (all of which have the same design) arrive to the queueing system according to a Poisson process with rate $\lambda$. For the single-type policy, type $k$ lots (i.e., lots consisting of wafers that have only type $k$ chips on them) arrive according to independent Poisson processes with rate $f_k\lambda$, for $k = 1, ..., K$, where $f_k$ is the fraction of lots that are of type $k$. In the next section we derive the fraction $f_k$ that causes the average number of non-defective type $k$ chips produced to be in the same relative proportions as $n_1, ..., n_K$. This mix of entering lots will be assumed for the basis of comparison, and the resulting production policy will be referred to as the *single-type policy*.

Readers are referred to Chapter 3 of Kelly [4] for a definition and examples of networks of quasi-reversible queues. One possible scenario is that each workcenter consists of a number of identical machines, the machines at each workcenter have the same exponential processing time distribution, and individual lots are served FCFS (first-come first-served) at each workcenter. For the single-type policy, type $k$ lots have their own arbitrary deterministic route through the workcenters of the fab, and for the multi-type wafer policy, all lots have the same arbitrary deterministic route through the fab. As will be seen in Section 3, the only performance measures of the queueing network that are required in our study are the various departure processes.

For both the single-type and multi-type policies, it will be assumed that each lot of wafers is defective with probability $q$, independent of all other lots, and of the arrival, service, and routing information. This type of defect can occur at a batch operation, such as an oven, where the entire lot is processed at once. Each lot consists of $L$ identical wafers, and a wafer in a non-defective lot is defective with probability $p$, independent of all other wafers in the lot. This type of defect can occur when wafer-by-wafer operations are performed.

Let $Y_{sk}$ be a random variable that takes on the value of one if a type $k$ chip allocated

to site $s$ of a non-defective wafer is non-defective, and takes on the value of zero otherwise. (When considering the problem of allocating rows of wafers, the variable $Y_{sk}$ can take on values larger than one.) Then the expected value and variance of $Y_{sk}$ will be given by $\mu_{sk}$ and $\sigma^2_{sk}$, respectively. It will be assumed that the random variables $Y_{sk}$ are mutually independent for $s = 1, ..., S$ and $k = 1, ..., K$. Let $Y_k = \sum_{s=1}^{S} Y_{sk}$ be the number of non-defective chips on a non-defective type $k$ wafer (a wafer consisting only of type $k$ chips). The random variable $Y_k$ has expected value and variance given by $\mu_k$ and $\sigma^2_k$, respectively, where $\mu_k = \sum_{s=1}^{S} \mu_{sk}$ and $\sigma^2_k = \sum_{s=1}^{S} \sigma^2_{sk}$. Many studies (see Stapper [8] and Albin and Friedman [1], for example) have verified that the defects generated in wafer fabrication tend to cluster, causing the variance of the random variable $Y_k$ to be very high, sometimes much larger than the mean.

The assumption of independence across sites of a wafer is not crucial to our stochastic analysis; however, it does allow for a convenient comparison between the single-type and multi-type policies. Moreover, this does not appear to be an overly stringent assumption, especially considering that the effect of an entirely defective wafer has already been captured.

Recall that the decision variable for the wafer design problem is $x_{sk}$, which is the fraction of site $s$ allocated to type $k$ chips. Let $\bar{N}(t)$ (respectively, $\bar{F}(t)$) be the number of sets of non-defective chips that have departed from the fab during the time interval $[0, t]$ under the single-type (respectively, multi-type) policy. The process $\bar{F}(t)$ is design-dependent, and the wafer design problem is to choose $x_{sk}$ to maximize $\lim_{t\to\infty} t^{-1} E[\bar{F}(t)]$. Our goal is to solve the wafer design problem, and to compare $\lim_{t\to\infty} t^{-1} E[\bar{F}(t)]$ under an optimal design with $\lim_{t\to\infty} t^{-1} E[\bar{N}(t)]$, which is the corresponding performance measure under the single-type policy.

## 2. Deterministic Analysis

In this section, a deterministic analysis is undertaken that ignores any variablility

in the production process and in the yield process, and looks only at long-run average values. We begin with the single-type case, and then proceed to the multi-type case. The single-type case is considered primarily as a basis for comparison against the multi-type case.

In the single-type case, each lot of wafers consists of only a single chip type. In order to analyze the performance of this case, a product mix must be specified; that is, the proportion of lots that are released into the fab that are of a given chip type. Recall that $n_k$ non-defective type $k$ chips are required to form a set, and the average number of non-defective chips on a non-defective type $k$ wafer is $\mu_k$ for $k = 1, ..., K$. Therefore, on average, it takes $n_k/\mu_k$ non-defective type $k$ chips to form a set, and thus the average fraction $f_k$ of starting lots that are of type $k$ should be

$$f_k = \frac{n_k/\mu_k}{\sum_{j=1}^{K}(n_j/\mu_j)}. \tag{2.1}$$

Notice that the average total number of non-defective wafers of all types that it takes to form a set is $\sum_{k=1}^{K}(n_k/\mu_k)$.

In the multi-type case, the product design problem can be formulated as the linear program (LP)

$$\max_{x_{sk}} \min_{1 \le k \le K} \{\frac{1}{n_k} \sum_{s=1}^{S} \mu_{sk} x_{sk}\} \tag{2.2}$$

$$\text{subject to } \sum_{k=1}^{K} x_{sk} = 1 \text{ for } s = 1, ..., S, \tag{2.3}$$

$$x_{sk} \ge 0, \tag{2.4}$$

where $x_{sk}$ denotes the fraction of site $s$ that is allocated to type $k$ chips.

The multi-type case will be further distinguished between the *multiplicative* case and the *non-multiplicative* case. The expected yields $\mu_{sk}$ will be called multiplicative if there exists a set of $S$ nonnegative constants $c_s$ summing to one such that

$$\mu_{sk} = c_s \mu_k \text{ for all } s = 1, ..., S; k = 1, ..., K. \tag{2.5}$$

If condition (2.5) does not hold, then the yields will be referred to as non-multiplicative. Since $\sum_{s=1}^{S} c_s = 1$, it follows that the term $\mu_k$ appearing in (2.5) is the expected yield for a non-defective type $k$ wafer, which was defined in Section 1.

The two factors that affect the expected yield $\mu_{sk}$ are the *circuit density* of type $k$ chips and the *location* of site $s$ on the wafer. As the circuitry of a chip becomes more dense, it becomes more difficult to reliably produce the chip. Also, for all chip types, there are certain parts of the wafer that are more apt to yield non-defective chips. The yield versus location surface is donut-shaped, with the lowest probability of success occuring in the middle and at the periphery of the wafer. This yield surface tells us that some sites achieve inherently higher yields than other sites, regardless of the chip type. Condition (2.5) effectively assumes that these two factors, circuit density and location, behave in an independent fashion in determining the resulting yield of a type $k$ chip in site $s$. Although this condition was thought to hold in our setting, it is clear that one wants to collect data and test this assumption when possible. See Stapper [7] and references therein for more on the complicated relationship between yield and location.

When (2.5) holds, then LP (2.2)-(2.4) can be rewritten as

$$\max_{y, x_{sk}} \; y \tag{2.6}$$

$$\text{subject to } \; y - \frac{\mu_k}{n_k} \sum_{s=1}^{S} c_s x_{sk} \leq 0 \;\; \text{for } k = 1, ..., K, \tag{2.7}$$

$$\sum_{k=1}^{K} x_{sk} = 1 \;\; \text{for } s = 1, ..., S, \tag{2.8}$$

$$x_{sk} \geq 0. \tag{2.9}$$

Let $\pi = (\pi_1, ..., \pi_K)$ and $\gamma = (\gamma_1, ..., \gamma_S)$ be the dual variables corresponding to constraints (2.7) and (2.8), respectively. Then the dual LP is

$$\min_{\pi_k, \gamma_s} \sum_{s=1}^{S} \gamma_s \tag{2.10}$$

$$\text{subject to } \gamma_s \geq \frac{\mu_k c_s \pi_k}{n_k} \;\; \text{for } s = 1, ..., S; k = 1, ..., K, \tag{2.11}$$

$$\sum_{k=1}^{K} \pi_k = 1, \tag{2.12}$$

$$\pi_k \geq 0. \tag{2.13}$$

If the constraints (2.11) are summed over $s = 1, ..., S$, then we obtain the relaxed dual LP

$$\min_{\pi_k, \gamma_s} \sum_{s=1}^{S} \gamma_s \tag{2.14}$$

$$\text{subject to } \sum_{s=1}^{S} \gamma_s \geq \frac{\mu_k \pi_k}{n_k} \text{ for } k = 1, ..., K, \tag{2.15}$$

$$\sum_{k=1}^{K} \pi_k = 1. \tag{2.16}$$

$$\pi_k \geq 0, \tag{2.17}$$

which can be rewritten as

$$\min_{\pi_k} \max_{1 \leq k \leq K} \frac{\mu_k \pi_k}{n_k} \tag{2.18}$$

$$\text{subject to } \sum_{k=1}^{K} \pi_k = 1, \tag{2.19}$$

$$\pi_k \geq 0. \tag{2.20}$$

The solution $\pi^*$ to this LP is

$$\pi_k^* = \frac{n_k}{\mu_k} \Big( \sum_{j=1}^{K} \frac{n_j}{\mu_j} \Big)^{-1} \text{ for } k = 1, ..., K, \tag{2.21}$$

and thus the optimal objective function value to the relaxed dual LP is $\left( \sum_{k=1}^{K} (n_k / \mu_k) \right)^{-1}$. Therefore, an optimal solution to the original dual LP (2.10)-(2.13) is $\pi^*$ and

$$\gamma_s^* = c_s \Big( \sum_{k=1}^{K} \frac{n_k}{\mu_k} \Big)^{-1} \text{ for } s = 1, ..., S. \tag{2.22}$$

Since $\pi_k^* > 0$ for $k = 1, ..., K$, the $K$ primal constraints (2.7) are all binding in the optimal solution. Since the set of equations

$$\frac{\mu_k}{n_k} \sum_{s=1}^{S} c_s x_{sk} = \Big( \sum_{k=1}^{K} \frac{n_k}{\mu_k} \Big)^{-1} \text{ for } k = 1, ..., K, \tag{2.23}$$

$$\sum_{k=1}^{K} x_{sk} = 1 \text{ for } s = 1, ..., S, \tag{2.24}$$

are redundant, there are in general many solutions to the primal LP. Thus we have proven the following.

**Proposition 1.** *Under condition (2.5),*

*(a) the optimal objective function value of the LP (2.2)-(2.4) is* $\left(\sum_{k=1}^{K}(n_k/\mu_k)\right)^{-1}$;

*(b) all the constraints in (2.7) are binding; and*

*(c) the solution to (2.2)-(2.4) is not unique.*

Notice that (a) tells us that the optimal number of sets from a non-defective wafer is equal to $\left(\sum_{k=1}^{K}(n_k/\mu_k)\right)^{-1}$. Recall that under the single-type policy under product mix (2.1), the average number of non-defective wafers required to obtain a set was $\sum_{k=1}^{K}(n_k/\mu_k)$. Thus, under the multiplicative yield assumption, there appears to be no advantage in using multi-type wafers. This is because the relative expected yields of the various chip types behave similarly across sites. However, this deterministic analysis ignores the variabiliy in the production process and the queueing of chips that takes place before they are assembled into sets. The next section will take a closer look at this problem, and will indeed show that there is a significant difference between the single-type and multi-type policies. Furthermore, the analysis in the next section will aid us in choosing which of the many solutions to the LP (2.2)-(2.4) will yield the best design.

Before finishing this section, let us return to the original LP under the non-multiplicative assumption that (2.5) does not hold. In this case, certain chip types can perform better in some sites than other sites relative to other chip types. Observe that

$$x_{sk} = f_k = \frac{n_k/\mu_k}{\sum_{j=1}^{K}(n_j/\mu_j)} \text{ for } s = 1, ..., S; k = 1, ..., K, \tag{2.25}$$

is feasible for LP (2.2)-(2.4) and achieves the objective function value $\left(\sum_{j=1}^{K}(n_j/\mu_j)\right)^{-1}$, which is the same performance as the single-type case. Thus,

**Proposition 2.** *The optimal objective function value of LP (2.2)-(2.4) under the non-multiplicative case is greater than or equal to the corresponding performance measure under the single-type policy.*

## 3. Stochastic Analysis

The stochastic analysis in this section assumes that the multiplicative assumption (2.5) holds. However, in the non-multiplicative case, a similar analysis can be used to compare the performance of the optimal multi-type policy (i.e., the solution to (2.2)-(2.4)) to the single type policy. If there happens to be multiple solutions to the LP in the non-multiplicative case, then the analysis can also be used to help decide which solution will lead to better performance.

3.1 **Single-Type Policy.** We begin this section by analyzing the single-type policy. Recall that under this policy, lots of type $k = 1, ..., K$ enter the fab according to independent Poisson processes with rate $f_k\lambda$, where $f_k$ was defined in (2.1). Since the fab is an open network of quasi-reversible queues, it follows that the departure stream $A_k(t)$ of non-defective type $k$ lots from the fab is Poisson with rate $(1-q)f_k\lambda$. Our assumptions in Section 2 also imply that the departure process of non-defective type $k$ wafers is a compound Poisson process $D_k(t) = \sum_{i=1}^{A_k(t)} W_i$, where $W_i, i = 1, ..., A_k(t)$, are independent and identically distributed (iid) binomial random variables with parameters $L$, the lot size, and $1-p$, the probability of a non-defective wafer. Let $N_k(t)$ be defined by

$$N_k(t) = \frac{\sum_{i=1}^{D_k(t)} Y_k^{(i)}}{n_k}, t \geq 0, \text{ for } k = 1, ..., K, \tag{3.1}$$

where $Y_k^{(i)}, i = 1, ..., D_k(t)$, are iid random variables having the same distribution as $Y_k$. For $k = 1, ..., K$, it follows that $E[N_k(t)] = at$ and $\text{Var}[N_k(t)] = s_k^2 t$, where

$$a = \frac{(1-q)\lambda L(1-p)}{\sum_{j=1}^{K}(n_j/\mu_j)} \tag{3.2}$$

and

$$s_k^2 = \frac{(1-q)\lambda}{\mu_k n_k \sum_{j=1}^{K}(n_j/\mu_j)} \left( L(1-p)\sigma_k^2 + \mu_k^2 L(1-p)p + L^2(1-p)^2\mu_k^2 \right). \tag{3.3}$$

Furthermore, notice that the $K$ processes $\{N_k(t), t \geq 0\}, k = 1, ..., K$, are mutually independent.

Recalling that $n_k$ type $k$ chips are required to form a set, it follows that the number of sets produced by time $t$ is $\min_{1\le k\le K}\lfloor N_k(t)\rfloor$, where $\lfloor x\rfloor$ denotes the integer part of $x$. For simplicity, we choose to analyze the process $N^*(t) = \min_{1\le k\le K} N_k(t)$, which will lead to a negligible error since

$$|\min_{1\le k\le K}\lfloor N_k(t)\rfloor - N^*(t)| \le 1 \text{ for all } t \ge 0. \tag{3.4}$$

It follows by the central limit theorem for renewal processes that, for $k = 1, ..., K$,

$$\frac{N_k(t) - at}{\sqrt{t}} \Rightarrow X_k \text{ as } t \to \infty, \tag{3.5}$$

where $X_k \sim N(0, s_k^2)$, i.e., $X_k$ is normally distributed with mean zero and variance $s_k^2$, and $\Rightarrow$ denotes convergence in distribution. Thus,

$$\lim_{t\to\infty} Pr(N^*(t) \le x\sqrt{t} + at) = 1 - \lim_{t\to\infty}\prod_{k=1}^{K} Pr(N_k(t) > x\sqrt{t} + at) \tag{3.6}$$

$$= \left(1 - \prod_{k=1}^{K}(1 - \Phi(\frac{x}{s_k}))\right), \tag{3.7}$$

where $\Phi$ is the cumulative distribution function of the standard normal. Thus we can analyze the asymptotic performance of the single-type policy by studying the minimum of $K$ independent normal random variables; this point will be pursued further in Section 3.4.

Notice that $s_k^2$ in (3.3) decreases as $n_k$ increases, and thus the chip types that have relatively few chips in a set will tend to cause the burstier, or more variable, departure streams, since these chip types will not be produced very frequently.

**3.2. Multi-Type Policy.** A similar analysis will now be applied to the multi-type policy. By the quasi-reversibility assumption, the departure process of non-defective lots from the queueing network is a Poisson process $A(t)$ with rate $\lambda(1-q)$. Therefore, the departure process of non-defective wafers is given by the compound Poisson process $D(t) = \sum_{i=1}^{A(t)} W_i$, where $W_i, i = 1, ..., A(t)$, are again binomially distributed iid random variables with parameters $L$ and $1-p$. Thus,

$$E[D(t)] = \lambda(1-q)L(1-p)t \tag{3.8}$$

and

$$\mathrm{Var}[D(t)] = \lambda(1-q)t[L(1-p)p + L^2(1-p)^2]. \tag{3.9}$$

Recall that the wafer design is defined by the decisions $x_{sk}, s = 1, ..., S$, and $k = 1, ..., K$, where $x_{sk}$ is the fraction of site $s$ allocated to type $k$ chips. Define the random variables $V_{sk}$ by

$$V_{sk} = \frac{x_{sk}Y_{sk}}{n_k} \text{ for } s = 1, ..., S; k = 1, ..., K. \tag{3.10}$$

Then these mutually independent random variables, which are dependent upon the design, have

$$E[V_{sk}] = \frac{x_{sk}\mu_{sk}}{n_k} \text{ and } \mathrm{Var}[V_{sk}] = \frac{x_{sk}^2\sigma_{sk}^2}{n_k^2}. \tag{3.11}$$

Define $T_k = \sum_{s=1}^{S} V_{sk}$, so that

$$E[T_k] = \sum_{s=1}^{S} \frac{x_{sk}\mu_{sk}}{n_k} \text{ and } \mathrm{Var}[T_k] = \sum_{s=1}^{S} \frac{x_{sk}^2\sigma_{sk}^2}{n_k^2}. \tag{3.12}$$

Let $F(t) = (F_1(t), ..., F_K(t))$ be defined by $F_k(t) = \sum_{i=1}^{D(t)} T_k^{(i)}$, where $T_k^{(i)}$ are iid random variables distributed as $T_k$. Then $\min_{1 \le k \le K} \lfloor F_k(t) \rfloor$ is the number of sets produced by time $t$. It follows that $E[F_k(t)] = b_k t$, $\mathrm{Var}[F_k(t)] = \Sigma_{kk} t$ for $k = 1, ..., K$, and $\mathrm{Cov}[F_j(t), F_k(t)] = \Sigma_{jk} t$ for $j \neq k$, where

$$b_k = \lambda(1-q)L(1-p)\left(\sum_{s=1}^{S} \frac{x_{sk}\mu_{sk}}{n_k}\right), \tag{3.13}$$

$$\begin{aligned}\Sigma_{kk} = {} & \lambda(1-q)L(1-p)\left(\sum_{s=1}^{S} \frac{x_{sk}^2\sigma_{sk}^2}{n_k^2}\right) \\ & + \lambda(1-q)[L(1-p)p + L^2(1-p)^2]\left(\sum_{s=1}^{S} \frac{x_{sk}\mu_{sk}}{n_k}\right)^2,\end{aligned} \tag{3.14}$$

and

$$\Sigma_{jk} = \lambda(1-q)[L(1-p)p + L^2(1-p)^2]\left(\sum_{s=1}^{S} \frac{x_{sj}\mu_{sj}}{n_j}\right)\left(\sum_{s=1}^{S} \frac{x_{sk}\mu_{sk}}{n_k}\right). \tag{3.15}$$

Recall that one of the goals of this section is to choose, for the multiplicative yield case (2.5), an effective design $x_{sk}$ from among the many optimal solutions to the dual LP

(2.2)-(2.4). In the rest of this section, we will restrict ourselves to designs $x_{sk}$ that solve LP (2.2)-(2.4) under condition (2.5). However, as noted earlier, a similar analysis can be performed under the non-multiplicative yield case.

**Proposition 3.** *If $x_{sk}$ solves LP (2.2)-(2.4), then $b_k$ defined in (3.13) satisfies $b_k = a$ for $k = 1, ..., K$, where $a$ is defined in (3.2).*

**Proof.** This follows directly from Proposition 1(a), (2.1), (3.2), and (3.13). ∎

Let $b = (b_1, ..., b_k)$, which equals $(a, ..., a)$ by Proposition 3, and we have, by the central limit theorem for renewal processes,

$$\frac{F(t) - bt}{\sqrt{t}} \Rightarrow Z \text{ as } t \to \infty, \tag{3.16}$$

where $Z \sim N(0, \Sigma)$, which is a multivariate normal random variable with mean zero and covariance matrix $\Sigma$, where $\Sigma$ is defined in (3.14)-(3.15).

Letting $F^*(t) = \min_{1 \le k \le K} F_k(t)$ and denoting the $k$th component of $Z$ by $Z_k$, we have

$$\lim_{t \to \infty} Pr(F^*(t) \le x\sqrt{t} + at) = 1 - Pr(\min_{1 \le k \le K} Z_k > x). \tag{3.17}$$

Thus, the asymptotic performance of the multi-type policy can be analyzed by studying the minimum of $K$ dependent normal random variables.

**3.3. A Comparison of Policies.** We begin this section by stating a generalization of Slepian's lemma (see Slepian [6]) that is due to Kahane [3].

**Proposition 4.** *Let $X = (X_k)$ and $Z = (Z_k)$ be normal $K-$dimensional vectors such that*

$$\begin{aligned} &E[X_j X_k] \le E[Z_j Z_k] \text{ if } (j,k) \in A, \\ &E[X_j X_k] \ge E[Z_j Z_k] \text{ if } (j,k) \in B, \\ &E[X_j X_k] = E[Z_j Z_k] \text{ if } (j,k) \notin A \vee B. \end{aligned} \tag{3.18}$$

*Let $f(x) = f(x_1, ..., x_k)$ be a function defined on $R^K$ with second derivatives satisfying*

$$\frac{\partial f}{\partial x_j \partial x_k} \geq 0 \text{ if } (j,k) \in A,$$
$$\frac{\partial f}{\partial x_j \partial x_k} \leq 0 \text{ if } (j,k) \in B. \tag{3.19}$$

*Then $E[f(X)] \leq E[f(Z)]$.*

With the aid of the Slepian inequality, we will provide a multi-type policy that outperforms the single-type policy.

**Proposition 5.** *For $j \neq k$, $\Sigma_{jk} \geq 0$ and is independent of the design $x_{sk}$.*

Proof. This result follows from Proposition 1(b) and (3.15). ∎

Let us again consider the design

$$x_{sk} = \frac{n_k/\mu_k}{\sum_{j=1}^{K}(n_j/\mu_j)} \tag{3.20}$$

defined in (2.25).

**Proposition 6.** *For the design $x_{sk}$ defined in (3.20), the corresponding covariance matrix satisfies $\Sigma_{kk} \leq s_k^2$ for $k = 1, ..., K$.*

**Proof.** From (3.3) and (3.14) we have

$$\Sigma_{kk} - s_k^2 = \frac{\lambda(1-q)[L(1-p)p + L^2(1-p)^2]}{\sum_{j=1}^{K}(n_j/\mu_j)}\left(\frac{1}{\sum_{j=1}^{K}(n_j/\mu_j)} - \frac{1}{(n_k/\mu_k)}\right)$$
$$+ \lambda(1-q)L(1-p)\left(\sum_{s=1}^{S}\left(\frac{x_{sk}^2\sigma_{sk}^2}{n_k^2}\right) - \frac{\sigma_k^2}{\mu_k n_k \sum_{j=1}^{k}(n_j/\mu_j)}\right). \tag{3.21}$$

The first term on the righthandside is non-positive since $\sum_{j=1}^{K}(n_j/\mu_j) \geq n_k/\mu_k$. Under design (3.20),

$$\sum_{s=1}^{S}\left(\frac{x_{sk}^2\sigma_{sk}^2}{n_k^2}\right) - \frac{\sigma_k^2}{\mu_k n_k \sum_{j=1}^{k}(n_j/\mu_j)}$$
$$= \frac{\sigma_k^2}{\mu_k^2 \sum_{j=1}^{k}(n_j/\mu_j)}\left(\frac{1}{\sum_{j=1}^{K}(n_j/\mu_j)} - \frac{1}{(n_k/\mu_k)}\right) \leq 0, \tag{3.22}$$

and hence the second term on the righthand side of (3.21) is also non-positive. ∎

**Proposition 7.** *Under design (3.20),*

$$\lim_{t\to\infty} \frac{E[F^*(t)]}{t} \geq \lim_{t\to\infty} \frac{E[N^*(t)]}{t}. \tag{3.23}$$

**Proof.** This follows from central limit theorems (3.5) and (3.16), and Propositions 4-6, where, in Proposition 4, the set $A$ corresponds to the diagonal elements and the set $B$ corresponds to the off-diagonal elements, and $f(x_1, ..., x_K) = \min(x_1, ..., x_K)$. ∎

Although design (3.20) outperforms the single-type policy, it is not necessarily a desirable multi-type design. By Propositions 4 and 5, it is clear that in order to find an effective design, one wants to make the diagonal terms $\Sigma_{kk}$ of the covariance matrix of $F(t)$ as small as possible. In particular, since $n_k^{-1} \sum_{s=1}^{S} x_{sk}\mu_{sk}$ is a constant for all designs that solve LP (2.2)-(2.4), it follows from (3.14) that the terms $n_k^{-2} \sum_{s=1}^{S} x_{sk}^2 \sigma_{sk}^2, k = 1, ..., K$ need to be made as small as possible.

We now propose two mathematical programs (that differ only by their objective functions) that should lead to effective, not optimal, designs. Both programs have the constraint set (2.4) and (2.23)-(2.24), which has the same form as the constraints of a transportation problem; this constraint set guarantees that a feasible solution will be an optimal solution to LP (2.2)-(2.4). The two objectives are

$$\min_{x_{sk}} \sum_{k=1}^{K} \sum_{s=1}^{S} \frac{\sigma_{sk}^2 x_{sk}^2}{n_k^2} \tag{3.24}$$

and

$$\min_{x_{sk}} \max_{1\leq k\leq K} \sum_{s=1}^{S} \frac{\sigma_{sk}^2 x_{sk}^2}{n_k^2}. \tag{3.25}$$

Although objectives (3.24) and (3.25) may both lead to an effective design, objective (3.25) will lead to a problem with non-linear constraints, which will be harder to solve than the quadratic program generated by objective (3.24). It should be noted that under the popular assumption that $Y_{sk}$ are Bernoulli random variables for all $s = 1, ..., S$ and

$k = 1, ..., K$ (and the yield is multiplicative), then $\sigma_{sk}^2 = \mu_{sk}(1 - \mu_{sk}) = c_s\mu_k(1 - c_s\mu_k)$, and

$$\sum_{s=1}^{S} \frac{\sigma_{sk}^2 x_{sk}^2}{n_k^2} = \frac{\mu_k}{n_k^2} \sum_{s=1}^{S} c_s x_{sk}^2 - \frac{\mu_k^2}{n_k^2} \sum_{s=1}^{S} c_s^2 x_{sk}^2. \tag{3.26}$$

If the variables $x_{sk}$ were all integer, the first term on the righthand side of (3.26) would be a constant by Proposition 1, but the second term would still depend on $x_{sk}$, and hence the math programs are not trivial in this case.

**3.4. Asymptotic Performance Analysis.** By (3.7) and (3.17), it follows that an asymptotic ($t \to \infty$) analysis of $N^*(t)$ and $F^*(t)$, which are the cumulative number of sets produced up to time $t$ for the various policies, can be performed by studying the minimum of $K$ (possibly dependent) normal random variables. In particular, Clark [2] has developed an iterative approximation (it is exact for $K = 2$) technique to calculate the moments of the minimum of $K$ normal random variables. Using his approximation, we can estimate $E[\min(X_1, ..., X_K)]$ and $E[\min(Z_1, ..., Z_K)]$ and, for the sake of concreteness, let us denote these estimates by $c_x$ and $c_z$, respectively. It follows by (3.5) that

$$\frac{N^*(t) - at}{\sqrt{t}} \Rightarrow \min(X_1, ..., X_K) \text{ as } t \to \infty, \tag{3.27}$$

and thus for large $t$, we can estimate $E[N^*(t)]$ by

$$c_x\sqrt{t} + at. \tag{3.28}$$

Similarly, $E[F^*(t)]$ can be estimated by

$$c_z\sqrt{t} + at. \tag{3.29}$$

Notice that the quantity $at$ appearing in (3.28) and (3.29) is the deterministic estimate of $E[N^*(t)]$ and $E[F^*(t)]$ if there was no variability in the production system. Since $c_x$ and $c_z$ will be negative (recall that $X$ and $Z$ have mean zero), it follows that the naive deterministic analysis overestimates the actual cumulative production of sets. The

difference in expected cumulative production between a multi-type policy and the single-type policy grows in proportion to $\sqrt{t}$, and is approximated by

$$(c_x - c_z)\sqrt{t}. \tag{3.30}$$

Thus, although the average production rates of the two policies both converge to $a$, the difference in the cumulative production between the two policies diverges.

## 4. A Numerical Example

We now present a numerical example that is based on disguised, but representative, data from an actual facility. As mentioned in the Introduction, the problem currently being addressed is to allocate rows of a wafer to various chip types. Our example has $K = 4$ chip types and ten rows, but to keep consistent with past notation and terminology, we will refer to the rows as sites, and thus $S = 10$. Readers should interpret $x_{sk}$ as the fraction of a particular row that is allocated to type $k$ chips.

A set is made up one type 1 chip, one type 2 chip, three type 3 chips, and twenty type 4 chips, and so $n_k = (1, 1, 3, 20)$ for $k = 1, ..., 4$. It is not uncommon for the universal "80-20" rule to hold here: 25% of the chip types account for 80% of the set. The multiplicative yield assumption (2.5) holds and the related data is

$$c_s = (\ .05 \quad .05 \quad .07 \quad .09 \quad .09 \quad .11 \quad .13 \quad .13 \quad .14 \quad .14\ ) \tag{4.1}$$

and

$$\mu_k = (\ 36 \quad 72 \quad 48 \quad 60\ ). \tag{4.2}$$

Recall that $\mu_{sk} = c_s \mu_k$ can be greater than one since $s$ indexes an entire row of the wafer. The lot size $L = 20$, the probability of a defective lot is $q = .05$, the probability of a defective wafer in a non-defective lot is $p = .12$, and the arrival rate $\lambda$ of wafers, which does not affect our analysis, equals one. Finally, the transpose of the matrix of variances

$\sigma^2_{sk}$ is given by

$$\sigma^2_{sk} = \begin{pmatrix} 2 & 40 & 45 & 4 & 40 & 12 & 4 & 24 & 3 & 38 \\ 30 & 4 & 42 & 15 & 44 & 12 & 18 & 15 & 20 & 4 \\ 10 & 18 & 30 & 2 & 44 & 4 & 18 & 20 & 20 & 4 \\ 14 & 14 & 4 & 20 & 40 & 40 & 4 & 4 & 6 & 20 \end{pmatrix}^T . \tag{4.3}$$

The calculation (2.1) of the fraction of starting lots $f_k$ of type $k$ for the single-type policy is given by

$$f_k = (\, 4/63 \quad 2/63 \quad 9/63 \quad 48/63 \,). \tag{4.4}$$

Five different policies were tested in our simulation experiment. The first policy is the single-type policy (denoted by SINGLE in Table I), where type $k$ lots are released into the fab according to the product mix in (4.4). The other four policies are multi-type policies. One is the proportional design (PROP) (2.25), where $x_{sk} = f_k$, for $s = 1, ..., 10$ and $k = 1, ..., 4$. Recall that this design was proven superior to the single-type design in Proposition 6. The last three designs were derived by solving mathematical programs. The design that solves the quadratic program (3.24), (2.4), (2.23)-(2.24) will be denoted by MINSUM, and the design that solves the program (3.25), (2.4), (2.23)-(2.24) will be denoted by MINMAX. The last design solves

$$\max_{x_{sk}} \sum_{k=1}^{K} \sum_{s=1}^{S} \frac{\sigma^2_{sk} x^2_{sk}}{n^2_k} \tag{4.5}$$

subject to (2.4). (2.23)-(2.24). This design, which will be denoted by MAXSUM, was tested because it should lead to an ineffective design, among the designs that solve, (2.4), (2.23)-(2.24), thus enabling us to assess the range of performance among different multi-type wafer designs.

For all five policies tested, 100 independent runs were made, each consisting of 2000 time units. For each policy, the cumulative production at times $t = 250, 500, 1000, 1500,$ and 2000 were collected, along with 95% confidence intervals. The results are displayed in Table I. In Table II, we present the percentage improvements in performance of the multi-type policies over the single-type policy at time $t = 2000$. The cumulative production at

time 2000 is also expressed in this table as a percentage of the ideal production rate that would occur in a deterministic environment.

| POLICY | CUMULATIVE PRODUCTION AT TIME | | | | |
|---|---|---|---|---|---|
| | $t = 250$ | $t = 500$ | $t = 1000$ | $t = 1500$ | $t = 2000$ |
| SINGLE | 6789(±356) | 15658(±467) | 33146(±687) | 49894(±969) | 67792(±1150) |
| MAXSUM | 8586(±214) | 17897(±290) | 36526(±396) | 55116(±529) | 73580(±618) |
| PROP | 9229(±126) | 18671(±179) | 37240(±259) | 55834(±303) | 74325(±335) |
| MINMAX | 9476(±138) | 18815(±193) | 37425(±253) | 56275(±300) | 75251(±350) |
| MINSUM | 9526(±133) | 18871(±171) | 37633(±225) | 56375(±296) | 75342(±354) |

**TABLE I. Cumulative Production of Sets**

The results show that the two proposed multi-type policies offer an 11% improvement in cumulative production at time $t$ =2000. Notice that percentage improvements are much higher for lower time values, and are as high as 40% at time $t = 250$. Although our two proposed policies, MINMAX and MINSUM, both outperform MAXSUM, the percentage difference in cumulative production among the multi-type policies is relatively small. Thus, it appears that any solution to the LP (2.2)-(2.4) in the non-multiplicative case will be effective in increasing the production of sets. Therefore, among these policies, the designer should probably allow other factors, such as ease of design, to dictate the wafer design decision. In particular, the MINMAX and MINSUM solutions are interior points of the constraint set, and so have mostly non-zero elements, whereas other less effective objectives, such as MAXSUM (or MAXMIN) lead to more easily implemented extreme point solutions.

| POLICY | PERCENTAGE IMPROVEMENT OVER SINGLE-TYPE POLICY | PRODUCTION AS A PERCENTAGE OF UPPER BOUND |
|---|---|---|
| SINGLE | — | 88.7% |
| MAXSUM | 8.5% | 96.3% |
| PROP | 9.6% | 97.2% |
| MINMAX | 11.0% | 98.5% |
| MINSUM | 11.1% | 98.6% |

**TABLE II. Comparison of Cumulative Production of Sets at Time t=2000.**

However, the performance of the multi-type policies seem even more impressive when expressed as a function of the deterministic upper bound. The single-type policy achieves production at only 71% of the deterministic rate at time $t = 250$ and only 88.7% of the deterministic rate at time $t = 2000$, whereas the MINSUM policy produces at 98.6% of the deterministic rate at $t = 2000$. In this light, the improvement from the worst to the best multi-type policy (96.3% to 98.6%) is quite significant.

Although our analysis has been restricted to the case of one multi-type wafer design, an important consideration is the marginal improvement in performance that can be attained if several distinct wafer designs were used to satisfy the demand for sets. If the dynamic release decisions of the various multi-type wafers does not depend on any state-dependent information such as the resulting yields, then the upper bound on the performance of such a policy is still the deterministic production rate. Therefore, we conclude from Table II that very little can be gained by using an additional wafer design, and thus it is unlikely that the cost of a second mask set would be covered by the marginal improvement in production. Moreover, recall that with electron beam technology, there are no mask sets,

and a different multi-type design could be used for each lot of wafers entering the fab. Table II suggests that very little production will be sacrificed if a single multi-type wafer is used to satisfy demand with this technology, unless a more intelligent (dynamic, state-dependent) release policy were implemented with several (or many) distinctly designed wafers.

| TIME | SIMULATED PRODUCTION | PREDICTED PRODUCTION | PERCENTAGE OVER |
|---|---|---|---|
| $t = 250$ | 6789($\pm$356) | 7152 | 5.3% |
| $t = 500$ | 15658($\pm$467) | 15711 | 0.3% |
| $t = 1000$ | 33146($\pm$687) | 33412 | 0.8% |
| $t = 1500$ | 49894($\pm$969) | 51440 | 3.1% |
| $t = 2000$ | 67792($\pm$1150) | 69638 | 2.7% |

**TABLE III. Simulated vs. Predicted Production for the SINGLE Policy.**

The large gap between the performance of the single-type policy and the deterministic upper bound contradicts the apparently intuitive balanced flow assumption that "what goes in must come out". Therefore, it is very important to be able to estimate $E[N^*(t)]$ and $E[F^*(t)]$ for various times $t$. In Tables III and IV, we estimate the expected value of the cumulative production at various times for our worst (SINGLE) and best (MINSUM) policy using (3.28) and (3.29), and compare them to the values obtained via simulation. We conclude that the method of estimation described in Section 3.4 is quite accurate in estimating the cumulative production, with most of the estimates falling within $2-3\%$ of the observed values. Since this is an asymptotic estimation procedure, it is not surprising that the estimates are higher than the observed values.

| TIME | SIMULATED PRODUCTION | PREDICTED PRODUCTION | PERCENTAGE OVER |
|---|---|---|---|
| $t = 250$ | 9526($\pm$133) | 9597 | 0.7% |
| $t = 500$ | 18871($\pm$171) | 19215 | 1.8% |
| $t = 1000$ | 37633($\pm$225) | 38457 | 2.2% |
| $t = 1500$ | 56375($\pm$296) | 57704 | 2.4% |
| $t = 2000$ | 75342($\pm$354) | 76953 | 2.1% |

**TABLE IV. Simulated vs. Predicted Production for the MINMAX Policy.**

Although the initial motivation of this study was to find an effective wafer design, perhaps the most important use of the analysis is to estimate the cumulative production in a highly variable manufacturing environment where finished goods are assembled from a variety of different parts.

**Acknowledgements**

The authors would like to thank Jihong Ou for performing the numerical calculations in Section 4, and acknowledge helpful discussions with Chacko Abraham, Dan Conners, and Steve Graves. The second author was partially supported by an IBM/University Manufacturing Systems Research Grant.

## References


[1] Albin, S. L. and Friedman, D. J. (1988). The Impact of Non-Poisson Defect Distributions in IC Fabrication. To appear in *Management Science*.

[2] Clark, C. E. (1961). The Greatest of a Finite Set of Random Variables. *Operations Research* 9, 145-162.

[3] Kahane, J.-P. (1986). Une Ineqalite du Type de Slepian et Gordon sue les Processus Gaussiens. *Israel Journal of Mathematics* 55, 109-110.

[4] Kelly, F. P. (1979). *Reversibility and Stochastic Networks*. John Wiley and Sons, New York.

[5] Singh, M. R., Abraham, C. T., and Akella, R. (1988). Planning for Production of a Set of Components When Yield is Random. *Fifth IEEE CHMT International Electronic Manufacturing Technology Symposium Proceedings*.

[6] Slepian, D. (1962). The One-Sided Barrier Problem for Gaussian Noise. *Bell System Technical Journal* 41, 463-501.

[7] Stapper, C. H. (1985). The Effects of Wafer to Wafer Defect Density Variations on Integrated Circuit Defect and Fault Distributions. *IBM Journal of Research and Development* 29, 87-97.

[8] Stapper, C. H. (1986). The Defect-Sensitivity Effect of Memory Chips. *IEEE Journal of Solid-State Circuits* SC-21, 193-198.