Allocation of Chips to Wafers in a Production Problem of Semiconductor Kits

Sridhar Seshadri; J. G. Shanthikumar

Operations Research, Vol. 45, No. 2 (Mar. - Apr., 1997), 315-321.

Stable URL:
http://links.jstor.org/sici?sici=0030-364X%28199703%2F04%2945%3A2%3C315%3AAOCTWI%3E2.0.CO%3B2-M

Operations Research is currently published by INFORMS.

ALLOCATION OF CHIPS TO WAFERS IN A PRODUCTION PROBLEM OF SEMICONDUCTOR KITS

SRIDHAR SESHADRI

Leonard N. Stern School of Business, New York University, New York

J. G. SHANTHIKUMAR

University of California, Berkeley, California

(Received July 1993; revision(s) received June 1994 and August 1994; accepted December 1994)

The problem of maximizing the production of *good sets* of semiconductor chips under random yield is reexamined in this paper. (A set of semiconductor chips is called a semiconductor kit.) This problem has been considered by Avram and Wein (1992) and Singh et al. (1988). To solve this problem we show that under certain combinations of assumptions the production process can be replaced by a black box. The use of the black box model considerably simplifies the analysis and reduces the simulation effort required for carrying out parametric analysis of the proposed solution procedure. The model includes that of Avram and Wein, and we extend their results to more general settings and strengthen their conclusions. Using the black box model, it is shown that the strategy of placing different types of chips on a single wafer gives larger yield of kits in a stochastic sense than the traditional method of placing single types of chips on a wafer. We compare the production of kits under different chip design and lot release policies and also carry out a parametric analysis with respect to factors such as set proportions and yield.

In the semiconductor industry customers typically order *sets* of different chips called a semiconductor kit. In this paper we consider the problem of maximizing the production of these kits under random yield. The problem is of practical interest as yields in semiconductor fabrication can be as low as 20% when making new or custom-built chips, and random failures of any one type of chip can disrupt the delivery of complete sets. The chips that make up the kit are manufactured in three steps, namely wafer preparation, wafer fabrication, and assembly and testing. We model the second step of this production process, because the lead time, the complexity of processes, as well as the investment in facilities are the greatest in wafer fabrication. Wafers containing anywhere from 20 to several hundred chips are manufactured in a semiconductor fab. While in the past only a single type of chip could be produced on a wafer, due to technological advances in the area of semiconductor fabrication it has now become possible to design and produce the different chips required in a kit on the same wafer. The design in which only a single type of chip is made on a wafer will be called a *single type* design, and the one in which several types of chips are made on the same wafer will be called a *multitype* design. Avram and Wein and Singh et al. showed that the multitype design can help counter the randomness due to yield variations in the fab by making chips of different types fail together—thus increasing the production of complete kits in a stochastic sense. This was a valuable contribution because any improvement in the time to execute an order for sets can lead to significant competitive advantage in this industry.

The timing and sequence of release of lots of wafers into the fab are termed the *release* or *input control* policy. When the different types of chips required for the kit are produced in the same fab using the single type design, the yield of kits will be influenced by the release control policy. Avram and Wein, using a quasi-reversible queueing network as the model of the fab, showed for a particular release control policy that the expected number of kits produced by a fixed time t is larger for the multitype design than the expected yield from the single type design. One of the objectives of this paper was to understand how the single type design would compare with the multitype design, when produced using different release control policies. To study the effect on kit yield due to different release control policies it is important to understand how a given stream of inputs to the fab gets transformed into streams of outputs from the fab. In Section 1, we show that the input streams get almost exactly transformed into output streams from the fab, i.e., the fab can be treated as a black box under some generally valid assumptions. In Section 2 we provide definitions and summary of the modeling assumptions. In Section 3, using the black box model of the fab, we extend Avram and Wein's results to more general settings and strengthen their conclusions by obtaining stochastic relationships between the outputs of kits under the two designs and different release control policies.

A recent use of the black box model can be found in Connors and Yao (1996). While Connors and Yao's objective is different, namely to minimize the input of wafers into the fab subject to meeting a given demand either with

Subject classifications: Inventory/production: random yields: product design. Industries: semiconductor. Probability: Stochastic model applications.
Area of review: MANUFACTURING, OPERATIONS AND SCHEDULING.

0030-364X/97/4502-0315 $05.00

a specified probability or to exceeding a given number of kits in expectation, the fab is modeled by them as a black box in a manner similar to ours, and they use the notion of associativity to lower bound the probability of obtaining kits as done in Claim 4 of this paper. The black box modeling approach not only simplifies the task of analyzing different release policies but also reduces the simulation effort needed for quantifying the value of the multitype design. Based on our results, we conclude that a multitype design is especially advantageous when the yields are low and the set proportions are more uneven.

1. PROCESS AND YIELD MODELS EXPLAINED AND COMPARED

In this section we briefly describe the yield models used in this paper, the modeling of the production processes as a black box, and the release controls investigated by us and compare our approach to the modeling methods used in Avram and Wein. For a more detailed description of the production process and the modeling issues, the reader is referred to Avram and Wein's paper and the references cited in this section. The major modeling assumptions along with the notation and the release controls are summarized in the next section.

Yield Models

The yield of good chips from the fab depends on lot failures, wafer failures, and chip failures. In this paper we model the lot and wafer failures as Bernoulli processes and assume that these failures are independent of the type of chips, as done in Avram and Wein. The latter assumption is valid if the lot and wafer failures are due to factors such as mishandling, intrusion of dirt, or machine failures, but also see the remark following Claim 5. When a lot or wafer fails, all the chips in the lot or wafer are assumed to be defective. A wafer has several *chip sites*. There is consensus that chip failures will depend both on the type of chip and the site (see, for example, Chapter 14 in Sze 1988, Ferris-Prabhu et al. 1987, Michalka et al. 1990, and Stapper 1985). Avram and Wein use a *multiplicative yield* model in their analysis to capture this twofold effect. In their model the expected yield is expressed as the product of two factors, one due to the site and the other due to chip type. We allow both a random yield model as well as the multiplicative yield model in our analysis (see *A0* below for more details about the two yield models). In the random yield model, the yield per wafer of chip type i is a random variable Y_i and is assumed to be independent of the yields of other types of chips placed on the same wafer. However as explained in Claim 4, the stochastic order relations can be extended when the yields of different types of chips on the same wafer are positively correlated. It should be noted that defect clustering on a wafer is still not fully understood and is the subject of ongoing investigation. The basic problem in developing a realistic defect model lies in correctly capturing the dependency between failures at different chip sites. By using a site factor only part of the effect of location on chip failure gets modeled. This has implication for determining "good" multitype designs.

Release Control

The first dimension of release control is deciding what should be the long run (time) average of the different types of chips released to the fab. Avram and Wein assume that the fractions of different types of chips released into the fab are such that the expected numbers of different types of good chips produced will be in the proportion required for making up the kit. Such an input policy is said to be *failure proportioned*. The failure proportioned input policy is adopted by us, too; see *A1* in the summary of assumptions. Singh et al. (1988) show that adopting the failure proportioned policy need not be optimal for the static problem of producing a *single* semiconductor kit. In the dynamic setting, when say several hundred kits are required, it is still an open question what should be the right mix entering the fab, especially when feedback regarding yields can be obtained in a timely manner. In what follows, a site is usually a row of the wafer and fractional allocation of a chip type to a site is allowed for analytical ease and is not too unrealistic.

The timing of inputs into a wafer fab has two dimensions: (i) deciding when to release a lot and (ii) deciding which type of wafer will be released next. For the first part Avram and Wein assume that the arrival stream of work to the fab can be represented by a Poisson process. For the second part, they assume that in the case of the single type design the arrival processes of different chip types to the fab are independent Poisson processes. This is termed a *random release policy* in our paper. In contrast, firstly, we do not assume anything about the composite arrival stream of work into the fab and instead work with a given distribution of output from the fab. Secondly, when dealing with the single type design, in addition to the random release policy, we permit *cyclic release* where wafers containing the different types of chips are released cyclically in sequence and in the proportion required to form a set. Another aspect of release control is deciding whether the control will be at the level of lots or at the level of wafers. When the control is at the level of lots, the lot will consist of the same type of wafers, whereas when the control is at the level of wafers the lot can be made up of different types of wafers. For the single type design, wafers usually consisting of the same type of chip are made into lots for input into the fab. In the first four claims in Section 3, we assume that the release control is at the level of wafers. In Claim 5 we examine the case when the control is exercised at the level of lots.

Model of the Production Process

A motivation for this paper was to understand how the multitype designs would perform when compared to the single type design produced using cyclic release. In order to study the impact of cycling the inputs, we had to model

how a given stream of inputs gets transformed into a stream of outputs from the fab. It turns out that under certain combinations of assumptions, it is possible to assume that the input stream of wafers gets mapped almost exactly into the output stream of wafers from the fab. We now describe two alternate sets of assumptions regarding the production process that are used in this paper to justify this mapping. First assume that the processing sequence and processing times for all designs are identical. (Avram and Wein, too, assume that the multitype design can be produced in a manner similar to the one used to produce the single type design; i.e., in comparison to the single type design it is assumed that the multitype wafer can be produced in about the same time, using similar production facilities, and the production processes will give similar yields.) Under this assumption, if the lots were processed as one unit at all stations then the stream of outputs under different designs will be the same. If the lots are not processed as one unit, but if there are single machines at each work center (*or* multiple machines with deterministic processing times), and the first in first out (FIFO) sequencing policy is used at all work centers, then it can be shown that the input stream gets mapped exactly into the output stream. The assumption of single machines at work centers, or of deterministic processing times with multiple machines, is made to prevent overtaking of a lot by a lot that was injected later into the fab. Excepting the assumption on sequencing, these assumptions are almost always met in practice. The assumptions are summarized in *A3*, *A4*, and *A5* (i) & (ii).

A second set of assumptions is that (a) we are concerned about the yield of complete kits over a long period of time and (b) the number of wafers (or lots) in the fab at any point in time is negligible compared to what is produced over a long period of time; see *A5* (iii). It then follows that the inputs of wafers are mapped into outputs until some point of time t, with an error equal to the number of wafers in the fab at the time t. (As the kit size is fixed, the error when measured in terms of kits will also be small; see Claim 2 for an application of this reasoning.) This set of assumptions dovetails with a requirement in Claim 4 that the number of chips produced is large enough for the normal approximation to hold good; see *A2*. Under the second set of assumptions, excepting process yields we need not be concerned whether the production processes are the same for making the two designs. These assumptions are seen to include the situation when the fab processes are modeled using a quasi-reversible queueing network, like in Avram and Wein. Either of these sets of assumptions allows us to assume that the inputs are almost exactly mapped into outputs and thus the fab itself appears to be a black box. The quasi-reversible network is also a black box in this sense but with restrictions placed on the input and fab production processes. Connors and Yao, using an approach similar to ours, work directly with the yield model, bypassing the fab production processes.

2. DEFINITIONS AND DESCRIPTION OF THE MODELING ASSUMPTIONS

Definitions and Notation

L = Lot size,
P(Lot is good) = p_L independent of all else,
W = Number of chips per wafer,
P(Wafer is good | Lot is good) = p_W,
Y_i = Random yield of chips from a wafer containing chip type i,
Number of chip types = k,
Set requirement proportions = $\{f_1 : f_2 : \ldots : f_k\}$ and f_i are integers,
$M(t)$ = Number of lots produced by time t,
$N(t)$ = Number of wafers produced by time t,
$N_g(t)$ = Number of good wafers produced by time t,
$N_{gi}(t)$ = Number of good wafers of type i produced by time t,
$N_c(t)$ = Number of chips produced by time t,
$N_{ci}(t)$ = Number of chips of type i produced by time t,
$S^{\alpha}(t)$ = Number of complete kits produced by time t under policy α,
$p_i := EY_i/W$ is the yield of chip type i expressed as a fraction,
μ_j = Yield factor of row j on a wafer,
t_i = Yield factor of a type i chip, and
nrow = Number of rows in a wafer

The semiconductor kit required by the customer is the set of k types of chips in the proportions $\{f_1 : f_2 : \ldots : f_k\}$. We assume that inspection is done after all processing has been completed in the fab. $N(t)$ is the number of wafers, good or bad, produced until time t, and $N_g(t)$ is the number of wafers out of $N(t)$ that are not completely defective. $N_c(t)$ and $N_{ci}(t)$ denote the number of good chips produced from good wafers. When necessary we use $\bar{N}_{ci}(t)$ to denote the total number of type i chips, whether good or bad, obtained from the $N_g(t)$ good wafers.

Summary of Modeling Assumptions

Six assumptions are used as stated below:

A0. *Either the random yield model or the multiplicative Bernoulli yield model is assumed to hold* (*also see the references cited in Section* 1 *and Avram and Wein for a more detailed description of the yield models and the modeling implications*).

(a) **Random yield model:** (i) The yield per wafer of chip type i is a random variable Y_i. It is assumed that the yields of chip type i from different wafers are independent and identically distributed. (ii) When a proportion r_i of the chip sites in each row is allotted to chip type i, the yield of type i chips from the wafer is independent of the yields of other types of chips placed on the same wafer, and has mean $r_i E(Y_i)$ and variance $r_i \mathrm{Var}(Y_i)$.

(b) **Multiplicative Bernoulli yield model:** Given that a wafer is good, the probability that a chip of type i located

in row j will be good is given by $\mu_j t_i$. The μ_js are termed the *site factor* and t_is the *type factors.* For carrying out the failure proportioned allocation, p_i, the average yield per wafer of chip type i, expressed as a fraction is given by the average of the μ_js times the type factor, i.e., $p_i = \Sigma_j(\mu_j t_i)/nrow$.

A1. *In all policies, the relative fractions of chip types input into the fab are maintained as:*

$$\left(\frac{\frac{f_1}{p_1}}{\sum_i \frac{f_i}{p_i}}, \frac{\frac{f_2}{p_2}}{\sum_i \frac{f_i}{p_i}}, \ldots, \frac{\frac{f_k}{pk}}{\sum_i \frac{f_i}{p_i}} \right) = (r_1, r_2, \ldots, r_k).$$

This allocation scheme is said to be *failure proportioned.* For the random release policy the fractions are maintained in the expected sense, for the cyclic release policy the ratios are maintained in each lot if wafers are cycled into the fab, or over one or more cycles of release if lots of single type wafers are cycled into the fab. For the multitype wafer, we attempt to maintain the ratio in each wafer.

A2. *The number of lots produced by time t, $M(t) \uparrow \infty$ with t.*

A3. *The lot size, L is such that Lr_i is an integer for all i.*

A4. *The internal control policies in the fab are of open loop type, and do not use information on the type of chip or type of wafer. The processing sequence and times of all wafer types, including a multitype design, are identical (have identical distributions when the processing times are random variables).*

A5. (i) *Lots are processed as one unit, i.e., lot by lot. or*

(ii) *The fab consists of single machines at each work center (or multiple machines with deterministic processing times), and the FIFO sequencing policy is used at all work centers. or*

(iii) *We are concerned with the yield of kits over a long period of time t, and if $R^a(t)$ were the number of lots still in the fab at time t, a = I, II, III, then:*

$$|R^a(t) - R^b(t)|/M^a(t) \to 0 \quad as\ t \uparrow \infty,\ a, b = I, II, III.$$

Policies Compared

Three different policies denoted as I, II, and III are compared in the analysis and the simulations. These policies combine the wafer design and input control decision and are described below:

I. STRR or Single type random release policy, where the probability that the n^{th} wafer released into the fab has type i chips is given by r_i, $i = 1, 2, \ldots, k$.

II. STCR or Single type cyclic release policy, where the k types of wafers are released in sequence in the proportion $\{r_1 : r_2 : \ldots : r_k\}$.

III. MTFP or Multitype failure proportioned policy, in which the fraction of chip type i allocated to site s on the wafer is given by r_i. In this policy the allocation is the same for all rows.

3. ANALYSIS

The policy numbers I, II, and III will be used as superscripts to indicate which policy is being examined. As explained in the previous section, we shall be working with a given distribution of lots produced by time t, $M(t)$. The first observation is that:

Claim 1. *Under A3, A4, and A5 we can "ignore" the lot failures and work only with wafer failures for computing the output of kits produced using the three policies.*

Proof. First we note that wafers of different types are assumed to make up a lot under the release policies described above. Further release control determines the composition of a lot by *A3.* (However, the assumption that wafers of different types and not lots of different types are sequenced into the fab is made for ease of exposition and can be eliminated as explained in Claim 5.) By *A4* and *A5* (i) or (ii) we can assume that the input sequence of wafers is almost exactly mapped into the output sequence of wafers. This mapping is by assumption under *A4* and *A5*(i) and follows from *A5*(ii) because no overtaking is possible at any of the work centers. *A5*(iii) states that for large t, the fraction of lots still left in the fab will be very small compared to those already produced; which allows us to assume with very little error that the input stream has been exactly mapped into the output stream of wafers. Coming to the claim:

For policy *III* ignoring lot failures and working with only good lots for computing the output of kits introduces no errors. For policy *I*, as the input stream is mapped into the output stream, the random release policy ensures that working with good lots is as good (in the stochastic sense) as working with all lots for calculating the yield of good chips of different types. For policy *II*, the integrality of Lr_i guarantees that working with good lots is permissible if the processing is done lot by lot. If the lots are broken up while processing, we shall still assume that they get reconstituted at each step of processing, and also that the lot failure probability is applicable to the original lot of wafers input to the fab. *A5*(ii) or *A5*(iii) can then be used to justify the claim for policy II. □

The assumption that Lr_i is an integer is quite critical to the proof of the above claim and is made because it simplifies the subsequent proofs. The more practical case, in which lots made up of the same type of wafer are cycled into the fab, is examined in Claim 5. In the remainder of the claims we shall be assuming that $N_g(t)$ is the same under all policies because the wafer failures do not depend on the type of chip and as inputs are almost exactly mapped into outputs. An example of the use of *A5*(iii) is given below which makes the mapping argument more precise.

Claim 2. *If $p_i = 1$ for all chip types, i.e., when there are no chip failures, then $S^{III}(t) \geqslant_{st} S^I(t)$, where $\geqslant_{st}$ stands for larger in the usual stochastic order.*

Proof. This follows from the definition of $\geq_{st}$ (see Stoyan for example) and:

$$P(S^{III}(t) \geq x) = P\left(N_g^{III}(t) \geq \left(\sum_i f_i\right)x\right)$$

$$\approx P\left(N_g^{I}(t) \geq \left(\sum_i f_i\right)x\right) \geq P(S^{I}(t) \geq x).$$

The proof under *A5*(iii) can be modified as follows:

$$P(S^{III}(t) \geq x)$$

$$= P\left(N_g^{III}(t) \geq \left(\sum_i f_i\right)x\right)$$

$$\geq P\left(N_g^{I}(t) \geq \left(\sum_i f_i\right)x + |N_g^{I}(t) - N_g^{III}(t)|\right)$$

$$\geq P\left(S^{I}(t) \geq x + |N_g^{I}(t) - N_g^{III}(t)| \Big/ \left(\sum_i f_i\right)\right), \quad \text{and}$$

$$\left|P\left(S^{I}(t) \geq x + |N_g^{I}(t) - N_g^{III}(t)| \Big/ \left(\sum_i f_i\right)\right) - P(S^{I}(t) \geq x)\right| \to 0$$

as $t \uparrow \infty$ by *A2* and *A5*(iii). □

Separate proof will not be given under *A5*(iii) in the remaining claims, as the ideas are similar to the ones used in the above example.

Claim 3. *If $p_i = 1$ for all chip types, i.e., when there are no chip failures, then $S^{III}(t) \geq_{st} S^{II}(t)$.*

Proof. This follows from observing that if we first allocate *n wafers* to the k types in the proportions $\{r_i\}$ and get x kits out of the good wafers of each type, then we will also get x kits by selecting the same number of good wafers from n and then allocating them to chip types. To see this, let x_i be the number of type i chips produced under policy *II*. As there are no chip failures, the total number of good chips produced under *III* will be $\Sigma_i x_i$. Then, kits produced under *II* $= \min_i\{x_i/f_i\} \leq \Sigma_i x_i/\Sigma_i f_i =$ kits produced under policy *III*. □

The stochastic order relations we could establish in the presence of chip failures are weaker.

Claim 4. *For large t, $S^{III}(t) \geq_{icv} S^{II}(t) \geq_{icv} S^{I}(t)$, where the relationship $\geq_{icv}$ denotes stochastically larger in the increasing concave sense.*

Proof. The basic idea of the proof is to reduce the problem to one of comparing two multivariate normal distributions with independent components and equal means, but with one having larger variances than the other. The result will then follow from an application of the generalized Slepian inequality used by Avram and Wein, Proposition 4. We will assume throughout that the normal approximation to the distribution of the number of chips produced by time t holds due to *A2 and* the finiteness of the second moment of the corresponding random variables because of *A0*. Therefore we assume that there exists a constant $c(t)$ such that the distribution of $N_{ci}(t)/c(t)$ is very nearly normal for *all* policies. Let $Z_{ci}^{a}(t)$, $a = I, II, III$; $i = 1, 2, \ldots, k$ be normally distributed random variables such that:

if a = I or II: $Z_{ci}^{a}(t)$, $i = 1, 2, \ldots, k$ are independent with mean and variance given by $E(N_{ci}^{a}(t))/c(t)$ and $\mathrm{Var}(N_{ci}^{a}(t))/c(t)^2$.

if a = III: $Z_{ci}^{III}(t)$, have the multinormal distribution with means $E(N_{ci}^{III}(t))/c(t)$, and product moments $E(N_{ci}^{III}(t)N_{cj}^{III}(t))/c(t)^2$.

First consider policies II and III. Fix the number of wafers produced as n/W. Then $Z_{ci}^{III}(t)$, $i = 1, 2, \ldots, k$ are associated random variables. To prove this assume without loss of generality, $\tau = x/r_i \geq y/r_j$. Let $\bar{N}_{ci}^{III}(t)$ be the number of chips, good or bad, of type i produced by time t. Then,

$$P(\bar{N}_{ci}^{III}(t) > x;\ \bar{N}_{cj}^{III}(t) > y | WN(t) = n)$$

$$= \sum_{z>\tau} \binom{n}{z} p_W^z (1 - p_W)^{n-z}$$

$$\geq \left[\sum_{z>\tau} \binom{n}{z} p_W^z (1 - p_W)^{n-z}\right]$$

$$\times \left[\sum_{z>y/r_j} \binom{n}{z} p_W^z (1 - p_W)^{n-z}\right]$$

$$= P(\bar{N}_{ci}^{III}(t) > x | WN(t) = n)$$

$$\times P(\bar{N}_{cj}^{III}(t) > y | WN(t) = n).$$

Therefore the covariance between $\bar{N}_{ci}^{III}(t)$'s is nonnegative. Using the theorem due to Pitt (see 5.1.1 in Tong 1990), the normal approximations to the random variables $\bar{N}_{ci}^{III}(t)/c(t)$'s are associated. Using the property that increasing functions of associated random variables are associated (see, for example, property P_3 on p. 30 of Barlow and Proschan 1981) and also the fact that under both the yield models the yield of good chips is an increasing function of the total number of chips (good or bad), we obtain that the random variables $Z_{ci}^{III}(t)$, $i = 1, 2, \ldots, k$, are associated for fixed n.

The means of $Z_{ci}^{III}(t)$ and $Z_{ci}^{II}(t)$ are the same (because of the use of the failure proportioned input policy), but the latter has a larger variance due to the *lot sizing* effect on a wafer, i.e.,

$$\mathrm{Var}(N_{ci}^{III}(t)) = r_i^2 E(Y_i)^2 \mathrm{Var}(N_g^{III}(t)) + r_i E(N_g^{III}(t)) \mathrm{Var}(Y_i)$$

$$\approx r_i^2 E(Y_i)^2 \mathrm{Var}(N_g^{II}(t)) + r_i E(N_g^{II}(t)) \mathrm{Var}(Y_i)$$

$$\leq r_i E(Y_i)^2 \mathrm{Var}(N_g^{II}(t)) + r_i E(N_g^{II}(t)) \mathrm{Var}(Y_i) = \mathrm{Var}(N_{ci}^{II}(t)).$$

The associatedness permits us to work with independent $Z_{ci}^{III}(t)$, because the minimum of associated random variables is stochastically larger compared to the minimum of independent random variables with the same marginal distributions; for example, see Theorem 3.2 of Barlow and Proschan. Next we use the fact that any increasing concave function of the number of kits is a concave function of the

Table I
Variance of Good Type i Chips Given N Lots are Produced

Policy	Variance
III	$Np_L(r_iLp_W\text{Var}(Y_i) + r_i^2E(Y_i)^2Lp_W(1-p_W)) + Np_L(1-p_L)(r_iLp_WE(Y_i))^2$
II	$Np_L(r_iLp_W\text{Var}(Y_i) + r_iE(Y_i)^2Lp_W(1-p_W)) + Np_L(1-p_L)(r_iLp_WE(Y_i))^2$
I	$Np_L(r_iLp_W\text{Var}(Y_i) + E(Y_i)^2Lr_ip_W(1-r_ip_W)) + Np_L(1-p_L)(r_iLp_WE(Y_i))^2$
II-L	$r_iNp_L(Lp_W\text{Var}(Y_i) + E(Y_i)^2Lp_W(1-p_W)) + Nr_ip_L(1-p_L)(Lp_WE(Y_i))^2$
I-L	$r_iNp_L(Lp_W\text{Var}(Y_i) + E(Y_i)^2Lp_W(1-p_W)) + Nr_ip_L(1-r_ip_L)(Lp_WE(Y_i))^2$

number of good chips of different types, i.e., $f(\cdot)$ is concave and increasing and $g(\cdot)$ is concave implies:

$$f(g(\lambda x_1 + (1-\lambda)x_2)) \geq f(\lambda g(x_1) + (1-\lambda)g(x_2))$$
$$\geq \lambda f(g(x_1)) + (1-\lambda)f(g(x_2)).$$

The first inequality follows by concavity of $g(\cdot)$ and because $f(\cdot)$ is increasing. The second inequality follows from the concavity of $f(\cdot)$. So an application of the Slepian inequality shows that the expected value of any increasing concave function of the number of kits is larger under policy III compared to policy II. The increasing concave ordering now follows from the definition of $\leq_{icv}$; see Stoyan for example. A similar proof can be given when there is positive dependence between the yields of different types of chips placed on the same wafer. The proof can be shown to extend to the case when lots and *not* wafers are sequenced cyclically into the fab.

To compare policies I and II, observe that given n wafers are produced, the probability of getting x kits under policy I is smaller than when the good wafer of different types are generated *independently* using binomial trials with success probabilities r_ip_W. These independent processes generating the good wafers have the same means but larger variances compared to the processes that generate the good wafers under policy II. This leads to the desired conclusion just as in the previous case. □

The arguments are similar when lots and not wafers are randomly injected into the fab. Consider two more policies: (i) STRRL: A single type random release policy, under which the probability that the n^{th} *lot* released into the fab has type i chips is given by r_i, $i = 1, 2, \ldots, k$; and (ii) STCRL: A single type cyclic release policy, where each lot is to be made up of only a single type of wafer and the k types of *lots* are released in sequence in the proportion $\{r_1:r_2:\ldots:r_k\}$. These two policies will be denoted by the superscripts I-L and II-L.

Claim 5.

(1) $S^{III}(t) \geq_{icv} S^{II}(t) \geq_{icv} S^{II\text{-}L}(t) \geq_{icv} S^{I\text{-}L}(t)$,

(2) $S^{III}(t) \geq_{icv} S^{II}(t) \geq_{icv} S^{I}(t) \geq_{icv} S^{I\text{-}L}(t)$,

and if $(1-p_L)L \leq 1$ *then*

(3) $S^{III}(t) \geq_{icv} S^{II}(t) \geq_{icv} S^{II\text{-}L}(t) \geq_{icv} S^{I}(t) \geq_{icv} S^{I\text{-}L}(t)$.

Proof. Consider (1) and (2) first. Because of Claim 4, we need to prove only the last two relationships in (1) and the last relationship in (2). The variances of good chips under the different policies are given in Table I. From the table we see that the variance of good chips increases as we move from policy II to policy II-L to policy I-L. As done in Claim 4, we argue that the probability of obtaining a given number of kits under I-L is smaller than when the good lots of different types are generated independently using binomial trials with success probabilities r_ip_L. The use of these two facts and of a multinormal approximation as well as an application of the Slepian inequality leads to (1). The arguments for proving (2) are similar. The comparison of policies I and II-L is difficult because the direction of the stochastic inequality will depend on the magnitudes of the two lot sizing effects, one due to lot failures and the other due to wafer failures. However by comparing the variances in Table I, we see that if $(1-p_L)L$ is smaller than one, then the variance of good chips under II-L is smaller than the variance under I. This observation leads to (3). □

Remark. The formulae given in Table I may be used to compare the different designs and release control policies even when the lot and wafer failure probabilities are different for different designs and/or depend on the chip type. The restriction in their use is that if the variance under the multitype design exceeds that under the single type design, we cannot directly conclude that the single type design is better.

Parametric Analysis

We can carry out a parametric analysis based on the results shown in Table I to determine the factors that favor the use of a multitype design. We carry out the analysis with respect to lot and wafer failure probabilities as well as the set proportions. Wafers are usually released in lots; therefore we restrict the analysis to policies III and II-L. Denote the variance under these policies to be σ^2_{III} and $\sigma^2_{II\text{-}L}$, respectively. From Table I, we obtain:

$$\sigma^2_{III}/\sigma^2_{II\text{-}L} = \frac{\text{var}(Y_i)/E(Y_i)^2 + r_i(1-p_W) + r_i(1-p_L)p_WL}{\text{var}(Y_i)/E(Y_i)^2 + (1-p_W) + (1-p_L)p_WL}.$$

In practice the lot size is around 20 wafers per lot; therefore the last term in the numerator as well as the denominator of this expression will tend to dominate. This shows that in practice, the higher the lot failure probability, the greater will be the improvement in kit yield from using a

multitype design. A similar analysis shows that if one type of chip is predominantly required in making up the set, then too the multitype design will be favored. The wafer failure probability is possibly the least important of the three variables.

Claim 6. *In the absence of wafer failures and when site failures are independent of one another,* $S^{III}(t) \stackrel{d}{=} S^{II}(t) \stackrel{icv}{\geq} S^{I}(t)$; *where* $\stackrel{d}{=}$ *stands for equality in distribution.*

Proof. Only the first assertion needs to be proved. But when there are no wafer failures, we are concerned only with chip failures at the sites. Given N wafers are produced, the number of type i chip allocated to site j is the same under policies *II* and *III* for all (i, j). The independence assumption is then used to justify the claim. □

Remarks. (1) The cases when there are no chip failures and the one in which there are no wafer failures in some sense are at the extreme ends of a spectrum. Claims 3 and 6 show that the multitype design is better than the single type design for these cases. Claim 4 shows that the result holds for some in-between cases as well.

(2) Assumptions *A3* and *A4* are restrictive. An open question (based on *A4*) is whether closed loop control will improve the performance of the cyclic release policy?

(3) Assumption *A0* is convenient for making comparisons, but maintaining the proportions r_i may not be possible due to the integer number of locations per site. Secondly, the proportional allocation scheme need not result in the best multitype chip design. We reason that chip types required in relatively smaller proportions will need better protection against failures when the number of wafers produced is small. This can be justified for example when only one set is required, consisting of one chip of type 1 and 100 of type 2. Determining the optimal allocation scheme for a multitype design is difficult as the optimal proportions will depend on the number of kits required, and thus will be related to the distribution of the wafers put out by the fab (*cf.* Singh et al.'s results). The paper by Connors and Yao deals with some of these issues.

(4) We have not discussed *different* multitype designs in this paper. Avram and Wein in their paper conclude that the scope for improving the yield from a wafer through allocation of chip types to sites may be marginal unless specific clusters of defects can be identified to group the location of chip types. Our experience, which is based on simulating the black box model of the fab and using the multiplicative yield model, is similar. Details of the simulations and the analysis can be had by writing to the authors.

4. CONCLUSIONS

The result from the analysis in Section 3 is that the multitype designs are better than the single type design. However the parametric analysis indicates that:

- For a stabilized design and process, implying low lot and wafer failure rates, the cyclic release model may perform well enough to offset the advantage of multitype design due to additional costs involved in producing the multitype wafers.
- When the set proportions are more uneven, the multitype design offers greater scope for improving the yield.

ACKNOWLEDGMENTS

Sridhar Seshadri was partly supported by grants to the University of California, Berkeley from the California State MICRO program and the Semiconductor Research Corporation. J. G. Shanthikumar was supported in part by Sloan Foundation grants for the study on "Competitive Semiconductor Manufacturing." We thank the anonymous referee who brought the Connors and Yao paper to our attention. The detailed comments from the associate editor and a referee, and suggestions from the area editor helped improve the presentation considerably. Their comments also led to a revision of our assumptions and greater generality of the results.

REFERENCES

AVRAM, F. AND L. WEIN. 1992. A Product Design Problem in Semiconductor Manufacturing. *Opns. Res.* **40,** 5, 986–998.

BARLOW, R. E. AND F. PROSCHAN. 1981. *Statistical Theory of Reliability and Life Testing: Probability Models*. To Begin With, Silver Spring, MD.

CONNORS, D. P. AND D. D. YAO. 1996. Methods for Job Configuration in Semiconductor Manufacturing. *IEEE Trans. Semiconductor Manufacturing*, **9,** 3, 401–411.

FERRIS-PRABHU, A. V., L. D. SMITH, H. A. BONGES, AND J. K. PAULSEN. 1987. Radial Yield Variations in Semiconductor Wafers. *IEEE Circuits and Devices Magazine.*

MICHALKA, T. L., R. C. VARSHNEY, AND J. D. MEINDL. 1990. A Discussion of Yield Modeling with Defect Clustering, Circuit Repair, and Circuit Redundancy. *IEEE Trans. Semiconductor Manufacturing,* **3,** 3, 116–127.

SINGH, M. R., C. T. ABRAHAM, AND R. AKELLA. 1988. Planning for Production of a Set of Components When Yield is Random. Fifth IEEE CHMT International Electronic Manufacturing Technology Proceedings.

STAPPER, C. H. 1985. The Effects of Wafer to Wafer Defect Density Variations on Integrated Circuit Defect and Fault Distributions. *IBM Res. Develop.* **29,** 1, 87–97.

STOYAN, D. 1983. *Comparison Methods for Queues and Other Stochastic Models*. John Wiley & Sons, New York.

SZE, S. M. 1988. *VLSI Technology (Second Ed.)*. McGraw-Hill, New York.

TONG, Y. L. 1990. *The Multi-variate Normal Distribution*. Springer-Verlag, New York.