Manufacturing Education at San Jose State University. These activities wil be further developed at SJSU and at other universities; the extent and quality of development will be determined by the demand and support of the semiconductor manufacturing industry.

## References


[1] D. K. Koska and J. D. Romano, "Countdown to the Future: The Manufacturing Engineer in the 21st Century," Society of Manufacturing Engineers Rep., Dearborn, MI, 1988.

[2] *1989 Annual Report, ABET*, New York, NY.

[3] L. Fuller, in *Proc. 8th Biennial University Government Industry Microelectronics Symp.*, 1989, p. 108.

[4] ——, in *Proc. 7th Biennial University Government Industry Microelectronics Symp.*, 1987, p. 147.

[5] *1990 SRC Source Book*, Higher Education Publication Inc., Falls Church, VA.

[6] C. Skinner, *Channel Magazine*, SEMI, Mt. View, CA, Feb./Mar. 1991.

[7] *Engineering Projects Program*, School of Engineering, San Jose State University, San Jose CA.


# On the Relationship Between Yield and Cycle Time in Semiconductor Wafer Fabrication

Lawrence M. Wein

***Abstract*—We derive a relationship between the mean amount of time wafers spend in the fab and the mean production rate of nondefective die, assuming the number of defects per die is a Poisson random variable whose mean varies linearly with the amount of time the wafer spends in the fab.**

## I. Introduction

An overriding concern in the semiconductor industry is *yield*, which is the mean proportion of die on a wafer that can ultimately be sold. It is well recognized that particulate contamination has a detrimental effect on yield (see, for example, Osburn *et al.* [5]), and that the amount of particulate contamination on a wafer increases as the wafer spends more time in the clean room, particularly when there is no surface preparation or cleaning between process steps. These two facts suggest that an increase in manufacturing *cycle time*, which is the amount of time a wafer spends in the clean room, should lead to a reduction in yield. However, little or no data has been published to back up this hypothesis, perhaps because of the sensitive nature of the data, or because of the intricate interdependence between yield and cycle time.

Despite the lack of published data, this paper investigates the implications of the premise that longer manufacturing cycle times lead to lower yields. We develop a simple mathematical model that incorporates the random effects of both yield and queueing. In particular, we assume that the mean number of defects per die on a wafer is a linear function of the wafer's cycle time, and that defects occur independently on each die according to a Poisson model, which requires that defects are uniformly and randomly distributed. In order to measure cycle time, we model the clean room as a simple single-server queueing system with exponential interarrival times and service times. Although the same qualitative results could be obtained by employing a more realistic queueing model (for example, modeling the clean room as a network of queues, as in Chen *et al.* [1]) and a more sophisticated yield model (such as the negative bionomial model or a time-dependent model; see Cunningham [2] for a recent survey of yield models), the analysis would become much more tedious. Consequently, our model may be too simplistic to reliably predict system performance in wafer fabs; however, the implications of our study still have considerable practical significance.

Based on the previous assumptions, this paper derives a closed form relationship between mean cycle time and *throughput rate*, which is defined as the mean number of nondefective die produced per unit of time *divided* by the number of die on a wafer; the throughput rate is defined in terms of wafers (rather than die) in order to maintain consistent units between the input and output of the queueing system. When yield does not depend on cycle time, queueing theory predicts the highly nonlinear relationship (pictured in Fig. 1) between mean cycle time and throughput rate, where the throughput rate equals the yield times the *start rate* of wafers, which is the rate at which wafers are released into the fab per unit of time. Fig. 1 shows that mean cycle time is an increasing convex function of throughput rate, and increases without bound as the throughput rate approaches the system's capacity. The term *capacity* is defined to be the throughput rate that would be achieved if all workstations (or at least the bottleneck workstations) were continuously busy.

However, when yield depends on cycle time, as we assume in our model, the nature of the mean cycle time versus throughput rate curve is fundamentally different in character than the classical curve pictured in Fig. 1; for two examples of our curve, the reader is referred to Fig. 2 in Section III. In particular, *mean cycle time is not an increasing function of the throughput rate*; increasing the start rate of wafers beyond a certain point leads to large cycle times, which in turn makes it difficult to produce nondefective die. We also derive the capacity of the queueing system, which is the maximum achievable throughput rate, by solving a simple optimization problem. The start rate of wafers that achieves the maximum throughput rate is also found, as are the resulting yield and mean cycle time.

Our analysis implies that if longer manufacturing cycle times cause lower yields of nondefective die (as is widely believed), then loading a fab beyond a certain critical level will lead to a reduction in the output rate of nondefective die. Thus, the detrimental effect that long manufacturing cycle times have on yield should be more thoroughly assessed and if this effect is substantial, then it should be taken into account by fab managers when they determine their rate of wafer starts. Finally, it is important to point out that the qualitative nature of the curves in Fig. 2 *holds for any manufacturing facility where an increase in cycle time or WIP inventory causes a reduction in the resulting yield*. The underlying cause for this reduction may not be due to particulate contamination, but to the delay in process feedback, the lack of worker accountability, the slow rate of learning and other factors that are cited by proponents of just-in-time (JIT) manufacturing (see Schonberger [7],

Manuscript received January 28, 1991; revised June 12, 1991. This work was partially supported by a grant from the Leaders for Manufacturing Program at MIT, an IBM/University Manufacturing Systems Research Grant, a grant from Texas Instruments, and National Science Foundation grant Award No. DDM-9057297.

The author is with the Sloan School of Management, Massachusetts Institute of Technology, Room E53-343, Cambridge, MA 02142.

IEEE Log Number 9106756.

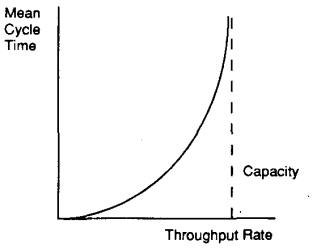


Fig. 1. Tradeoff when yield does not depend on cycle time.

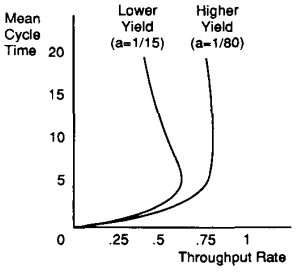


Fig. 2. Tradeoff when yield depends on cycle time.

Schmenner [6] or, for a case study in wafer fabrication, Martin-Vega *et al.* [4]).

## II. The Model

The wafer fab will be modeled as a single-server queueing system, where wafers arrive according to a Poisson process with rate $\lambda$, which will be referred to as the start rate. Service times are exponentially distributed with rate $\mu$, where $\lambda < \mu$ in order to assure system stability. Let the *cycle time* $W$ be a random variable denoting the amount of time a customer spends in the queueing system under steady-state conditions. It is well known (see, for example, Kleinrock [3]) that $W$ is exponentially distributed with mean $(\mu - \lambda)^{-1}$.

The classic Poisson yield model assumes that the number of defects on each die of a wafer is an independent Poisson random variable with mean $d$. The yield is defined as the probability a die contains no defects, which is $e^{-d}$, independently of all other die on the wafer. We alter this model by assuming that the mean number of defects per die on a wafer is a linear function of the length of time the wafer spends in the queueing system. In particular, we let $d = aW$, where $W$ is the steady-state cycle time defined above and the *defect rate* $a$ represents the rate at which defects occur on the die per unit of time. The primitive data of the model are the start rate $\lambda$, the service rate $\mu$ and the defect rate $a$, although the start rate can be thought of as being at the discretion of the fab manager.

Suppose each wafer contains exactly $D$ die and let $Y_W$ be the number of nondefective die on a wafer whose cycle time is $W$. Then $Y_W$ is a binomial random variable with parameters $D$ and $e^{-aW}$. If we define the random variable $Y$ to be the number of nondefective die on a wafer under steady-state conditions, then the expected value of $Y$ is

$$E[Y] = \int_0^\infty (\mu - \lambda) e^{-(\mu - \lambda)W} D e^{-aW} \, dW \tag{1}$$

$$= \frac{D(\mu - \lambda)}{\mu - \lambda + a}. \tag{2}$$

TABLE I
A Numerical Example at Two Different Defect Rates

| Defect Rate | Start Rate | Throughput Rate | Mean Cycle Time | Yield |
|---|---|---|---|---|
| 1/15 | 0.40 | 0.360 | 1.67 | 0.900 |
| 1/15 | 0.60 | 0.514 | 2.50 | 0.857 |
| 1/15 | 0.80 | 0.600 | 5.00 | 0.750 |
| 1/15 | 0.90 | 0.540 | 10.0 | 0.600 |
| 1/15 | 0.95 | 0.407 | 20.0 | 0.429 |
| 1/80 | 0.40 | 0.392 | 1.67 | 0.980 |
| 1/80 | 0.60 | 0.582 | 2.50 | 0.970 |
| 1/80 | 0.80 | 0.753 | 5.00 | 0.941 |
| 1/80 | 0.90 | 0.800 | 10.0 | 0.889 |
| 1/80 | 0.95 | 0.760 | 20.0 | 0.800 |

Let $\bar{Y}$ denote the yield, which is defined to be the mean proportion of nondefective die on a wafer. Then $\bar{Y} = E[Y]/D = (\mu - \lambda)/(\mu - \lambda + a)$. As mentioned earlier, we will define the throughput rate $T$ as the mean number of nondefective die produced per unit of time *divided by* the number of die per wafer. Thus,

$$T = \lambda \bar{Y} = \frac{(\mu - \bar{W}^{-1})\bar{W}^{-1}}{\bar{W}^{-1} + a}. \tag{3}$$

Solving for $\bar{W}$ in terms of $T$ in (3) gives

$$\bar{W} = \frac{\mu - T \pm \sqrt{(\mu - T)^2 - 4aT}}{2aT}. \tag{4}$$

Equation (3) or (4) can be used to plot the curve describing the important relationship between the mean cycle time $\bar{W}$ and the throughput rate $T$. The fab's *capacity*, which is its maximum achievable throughput rate, is derived by finding the value of $\bar{W}$ that maximizes the right side of (3). This calculation yields

$$\bar{W}^* = \frac{1}{\mu} + \sqrt{\frac{a + \mu}{a\mu^2}}, \tag{5}$$

and the resulting capacity is

$$T^* = \frac{\mu^2 \sqrt{a(a + \mu)}}{2a^2 + 2a\mu + (2a + \mu)\sqrt{a(a + \mu)}}. \tag{6}$$

Furthermore, the unique start rate of wafers that achieves this capacity is

$$\lambda^* = \frac{\mu + T^*}{2}, \tag{7}$$

which can be expressed solely in terms of the primitive parameters $a$ and $\mu$ by substituting (6) into (7).

## III. A Numerical Example

In order to gain a better understanding of the implications of this analysis, we will display some numerical calculations. We let the service rate $\mu$ equal one wafer per unit of time (time is in arbitrary units), so that the start rate $\lambda$ also represents the fraction of time the server is being utilized over the long run. For the case where the defect rate equals 1/15 defects per die per unit of time, Table I contains five values of the start rate and the corresponding values for the three dependent variables. Of course, as the start rate increases, the mean cycle time increases and the yield decreases. The system capacity, which is the maximum value of the throughput rate, is 0.6, which is achieved when the system is 80% utilized.

At utilization levels higher than 80%, the cycle time becomes too large, which results in lower yield and hence lower throughput.

Table I also contains similar calculations when the defect rate is lowered to 1/80. Since defects are appearing on the wafer at a slower rate, this case results in higher yield. The capacity of this system is 0.8, which is achieved by utilizing the system 90% of the time. The curve of mean cycle time versus throughput rate for both of these examples is displayed in Fig. 2. In both curves, the throughput rate decreases at higher congestion levels, although the effect is more pronounced in the higher defect rate (i.e., lower yield) case.

### ACKNOWLEDGMENT

This paper was partially stimulated by a conversation with Ivan Darius of Texas Instruments.

### REFERENCES

[1] H. Chen, J. M. Harrison, A. Mandelbaum, A. van Ackere, and L. M. Wein, "Empirical evaluation of a queueing network model for semiconductor wafer fabrication," *Operations Research*, vol. 36, pp. 202–215, 1988.

[2] J. A. Cunningham, "The use and evaluation of yield models in integrated circuit manufacturing," *IEEE Trans. Semicond. Manufact.*, vol. 3, Aug., pp. 60–71, 1990.

[3] L. Kleinrock, *Queueing Systems, Volume I: Theory*. New York: Wiley, 1975.

[4] L. A. Martin-Vega, M. Pippin, E. Gerdon, and R. Burcham, "Applying just-in-time in a wafer fab: A case study," *IEEE Trans. on Semicond. Manufact.*, vol. 2, pp. 16–22, 1989.

[5] C. M. Osburn, H. Berger, R. P. Donovan, and G. W. Jones, "The effects of contamination on semiconductor manufacturing yield," *The Journal of Environmental Sciences*, pp. 45–57, Mar./Apr. 1988.

[6] R. W. Schmenner, "The merit of making things fast," *Sloan Management Review*, vol. 30, pp. 11–17, 1988.

[7] R. J. Schonberger, *Japanese Manufacturing Techniques: Nine Hidden Lessons in Simplicity*. New York: The Free Press, 1982.

## Real-Time Control of Multiproduct Bulk-Service Semiconductor Manufacturing Processes

John W. Fowler, Don T. Phillips, and Gary L. Hogg

*Abstract*—**This paper demonstrates how knowledge of future arrivals can be used to improve control of multiproduct bulk-service semiconductor manufacturing processes. The objective of the research reported herein is to reduce the average time that lots spend waiting to be processed. A review of the current literature reveals that several researchers have dealt with the control of bulk-service queueing systems; however, only one paper has addressed the use of knowledge of future arrivals and it only considered the single product case. This research reexamines the single-product-single-tube case and then explores the multiple-products single tube case. For both cases, a control strategy is devised and evaluated through the use of systems simulation. The steady-state performance of each control strategy is then compared to the steady-state performance of the theoretically optimal control strategy not considering the timing of any future arrivals (i.e., a Minimum Batch Size strategy). The experimental results indicate that the control strategies developed in this paper perform well under a wide variety of conditions.**

Manuscript received March 18, 1991; revised November 19, 1991. This work was partially supported by Semiconductor Research Corporation (SRC), SEMATECH, Advanced Micro Devices (AMD), and the Department of Industrial Engineering at Texas A&M University.

J. W. Fowler is with SEMATECH, 2706 Montopolis Drive, Austin, TX 78741.

D. T. Phillips and G. L. Hogg are with the Department of Industrial Engineering, Texas A&M University, College Station, TX 77843.

IEEE Log Number 9106757.

### I. INTRODUCTION

Manufacturing integrated circuits is an extremely complex and challenging endeavor. The total process involves a mixture of single-wafer, single-lot, and multi-lot processes employing various technologies. This is further complicated by the re-entrant flow of lots through the same processes as the layers of the wafer are fabricated. This paper focuses on a subset of the total problem, the *bulk-service* or *batch* (multi-lot) elements. These operations are performed in furnace *tubes*. The goal is to provide methods which will serve as subprocedures within a global wafer fabrication control methodology.

When an appropriate number of lots are available and a tube capable of processing that step is available, a batch is formed and a processing cycle is initiated. If there is a batch which equals the tube capacity (i.e., a full load) at a given process step and a capable furnace tube is available, there is no benefit in delaying the start of the batch. However, if a tube is available and less than a full load waiting, a nontrivial decision must be made to start the partial batch or to wait for additional lots. This decision is made when lots arrive to the batch processing area or when a furnace tube becomes available. The rule set used in this decision process is referred to as a *control strategy*.

A *lot* of wafers undergoes several batch processing steps in the fabrication process. It is not unusual for a single furnace tube to be capable of processing more than one step (herein called a *product*), thus creating the need to consider the impact of multiple products on the control problem.

Most semiconductor manufacturers have shop-floor control systems which provide a means to predict, with reasonable accuracy, the timing of future arrivals to batch processing steps. This information should be useful in making a decision of whether or not to start processing a partial load. Currently, this information is often used in a qualitative manner, but its potential is far from being fully exploited. This research seeks to add mathematical rigor and formalism to the use of this information and also seeks to determine the effect of errors in the prediction of the timing of future arrivals.

The semiconductor industry is very much driven by cycle time. Therefore, the objective of this research is to use arrival information to develop and test a control strategy for multiproduct batch processes that will reduce the average cycle time per lot. The performance measure actually used to compare alternative strategies is the average time lots spend in queue waiting to be processed. Since the processing times are usually constant, minimizing the average time lots spend in queue is essentially equivalent to minimizing the average cycle time. This research is more fully documented in Fowler [1].

### II. REVIEW OF RELEVANT LITERATURE

#### A. Research Using Only the Current State of the System

Chaudhry and Templeton [2] devote an entire book to bulk arrival and bulk service queues and Crabill *et al.* [3] provide a bibliography of research on the optimal design and control of queues.