EMPIRICAL EVALUATION OF A QUEUEING NETWORK MODEL FOR SEMICONDUCTOR WAFER FABRICATION

HONG CHEN, J. MICHAEL HARRISON, AVI MANDELBAUM, ANN VAN ACKERE, and LAWRENCE M. WEIN

Stanford University, Stanford, California

(Received December 1986; revision received May 1987; accepted October 1987)

This paper concerns performance modeling of semiconductor manufacturing operations. More specifically, it focuses on queueing network models for an analysis of wafer fabrication facilities. The congestion problems that plague wafer fabrication facilities are described in general terms, and several years' operating data from one particular facility are summarized. A simple queueing network model of that facility is constructed, and the model is used to predict certain key system performance measures. The values predicted by the model are found to be within about 10% of those actually observed. These results suggest that queueing network models can provide useful quantitative guidance to designers of wafer fabrication facilities, and we discuss refinements and extensions of our elementary model that are likely to be important in other settings. However, an even more important benefit to be gained from queueing theory is the simple qualitative point that congestion and delay in wafer fabrication are caused by variability in the operating environment. To significantly reduce manufacturing cycle times, one must reduce that variability.

This paper is concerned with rough-cut performance modeling of semiconductor manufacturing operations. Attention is restricted to the wafer fabrication stage of integrated circuit (IC) production, and the performance measures of primary interest are the average throughput rate and manufacturing cycle time. We describe a modeling study and a data analysis aimed at empirical validation of the model constructed, hoping to build confidence in analytical queueing models, as opposed to simulation studies, and thus pave the way for their use in applied work. This introductory section contains a brief and highly selective description of IC manufacturing, an account of the congestion problems that plague wafer fabrication, an overview of performance modeling, and an outline of the remainder of the paper.

An integrated circuit, commonly referred to as an IC or a semiconductor chip, is a complex device that consists of miniaturized electronic components and their interconnections. The production of IC's is accomplished in a four-stage process that begins with raw wafers of silicon or, less commonly, gallium arsenide. Wafers are grouped in *lots*, the members of which travel together in a standard container and are destined for conversion to the same final product. The lot size, usually between 20 and 100 wafers, differs from one production facility to another and may differ from one product to another within the same facility.

The first stage of IC production is called wafer processing or wafer fabrication. It is conducted in a so-called clean room, where special means are employed to maintain a low density of airborne particles. The term *wafer fab* is commonly used to mean a clean room in which wafer fabrication is conducted. Here the intricate miniature circuitry for a number of identical chips is created on each wafer. The individual chips-to-be are referred to as *dice*. The circuitry is created by a lengthy and complex process (described later), and the number of dice per wafer may vary from just a few to many hundreds. Wafer fabrication requires a long sequence of processing steps and involves many separate pieces of equipment, through which lots of wafers are routed in the traditional job shop fashion.

In the second stage of IC production, commonly referred to as wafer probe, the following occurs: (a) the individual dice on a wafer are tested for functionality by delicate electrical probes, (b) dice that fail to meet specifications are marked with an ink dot, (c) the wafers are scored and broken into separate individual dice, and (d) the defective dice are discarded. In the third stage of production, called assembly, electrical leads are connected to the individual dice, which are then encapsulated in plastic or ceramic shells called *packages*. In the fourth stage of production, packaged chips are subjected to a final functional test and burn-in.

Perhaps the greatest single determinant of economic

Subject classification: 683 queueing applications, 697 queueing networks.

0030-364X/88/3602-0202 $01.25

success for an IC manufacturer is the total process yield, that is, the fraction of individual dice that survives all stages of production and testing to emerge as salable packaged chips. Total yield may be 80% or higher for relatively simple circuits produced with mature technologies, but figures below 10% are not uncommon for large, highly integrated products in the early stages of production.

The wafer fabrication stage dominates the economics of IC production, and it is here that semiconductor manufacturers concentrate their research and development efforts: wafer fabrication is probably the most complex manufacturing process in the world today, and it requires an enormous investment in plant and equipment. Because capital costs are high and variable processing costs are relatively low, high utilization of wafer fabrication equipment is a generally accepted goal in the semiconductor industry. Most wafer fabs are operated on a three-shift basis for either 5 or 7 days per week, but the amount of time spent to actually process wafers is limited by several factors, such as preventive maintenance, setup, absence of qualified operators, end of shift effects (see Section 2), and frequent episodes of unscheduled downtime. Some of this unscheduled downtime is due to the literal failure of equipment, but "process tuning" is often a more important downtime category. If a manufacturer reduces any of these sources of equipment unavailability, or the time required for actually processing wafers on any given piece of equipment, a higher throughput rate, and hence, a lower unit cost can be achieved, provided that process yields are not adversely affected.

For purposes of this paper, a piece of equipment is *idle* if it is available for processing but is starved for work. Equivalently, the equipment is idle if it is neither processing wafers nor rendered unavailable for one of the reasons enumerated above. The *idleness rate* for a piece of equipment is defined as the overall fraction of working hours that it spends in the idle condition. The conventional wisdom among semiconductor manufacturers is that the idleness rate for critical fabrication equipment should be no larger than 10%. Given this, it will come as no surprise to readers familiar with queueing theory that wafers spend most of their time waiting rather than being processed. To set terminology, let us define the *cycle time* for a lot of wafers as the total number of working hours that elapse between its entry into the clean room and its exit. This same quantity will occasionally be called the *manufacturing cycle time* or *manufacturing interval* for wafer fabrication.

To get a feel for the magnitude of queueing effects in IC manufacturing, let us consider a wafer fab dedicated to production of a single, reasonably complicated product (say, a VLSI microprocessor). Production of these circuits involves a total of perhaps 200 distinct fabrication steps, many minor in character. If we add the times required for all of these operations, taking reasonable account of such overhead factors as loading, unloading and operator orientation time, the total might come to 120 hours, which amounts to 1 working week with three-shift operations 5 days per week. Under such circumstances, the average throughput time for wafer fabrication would typically be 5–10 weeks. In the semiconductor industry it is common to describe this state of affairs by saying that the actual-to-theoretical ratio is between 5 and 10, or that the manufacturing interval is 5–10 times the theoretical.

The total manufacturing interval for VLSI circuits, considering all stages of production, may be as long as 4 months, and this is widely recognized as a major problem for semiconductor manufacturers. In the case of customized products, the nature of the problem is obvious, since the order lead-time imposed on customers must be at least as large as the total manufacturing interval. On the other hand, standardized products can be made to stock, but here again, long manufacturing intervals cause trouble because production must be based on forecasts of market demand many months in the future, and major demand shifts are commonplace. Moreover, product life cycles are short in the semiconductor industry, so the risk of obsolescence for finished goods inventory is always present. Finally, there is an established negative correlation between manufacturing interval and yield in wafer fabrication, which provides another strong motivation for reduction of throughput times.

Thus, it is essential that the designer of a wafer fabrication line has a means to predict key performance measures, including average throughput time, given only processing system characteristics that are known or can reasonably be estimated before the system goes into operation. The most obvious means to generate such performance predictors is the Monte Carlo simulation, but experience in other areas suggests that mathematically tractable queueing network models, although less flexible than simulation models and based on apparently restrictive assumptions, are far easier to use, generate more qualitative insight with respect to essential system relationships, and are accurate enough to provide quantitative guidance to system designers. Thus, we shall focus our attention

on such models. The potential role of queueing network models for the analysis of manufacturing systems has been recognized before, most notably in the influential paper of Solberg (1977), but such models are not incorporated, thus far, in accepted engineering practice.

Solberg (1983) also noted the general lack of sophistication in manufacturing system design, and his remarks apply even more forcefully to the semiconductor industry, where manufacturing systems engineering is still in its infancy. Nonetheless, there are recent signs of interest in performance modeling of wafer fabrication. Dayhoff and Atherton (1984, 1986a, b, 1987) describe the potential relevance of simulation methodology for analysis of wafer fab operations, and successful wafer fab simulation studies are reported by Spence and Welter (1987) and Burman et al. (1986). The last paper surveys and summarizes the experience of an operations research team involved in the analysis of AT&T wafer fab operations, and discusses queueing network models of the type under consideration here.

The term "system design" has been used frequently in this introduction, and for many readers those words may conjure up a picture of engineers laying out a new physical facility. However, most wafer fab design activity is actually aimed at reconfiguration of existing facilities. Such reconfigurations may occur several times yearly, so rough-cut design tools of the type discussed here have enormous potential value.

The description of wafer fabrication given earlier was implicitly oriented toward production facilities. Research and development laboratories (hereafter referred to as R&D labs or facilities) constitute a second, closely related class of wafer fabs. Wafer processing in such facilities is aimed at the development of new products or processes, as opposed to the production of salable chips, but the equipment, operating procedures, and process flow are essentially the same as in a production facility. In this paper, a particular wafer fab will be analyzed in some detail. It happens to be an R&D lab, but we believe that the differences between R&D facilities and production facilities are relatively unimportant for performance modeling purposes. That is, the same model structure applies in both cases, although the parameter values that characterize the fab may be quite different. This issue will be discussed further in Section 2.

Section 1 describes both the general character of IC wafer fabrication and the particular wafer fab facility alluded to above. In Section 2 we lay out an apparently naive queueing network model of the fab, and the performance predictions of that model are compared against actual observed performance. For the particular fab facility that we studied, a simple queueing model predicts aggregate performance characteristics with surprising accuracy, but there are several refinements or generalizations of the model that are likely to be important in other settings. Those extensions are discussed in Section 3, along with other potential directions for future research.

1. Wafer Processing and the TRC Silicon Fab

This section contains some general information about the operations involved in wafer fabrication, plus specific information about the one wafer fab facility we studied in detail. Readers are referred to Sze (1983) or Gise and Blanchard (1986) for a detailed description of semiconductor wafer fabrication. As stated earlier, wafer fabrication is done in a clean room, which is typically divided into U-shaped bays. A bay generally contains a major piece of equipment on which a basic operation is performed, plus ancillary equipment or facilities involved in closely related operations. Operators process lots on the inside of the U, and most equipment is positioned so that maintenance can be performed on the outside of the U, which is outside the clean room.

Each lot entering the clean room has an associated process flow, often called a recipe, that consists of precisely specified *operations* executed in a prescribed sequence on prescribed pieces of equipment. If all goes well, this exact sequence of operations is performed, but sometimes inspections reveal that an operation was not executed to specification, in which case, some or all of the wafers in the lot are either scrapped or reworked. This latter phenomenon introduces a stochastic element to the routing of lots. Integrated circuit fabrication involves the creation of multiple layers on a silicon wafer, and the operations involved in the creation of each successive layer are essentially the same, so lots can, and typically do, return repeatedly to some pieces of equipment.

1.1 Equipment Categories and Major Processing Steps

The wafer fab we studied in detail is the Hewlett-Packard Technology Research Center Silicon Fab (hereafter referred to as the TRC fab), which is a relatively large R&D facility in Palo Alto, California. The TRC fab contains 52 processing centers, which are listed and described in Table I. (The official TRC equipment list is substantially longer than this, but the equipment listed in Table I accounts for all but a small fraction of total queueing and processing time.)

Table I
TRC Processing Equipment

Name of Equipment	Operation	Description	Name of Equipment	Operation	Description
MCLN	Deposition	Clear bench at sputter	LPCLN	Deposition	Clean bench for LPCVD tubes
SLSP	Deposition	Sloan sputter	TU72	Deposition	Low pressure CVD tube
SPUT	Deposition	Perkin-Elmer 4400 sputter	TU73	Deposition	Low pressure SINI CVD tube
PLM5L	Deposition	Plasma enhanced CVD lower tube	TU74	Deposition	Low pressure SI02 CVD tube
			TU84	Deposition	ASM-LTO-UPCVD system
PLM5U	Deposition	Plasma enhanced CVD upper tube	PHSOG[a]	Deposition	Trilayer bottom and spin on glass
CLEAN	Deposition	Clean wet bench for OXI/DIFF tubes	PHPNS	Lithography	Pre-bake/negative spin resist
			PHPPS[b]	Lithography	Pre-bake/positive spin resist
TU11	Deposition	Metal alloy tube	PHGCA[c]	Lithography	Two GCA align/developers
TU13	Deposition	Field oxidation tube	PHPED[d]	Lithography	Two Perkin-Elmer align/developers
TU21	Deposition	Field oxidation tube			
TU24	Deposition	P predeposition bipolar/ MOS tube	PHDI	Lithography	Develop inspect
			PHHB	Lithography	Hardbake station
			PHBI	Lithography	Bake inspect
TU31	Deposition	Special tube	PHFI	Lithography	Final inspect
TU32	Deposition	Source drain doping tube	PLM3	Etching	Plasma reactor-therm "RIE"
TU33	Deposition	Source drain doping tube	PLM4	Etching	AME 8100
TU34	Deposition	Field oxidation tube	PLM6	Etching	Plasma etcher for aluminum
TU41	Deposition	N-well drive-in tube	PLM8	Etching	Oxide/nitride dry TEK etch
TU42	Deposition	N-well drive-in tube	PLM21	Etching	Plasma II reactor
TU43	Deposition	Annealing for silicides	PLM22	Etching	Plasma II etch
TU44	Deposition	P-well drive-in tube	PHWET	Etching	Wet etch station
TU51	Deposition	MOS gate oxidation tube	PHPLO	Etching/ resist strip	Etchers and strip/clean for plasma etch
TU52	Deposition	MOS gate oxidation tube			
TU53	Deposition	MOS gate oxidation tube	PHCLN	Photoresist strip	Strip/clean for wet etch
TU54	Deposition	MOS gate oxidation tube			
TU61	Deposition	Field oxidation tube	IMP1	Ion implantation	Old ion implanter
TU62	Deposition	Field oxidation tube			
TU63	Deposition	Field oxidation tube	IMP2	Ion implantation	New ion implanter
TU64	Deposition	Field oxidation tube			
TU71	Deposition	Alloy tube			

[a] "PHSOG" includes three equipments: "SOG1," "SOG2" and "SOG3."
[b] "PHPPS" includes two equipments: "POS SPIN1" and "POS SPIN2."
[c] "PHGCA" includes two equipments: "GCA1" and "GCA2."
[d] "PHPED" includes two equipments: "PE1" and "PE2."

As noted in the table, four of those centers contain several identical pieces of equipment, and for reasons that will be explained later (see Subsection 1.3), each piece of equipment will be treated as a separate service station in our queueing network model of the fab. Individual pieces of equipment will be referred to as *stations* or *machines* in the remainder of this paper. Each piece of equipment is associated with one of the five generic operations described in the following paragraphs. In creating any single layer of an integrated circuit, these five operations, or perhaps some subset of them, are executed in the order listed below. Figure 1 gives a pictorial representation of the operations required to create a single layer, and readers may find it helpful to reference this figure when reading the descriptions.

Deposition. A thin film of material is deposited to form a layer of the integrated circuit. The wafers must be cleaned immediately before this operation (that is, within a specified time before the operation is performed) to avoid particle contamination. The deposition technologies used by TRC are: (1) oxidation, (2) chemical vapor deposition (CVD), (3) spin on glass (SOG), and (4) physical vapor deposition (sputtering). The TRC fab has one sputtering bay, two CVD bays and six oxidation bays.

Lithography. Also called photolithography, masking or patterning. The wafer is coated with a light-sensitive material called photoresist, which is then exposed to ultraviolet light through a mask that contains a pattern reflecting the intended geometry of the circuit. The

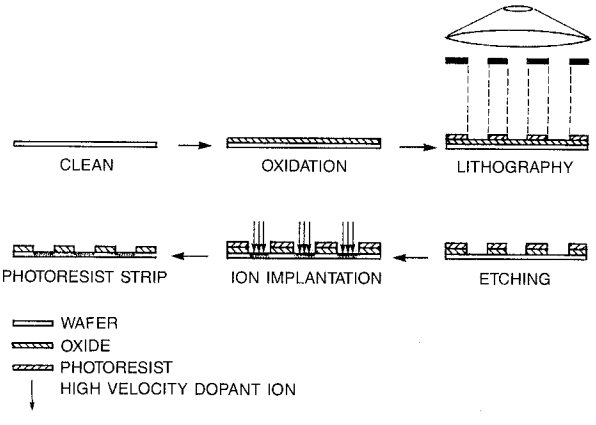

Figure 1. Illustrative process sequence for fabrication of a single oxide layer.

exposure step, generally referred to as photoexpose, is the most complex and delicate operation in wafer fabrication, and it involves the most expensive equipment found in the fab. At this stage of wafer fabrication one encounters the most common and important type of rework: if inspection shows that the pattern exposed in the photoresist does not meet specifications (because of mask misalignment, for example), then the entire photoresist layer must be stripped away and the wafer prepared for a repetition of lithography operations. If the pattern is acceptable, then exposed photoresist is removed by a developer, some minor additional operations are performed, and the wafer goes forward with a protective covering of hardened, unexposed photoresist covering its surface selectively in a pattern dictated by the mask. TRC uses two types of machines for photoexpose operations, GCA steppers and Perkin-Elmer aligners, of which there are two each. The TRC fab has four lithography bays, each contains one photoexpose machine.

Etching. Circuits are defined by etching away the portion of the deposited layer that is not protected by photoresist. There are two etching technologies used in the TRC fab; wet etching and plasma etching. The latter (more precise) technology is more heavily used, and the TRC has six plasma etchers.

Ion Implantation. In order to change the electrical properties of the surface not protected by photoresist, the wafer is exposed to accelerated ions which are implanted to a predetermined depth and concentration. Boron, phosphorus, and arsenic ions are most frequently implanted because even a small number will dramatically alter the electrical properties of the underlying silicon. The TRC fab has two ion implanters.

Photoresist Strip. Finally, the pattern of photoresist that remains on the wafer after lithography is removed (stripped) using a process similar to etching.

In addition to these five types of operations, many cleaning, measurement and inspection operations are performed throughout the fab. The cleaning operations prevent contamination of the wafer, and the inspection and measurement operations are performed to identify defective wafers, which are then scrapped or reworked.

Different kinds of integrated circuits require different processing steps during fabrication, and thus, have different process flows through the fab. Production facilities (as opposed to R&D labs) usually make more than one product, and processing technology changes quite often, so there is substantial diversity in product routing when the operations of any given fab are examined over a period of several years.

As stated earlier, an individual lot of wafers may visit one or more pieces of equipment repeatedly during the course of wafer fabrication. In particular, it is usual to use the same photoexpose machine (stepper or aligner) in the creation of each layer. Ion implanters and inspection stations may also be visited repeatedly. Deposition equipment and plasma etchers are often dedicated to a single operation in a single process; it is, therefore, common for them to be visited only once during the entire fabrication sequence.

1.2. Production Versus Research and Development

As stated previously, there are two types of wafer fabs; production facilities and R&D labs use the same types of equipment to execute essentially the same operations. However, there are important differences between these facility types, the most obvious derive from the experimental nature of R&D lab operations. The new equipment commonly found in R&D labs may fail often, and it is frequently shutdown for redesign or recalibration. Also, the fabrication processes used in R&D labs are often poorly understood, which leads to frequent changes of recipe and recalibration of equipment. In contrast, the processes used in production facilities are relatively stable, and equipment is often dedicated to a single task. There are additional differences that may not be so obvious; because our study focuses on modeling and analysis of a research and development laboratory, it is important that those additional differences be clearly understood.

For ease of exposition, consider an R&D lab completely dedicated to the development of a new

12-layer process technology to be used in the production of a particular product or family of products (such as microprocessors or memory circuits). During the development phase, engineers experiment with various operation times, machine settings, and so forth, to assure that the new process produces wafers with all the desired properties. In the course of this experimentation, engineers may submit lots of wafers on which all 12 layers are to be created, on which only a few layers are to be created, or on which only a few operations are to be performed. Thus, even though the fab is, in a sense, dedicated to a single process, it must process lots that require anything from 2 to 200 operations. A typical R&D lab actually works on the development of several new processes simultaneously, so the diversity of lots entering the fab for processing may be staggering. In building a queueing network model of such a facility, it is desirable to aggregate lots into a workable number of customer types because explicitly representing all known distinctions among incoming lots leads to a model of prohibitive complexity. This process of aggregation ultimately causes models of production fabs and development fabs to be quite similar. In our study of the TRC fab, we adopted the aggregated lot categories used by TRC management for internal communication, as listed in Table II. Readers will see that the critical characteristics of a lot for purposes of this categorization scheme are: (a) the total number of operations required, and (b) the type of machine used for photoexpose operations (GCA stepper or Perkin-Elmer aligner). The special attention given to lithography equipment in this categorization scheme derives from the central and sensitive role of these operations in a fabrication process. The letters S and L that appear in the lot category abbreviations stand for *short* and *long*, respectively.

Table II
TRC Lot Categories (PHGCA and PHPED Are Described in Table I)

Name	Brief Description
MLOT	Monitor lots are nonproduct lots used to monitor the operating performance of the equipments
LGCA	All non-MLOTs using PHGCA and having at least 35 operations
LPED	All non-MLOTs using PHPED and having at least 35 operations
SGCA	All non-MLOTs using PHGCA and having less than 35 operations
SPED	All non-MLOTs using PHPED and having less than 35 operations
SOTHER	All non-MLOTs using neither PHGCA nor PHPED and having less than 35 operations

Monitor lots (the first category in Table II) are used to monitor the operating performance of equipment in the TRC fab. That is, wafers from these lots are processed periodically on certain pieces of equipment to ensure that the equipment is properly calibrated and performing to specification. They are nonproduct lots, but they will be treated as an ordinary customer type in our queueing network model because their processing consumes machine capacity like any other lot category.

There are several other categories of lots processed in the TRC fab that do not appear in Table II, but they account for just a small fraction of the facility's business and have very little impact on equipment utilization or the throughput times experienced by other lot categories. We emphasize that many other schemes could be used to categorize lots in the TRC fab, but the fab management is familiar and comfortable with the one described in Table II. In modeling a production facility, it would be most natural to define one customer type for each of the finished products made in the fab.

Previous paragraphs emphasized input diversity as a feature that distinguishes R&D fabs from production fabs. Another distinguishing feature of an R&D facility is the important role played by *engineering hold time*. For example, in the TRC fab, most lots are associated with one engineer or manager who may specify points in advance during the process at which he or she would like to inspect the wafers before further operations are performed. At such a point, an operator will put the lot "on hold", and it remains in that status until the responsible individual appears, does whatever is necessary, and releases the lot for further processing. Also, an operator occasionally puts a lot on hold when some anomalous result is obtained in testing, and the lot does not move further until the responsible engineer or manager makes a dispensation. Hold times may vary from minutes to months, and on average, engineering hold time accounts for about 35% of the total time that lots spend in the TRC fab.

1.3. Flow Control in the TRC Fab

The TRC fab is a classic job shop operation, in which lots of diverse character are routed through a collection of general purpose work centers for execution of prescribed operations. The performance of a job shop is affected by the flow control mechanisms employed, that is the input control, routing and sequencing rules built into the shop's operating systems. In the case of TRC, there are no routing issues worthy of discussion because operations must be performed in a particular

order and engineers specify precisely which piece of equipment to use for each operation on virtually all lots. (In principle, two GCA steppers or two ion implanters may be interchangeable, but reproducibility is difficult to achieve with state-of-the-art processes, and this consideration leads to specification of a unique machine for each operation.) For this reason we ultimately represent each machine as a separate single-server station in our queueing model of the fab, rather than representing a group of apparently identical machines as a multiserver station.

Throughout the period covered by our database, TRC management used the following mechanism to restrict input of nonmonitor lots to the fab: each research team that generated work for the fab was allocated a maximum number of lots that could be in process at any given time; the sum of these allocations is 80 lots. During most of the period covered by our database, the number of nonmonitor lots in the fab remained well below the theoretical maximum of 80, as we will discuss in Subsection 2.2. Monitor lots were injected into the fab on an open-loop basis, and the average number of monitor lots in the system was in the range of 10 to 15.

To sequence lots through individual machines (choosing a particular lot from a queue to be processed next), TRC does not use a rigid first-in-first-out (FIFO) discipline, although FIFO is the default rule at every workstation. On the other hand, the scheduling system used at TRC is not aimed at optimization of any intrinsic efficiency measure, such as average throughput rate or average throughput time. Rather, as in most job shops, the priorities used to sequence lots through each workstation at TRC are based partly on the perceived urgency associated with the lots queued at that station. Some management personnel humorously refer to this scheduling system as the "most influential engineer rule."

2. A Naive Queueing Network Model

Next we develop a simple queueing network model of the TRC fab and then the performance predictions of the model are compared against actual observed performance. Our model is rooted in the now-classical theory of product-form queueing networks, as described by Kelly (1979) or Baskett, Chandy, Muntz, and Palacios (1975). The latter paper, referred to hereafter as BCMP, serves as our standard reference, and there are two points about the BCMP theory that require particular mention. First, we wish to consider a network model in which arriving customers are divided into a number of distinct "types" that differ with respect to routing. The *route* of a customer is defined as the sequence of service stations and delay nodes (see below) visited by that customer, listed in the order they are visited. It is possible that one customer will repeatedly visit a given node or station, and we want to allow stochastic routing. More specifically, we wish to consider a model in which routes for customers of any given type are drawn from a general distribution that characterizes that type; the routes of individual customers are statistically independent. This level of generality with regard to routing may be accommodated in the BCMP theory by defining "customer classes" appropriately, provided that certain assumptions are satisfied regarding the service requirements of customers at the various nodes of the network. (It is crucial to define an arbitrarily large number of customer classes in the BCMP theory.) Roughly speaking, one must define a different customer class for each combination of customer type and routing history. Because we are interested only in the results of the BCMP theory, as opposed to the mathematical derivation of those results, it is not necessary to spell out what the customer classes are in our particular model, or even how many classes there are. The only routing data that occur in the formulas of ultimate interest are the expected number of visits to each service station by customers of each type; this is a result of startling and unexpected simplicity, and it is very much dependent on the other assumptions of the BCMP theory (like exponential service time distributions).

A second point that deserves special mention concerns the representation of engineering hold time in a queueing network model. We envision engineering holds as visits to a delay node, also called an infinite-server node or ample-server node in the literature of queueing theory. The essential feature of a delay node is that the duration of a customer's stay is unaffected by the number of other customers who occupy the node simultaneously, and in the BCMP theory the probability that a customer will visit a delay node, and the distribution of delay time or hold time given such a visit, depend on the customer's type and stage of completion in a more or less arbitrary fashion. Again, this is accomplished by defining customer classes appropriately. In the end, it is only the expected total time spent at delay nodes by customers of different classes that figures in the formulas of primary interest, so one need not spell out the holding time distributions at delay nodes, or even how many delay nodes the network contains. For the sake of concreteness, we speak in terms of a single delay node at which all engineering holds are imagined to occur.

The following exposition begins with a specification of inputs to and outputs from our queueing network model of the TRC fab. The formulas used to compute outputs from inputs are specified, and then we discuss the assumptions and theory that justify these formulas.

2.1. The Basic Mixed Network Model

Lots of wafers play the role of customers in our model of the TRC fab, and the six lot categories listed in Table II play the role of customer types. The 57 pieces of equipment listed in Table I are identified as single-server nodes or stations in our network model. The service stations are indexed by $i = 1, \ldots, 57$, and the customer types or lot categories by $j = 1, \ldots, 6$, with $j = 1, \ldots, 5$ corresponding to the various categories of nonmonitor lots and $j = 6$ designating monitor lots. In light of the operating discipline described in Subsection 1.3, we adopt what BCMP call a mixed network model, closed with respect to nonmonitor lots and open with respect to monitor lots. That is, in our model of the TRC fab, the total number of nonmonitor lots remains constant, with new lots inserted only as old lots complete processing, whereas monitor lots are assumed to be inserted on an open-loop basis. The parameters of our network model, to be estimated from TRC operating data, are

N = the fixed population size for nonmonitor lots;
p_j = the fraction of all nonmonitor lots processed through the fab of type j ($j = 1, \ldots, 5$);
α_6 = the average input rate (or arrival rate, or throughput rate) for monitor lots expressed in lot starts per hour;
v_{ij} = the average number of visits to service station i by lots of category j ($i = 1, \ldots, 57$ and $j = 1, \ldots, 6$);
H_j = the average total hours of engineering hold time experienced by lots of category j ($j = 1, \ldots, 6$);

and

μ_i = the *effective* service rate at service station i, defined as the average number of operations completed per hour of *nonidle* time ($i = 1, \ldots, 57$).

On the other hand, the output quantities (predicted performance measures) to be determined from the model are

α = the overall average throughput rate for nonmonitor lots, expressed in lot starts per hour;

and

T_j = average cycle time for lots of category j ($j = 1, \ldots, 6$).

As readers will see shortly, the throughput rate α is determined by an iterative procedure. Given a trial value for α, we define

$$\alpha_j = p_j \alpha \quad (j = 1, \ldots, 5), \tag{1}$$

interpreting α_j as the average rate at which lots of category j are processed through the fab. Similarly, it is convenient to define

$$\lambda_{ij} = \alpha_j v_{ij} \quad (i = 1, \ldots, 57 \text{ and } j = 1, \ldots, 6), \tag{2}$$

$$\lambda_i = \sum_{j=1}^{J} \lambda_{ij} \quad (i = 1, \ldots, 57), \tag{3}$$

and

$$\rho_i = \lambda_i / \mu_i \quad (i = 1, \ldots, 57). \tag{4}$$

Obviously λ_{ij} represents the average number of type j visits to station i per hour; λ_i is the average number of services performed at station i per hour, and $1 - \rho_i$ represents the average fraction of time that server i spends in the idle condition. In the standard terminology of queueing theory, ρ_i would be called the utilization rate for server i, but in our context that term is somewhat misleading because "utilization" includes both processing time and downtime. If the network is stable, one must have $\rho_i < 1$ for each $i = 1, \ldots, 57$, and only values of α that respect this requirement will be considered. In addition to the primary output variables identified earlier, it is convenient to identify the intermediate quantities

L_{ij} = average number of type j customers present at station i;
L_i = average total number of customers present at station i;
W_i = average throughput time per customer visit to station i (averaged over all customer types);

and

δ_j = average number of type j customers at the delay node (on hold).

The essence of our network model lies in the following equations:

$$L_i = \frac{\rho_i}{1 - \rho_i}, \tag{5}$$

$$L_{ij} = \left(\frac{\lambda_{ij}}{\lambda_i}\right) L_i, \tag{6}$$

$$\delta_j = \alpha_j H_j, \tag{7}$$

and

$$\sum_{j=1}^{5}\left(\sum_{i=1}^{57} L_{ij} + \delta_j\right) = N \quad \text{(given)}. \tag{8}$$

The left side of (8) strictly increases as a function of α, as expected on intuitive grounds, and suggests the following iterative scheme for computing α.

Algorithm. Guess a trial value for the throughput rate α and use formulas 1 to 7 to determine the corresponding L_i, L_{ij} and δ_j values. Now calculate the quantity on the left side of (8). If the result is sufficiently close to N, terminate. Otherwise, adjust the value of α upward or downward as appropriate and repeat.

Finally, to determine the average cycle times for various lot categories, given the value of α, we use equations

$$W_i = \frac{L_i}{\lambda_i}, \tag{9}$$

and

$$T_j = \sum_{i=1}^{I} v_{ij} W_i + H_j. \tag{10}$$

Our model requires more than 400 numbers as input data, but the visit frequencies v_{ij} account for about 85% of that total, and these parameters constitute the most basic sort of physical process information. Visit frequencies must be estimated (it is really more appropriate to speak of counting than of estimating) in any sort of meaningful capacity analysis, and for many facilities, reliable historical values are available in a machine processable data base. Similarly, it is hard to imagine any meaningful characterization of processing system capabilities that is less burdensome than estimating the effective service rate μ_i for each service station or work center worthy of inclusion in the network model.

One of the most important features of our modeling approach is the use of a mixed network model, closed with respect to nonmonitor lots, in which N is viewed as a design parameter and α is viewed as a performance characteristic. We use this approach because it reflects the reality of decision making and managerial control at TRC (and many other fabs): managers determine the total work-in-process inventory level N by direct action, and then they are forced to let the chips fall where they may (no pun intended) with regard to the throughput rate α.

The justification or motivation of Formulas 1–10 involves a three-stage argument. First, consider an open network with no server breakdowns, multiple customer types and a general routing distribution for each type. Assume that each type j arrives according to a Poisson process with given average arrival rate α_j, that service times at each station i are exponentially distributed with average service rate μ_i, and that a first-in-first-out discipline is employed at each station. Then Formulas 2–7, 9 and 10 all hold exactly, as one can deduce from the results of BCMP. Second, Whitt (1984) shows by numerical examples and by mathematical arguments that open network performance relationships give good approximations to closed network behavior if either the number of stations or the number of customers in the closed network is large, and his arguments strongly suggest that the same will be true for the mixed network case of interest to us here. To use the open network formulas in this setting, we view the throughput rate α as an unknown, use (1) and the product of mix proportions p_j to define arrival rates α_j in terms of α, and then impose the closure requirement (8) as suggested by Whitt (1984). Equation 8 involves a critical piece of additional data, the fixed population size for the closed part of the network, and it allows us to deduce the value of α by means of the simple iterative procedure outlined above.

Third and finally, the performance relationships 1–10 must be modified or reinterpreted to account for service interruptions; the events that render a service station unavailable to customers for any reason. In the mixed network model under consideration, we allow each station i to have a general (not necessarily exponential) service time distribution with mean m_i. Further suppose that interruptions occur according to a Poisson process at average rate β_i per unit time while server i is working, and interruptions do not occur when the server is idle. Finally, assume that the duration of service interruptions at station i has a general (not necessarily exponential) distribution with mean r_i. Such a system can be reduced by a standard trick to an equivalent network without service interruptions, cf. Vinod and Altiok (1986). The equivalent network has a general service time distribution at each station that is different from the original service time distribution. Specifically, define the *effective service time* of a customer as the actual service time *plus* the total duration of all interruptions that occur during that service. (For concreteness, we assume that interruptions have no effect on customers except to delay completion of service.) Then, the effective service times of customers entering station i are independent and identically distributed random variables, and the

effective service rate at station i is

$$\mu_i = [m_i(1 + \beta_i r_i)]^{-1}. \tag{11}$$

The right side of (11) represents the long-run average number of services completed at station i per unit of nonidle time, so (11) agrees with our original definition of μ_i. If the effective service time distributions are all exponential, then the formulas that characterize a network without service interruptions apply equally to the network with interruptions, provided that μ_i is reinterpreted as the *effective* service rate at station i. More generally, if the *variability* reflected in the effective service time distribution at each station is consistent with an assumption of exponentiality (that is, the standard deviation is approximately equal to the mean), one still gets a good approximation.

As we explain next, a reliable histogram for service times or effective service times, cannot be obtained from the TRC database, although it reliably estimates the overall average effective service rate. (The distinction here is between determining an entire distribution and its mean.) Thus, it was impossible to verify or refute the exponentiality assumptions that play such a key role in the BCMP theory, but we plunged forward with a model based on that theory because of its appealing simplicity. Similarly, given the exponentiality assumptions, we could have used the formulas developed by BCMP and others for *exact* analysis of a mixed network, but we adopted the approximation scheme proposed by Whitt (1984) because it leads to predictive performance relationships that are easier to understand, communicate and implement.

2.2. Parameter Estimation and Modification of the Basic Model

The TRC fab has been supported by computerized information systems for many years, and data that describe fab operations from January 1, 1983 to August 15, 1985 were analyzed in our study. During this period, the fab ran three shifts per day, 5 days per week. We began our data analysis by transforming the time scale of each relevant database to remove weekend and holiday time from consideration, so the term "hours" should be interpreted hereafter to mean "working hours."

A major virtue of the simple mixed network model described above is that it requires only parameter values for which relatively reliable estimates were obtained. The visit frequencies v_{ij}, the product mix proportions $p_1, \ldots, p_5$ for nonmonitor lot categories, and the input rate α_6 for monitor lots were all estimated in the obvious way from counting data, using average values for lots that completed processing during the 30-month period covered in our database. The engineering hold time parameters H_j were estimated by similar historical averages.

Recall that our queueing network model of the TRC fab has 57 service stations. As mentioned in Subsection 1.1, the real fab contains a number of other minor work centers, which collectively account for about 10% of all processing time and queueing time. To include them in the model would have greatly complicated data collection and data analysis, so they are simply excluded by definition from the system under study. To reflect that these minor work centers were deleted from the model, N was estimated by the average number of nonmonitor lots present in the fab minus the average number occupying the deleted work centers. Over the 30-month period covered by our database, the average total population size for nonmonitor lots was 68. Of these, an average of two lots occupied work centers deleted in the queueing model, so a parameter value of $N = 66$ was adopted. In the end, the throughput time predictions of the model will be compared against observed average throughput times exclusive of time spent at the deleted work centers.

Our final estimation problem concerns the effective service rates μ_i, and the estimates that were ultimately used in our test of the model appear in the last column of Table III. As the format of that table suggests, these numbers were built up by estimating separately, for each service station i, the average number of operations completed per hours over the period covered by our database, and the overall fraction of the period that the station spent idle. In accordance with our earlier definition, the effective service rate was then estimated by

$$\mu_i = \frac{\mathscr{A}}{1 - \text{idleness rate at station } i}. \tag{12}$$

where $\mathscr{A}$ = average number of operations performed at station i per hour. The numerator on the right side of (12) was easy to determine because this requires only counting data, but estimation of the overall idleness rate required that lot tracking data be combined with equipment status data (recall that for us "busy time" includes both downtime and processing time), and there were also serious data entry errors in both databases. Those data entry errors make for uncertainty as to when the processing of one lot leaves off and the processing of the next begins, and for uncertainty as to which sort of failure or interruption renders a service station unavailable, but in the end, it

Table III
Computation of Effective Service Rates

Equipment Name	Number of Operations	Operations per 24 Hours	Idleness Rate	Effective Service Rate (Lots per Hour)	Equipment Name	Number of Operations	Operations per 24 Hours	Idleness Rate	Effective Service Rate (Lots per Hour)
MCLN	1539	2.35	94.99	1.95	TU73	1257	1.92	20.94	0.10
SLSP	371	0.57	58.47	0.06	TU74	755	1.15	27.82	0.07
SPUT	2330	3.55	61.94	0.58	TU84	648	0.99	40.06	0.07
PLM5L	587	0.88	25.32	0.07	SOG1	138	0.21	85.00	0.06
PLM5U	157	0.25	33.71	0.02	SOG2	138	0.21	85.00	0.06
CLEAN	8554	13.03	57.96	1.29	SOG3	138	0.21	85.00	0.06
TU11	166	0.25	90.85	0.12	PHPNS	5546	8.45	61.08	0.90
TU13	143	0.22	66.24	0.03	POS SPIN1	2260	3.44	40.00	0.24
TU21	230	0.35	67.20	0.04	POS SPIN2	2260	3.44	40.00	0.24
TU24	58	0.13	89.40	0.05	GCA1	1482	2.26	10.00	0.10
TU31	152	0.23	49.55	0.02	GCA2	1482	2.26	10.00	0.10
TU32	219	0.33	79.25	0.07	PE1	1788	2.72	40.00	0.19
TU33	384	0.59	73.79	0.09	PE2	1788	2.72	40.00	0.19
TU34	276	0.42	51.49	0.04	PHDI	6345	9.67	48.64	0.78
TU41	314	0.48	81.57	0.11	PHHB	5646	8.60	68.66	1.14
TU42	786	1.20	62.75	0.13	PHBI	420	1.00	88.48	0.36
TU43	209	0.32	56.71	0.03	PHFI	4667	7.11	62.64	0.79
TU44	275	0.58	61.65	0.06	PLM3	1032	1.57	52.29	0.14
TU51	1008	1.54	48.05	0.12	PLM4	356	0.54	55.73	0.05
TU52	1118	1.70	33.85	0.11	PLM6	115	0.36	35.88	0.07
TU53	750	1.43	41.99	0.10	PLM8	124	0.62	49.55	0.15
TU54	362	0.55	68.65	0.07	PLM21	915	1.39	16.16	0.21
TU61	224	0.34	64.98	0.04	PLM22	368	0.56	21.00	0.09
TU62	524	0.80	44.60	0.06	PHWET	15363	23.41	53.57	2.10
TU63	446	0.68	52.79	0.06	PHPLO	4951	7.54	67.72	0.97
TU64	339	0.52	61.98	0.06	PHCLN	6193	9.44	68.23	1.24
TU71	1865	2.84	66.16	0.35	IMP1	1422	3.19	42.65	0.23
LPCLN	612	0.93	92.71	0.53	IMP2	699	2.42	25.03	0.13
TU72	1015	1.55	33.54	0.10					

was felt that reliable estimates for the idleness rates were obtained. That is, the TRC data proved adequate to estimate the overall average effective service rate at each station, but any finer statistical characterization, such as second moment information or complete histograms, would have been impossible to obtain. For a full account of the difficulties encountered in parameter estimation, and the measures taken to circumvent those difficulties, readers are referred to the technical report by Chen et al. (1986).

Note that four processing stations in the TRC fab (PLM6, PLM8, PLM21 and PLM22) are operational only one shift per day, and three others (SPUT, PLM5L and PLM5U) are operational two shifts per day. When Formula 5 was applied to any of these one- or two-shift stations, which amounts to fitting a steady-state $M/M/1$ model to the station, the model grossly overestimated average WIP. On the other hand, it was found that deterministic station models, which assume that inventory builds in a deterministic linear fashion over the off-shift period at a rate specified by the input rate to the station, and then dissipates in a linear fashion once operation is begun, grossly underestimate average WIP at such stations. For the one- and two-shift stations, we replaced (5) by a simple procedure that mixes these two extreme alternatives. See Chen et al., p. 61, for details.

2.3. Comparison of Predicted and Observed Performance

To repeat, our simple queueing network model generates a predicted average throughput rate for the closed part of the network (the nonmonitor lots), and a predicted average cycle time for each lot category. The average cycle time predicted for each lot category is displayed in Table IV, along with the actual observed values. As explained in the previous subsection, the actual cycle time values reported in Table IV exclude time spent at minor work centers omitted in our network model.

Table IV
Evaluation of the Mixed Queueing Network Model

Lot Category	Total Throughput Time (Hours)			Total Exclusive of Engineering Hold Time (Hours)		
	Observed	Predicted	Error (%)	Observed	Predicted	Error (%)
MLOT	57.00	63.38	11	52.00	58.38	12
LGCA	1385.88	1309.35	−6	829.56	753.03	−9
LPED	779.76	843.29	8	441.42	504.95	14
SGCA	372.31	355.17	−5	202.61	185.47	−8
SPED	237.14	234.65	−1	123.56	121.07	−2
SOTHER	166.67	167.54	1	84.07	85.14	1
Non-MLOT	507.75	498.40	−2	288.44	279.16	−3

The predicted throughput rate for nonmonitor lots is 0.104 lots per hour, which is extremely close to the observed throughput rate of 0.107 lots per hour. As for total throughput time, the predicted value is within 11% of the observed value for the six lot categories, and within 8% for all categories of nonmonitor lots. This comparison overstates the accuracy of the model because observed engineering hold time is directly incorporated in predicted total throughput time. A more meaningful comparison is between predicted and observed total throughput time *exclusive of engineering hold time.* On this basis the fit is less precise overall, but the maximal error is 14%. These results are consistent with past studies of computer system performance by means of queueing network models. For example, according to Lazowska et al. (1984), page 14, "a large body of experience indicates that queueing network models can be expected to be accurate to within 5–10% for [throughput rates] and to within 10–30% for [cycle times]." Similarly, in the manufacturing systems domain, Solberg (1977) found that the throughput rate predictions of a queueing network model agreed with the results of a detailed simulation to within 2.2%.

Throughout this paper, discussion is restricted to network-level performance characteristics, such as average total cycle time for a given lot category, as opposed to measures of throughput or congestion for individual stations. If one is concerned only with such aggregate characteristics, our simple model is quite adequate, which is remarkable in light of all the crude approximations involved. In contrast, and as anticipated, the station-level predictions of the model are typically bad, and it was not deemed useful to report them here.

Issues of parameter estimation and model validation are treated at some length in performance analysis textbooks, such as Lavenberg (1983), Lazowska et al., MacNair and Sauer (1985) and Sauer and Chandy (1981). In a perusal of those textbooks, one is struck by the wealth of practical experience and modeling insight that has accumulated in the area of computer and communication systems. It is our hope that the present study contributes to and helps stimulate the development of a corresponding body of knowledge for manufacturing systems generally, and for semiconductor manufacturing specifically.

4. Directions for Future Research

In Section 2 we described a simple queueing network model that yields quantitative performance predictions and requires only first moment information, or average rates of occurrence, as input. Although the model performs well for the one wafer fab we studied in detail, it is important to understand the limitations of the elementary theory on which it is based, and the refinements and extensions of the theory that are likely to be important in other settings.

A central issue in modeling manufacturing systems as queueing networks is the representation of equipment failures, or in more general terms, server interruptions. A particular failure mode, referred to hereafter as *simple server breakdown*, that is relatively easy to incorporate in a conventional queueing model assumes the following: failures occur only when ordinary services are in process, the failures occur in Poisson fashion, and repair times are independent of previous processing history. In this situation, the model with server breakdown can be mapped into an equivalent model without breakdown, using an effective service time distribution that accounts for both the actual service time and for delays due to breakdown. This is precisely the rationale that we used to justify our model of the TRC fab, without any attempt to verify the hypotheses.

An equivalent representation of simple server breakdown is obtained by viewing failures as a second class

of customers who: (a) arrive according to a Poisson process that runs only when ordinary customers are being served, and (b) are served on a preemptive priority basis. Unfortunately, the server interruptions that occur in real manufacturing systems are often of a more complicated, or at least different, character. In a wafer fab, it is not true that downtime events occur only when lots are being processed. Also, some types of server interruptions (like testing and requalification) may be best represented as nonpreemptive priority customers, and others (like certain types of preventive maintenance) may be best represented as a second class of customers with *lower* priority than ordinary processing tasks.

Another type of server interruption that is important to wafer fabrication, and in other manufacturing operations as well, is scheduled off-periods. For example, as mentioned in Subsection 2.2, there are several stations in the TRC fab that only operate one or two shifts per day, while most stations operate three shifts. At such a station, stochastic variability may be less important as a source of delay than simple customer build-up during the off-periods, and thus, conventional queueing models may fail to capture the essence of system behavior. For the case of a single-server station operating in isolation, Federgruen and Green (1987) analyzed the effect of both scheduled off-periods and random server breakdowns, but much remains to be done in the development of tractable network models, or network approximations, that adequately account for server interruptions.

Another major shortcoming of simple queueing models is that they impose restrictive distributional assumptions. There is, however, a growing literature on approximate analysis of queueing networks that allows general distributions, and this is a potentially important body of work for manufacturing system applications. For a discussion of its content, consider an open network with general renewal input processes, and general service time distributions (recall that simple server breakdown can be accommodated in this framework). Whitt (1983a, b) and his coworkers at AT&T Bell Laboratories developed a performance analysis software package, called the Queueing Network Analyzer (QNA). This package estimates system-level characteristics of such networks from the following station-level approximation (we shall not describe the most general form of the approximation):

$$\mathrm{EW} = \frac{\tau\rho(c_a^2 + c_s^2)}{2(1-\rho)}, \tag{13}$$

where EW is the average customer waiting time (before the customer begins service) at any particular station, $1 - \rho$ is the idleness rate at the station, τ is the average (effective) service time, c_a is the coefficient of variation for interarrival times, and c_s is the coefficient of variation for (effective) service times. This formula shows explicitly the dependence of customer delay on the variability parameters c_a and c_s, and the essence of Whitt's approach to network analysis lies in his method for estimating c_a from more primitive information. In studying the TRC fab, we considered using a two-moment approximation, but the TRC database was not adequate to estimate the coefficients of variation. In the absence of such data problems, Equation 13 could be used to refine the mixed network model described in Section 2. Specifically, the general philosophy espoused by Whitt (1984) leads to the following procedure. First, add τ to the right side of (13) to get an estimate of the average total delay per customer visit to the station in question. Next, using Little's Law, multiply that quantity by the arrival rate λ to get a two-moment estimate of the average total population L at the station, and use this formula in place of (5). Readers may wish to verify that the more general formula for L reduces to (5) in the case of exponential interarrival and service times ($c_a = c_s = 1$), as it should.

In Bitran and Tirupati (1986), a study that was motivated by a concern with modeling wafer fabrication, the authors argue that Whitt's method does not work well for networks with multiple customer types and deterministic routing, and suggest an alternative approach to the estimation of c_a that seems to give better results for such systems. Stimulated by this work, further research on two-moment approximations for complex queueing networks is currently underway, and this promises to become an important literature for manufacturing systems analysis.

A characteristic of wafer fabrication that departs significantly from the basic structural assumptions of conventional queueing models is temporal linkage of operations, such as cleaning and deposition. A lot arriving at a clean station may find that station idle but it still may not be processed immediately. Rather, the operator waits to be sure that a furnace tube will be available for the deposition operation that must immediately follow the cleaning. The lot in question is physically queued at the clean station, but it is actually waiting for a furnace tube to become available. Conventional queueing models do not allow for this feature, and thus its analysis represents a potentially interesting topic for future research.

Finally, it is important to recognize that lot sizes in wafer fabrication are largely discretionary, and the

choice of a particular lot size can substantially influence the severity of queueing effects in a fab. For example, the management of a production fab may choose to start n lots per day of 40 wafers each or $2n$ lots per day of 20 wafers each, and the average manufacturing cycle time experienced under those two policies may be quite different. The influence of lot size on cycle time was discussed in general terms by Karmarkar (1987) and explored in the scientific context of wafer fab by Spence and Welter, but much remains to be done on this subject.

Acknowledgment

This research was partially supported by National Science Foundation grant ECS-8603857, and by the Semiconductor Research Corporation under grant number 84-01-046 (Manufacturing Science for VLSI, Center for Integrated Systems, Stanford University). We are indebted to John Shott for his invaluable technical assistance, to the Hewlett-Packard Corporation for agreeing to participate in the research project, and to Terry Harms, Ed Middlesworth and Susan Okada of the Hewlett-Packard Technology Research Center, without whose cooperation this work would not have been possible. Finally, we gratefully acknowledge the collegial support provided by David Burman of AT&T Technologies (ERC), who generously shared his group's experience in modeling and analysis of wafer fab operations.

References

BASKETT, F., K. M. CHANDY, R. R. MUNTZ AND F. G. PALACIOS. 1975. Open, Closed and Mixed Networks of Queues with Different Classes of Customers. *J. Assoc. Comput. Mach.* **22,** 248–260.

BITRAN, G. R., AND D. TIRUPATI. 1986. Multiproduct Queueing Networks with Deterministic Routing: Decomposition Approach and the Notion of Interference, Working Paper 1764-86, Sloan School of Management, M.I.T.

BURMAN, D. Y., F. J. GURROLA-GAL, A. NOZARI, S. SATHAYE AND J. P. SITARIK. 1986. Performance Analysis Techniques for IC Manufacturing Lines. *AT&T Bell Labs. Tech. J. 65.*

CHEN, H., J. M. HARRISON, A. MANDELBAUM, A. VAN ACKERE AND L. M. WEIN. 1986. Queueing Network Models of Semiconductor Wafer Fabrication, Technical Report, Stanford Center for Integrated Systems.

DAYHOFF, J. E., AND R. W. ATHERTON. 1984. Simulation of VLSI Manufacturing Areas. *VLSI Design* (December), pp. 84–92.

DAYHOFF, J. E., AND R. W. ATHERTON. 1986a. Signature Analysis: Simulation of Inventory, Cycle Time and Throughput Trade-offs in Wafer Fabrication. In *IEEE Trans. Components Hybrids Mfg. Technol.* **CHMT-9,** 498–507.

DAYHOFF, J. E., AND R. W. ATHERTON. 1986b. Signature Analysis of Dispatch Schemes in Wafer Fabrication. In *IEEE Trans. Components Hybrids Mfg. Technol.* **CHMT-9,** 508–525.

DAYHOFF, J. E., AND R. W. ATHERTON. 1987. A Model for Wafer Fabrication Dynamics in Integrated Circuit Manufacturing. *IEEE Trans. Syst. Man Cybernet.* **SMC-17,** 91–100.

FEDERGRUEN, A., AND L. GREEN. 1987. Queueing Systems with Service Interruptions. *Opns. Res.* **34,** 752–768.

GISE, P., AND R. BLANCHARD. 1986. *Modern Semiconductor Fabrication Technology.* Prentice-Hall Reston, Englewood Cliffs, N.J.

KARMARKAR, U. S. 1987. Lot Sizes, Lead Times and In-Process Inventories. *Mgmt. Sci.* **33,** 409–418.

KELLY, F. P. 1979. *Reversibility and Stochastic Networks.* Wiley, New York.

LAVENBERG, S. S. 1983. *Computer Performance Modeling Handbook.* Academic Press, New York.

LAZOWSKA, E. D., J. ZAHORJAN, G. S. GRAHAM AND K. C. SEVCIK. 1984. *Quantitative System Performance.* Prentice-Hall, Englewood Cliffs, N.J.

MACNAIR, E. A., AND C. H. SAUER. 1985. *Elements of Practical Performance Modeling.* Prentice-Hall, Englewood Cliffs, N.J.

SAUER, C. H., AND K. M. CHANDY. 1981. *Computer Systems Performance Modeling.* Prentice-Hall, Englewood Cliffs, N.J.

SOLBERG, J. J. 1977. A Mathematical Model of Computerized Manufacturing Systems, paper presented at the 4th International Conference on Production Research, Tokyo (August 22–30).

SOLBERG, J. J. 1983. Mathematical Design Tools for Integrated Production Systems. In *Efficiency of Manufacturing Systems,* B. Wilson et al. (eds.). Plenum Press, New York.

SPENCE, A. M., AND D. J. WELTER. 1987. Capacity Planning of a Photolithography Work Cell in a Wafer Manufacturing Line. In *Proceedings 1987 IEEE International Conference on Robotics and Automation,* Vol. 2, pp. 702–708.

SZE, S. M. 1983. *VLSI Technology.* McGraw-Hill, New York.

VINOD, B., AND T. ALTIOK. 1986. Approximating Unreliable Queuing Networks under the Assumption of Exponentiality. *J. Opnl. Res. Soc.* **37,** 309–316.

WHITT, W. 1983a. The Queuing Network Analyzer. *Bell Syst. Tech. J.* **62,** 2779–2815.

WHITT, W. 1983b. Performance of the Queueing Network Analyzer. *Bell Syst. Tech. J.* **62,** 2817–2843.

WHITT, W. 1984. Open and Closed Models for Networks of Queues. *AT&T Bell Labs. Tech. J.* **63,** 1911–1979.