A Production Planning Methodology for Semiconductor Manufacturing
Based on Iterative Simulation and Linear Programming Calculations

Yi-Feng Hung and Robert C. Leachman

Abstract—We introduce a methodology for automated production planning of semiconductor manufacturing based on iterative linear programming (LP) optimization and discrete-event simulation calculations. The LP formulation incorporates epoch-dependent parameters for flow times from lot release up to each operation on each manufacturing route. LP-derived release schedules are used as input to the simulation model, from which statistics on flow times are collected and used to reformulate the LP model for a revised planning calculation. Iteration continues until satisfactory agreement between simulation and LP models is obtained. We demonstrate in experiments on an industry data set that a relatively small number of iterations is required to develop a production plan correctly characterizing future flow times as a function of factory load and product mix. The methodology makes possible automated production planning of semiconductor manufacturing on an engineering workstation.

I. INTRODUCTION

Increased importance of on-time delivery in the semiconductor industry has led to a need to improve production planning methodology and practices. At present, almost every semiconductor manufacturer utilizes the basic “explosion” logic of manufacturing requirements planning (MRP) systems to perform production planning calculations, briefly summarized as follows. In this methodology, required quantities of manufacturing lots to release are derived from desired output quantities by scaling the output according to prespecified flow time parameters and by shifting the time period of desired output back to a planned period for raw material release using prespecified flow time parameters. The resulting release schedule then may be further adjusted for reasons of capacity limitations, desirable lot sizes, practical ramp rates, etc., in order to derive the final production plan. Whether incorporated in home-grown spreadsheets or in commercial MRP-type planning software, the use of prespecified flow time parameters prevails. In most applications, the flow times are not only prespecified but assumed to be static over the entire planning horizon. However, it is well-known among researchers and practitioners that product flow times are uncertain, and that their distributions may shift with changes in the level of factory loading and/or changes in the product mix. Since production planning is the task of determining factory load and product mix over time, it can be difficult to ascertain the validity of a production plan derived using prespecified flow time parameters.

Another characteristic of production planning in the semiconductor industry is that planning is very time-consuming. Only a very few manufacturers have automated planning to the extent that an official plan can be regenerated once a week, say, over a weekend. More typical is the situation in which a planning cycle consumes one or several weeks, involving a number of management meetings to negotiate trade-offs and to obtain “buy-in” to the plan. This lengthy planning process inevitably means that quotations of delivery dates to customers must be made based on old and perhaps stale plans and/or on very sketchy supply-side information. It also means that a sizable proportion of production release are made in response to demand forecasts rather than actual customer orders.

With this on-going challenge of production planning in the industry as a background, the last ten years have seen rapid development of factory simulation software. Commercial software is now available and successfully applied in many companies to characterize wafer fab flow times as a function of product mix and variability in equipment availability. The decade also has witnessed rapid development of powerful linear programming software operating on work station computers. Problems with 150,000 constraints on 150,000 variables now can be routinely solved in a matter of hours.

In this paper we introduce a methodology exploiting these new software and hardware developments to provide an automated production planning capability specifically tailored for the semiconductor industry. The methodology involves iterative simulation and linear programming calculations to simultaneously establish future flow times as well as production quantities. We take advantage of the General Framework for production planning models proposed in [3] to develop a linear programming formulation that embeds many more
parameters than usual, thereby admitting dynamic flow times. We also take advantage of an aggregated simulation model that predicts flow times accurately yet economizes on computer run time. We demonstrate the practicality and accuracy of the methodology on actual industry data.

The outline of the paper is as follows. In Section II we introduce our terminology and provide an outline of the linear programming formulation of the planning problem. Details of the formulation are reserved for the Appendix. In Section III we discuss in detail the flow time parameters of the formulation. In Section IV we discuss the aggregated wafer fab simulation model. In Section V, we formally define the overall planning methodology integrating simulation and linear programming calculations. In Section VI, we discuss computational results from application of this methodology to industrial data. Finally, in Section VII, we provide our conclusions and suggestions for further research.

II. LINEAR PROGRAMMING FORMULATION OF THE PRODUCTION PLANNING PROBLEM

For simplicity of exposition, we restrict our attention to the planning of “front-end manufacturing,” i.e., wafer fabrication and electrical test. We further assume that demands to be loaded on the front end are net of die inventory and projected output of work-in-process (WIP). In principle, the methodology we introduce could be applied for planning the entire semiconductor manufacturing process. For linear programming formulations encompassing the allocation of WIP and inventory as well as planning the back-end manufacturing (i.e., device packaging and test), see [7].

A. Terminology

We use the following terminology. A front-end manufacturing facility is termed a wafer fab. A work station is a group of identical machines within the fab used to perform particular operations that add value to the raw material. A wafer is an individual unit of processing material that passes through the wafer fab. A blanket wafer fed into the wafer fab will emerge imprinted with hundreds of dice. Each die is an integrated circuit. The final operation of wafer fab is a wafer electrical test that uses a test machine to grade each individual die. The percentage of good dice on a finished wafer defines the die yield. Dice with the proper grade will be packaged (assembled) and tested in the back end to become final products of semiconductor manufacturing. A lot is a quantity of a single wafer type processed as a whole and traveling as a unit between work stations. An operation is the performance of a particular step of the required processing activity on a lot of wafers by a particular work station. The performance will consume a certain amount of time on a machine and on the lot. A route is a sequence of operations required to produce finished wafers.

Flow time (also known as production lead time, turn-around time, throughput time, or cycle time—the latter term most commonly used in the semiconductor industry) is the elapsed time between two events occurring to a lot of a particular product following a particular route. Total flow time is the overall time a lot spends on the shop floor. Flow time from release up to an operation is the age of a lot when it commences that operation. Flow time from an operation to finish is the additional duration to be spent in the shop before a lot is completed. Fig. 1 diagrams the flow time from release up to operation \( I \), flow time from operation \( I \) to output, and total flow time (from lot release to lot output) of a particular lot. The components of flow time include, in addition to actual processing time, the delays waiting in queues and the times to transport from the site where one operation is performed to the next. In semiconductor manufacturing, because the number of operations in production routes is large (hundreds), and because of the trade-off of waiting time in exchange for high equipment utilization in a factory of unreliable equipment, the total flow time is in general much longer than the total processing time.

B. Outline of the Basic LP Formulation for Front End Planning

This section will outline the formulation of an LP model for wafer fab production planning. This model was originally proposed in [6], and has seen application at Intel Corporation, Harris Corporation-Semiconductor Sector, Advanced Micro Devices, Inc. and Samsung Electronics Corp., Ltd. The assumptions underlying this model are as follows.

1) The activities of the model are the production activity on each of the wafer fab routes. Activity levels are measured in terms of the quantity of wafers released; the quantity of output is measured in terms of good die. Input-output relationships of such activities are time-phased and thus must be described by dynamic production functions [3]. For simplicity of exposition, we assume that each wafer type provides a single type of die, and thus wafer types and die types are synonymous. There can be alternative wafer fab routes for producing the same wafer type (die type).

2) The overall planning horizon is divided into planning periods in which demands, capacities and production rates are assumed to be held constant. The length of each planning period for each wafer fab facility may vary and is measured in terms of working days. The length of each
period also is measured in calendar days for the purpose of discounting cash flows in the objective functions.

3) A production variable is defined as a quantity of a particular wafer type to be released following a particular route during a planning period. An inventory variable is defined as the inventory level of a particular die type at the end of a planning period. A backorder variable represents the quantity of die demand that can not be satisfied on time at the end of a planning period.

4) The demands are expressed in terms of time-phased die output requirements and are assumed to be net of initial die inventory and net of equivalent die output of the initial work-in-process (WIP). These demands are divided into prioritized classes that are loaded onto front end facilities by incremental linear programming calculations. Demands in class 1 are loaded first, then demands in classes 1 and 2 are loaded subject to not exceeding backorder levels associated with class 1, etc. The limits on backorders are expressed as upper bounds on the back order variables. The formulation for all classes is the same, except for the values of the demands and the lower bounds on back order variables.

5) We assume production is rate-based, i.e., the release quantity in a particular period is to be distributed uniformly over the period.

6) Capacity constraints are enforced that require the total workload (expressed in machine-hours) on a machine type in each planning period to be less than the capacity (expressed in machine-hours) of the machine type in that period. The total workload includes the estimated machine-hours to flush initial WIP out of all routes as well as loads from planned releases. Workloads are estimated for each operation on each route using flow time parameters to be discussed.

7) As a horizon condition, we require that the wafer fab enter steady-state, whereby production releases on each route are required to follow some constant rate in all periods falling within one total flow time of the planning horizon. The planning periods that overlap the interval beginning one total flow time for a route before the planning horizon until the horizon are termed frozen periods with respect to that route. Demands from each class in the last planning period are assumed to continue at the same rate forever. Enforcing these constraints is accomplished by simple variable substitutions and the addition of an extra planning period placed after the given horizon, as detailed in the Appendix.

The basic form of the LP formulation can be broadly structured as follows.

Maximize the Discounted Sum of

\[ \text{die output revenue} - \text{raw material cost} - \text{die inventory holding cost} - \text{cost of backordered die demands} \]

Subject to

1) Resource Capacity. For each work station in each planning period,

\[ \text{(machine hours required to process new releases)} \]
\[ \leq \text{(available machine hours for processing activity)} \]
\[ - \text{(machine hours required to flush initial WIP)}. \]

2) Die Demands: For each die type and each period,

\[ \text{(die output during the period)} + \text{(inventory at the start of period)} - \text{(backorders at the start of period)} - \text{(inventory at the end of period)} + \text{(backorders at the end of period)} = \text{(demands during the period)}. \]

3) Variable Ranges: For each variable and each period,

\[ 0 \leq \text{(backorder variables)} \]
\[ \leq \text{(upper bound on backorder quantity)}, \]
and all other variables \( \geq 0 \).

The complete detailed formulation of the LP model is provided in the Appendix.

III. Expressing Loads and Output in Terms of Release Variables

The formulation requires one to express the output and the loads on resources in terms of the production release variables. To construct the constraints for die demands and for resource capacities outlined above, pre-specified flow times implicitly or explicitly must enter the formulation.

In almost all published linear programming formulations for production planning and in textbook presentations of MRP logic, the release quantity in a given planning period is mapped to output in some single period. That is, there is a flow time parameter for each planning period assumed to be applicable to all releases scheduled in that period. Similarly, the resource load at a processing step associated with the release quantity in a given period on a given route containing the step is modeled to occur entirely within some single planning period. See, for example, [1] or [5]. This conventional modeling assumption for flow times is depicted in Fig. 2. The upper time line in the figure shows a release schedule, while the lower time line shows the corresponding output schedule. Note that the boxes on the lower time line correspond one-to-one with those on the upper time line and are simply shifted by a fixed flow time parameter. A figure depicting the scheduled workload at a particular operation as a function of the release schedule would be similar.

In this paper we depart from conventional modeling of flow times in two respects. First, our flow time parameters are real-valued rather than integral, whereby releases in one planning period result in scheduled output and operation workloads that are spread over more than one planning period. Second, our flow time parameters are epoch-based rather than period-based; that is, we use parameters that apply specifically to the time points marking the beginning of each planning period. These epochs are measured in terms of the number of working
days since the beginning of the first planning period. The flow time parameters applicable to the epochs marking the beginning and end of each planning period will in general be distinct.

Fig. 3 illustrates our flow time parameters. Total flow times (from release to output) are specified that are applicable to the boundary epochs between consecutive planning periods. The arrows in the figure are used to identify the flow times applicable to output realized at the start and end of planning period 3; the arrows point to the corresponding epochs for wafer release. Note that the flow times applicable to output realized at the start and end of planning period 3 are different. As will be discussed, the value of each of these parameters is determined by examining the simulated flow times of lots which arrive at die bank at times close to the corresponding boundary epoch of the planning period.

In view of the assumption of uniform rates of wafer releases in each planning period, the flow time parameters can be used to determine the fraction of wafer releases in various periods that contribute output in a given planning period. For example, in the figure, it appears that approximately 80% of the wafer releases in period 2 and 30% of the wafer releases in period 3 contribute output in period 3.

Given a prespecification of these epoch-dependent flow times, the coefficients on production release variables in these inequalities are calculated as follows. We provide formulas that express the quantity of wafers passing through an arbitrary operation \( l \) in period \( p \) in terms of the release variables. Modifications to the formulas to express output in period \( p \) in terms of the release variables are discussed immediately afterwards. We generalize the presentation for the case of variable-length planning periods.

Indices:
- \( i \): index of route.
- \( p, q \): index of planning period, \( p = 1, 2, 3, \ldots, P \).
- \( \tau, \ell \): epoch, a point on the continuous time line beginning with time 0 at the start of the first planning period.

Parameters:
- \( \tau_p^i \): number of working days on route \( i \) from start of period \( 1 \) (time 0) until the end of period \( p \), \( p = 1, 2, 3, \ldots, P \), all \( i \).
- \( [\tau_{i,j}^-] \): smallest index \( p \) such that \( \tau_{i,j}^p > \tau \).
- \( F_{p,l}^i \): the expected flow time from wafer release to operation \( l \) occurring at epoch \( \tau_{i,l}^p \) of route \( i \).
- \( F_i^p \): the expected flow time from wafer release to finish occurring at epoch \( \tau_{i,j}^p \) of route \( i \).
- \( e_i^p \): the expected wafer yield from wafer release to operation \( l \) of route \( i \), effective for loading of the operation in period \( p \).
- \( e_{i,p,l} \): the expected wafer yield from wafer release to finish of route \( i \), effective for output in period \( p \).

Variables:
- \( X_{p,i}^q \): wafer release quantity for route \( i \) in period \( p \).

Derived Variables:
- \( Y_{p,i}^q \): wafer quantity consuming machine hours at operation \( l \) of route \( i \) in period \( p \), equal to a linear combination of the \( \{X_{q,i}^l\} \) variables, \( q = 1, 2, 3, \ldots, P \).
- \( Y_{p,i} \): wafer output quantity from route \( i \) in period \( p \), equal to a linear combination of the \( \{X_{q,i}^l\} \) variables, \( q = 1, 2, 3, \ldots, P \).

1) Calculating Coefficients for Expressing Load of Operation \( l \) of Route \( i \) in Period \( p \) in Terms of Release Variables: To calculate the coefficients for machine loading constraints for the linear programming model, we apply the following formulas to each operation \( l \) of each route \( i \) in each period \( p \).

For the workload of a certain operation \( l \) of route \( i \) in period \( p \), the two boundary epochs are \( \tau_{p-1}^i \) and \( \tau_p^i \). The respective flow times are \( F_{p-1,l}^i \) and \( F_p^i \). Then, the corresponding release epochs mapped from the \( \tau_{p-1}^i \) and \( \tau_p^i \) boundaries of load period \( p \) are \( \tau_{p-1}^i - F_{p-1,l}^i \) and \( \tau_p^i - F_p^i \), respectively. See Fig. 4.

Let

\[ p^- = [\tau_{p-1}^i - F_{p-1,l}^i]^+ \quad \text{and} \quad p^+ = [\tau_p^i - F_p^i]^+ \]

Case 1: The interval \( [\tau_{p-1}^i - F_{p-1,l}^i, \tau_p^i - F_p^i] \) is contained in some open interval between two time grid points, as illustrated in Fig. 4(a). Then \( p^+ \) and \( p^- \) are equal in this case.
IV. SIMULATION MODEL

The test data set used in this research was adapted from an actual wafer fab data set provided by Micron Technology, Inc. in 1989, reflecting the situation in a wafer fab of the company at the time. (Names of wafer types and machine types reported herein have been disguised from actual names at Micron.) The data set is appropriate for studying the effect of flow times on planning, since, at the time, the company was going through a major transition in product mix as it moved from the 256 K DRAM generation of products to the 1 Megabit DRAM generation, with total demands exceeding factory capacity. The test data set includes 10 types of products (wafers) with demands in two classes (booked customer orders and potential sales). Each product type follows a distinct route. Each route has between 86 and 187 operations involving 30 work stations. Defined for each operation is the required work station (resource type) and the processing time. Depending on the type of operation, the processing time may be expressed as time per wafer, time per lot, time per machine load, or some combination of such factors. Each work station (resource type) consists of a given number of identical machines operated in parallel. Also defined for each work station are mean time between failure (MTBF) and mean time to repair (MTTR) parameters applicable to each machine in the work station. These parameters do not account for all machine down times, but rather only unplanned machine down times, i.e., equipment failures. As discussed in Section V below, we conducted experiments in which machine availability is arbitrarily reduced (and variability is increased) to more realistic levels.

A simplifying assumption made in our simulations is that all operations are lot-based. Considering the various processing time factors in the original data set, the performance time of each operation was summarized by a single time per lot, assuming all operations of all products have the same lot size, 50 wafers. The processing time of each operation in the simulation is expressed as the time to process one lot. The general formula used is

\[ Y_{i,p}^l = \frac{e_{i,p}^l}{\tau_{i}^l - \tau_{i+1}^l} \times \text{time per wafer} \times \text{lot size} + \text{time per lot} \times \frac{1}{\text{number of lots per batch}}. \]

(In the LP model, the processing time used is per wafer, i.e., the time per lot divided by the lot size.) For any particular operation in the data set, the performance time of each operation was determined by a single time per lot, assuming all operations of all products have the same lot size, 50 wafers. The processing time of each operation in the simulation is expressed as the time to process one lot. The general formula used is

\[ Y_{i,p}^l = \frac{e_{i,p}^l}{\tau_{i}^l - \tau_{i+1}^l} \times \text{time per wafer} \times \text{lot size} + \text{time per lot} \times \frac{1}{\text{number of lots per batch}}. \]

For \( q = p^+ + 1, \ldots, p^{++1} \), the coefficient of \( X_{i,p}^j \) as we express \( Y_{i,p}^l \) is \( e_{i,p}^l \). That is, all the wafers released in such a period \( q \) will contribute workload in period \( p \).

\[ Y_{i,p}^{+} = \frac{\tau_{i}^{p^+} - (\tau_{i}^{p^+} - \tau_{i+1}^{p^+})e_{i,p}^l}{\tau_{i}^{p^+} - \tau_{i+1}^{p^+}} + \sum_{q=p^+}^{p^{++1}} e_{i,p}^l X_{q}^i + \frac{(\tau_{i}^{p+1} - \tau_{i}^{p+})e_{i,p}^l}{\tau_{i}^{p+1} - \tau_{i+1}^{p+}} X_{i}^{+}. \]

2) Calculating Coefficients for Expressing Output in Terms of Release Variables: To calculate the coefficients for output \( Y_{i,p}^l \) in terms of release variables, we can use the formulas for calculating \( Y_{i,p}^l \) above if we simply replace \( F_{i,p}^l \) and \( e_{i,p}^l \) by \( F_{i}^l \) and \( e_{i}^l \), respectively. To calculate the coefficients expressing output from release variables for each route of the linear programming model, we simply apply the above formulas with this substitution for each route \( i \) in each period \( p \).
The planning horizon in the simulation and LP planning models is 12 periods. Each planning period is a 30-working-day month. Each working day consists of 24 working hours. Planned production releases in each period are distributed uniformly over the period in the simulations.

The simulation programs used in this research are written in Objective-C [9], based on the BLOCS simulation software package [2]. Various subclass objects added to the BLOCS library were written to facilitate this research [4]. Simple first-in, first-out dispatching priorities were assumed throughout.

To collect flow times for each simulated lot, we include one variable per lot in the simulation that records the release time of the lot. Once an operation on a lot is initiated, the flow time from release to operation is computed from this variable and the current clock time and reported to a data collection object (an object written in Objective-C). The total number of variables required is equal to the number of lots-in-process. Hence a relatively modest amount of memory space is required for the simulation program. Since it would be coincidental to have lots in the simulation commence an operation exactly at a boundary epoch of a planning period, the method we use to estimate the flow time effective at the boundary is to linearly interpolate the flow times of lots that initiate processing immediately before and immediately after the boundary epoch.

V. ITERATIVE SCHEME

Ideally, flow times should be decision variables of a planning model since the planned resource workloads will determine the achievable flow times. The familiar trade-off curve between flow time and machine utilization predicted by queueing theory is shown in Fig. 5. As can be seen, the relationship between these two variables is not linear, suggesting that it would be very difficult to develop a practical analytical planning model incorporating both flow times and workloads as variables.

The approach to planning suggested by our research is to embed flow times predicted by simulation in an LP planning model. Using initial estimates of flow times, or perhaps historical average flow times, an LP production planning model is formulated as above and solved to generate a trial release schedule. We then run a simulation model using the release schedule from the LP as input, whereby the total wafer releases planned in each period are converted into lots and rounded up to the next integral lot quantity, and the simulated lot releases are spread uniformly over the period. During the simulation run, the statistics on flow times are collected. These statistics enable us to establish revised estimates of flow times that are used to reformulate the LP for the next LP planning run. We continue to iterate between the two models until satisfactory agreement in flow times is achieved. Fig. 6 depicts this iteration scheme.

VI. COMPUTATIONAL EXPERIMENTS

A. Deterministic Simulation Model

In the first computational experiment, we simulate a deterministic fab, i.e., one with constant machine availability and constant die yields. Results (in terms of flow time agreement) from applying the proposed LP—simulation iteration scheme to the Micron data set are plotted as the diamond-marked line of Fig. 7 and are tabulated in Table I. The table shows that the mean absolute deviation (MAD) in total flow times for the 10 products is steadily reduced, reaching a minimum in iteration 29, whose percentage MAD is only 0.953%. Note that, starting from quite poor estimates of flow times, the percentage MAD is only 5% after 5 iterations, probably acceptable for planning purposes.
That is, the percentage of over-schedule can be calculated similarly.

Let 
\[ l_{op} = \text{the LP-projected cumulative output point for product } i \text{ at the end period } p. \]
\[ so_{ip} = \text{the simulated cumulative output point for product } i \text{ at the end of period } p. \]

We define under-schedule as the shortfall of simulation output compared to LP output. The over-schedule is the other way around.

The output curves of LP and simulation models also can be compared. As far as the planning system is concerned, the key measure of feasibility of the plan is that the planned cumulative output point at each end point of a planning period is in fact achievable. We verify the achievability of the plan using the simulation model. In general, we expect the difference in total flow times between these two models times the production rate, since the cumulative release curves are the same for both LP and simulation. As long as we can obtain agreement between models in their total flow times, we expect to obtain agreement in output curves. To numerically verify this, we define the following measurements for comparing the output curves between LP solution and simulation.

Let
\[ l_{op} = \text{the LP-projected cumulative output point for product } i \text{ at the end period } p. \]
\[ so_{ip} = \text{the simulated cumulative output point for product } i \text{ at the end of period } p. \]

We define under-schedule as the shortfall of simulation output compared to LP output. The over-schedule is the other way around.

The percentage of under-schedule is the amount of under-schedule divided by the total production during the period. The percentage of over-schedule can be calculated similarly. That is,

\[ \% \text{ Under-Schedule at the End of Period } p = \frac{\sum_{i \in \text{all product}} \text{Max}(0, l_{op} - so_{ip})}{\sum_{i \in \text{all product}} l_{op}} \times 100\%. \]

\[ \% \text{ Over-Schedule at the End of Period } p = \frac{\sum_{i \in \text{all product}} \text{Max}(0, so_{ip} - l_{op})}{\sum_{i \in \text{all product}} l_{op}} \times 100\%. \]

\[ \% \text{ Over-Schedule at the End of Period } p = \frac{\sum_{i \in \text{all product}} \text{Max}(0, so_{ip} - l_{op})}{\sum_{i \in \text{all product}} l_{op}} \times 100\%. \]

\[ \% \text{ Under-Schedule at the End of Period } p = \frac{\sum_{i \in \text{all product}} \text{Max}(0, l_{op} - so_{ip})}{\sum_{i \in \text{all product}} l_{op}, p - 1} \times 100\%. \]

\[ \% \text{ Under-Schedule at the End of Period } p = \frac{\sum_{i \in \text{all product}} \text{Max}(0, l_{op} - so_{ip})}{\sum_{i \in \text{all product}} l_{op}, p - 1} \times 100\%. \]

\[ \% \text{ Under-Schedule at the End of Period } p = \frac{\sum_{i \in \text{all product}} \text{Max}(0, l_{op} - so_{ip})}{\sum_{i \in \text{all product}} l_{op}, p - 1} \times 100\%. \]

\[ \% \text{ Under-Schedule at the End of Period } p = \frac{\sum_{i \in \text{all product}} \text{Max}(0, l_{op} - so_{ip})}{\sum_{i \in \text{all product}} l_{op}, p - 1} \times 100\%. \]

\[ \% \text{ Under-Schedule at the End of Period } p = \frac{\sum_{i \in \text{all product}} \text{Max}(0, l_{op} - so_{ip})}{\sum_{i \in \text{all product}} l_{op}, p - 1} \times 100\%. \]

\[ \% \text{ Under-Schedule at the End of Period } p = \frac{\sum_{i \in \text{all product}} \text{Max}(0, l_{op} - so_{ip})}{\sum_{i \in \text{all product}} l_{op}, p - 1} \times 100\%. \]

\[ \% \text{ Under-Schedule at the End of Period } p = \frac{\sum_{i \in \text{all product}} \text{Max}(0, l_{op} - so_{ip})}{\sum_{i \in \text{all product}} l_{op}, p - 1} \times 100\%. \]
run in order to estimate mean flow times. Each simulation run requires one initial random value seed to generate one machine pattern. For this experiment, results from 16 runs of the simulation model are averaged to calculate mean flow times. In the LP formulation, we account for the expected deterioration of the performance of the iteration scheme. The best total flow time agreement of 5% is probably adequate for semiconductor manufacturing due to uncertainty in the LP formulation.

Table III displays the results comparing planned and simulated mean total flow times generated by the iteration scheme in the case of simulated random machine failures. The output curve comparison is shown in Table IV. As can be seen in the tables and as plotted as the square-marked line in Fig. 7, the introduction of random machine failures does not significantly deteriorate the performance of the iteration scheme. The best flow time agreement is achieved at iteration 36, whose mean total flow time difference is less than 0.8%. Agreement within 5% is achieved in 7 iterations.

Table V shows the MTBF, MTTR and the equipment availability factors for the eight most heavily utilized machines in the original Micron data set, as well as for three experiments in which machine failure rates are scaled upwards. Four of these work stations (mostly wafer electrical test stations) have very high-reliability machines for which scaling the failure rate has little effect on availability, while the other four work stations (dry etch, chemical vapor deposition and implant stations) include more unreliable machines. As can be seen, for the four work stations with unreliable equipment, availability falls from a range of 73%–90% down to a range of 47%–76% as the failure rate scaling factor is increased from 1.0 to 3.0. Table VI shows the results for agreement in flow times between LP and simulation models for these experiments. As can be seen, agreement within 5% is obtained within 5 or 6 iterations for failure rates up to three times the given values.

Thus the rate of convergence to expected flow times does not seem to deteriorate with increasing machine variability, at least within the range of failure rates tested here.

C. Stopping Rule for Iterative Calculation

In practice, a total flow time agreement of 5% is probably adequate for semiconductor manufacturing due to uncertainty
in die yields and machine availability, and due to the indivisibility of lots. For the deterministic case such agreement was obtained at the fifth iteration. For the machine failure experiments with MTBF parameters up to three times the given data, such agreement was realized in seven iterations or less. We can keep iterating more LP and simulation calculations to achieve better accuracy, but there is a limitation of achievable accuracy, about 1% for the experiments in the case of the Micron data set. We remark that if the entire Micron data set of 89 work stations were simulated, the percentage agreement likely would be substantially better, since there is little variance in the flow times of the many low-utilization work stations that were omitted in our experiments.

In the experiments described above, we do not start the iterative calculation with good estimates of future flow times. In practice, we could implement the usual “rolling horizon” approach to planning, whereby the planning calculation is performed at a fixed interval, such as once a week, and the planning horizon length is kept constant while rolling forward on the time axis from the previous planning time. Each plan would start with the latest factory information about WIP status and machine status. We could use the epoch flow time estimates from the previous planning calculation as good initial estimates of flow times in the current planning calculation. In such a case, it should take fewer iterations to obtain satisfactory flow time agreement.

D. Discussion of Case Study Results

Dynamic (i.e., time-varying) product mixes can make the work station utilizations time-varying over the planning horizon, in the case that the various products follow distinct routes. According to queueing theory, the time-varying utilizations will in turn cause time-varying flow times, stressing the importance of embedding time-varying flow time parameters in the planning model. This is borne out in our case-study results.

For the case of the deterministic factory, satisfactory agreement in flow times is achieved at iteration #5. The graph of planned work station utilizations (as a percentage of expected capacity) versus the period number for the results of this iteration is shown in Fig. 8. Each line represents one of the 10 most heavily loaded work stations. We can see the loads on work stations are varying along the planning horizon. The total flow times effective at the end of the period versus period number are graphed in Fig. 9. Each line in the graph represents one product (route). We can observe that the total flow times are indeed time-varying.

For the case of machine failures simulated according to the given MTBF data, satisfactory agreement in flow times is achieved at the seventh iteration. Fig. 10 shows the graph of planned work station utilizations (as a percentage of expected capacity) versus the period number for iteration #7 of this case, and Fig. 11 shows the total flow time effective at the end of the period versus period number. We can clearly see that both work station loads and flow times are time-varying for this case as well. In Fig. 11, we can observe that some routes (products) have relatively longer flow times in the middle of the planning horizon. The reasons for this can be deduced from the data provided in Tables VII and VIII. Table VII identifies the bottleneck work stations (i.e., the work stations that are fully utilized) in each period. Table VIII shows the number of visits to the bottleneck work station for each product (route). We can see that work station M22 is only visited by route Z. Therefore, changes in the utilization of M22 will not affect flow times of other routes. Work station M20 is only fully loaded in period 3. Even though it is used by 6 routes, it does not affect flow times over the whole horizon very much. On the other hand, work stations M9 and M28 are the dominant work stations in terms of effects on flow times. From Table VII, we can see the bottleneck work station is shifting from M9 to M28, and then from M28 back to M9 again as we move through the planning horizon. During the middle periods of the planning horizon, routes A, C, T, U have relatively longer total flow times. The reason is that they make more visits to M28 (the bottleneck work station during the middle periods) than the other routes. While M28 is the bottleneck, the more visits a route makes to M28, the longer delay there will be for lots on that route.

Table IX shows the ratio of average total flow time over the theoretical total flow time for iteration #7 of this machine failures case, where theoretical total flow time is defined as the sum over all operations on the route of the processing times (i.e., excluding all waiting times). The average of this ratio over all routes is 4.15. We remark that this ratio for the real fab could be much lower. The reason for such a high
value for this ratio in our simulations is because this test data set only contains the 30 highest utilization work stations and excludes the low utilization work stations (59 of the 89 total work stations), which in general will have small waiting times. The ratio of flow time (waiting time plus the processing time) of one operation performed by such a work station to the pure processing time will be very close to 1. Including these low utilization work stations in the manufacturing network would certainly decrease the ratio.

E. Operation Flow Times

Most published analytical models for capacity analysis are steady-state in nature, i.e., they implicitly assume zero flow times in the constraints for resource capacities. Moreover, the authors have observed that most semiconductor companies perform their capacity analyzes incorporating this assumption as well. That is, all capacity consumption by wafer releases is assumed to occur in the time period of release. To investigate the importance of representing operation flow times in the planning model, the above experiments with the Micron data set were repeated with LP formulations in which the operation flow times were fixed at 0, but total flow times were iteratively updated according to simulation results. The iterative scheme did not steadily converge in these experiments. Instead the output curves fluctuated between over-schedule and under-schedule disagreement. For re-entrant flow production such as semiconductor manufacturing, it seems to be essential to include operation flow times in capacity analysis of time-varying product mixes.

We remark that while the inclusion of such parameters makes the LP model more complicated to formulate, it does not make it appreciably more difficult to solve, since the number of variables and constraints in the model remains the same.

VII. IMPLICATIONS FOR INDUSTRIAL PRACTICE, EXTENSIONS AND FURTHER RESEARCH

The results of this research show that iterative planning calculations with an LP planning model and a simulation model can generate accurate production plans on industry data. The formulation of the LP model based on epoch-dependent flow times and the inclusion of distinct flow times for every individual operation performed by bottleneck resources seems necessary to achieve convergence to a valid production plan.

Very detailed simulation models of semiconductor manufacturing can require excessive amounts of computer execution time. However, it is not necessary to utilize a very large simulation model to obtain accurate flow time estimates. We have found that operations performed on low-utilization work stations may be replaced by fixed flow times without significant loss of accuracy. For the Micron data set described herein, we compared simulations of the model including 30 work stations to simulations of a model that included only 10 of these work stations. The 10 work stations retained
Fig. 11. Total cycle time versus period number (Iteration #7 of machine failure case).

were those whose operations had the highest variance in flow times—which coincided with the 10 work stations with the highest utilizations. All operations performed on the other work stations were replaced by fixed time delays equal to the mean flow times simulated in the 30 work-station model. Comparing the simulated results of the 30-work-station model to the results for the reduced model, the agreement in total flow times and in utilizations of the 10 work stations in common was within 0.5%, yet the simulation run time was reduced from 30 minutes to 6 minutes. See [4] for more details.

Considering the 59 work stations from the original Micron data set that were omitted in this research, putting fixed lags back in the simulation to account for them would increase simulation run time only very slightly, as hardly any events would be added to the discrete-event calculation. Thus a single simulation run of operation of the Micron fab over a 12-month period at a level of detail adequate to compute flow times requires about 6 minutes.

The computer system used in this research is a SUN SPARC IPC engineering work station. It takes about 20 minutes to formulate and solve the 10-product, 12-month LP formulation, and, as discussed above, it takes about 6 minutes to run one (reduced) simulation. If we use the simulation model with random machine failures, it takes about 1.6 hours to complete 16 simulation runs required for one iteration. Adding the CPU time to formulate and solve the LP, it will take approximately 2 hours for one iteration. In the tests that we have conducted, it takes no more than 7 iterations to get acceptable flow time agreement, say, within 5% for realistic cases of machine variability. The number of required iterations should go down if reasonably good initial estimates of flow times are available, and the percentage gap in agreement will go down as fixed lags are added in to account for the operations on low-utilization equipment. Thus it would appear that very accurate production plans for a wafer fab could be generated overnight on an engineering work station.

While the iterative scheme proposed herein is presented using discrete-event simulation to estimate mean flow times, an analytical model able to predict mean flow times with acceptable accuracy could be used in its stead. Such an approach could be attractive if the analytical model offered substantial reductions in computation time compared to simulation.

We have not discussed here the case in which there are

### TABLE VII

**Bottleneck Work Stations versus Period Number (Iteration #7 of Machine Failures Case, Original Failure Rates)**

<table>
<thead>
<tr>
<th>Period</th>
<th>Bottleneck Work Stations (Fully Loaded)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>M9 M22</td>
</tr>
<tr>
<td>2</td>
<td>M9 M22</td>
</tr>
<tr>
<td>3</td>
<td>M9 M20 M22 M28 M28</td>
</tr>
<tr>
<td>4</td>
<td>M9 M20 M22 M28 M28</td>
</tr>
<tr>
<td>5</td>
<td>M9 M20 M22 M28 M28</td>
</tr>
<tr>
<td>6</td>
<td>M9 M20 M22 M28 M28</td>
</tr>
<tr>
<td>7</td>
<td>M9 M20 M22 M28 M28</td>
</tr>
<tr>
<td>8</td>
<td>M9 M20 M22 M28 M28</td>
</tr>
<tr>
<td>9</td>
<td>M9 M20 M22 M28 M28</td>
</tr>
<tr>
<td>10</td>
<td>M9 M20 M22 M28 M28</td>
</tr>
<tr>
<td>11</td>
<td>M9 M20 M22 M28 M28</td>
</tr>
<tr>
<td>12</td>
<td>M9 M20 M22 M28 M28</td>
</tr>
</tbody>
</table>

### TABLE VIII

**Number of Visits to Bottleneck Machines**

<table>
<thead>
<tr>
<th>Product (Route)</th>
<th>M9</th>
<th>M20</th>
<th>M22</th>
<th>M28</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>3</td>
<td>6</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>C</td>
<td>3</td>
<td>6</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>E</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>F</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>G</td>
<td>3</td>
<td>5</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>H</td>
<td>3</td>
<td>6</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>I</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>J</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

### TABLE IX

**Ratio of Average to Theoretical Flow Times (Iteration #7 of Machine Failure Case, Original Failure Rates)**

<table>
<thead>
<tr>
<th>Product (Route)</th>
<th>Theoretical Total Flow Time</th>
<th>Average Total Flow Time</th>
<th>Ratio of Avg. to Theoretical</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>9.17</td>
<td>46.21</td>
<td>5.04</td>
</tr>
<tr>
<td>C</td>
<td>8.18</td>
<td>45.32</td>
<td>5.54</td>
</tr>
<tr>
<td>D</td>
<td>4.22</td>
<td>11.49</td>
<td>2.72</td>
</tr>
<tr>
<td>F</td>
<td>3.36</td>
<td>11.28</td>
<td>3.36</td>
</tr>
<tr>
<td>H</td>
<td>3.61</td>
<td>10.91</td>
<td>3.02</td>
</tr>
<tr>
<td>R</td>
<td>6.96</td>
<td>26.40</td>
<td>3.80</td>
</tr>
<tr>
<td>T</td>
<td>8.43</td>
<td>41.75</td>
<td>4.96</td>
</tr>
<tr>
<td>U</td>
<td>8.29</td>
<td>44.25</td>
<td>5.34</td>
</tr>
<tr>
<td>X</td>
<td>5.57</td>
<td>23.92</td>
<td>4.29</td>
</tr>
<tr>
<td>Z</td>
<td>4.34</td>
<td>14.88</td>
<td>3.43</td>
</tr>
<tr>
<td>Overall Average</td>
<td>6.21</td>
<td>25.77</td>
<td>4.15</td>
</tr>
</tbody>
</table>
alternative machine types with overlapping but distinct capabilities in terms of operations that can be performed by each type. This phenomenon is quite common in the semiconductor industry, where the typical fab may have in use two or more generations of a particular type of processing equipment. Extension of the LP formulation to model alternative machine types is discussed in [8].

We also have not considered the issue of planning safety stocks to cope with uncertainty in die yields. It would be of interest to incorporate some analysis of die yield uncertainties in the iterative scheme studied herein.

An expanded version of the LP planning model described herein has already been used at Harris Corporation—Semiconductor Sector as part of its "IMPReSS" company-wide production planning system. The expanded version is described in [7].

APPENDIX

LINEAR PROGRAMMING FORMULATION FOR CAPACITATED LOADING OF WAFER FABRICATION FACILITIES

Indices:
- $g$ : die type.
- $i$ : wafer fab route.
- $l$ : processing step (i.e., operation) on a wafer fab route. $l_i$ indexes the last step on wafer fab route $i$.
- $k$ : resource type (i.e., machine type).
- $p, q$ : planning period, $p = 1, 2, \ldots, P$. $P$ is the planning horizon. An extra period $P + 1$ is appended to the planning horizon whose length is equal to the flow time of the longest fab route.
- $r$ : demand class, $r = 1, 2, \ldots, R$.
- $G^r$ : set of all die types appearing in $r$-th demand class.
- $I^r$ : set of all wafer routes producing die types appearing in $G^r$.
- $K^r$ : set of all resource types loaded by routes in $I^r$.

Parameters:
- $a_{ikp}$ : average machine-hours of machine type $k$ used in operation $l$ of wafer fab route $i$, per wafer processed in period $p$.
- $c_{ik}$ : discounted incremental cost per wafer start on wafer fab route $i$.
- $\bar{c}_{kp}$ : hours of machine type $k$ available for processing activity in period $p$.
- $e_{kp}$ : average wafer yield of route $i$ in period $p$, i.e., the expected number wafer outs per wafer released.
- $d^r_{gp}$ : demands for die type $g$ in period $p$ for $r$-th demand class. $d^r_{g, P+1}$ is the demand rate (demand during period divided by the number of calendar days in period) of the last period times the length of the extra period. Demands for class $r$ include all demands in classes 1, 2, $\ldots$, $r - 1$ as well as demands in priority class $r$. All demands are net of projected output from work-in-process (WIP) and initial die inventory. $D^r_{gp}$ denotes the cumulative demand for class $r$ up to the end of period $p$.
- $b^r_{gp}$ : backorder cost of die type $g$ in period $p$ for $r$-th demand class.

$h_{gp}^r$ : inventory holding cost for die type $g$ in period $p$ for $r$-th demand class.
- $v_{gp}^r$ : discounted average revenue per die type $g$ in period $p$ for $r$-th demand class.
- $u_{gp}$ : quantity of good die type $g$ out per wafer out from final wafer fab route $i$ in period $p$.
- $f_{gp}$ : first time period possible to obtain output of die type $g$ from new wafer release, considering the flow times of routes that produce die type $g$.
- $z_p$ : first frozen period of wafer fabrication route $i$. The production rates in all periods after this period will be set equal to the rate in this period in order to satisfy the steady-state horizon condition.
- $s_p$ : earliest period number (nonpositive), in which current WIP would have started considering the assumed flow times.
- $w_p$ : number of working days in period $p$ at the plant associated with route $i$. ($w_i, P+1 = \text{longest flow time of any wafer fab route}$.)
- $X_p^i$ : equivalent wafer releases generating current WIP status of route $i$ considering the assumed flow times and wafer yields, defined in periods before the start of the planning horizon, $p = 0, -1, -2, \ldots, -s_p$.
- $\hat{B}_{gp}$ : upper bound on backorders of demands for die type $g$ in period $p$ in the LP for the $r$-th demand class LP, defined for $r = 2, \ldots, R$, as $\hat{B}_{gp} = B_{gp}^{r-1} + D_{gp}^r - D_{gp}^r$, where $B_{gp}^{r-1}$ is the value of the backorder variable in the solution of the LP for the $(r - 1)$-th demand class.

Variables:
- $X_p^i$ : release variable, i.e., the number of wafers to be released on wafer fab route $i$ in period $p, i \in I^r, p = 1, \ldots, z_p$.
- $I_{gp}$ : inventory level of die type $g$ at the end of period $t$, expressed in terms of the number of die, $g \in G^r$, $p = 0, -1, -2, \ldots, -s_p$.
- $B_{gp}$ : backorder level of die type $g$ at the end of period $p$, expressed in terms of the number of die, $g \in G^r$, $p = f_{gp}, \ldots, P + 1$.

1) Shorthand Notation for Wafer Releases in Frozen Periods:
- $X_p^i$ : wafer output from wafer fab route $i$ in period $p$. If $p > z_p$, then $p$ is one of the frozen periods during which the release rates are held constant. The release rate in this case is defined by $X_p^i$, i.e.,

$$X_p^i = \frac{w_p}{w_i, z_p} X_{z_p}^i;$$

otherwise,

$$X_p^i = X_p^i.$$

Shorthand Notations for Linear Combinations of Start Variables Obtained Using the Epoch-Dependent Flow Time Parameters (see Section III):
- $Y_p^i$ : wafer output from wafer fab route $i$ in period $p$ expressed as a linear combination of the $X_{ip}$ variables.
\[ \hat{Y}_{p,i} = \text{quantity of wafers arriving at step } l \text{ of wafer fab route } i \text{ in period } p, \text{ expressed as a linear combination of the } X_{p,i} \text{ variables.} \]

Let \( T_p^i \) = total flow time applicable at the end of period \( p \) of route \( i \).

In Case 1 as shown in Fig. 4(a),

\[
\hat{Y}_{ip} = \frac{\left( T_{p-1}^i - F_{p-1}^i \right) - \left( T_{p-1}^i - T_{p-1}^i \right)}{\left( T_{p-1}^i - T_{p-1}^i \right)} e_p \hat{y}_{p,i}.
\]

In Case 2 as shown in Fig. 4(b),

\[
\hat{Y}_{ip} = \frac{\left( T_{p-1}^i - F_{p-1}^i \right) - \left( T_{p-1}^i - T_{p-1}^i \right)}{\left( T_{p-1}^i - T_{p-1}^i \right)} e_p \hat{y}_{p,i} + \sum_{q=p-1}^{p} c_q^p \hat{X}_q^i + \frac{\left( T_{p-1}^i - F_{p-1}^i \right) - \left( T_{p-1}^i - T_{p-1}^i \right)}{\left( T_{p-1}^i - T_{p-1}^i \right)} e_p \hat{X}_q^i.
\]

Letting \( F_{p,i} \) denote the flow time applicable at the end of period \( p \) up to operation \( l \) of route \( i \), and letting \( e_p \) denote the wafer yield up to step \( l \) effective in period \( p \), we can calculate \( \hat{Y}_{p,i} \) using the same formulas above but replacing \( F_{p,i} \) with \( F_{p,i} \), and \( e_p \) by \( e_p \).

Note that the expressions for \( \hat{Y}_{p,i} \) when \( p^- < 1 \) are terms which involve the releases equivalent to the initial work-in-process (WIP) status. In this way, loads on resources from processing WIP are accounted for in the planning model.

2) LP Formulation (for \( r \)-th Demand Class): Maximize

\[
\sum_{g \in G^r} \sum_{i \in i} \sum_{p=1}^{P+1} u_{ip}^r u_{ip} \hat{Y}_{ip} - \sum_{i \in I^r} \sum_{p=1}^{P} c_{ip} \hat{X}_p
\]

\[
- \sum_{g \in G^r} h_{ip}^r l_{ip} - \sum_{g \in G^r} b_{ip}^r B_{ip}.
\]

Note: \( i \in g \) means wafer fab route \( i \) produces die type \( g \).

Constrains:

1) Resource Capacity:

\[
\sum_{i \in I^r} \sum_{l=1}^{L} a_{ilp} \hat{Y}_{p,i} \leq \bar{c}_{kp}, \quad p = 1, \ldots, P + 1, \quad \text{for all } k \in K^r.
\]

2) Die Demands:

\[
\sum_{i \in g} u_{ip} \hat{Y}_{ip} - I_{gp} + B_{gp} = \sum_{q=1}^{P} d_{gq}^p, \quad g \in G^r, \quad p = fp_g,
\]

\[
\sum_{i \in g} u_{ip} \hat{Y}_{ip} + I_{gp} - B_{gp} - I_{gp} + B_{gp} = d_{gq}^p, \quad g \in G^r, \quad p = fp_g + 1, \ldots, P - 1.
\]

\[
\sum_{i \in g} u_{ip} \hat{Y}_{ip} - B_{gp} + B_{gp} = d_{gq}^p, \quad g \in G^r, \quad p = P, P + 1.
\]

Note: \( i \in g \) means wafer fab route \( i \) produces die type \( g \).

The authors are grateful to Micron Technology, Inc. for supplying us with an excellent test data set.

REFERENCES


[9] Yi-Feng Hung was born in Tainan, Taiwan in 1961. He received the B.S. degree in industrial engineering from National Tsing Hua University, Hsinchu, Taiwan, in 1984. He received the M.S. degree in industrial engineering and the Ph.D. in industrial engineering and operations research from the University of California, Berkeley, in 1989 and 1991, respectively. He is an Associate Professor of Industrial Engineering at the National Tsing Hua University, Hsinchu, Taiwan. His research interests include production management, simulation applications and factory database design for semiconductor manufacturing. Dr. Hung currently works closely with the semiconductor companies in Science-based Industrial Park, Hsinchu, Taiwan on issues of manufacturing management.

Robert C. Leachman, for a photograph and biography, see this issue, p. 169.