HANDBOOK OF APPLIED OPTIMIZATION

Handbook of
Applied Optimization

EDITED BY

Panos M. Pardalos and Mauricio G. C. Resende

OXFORD
UNIVERSITY PRESS
2002

OXFORD
UNIVERSITY PRESS

Oxford New York
Athens Auckland Bangkok Bogotá Buenos Aires Calcutta
Cape Town Chennai Dar es Salaam Delhi Florence Hong Kong Istanbul
Karachi Kuala Lumpur Madrid Melbourne Mexico City Mumbai
Nairobi Paris São Paulo Shanghai Singapore Taipei Tokyo Toronto Warsaw
and associated companies in
Berlin Ibadan

Published by Oxford University Press, Inc.
198 Madison Avenue, New York, New York 10016
www.oup.com

Oxford is a registered trademark of Oxford University Press

Library of Congress Cataloging-in-Publication Data

Handbook of applied optimization / edited by Panos Pardalos and Mauricio Resende.
p. cm.
Includes bibliographical references and index.
ISBN 0-19-512594-0 (hardcover : alk. paper)
1. Mathematical optimization. 2. Programming (Mathematics)—Industrial applications.
I. Pardalos, P. M. (Panos M.), 1954- II. Resende, Mauricio G. C., 1955-

QA402.5 .H334 2001
519.3–dc21 00-068146

EDITORIAL AND PRODUCTION STAFF
Acquiring Editor: Sean Pidgeon
Development Editor: Sean Pidgeon
Assistant Project Editor: Merilee Johnson
Manufacturing Controller: Genieve Shaw
Book Designer: Joan Greenfield
Managing Editor: Matthew Giarratano
Publisher: Karen Day

1 3 5 7 9 8 6 4 2

Printed in the United States of America
on acid-free paper

To our parents,
Kalypso and Miltiades,
Renalva and Roberto (in memorium)

CONTENTS

PART ONE: ALGORITHMS

PART TWO: APPLICATIONS

PART THREE: SOFTWARE

Vollman, T. E., W. L. Berry, and D. C. Whybark. 1992. *Manufacturing Planning and Control Systems*. 3d ed. Burr Ridge, IL: Richard D. Irwin Inc.

Wagner, H. M. and T. Whitin. 1958. "Dynamic version of the economic lot size model." *Mgmnt Sci.* **5(1)**: 89–96.

18.6 Semiconductor Production Planning

Robert C. Leachman

ABSTRACT The planning of semiconductor manufacturing presents formidable challenges for the successful application of mathematical optimization. Formulation techniques are described that precisely model noninteger flow times, time-phased capacity consumption by reentrant process flows, proper horizon behavior, multiple priority classes of demands and multiple objectives, binning and substitution in product structures, and arbitrary distributions of initial work in process. We also formulate exact capacity analyses of nonhomogeneous resources, machine arrangement constraints, and alternative combinations of resources necessary to execute manufacturing operations. Brief histories are provided of key modeling contributions and of optimization applications to production planning in the industry.

1. Introduction

Semiconductors form the basis of electronics, the world's largest industry except for agriculture. The manufacture of integrated circuits is one of the most capital-intensive and complex manufacturing processes extant. A state-of-the-art wafer fabrication facility processing twentyfive thousand twelve-inch silicon wafers per month, each printed with integrated circuits of electrical switches as small as 180 nanometers in length, represents an investment of several billion U.S. dollars. Most of the processing equipment installed in this factory will need to be replaced within a single-digit number of years in order to keep up with the advances in process technology.

While most other capital-intensive process industries have long been fruitful areas for the application of mathematical optimization to production planning, regular application in semiconductor production planning was unknown until the second half of the 1980s, and did not become widespread until the second half of the 1990s. This lag compared to other process industries is due to the extraordinary challenges posed by semiconductor manufacturing to the formulation of a computationally tractable yet sufficiently accurate mathematical programming model.

The semiconductor manufacturing process consists of hundreds of process steps performed by tens of machine types on discrete manufacturing lots. The process flow is re-entrant in the sense that a given type of processing machine performs a number of steps in the process flow, interspersed with process steps performed by other machine types. Process flow times between different steps performed by the same machine type range from as little as hours to as long as many weeks. The overall flow time through a wafer fabrication process ranges from twenty to sixty days; the subsequent wafer testing, device assembly, burn-in, final testing, and packing processes add another five to fifteen days of flow time. Complexities for planning arising from the manufacturing process include steep learning curves on process yields, process times, flow times, and equipment efficiencies; uncertainties in yields, flow times, and capacities; nonhomogeneous machines and tooling; and the integrality of production lots.

Another aspect of complexity arises from the semiconductor product structure. Testing of production lots generates a distribution of quality-graded products known as bin splits. Products that are sold and bins of quality have a many-to-many relationship. There is an embedded optimization problem concerning how many (pretested) products should be fabricated and assembled, and how the consequent quality bins should be allocated, in order to best service demands. There are also complexities arising from the nature of the demands: there may be significant uncertainties in market demands; and there may be a great variety of demand types for each finished good, ranging from firm orders to flexible customer contracts to reliable forecasts to risky forecasts.

The sheer size of the planning problem in most semiconductor manufacturing companies is daunting: even a single wafer fabrication plant performs hundreds of process steps on hundreds of wafer types using dozens of machine and tool types. A typical company has thousands or even tens of thousands of finished goods types manufactured in dozens of manufacturing areas.

Planning in the industry is performed both incrementally and in a regenerative fashion. Incremental planning involves adding production to an existing plan in order to meet new demands, without other adjustment of the existing plan; regenerative planning involves a complete reassessment of the plan in light of revised demands or other changes in the input data. Incremental planning is of course a much smaller problem with a much shorter turnaround time, but regenerative planning offers the ability to more fully optimize production. We shall treat the more difficult regenerative planning problem in this article.

We remark that production planning is not the only potential application of optimization within the domain of semiconductor manufacturing management. There have also been efforts to apply it to establish short-term goals for factory floor operations (e.g., Fordyce et al., 1992; Kang, 1996), and, more recently, to apply it to the problem of capacity planning (e.g., Liu, 1998; Stray et al., 2000). However, we shall restrict our attention here to the production planning problem, in which production is scheduled subject to process flow times and machine capacities that are prespecified.

2. Formulation of Semiconductor Planning Models

We now discuss the key techniques for formulating a practical linear programming model for planning semiconductor production.

2.1. A Brief History of Key Contributions

Pioneering work on the formulation of practical optimization models for semiconductor production planning was carried out beginning in 1985 at the University of California at Berkeley, under the direction of Prof. R. C. Leachman. This effort continues to the present day. The Berkeley researchers have helped various semiconductor manufacturers sponsoring their research to implement software incorporating their optimization formulations, and these implementations have sometimes been referred to as the Berkeley Planning System (BPS). Working independently, researchers at IBM also made useful contributions beginning in the early 1990s, particularly to the capacity analysis of nonhomogeneous machines. Software incorporating these contributions has been utilized in IBM's microelectronics manufacturing management.

Key contributions are summarized as follows:

1. Variables of the planning model represent constant production rates in given planning periods, expressed either as rates of release into process flows or rates of output from process flows (Leachman, 1986, 1993).
2. Noninteger time lags are used to map, through continuous working time, the release rates into process output rates; inventory balance constraints are formulated at noninteger time points in order to insure mass conservation through continuous time (Hackman and Leachman, 1989).
3. Noninteger time lags also are used to map release rates to workload rates for process steps, which are summarized into total workload rates on capacitated resources (Leachman, 1986, 1993).

4. A steady-state horizon condition is formulated to mitigate undesirable horizon effects induced by the presence of reentrant flows in a finite horizon model (Leachman, 1993).
5. In the case of time-varying bottlenecks, an iterative simulation-optimization scheme has been developed to jointly establish flow times and production quantities (Hung and Leachman, 1996).
6. Binning and substitution at product structure junctions are exactly formulated (Leachman, 1986, 1993).
7. Iterative optimization of multiple demand priority classes is used to treat varying types of demands and market uncertainties. Multiobjective programming is used, with bounds on variables placed to ensure there is no switching between alternative optimal solutions (Leachman, 1993).
8. Capacity analysis of nonhomogeneous machines, of machine assignment constraints, and of alternative sets of machines and tools has been progressively refined (Leachman and Carmon, 1992; Fordyce and Sullivan, 1995; Lin, 1999).
9. Scheduling of the initial work in process is explicitly incorporated into the planning model while maintaining computational tractability (Lin, 1999).

2.2. *Preliminary Model*

We first consider an unrealistic, textbook-type LP formulation of semiconductor production planning. Subsequently, we introduce modifications that make the model more suitable for industrial application. For simplicity of exposition, we first consider a single-stage system, that is, demand items and items released to production are identical, hereafter simply referred to as products.

Wafer fabrication areas can often be modeled as a single-stage system with re-entrant flows. The process flows may include upward of six hundred process steps, but the product identity at the first and last step in each flow may be identical. When there are many products to be processed, it is impractical to formulate a planning model incorporating scheduling variables and balance constraints for each step. Instead, we first formulate a model including only variables representing the release of raw material for processing each product. Flow time parameters are employed to estimate subsequent workloads associated with the release of raw material at each downstream process step. The model includes capacity constraints for multiple resource types and inventory balance constraints on completed products.

Subscripts

t = time period $(t-1, t]$, $t = 1, 2, 3, ..., T$, where T is the planning horizon.

i = product, $i = 1, 2, 3, ..., N_I$.

j = processing step, $j = 1, 2, 3, ..., N_i$.

k = machine type or other scarce resource, $k = 1, 2, 3, ..., K$.

Parameters

d_{it} = demand for product i in period t.

a_{kij} = processing time per unit on machine type k when performing step j on product i.

L_i = estimated flow time from release of raw material until completion of product i.

L_{ij} = estimated flow time from release of raw material until step j is initiated on product i.

b_i = estimated yield from release of raw material until completion of product i.

b_{ij} = estimated yield from release of raw material until step j is initiated on product i.

c_{kt} = capacity (machine-hours) of machine type k in period t.

h_i = holding cost per unit per period of completed product i.

s_i = backorder cost per unit per period of completed product i.

r_i = avoidable cost (e.g., raw materials) per unit of product i released into production.

Variables

$x_{i,t}$ = quantity of raw material for the manufacture of product i to be released during period t. $x_{i,0}, x_{i,-1}, x_{i,-2}, ..., x_{i,-Li+1}$ must be given.

$I_{i,t}$ = inventory of completed product i at time t. $I_{i,0}$ must be given.

$BO_{i,t}$ = shortage of completed product i at time t. $BO_{i,0}$ must be given.

Formulation

$$\text{Minimize } \sum_{i=1}^{N}\sum_{t=1}^{T}(r_i x_{i,t} + h_i I_{i,t} + s_i BO_{i,t})$$

Subject to

$$\sum_{\tau=1}^{t} b_i x_{i,\tau-L_i} - I_{i,t} + BO_{i,t} = \sum_{\tau=1}^{t} d_{i,\tau} - I_{i,0} + BO_{i,0}, \quad i = 1, ..., N_I, \quad t = 1, ..., T.$$

$$\sum_{i=1}^{N_I}\sum_{j=1}^{N_i} a_{kij} b_{ij} x_{i,t-L_{ij}} \leqslant c_{kt} \quad k = 1, ..., K, \quad t = 1, ..., T.$$

$$x_{it} \geqslant 0, \; I_{it} \geqslant 0, \; BO_{it} \geqslant 0, \quad i = 1, ..., N_I, \quad t = 1, ..., T.$$

We have expressed a classical production planning formulation, allowing inventory and backorders with respect to given demands, and making production subject to multiple resource capacity constraints. Integer time lags between raw material release and resource consumption, and between raw material release and availability of finished product are incorporated.

It is straightforward to expand the formulation to model multistage systems, that is, to accommodate a product structure using inventory balance constraints that express mass conservation of intermediate products. Alternative process flows for the products also are easily incorporated.

2.3. Needed Modifications

We now point out weaknesses in the classical formulation that cause it to generate inaccurate or undesirable production plans for semiconductor manufacturing, and then introduce modifications to overcome these weaknesses.

ACCOMMODATING REAL-VALUED TIME LAGS

The first undesirable characteristic of the formulation that we shall deal with is the integrality of the time lags. The technique for formulation of an exact LP model incorporating noninteger time lags, first reported by Hackman and Leachman (1989), is summarized as follows. Let us view the release of raw material to manufacture product i as a continuous time function $x_i(t)$. Assuming the rate of release is uniform within each time period, we can relate the LP variables to this time function by observing that the value of $x_i(t)$ over the interval $(t-1, t]$ is x_{it}. That is, $x_i(t)$ is restricted to be a step function by the LP formulation. Similarly, we may view the demand for product i as a continuous time function $d_i(t)$, again assuming uniform rates within each given period. Using these continuous functions, we may rewrite the inventory balance constraint for product i as

$$\int_{-L_i}^{t-L_i} x_i(\tau)d\tau - I_i(t) + BO_i(t) = \int_{0}^{t} d_i(\tau)d\tau - I_i(0) + BO_i(0), \quad \text{all } t \geqslant 0, \quad i = 1, ..., N_I. \tag{1}$$

Here, $I_i(t)$ denotes the inventory level of product i at an arbitrary point t in continuous time, and $BO_i(t)$ denotes the shortage of product i at time t.

These integrals reduce to linear combinations of the x_{it} variables and the d_{it} parameters when the constraint is evaluated at any arbitrary point t, even when t is noninteger. For example, suppose $L_i = 0.7$ and $t = 2.5$. Then the first integral of (1) reduces to

$$0.7x_{i,0} + x_{i,1} + 0.8x_{i,2}.$$

Thus, constraints suitable for the application of linear programming are obtained even when the L_i parameters are fractional. It remains to determine at what values of t do the constraints (1) need to be enforced in order to insure inventory balance throughout continuous time. Hackman and Leachman (1989) show that if the constraints are enforced at all points at which either integrand may change rates, then inventory balance is assured throughout continuous time. In the case of (1), these are precisely the positive integers (i.e., the epochs when the rate of outflow from inventory, $d_i(t)$, may change), and the points of the form $(t - L_i)$ where t is a positive integer greater than or equal to L_i (i.e., the epochs when the rate of inflow to inventory may change). Note that the implementation of the constraints at the latter kind of points induces the formulation of inventory variables $I_i(t - L_i)$ in addition to the usual inventory variables $I_i(t)$ defined for the integer points.

Using the continuous time functions, we may also rewrite the capacity constraint for machine type k as

$$\int_{t-1}^{t} \sum_{i=1}^{N_I} \sum_{j=1}^{N_i} a_{kij} x_i(\tau - L_{ij}) d\tau \leqslant c_{kt}, \quad k = 1, ..., K, \quad t = 1, ..., T. \tag{2}$$

This constraint also reduces to a linear constraint in the $x_{i,t}$ variables when evaluated at a given value of t, even if the L_{ij} parameters are noninteger.

Given a large number of steps performed on a given machine type for each product, and given a large number of products, there are potentially many time points per period where the integrand may change rate. Strictly speaking, each interval of constant capacity consumption needs to be individually constrained by the resource capacity in order to ensure resource conservation throughout continuous time. While it is possible to formulate linear constraints for every interval between epochs at which the integrand may change rate, it may not be practical to do so. Moreover, as a practical matter, one cannot expect the epochs t at which the input rate $x_i(t)$ is adjusted to precisely map to epochs $t + L_{ij}$ at which the workload at step j, $x_i(t + L_{ij})$, changes rate, given the uncertainties in the production process and the level of detail at which the model is formulated.

Fortunately, the wide variety of values for the L_{ij} parameters in a practical case tends to result in level loads in each given planning period, especially when capacity is fully utilized and when flows are very re-entrant. As a practical approximation, it is suggested to simply enforce constraints (2) only for the intervals defined by the given planning periods $(t-1, t], \; t = 1, ..., T$. Using this approach, Hung and Leachman (1996) demonstrate very close agreement between machine utilizations and product flow times predicted by such an LP planning model and utilizations and flow times predicted by detailed, discrete-event simulation on actual industrial data.

Accommodating Factory Working Calendars and Variable-Length Planning Periods

Time lags as above are measured in terms of working time. The lengths of planning periods (in terms of working time) in the planning model may need to be different from unity. This is because of working holidays occurring in the planning periods, and/or because of a desire for increased granularity in early periods of the plan, for example, weekly periods over the first several months of the planning horizon, and monthly periods after that. A generalization of the approach in the previous section is required. Basically, the limits of integration appearing in equations (1) and (2) must be properly established in terms of epochs of continuous working time measured from the start of the planning horizon. See Hackman and Leachman (1989) for details. Moreover, shortage and holding cost parameters also may need to vary by period, appropriately scaled to reflect the lengths of the periods (Hung, 1991; Lin, 1999).

CONTROLLING HORIZON EFFECTS

In the basic finite-horizon LP formulation above, optimal values of the variables for the release rates of raw material into capacitated reentrant process flows exhibit a peculiar, undesirable behavior. Variables for product release in periods within one flow time of the horizon will be set equal to zero, since they do not contribute to demands included in the formulation. Variables for the release rates in periods just before this will typically have surprisingly large values. Such unreasonably large values are feasible since they do not have to compete for capacity with releases in subsequent periods. The tacit assumption that production is permanently terminated at the horizon T is the root cause of this undesirable behavior.

To overcome this problem, Leachman (1993) incorporates into the formulation a steady-state horizon condition. For each product i, the interval $[T - L_i, T]$ is termed the frozen interval for product i. Releases of product i in all time periods that intersect the frozen interval are constrained to be equal, that is, a single variable is utilized to represent releases in each of these periods (multiplied by an appropriate scaling constant if there are differences in period lengths). For the purposes of computing inventory balance in constraint (1) and resource capacity consumption in constraint (2), an additional time period is appended on to the planning horizon, with length equal to L_i working days. Demand for product i in this period is set to be the same rate as the average rate in the frozen period. Inventory and backorders of completed product i are measured in the objective function at both the start and the end of the extra period.

Formulations including this steady-state condition generate optimal plans that exhibit smooth production.

INCORPORATING MARKETING CONSIDERATIONS

The total demand for each product is an aggregate of components with varying priorities. One component corresponds to firm customer orders with promised delivery dates. Another may correspond to replenishment of inventory stocks supporting long-term customer contracts. Even within the component of product demand corresponding to market forecasts unrealized by on-hand customer orders or contracts, priorities are varying. One portion of the forecast may be fairly certain to be realized; the final portion may be more risky, that is, there is a significant probability that not that much demand will materialize.

Treating the entire demand for products with uniform, linear tasks for backorders and inventory is an oversimplification of the planning problem, leading to optimal solutions that may be undesirable to marketing management. That is, in a capacity-constrained situation, an optimal solution may deny capacity to confirmed customer orders for one product and award the capacity to risky forecasts for another.

Leachman (1993) proposes a structure of demand classes and iterative optimization calculations to cope with this situation. In this structure, the overall market demands for each product are categorized into R priority classes, with the understanding that demands in class $n+1$ must not inhibit the servicing of demands in class n as on-time as is feasible. Linear programs are formulated and solved for each class $r = 1, 2, ..., R$. Let $d_i^r(t)$ denote the rate at time t of demands in class r for product i. The demand rate used in Equation (1), the inventory balance constraint for the rth formulation, is

$$\sum_{s=1}^{r} d_i^s(t),$$

that is, demands are cumulative over classes. To this formulation one appends an upper bound on the back-order variables, specifically,

$$BO_{i,t} \leqslant BO_{i,t}^{r-1} + \int_0^t \sum_{s=1}^{r} d_i^s(\tau) d\tau - \int_0^t \sum_{s=1}^{r-1} d_i^s(\tau) d\tau, \quad \text{all } i, \quad \text{all } t,$$

where

$$BO_{i,t}^{r-1}$$

denotes the optimal value of the back orders of product i at time t from the $r-1$th formulation. This bound ensures that the maximum allowed back orders for the demand classes 1 through r equals the optimal back orders in classes 1 through $r-1$ plus the increment in cumulative demand between classes $r-1$ and r.

Note that the only difference between formulations for class $r-1$ and class r are the right-hand-side demand parameters, the bounds on the back-order variables, and possibly some objective function coefficients, if relative product priorities vary from class to class. Also note that the optimal solution to class $r-1$ is a feasible solution to class r, provided the values of the backorder variables are adjusted to reflect the increment in cumulative demands. Thus solving the series of R linear programs is not too much more computationally difficult than solving a single linear program. Leachman and Raar (1994) report that in industrial practice with very large planning problems (in excess of 100,000 rows), the time to solve four demand classes is about 2 to 2.5 times the time required to solve the first class.

Another marketing consideration concerns a policy choice between build-to-order and build-to-forecast. Where forecast accuracy is high and raw material costs low, it may be desirable to build to forecast in order to compress lead times offered to customers. But for products with significant raw material costs and unreliable demand forecasts, it may be more prudent to delay production until receipt of a customer order. In the case of semiconductors, back end production (assembly and test) is typically built to order, except for high-volume standard or commodity products. On the other hand, wafers are usually fabricated to forecast.

We may handle this consideration as follows. For each product in the overall product structure we suppose an input parameter is defined indicating whether the product is build-to-order or build-to-forecast. (In a rational case, if a product is declared build-to-order, then all child products are also build-to-order.) If the parameter indicates the product is build-to-order, then all production in the first planning period must be for order-based demands only. Planned production in subsequent periods can be for both orders and forecasts, assuming the plan is recalculated on a rolling horizon basis.

Suppose class $r-1$ is the last demand class corresponding to protection of customer commitments, and class r is the first class corresponding to forecasted demand. When the parameter indicates the product is build-to-order, upper bounds on the release variables x_{it} are introduced into the formulations for classes $r, r+1, \ldots, R$, as follows:

$$x_{i,t} \leqslant x_{i,t}^{r-1},$$

where

$$x_{i,t}^{r-1}$$

is the optimal value of the release variable for the formulation of class $r-1$.

In summary, marketing considerations can be conveniently incorporated into the planning model through the use of upper bounds on production and back-order variables in multiple optimization calculations over prioritized demand classes. This approach does not expand the formulation size, yet it enables the generation of much more desirable production plans.

Handling Product Structures Featuring Binning and Substitution

Product structures in most manufacturing environments consist almost exclusively of one-to-many relationships, that is, end-item demands may be uniquely sourced to intermediate products and raw materials. These environments are suitable for the application of materials requirements planning (MRP) logic to translate end-item demands into demands for intermediate products and raw materials. Semiconductor product structures may feature a considerable number of many-to-many relationships. These many-to-many relationships arise because of frequent design revisions and because of quality grading of tested products. In the semiconductor environment, optimization is required to correctly translate demands back through the product structure.

Design revisions lead to many-to-many relationships as follows. Suppose a new revision is offered of an original product. The revision offers higher performance, and it corrects some shortcomings in the design of the original product. Depending on the customer application, the revision may require some customers to revise their system in order to utilize the revision, while others may utilize the original product or the revision indiscriminately. New customers may choose to design their systems to solely utilize the revision. The semiconductor manufacturer then faces three kinds of demands: customers who insist on purchasing only the original product, customers who insist on purchasing only the revision, and customers willing to accept either original product or the revision. The last type of demand cannot be uniquely sourced.

Many semiconductor products undergo binning, whereby a distribution of quality-graded products emerges from testing a single manufacturing lot of source product. Each bin defines a specific range or performance for one or several electrical attributes of performance, for example, a bin definition might be of the form 'speed between 200 and 300 MHz and power consumption less than 100 mA'. The fractions of the source product lot falling into each bin of quality are known as bin splits. The bin splits are characteristic of the manufacturing process and must be regarded for planning purposes as prespecified (but probabilistic).

Each finished goods type has specific requirements for electrical performance. A particular bin is generally suitable for a number of different finished goods types. Conversely, there may be a number of bins whose attributes fulfill the requirements of a particular finished goods type. When a bin of higher quality is used to satisfy demands, this is known as substitution.

It is often the case that there are alternative testing procedures generating different bin distributions. For example, the same basic packaged device can be put through an elaborate testing procedure that generates a large set of bins, or through a simpler testing procedure that generates only lower-grade bins. The simpler testing procedure offers a reduction in test capacity consumption that could be attractive, depending on the mix of demands.

We now modify the formulation of Section 2.2 to admit a two-stage product structure including constraints modeling binning and substitution, as developed in Leachman (1993). New variables are introduced to allocate bins to releases of the second-stage products. To admit the case of testing and processing alternatives, product shall be taken to mean a particular device passed through a particular processing and testing alternative.

Subscripts

t = time period $(t-1, t]$, $t = 1, 2, 3, ..., T$, where T is the planning horizon.

i = first-stage product, $i = 1, 2, 3, ..., N_I$.

m = second-stage product, $m = 1, 2, 3, ..., N_M$.

n = quality bin, $n = 1, 2, 3, ..., N_B$. $m \, \varepsilon \, n$ denotes that demands for product m may be sourced using bin n. $n \, \varepsilon \, m$ denotes that bin n is suitable for meeting demands for product m.

j = processing step, $j = 1, 2, 3, ..., N_i$ for first-stage process flows, and $j = 1, 2, 3, ..., N_m$ for second-stage flows.

k = machine type or other scarce resource, $k = 1, 2, 3, ..., K$.

Parameters

d_{mt} = demand for product m in period t.

a_{kij} = processing time per unit on machine type k when performing step j on product i. (a_{kmj} is similar.)

f_{ni} = bin split to bin n from completed product i.

L_i = estimated lead time from release of raw material until binning of product i. (L_m is similar.)

L_{ij} = estimated lead time from release of raw material until step j is initiated on product i. (L_{mj} is similar.)

b_i = estimated yield from release of raw material until binning of product i. (b_m is similar.)

b_{ij} = estimated yield from release of raw material until step j is initiated on product i. (b_{mj} is similar.)

c_{kt} = capacity (machine-hours) of machine type k in period t.

h_m = holding cost per unit per period of completed product m.

s_m = shortage cost per unit per period of completed product m.

r_i = avoidable cost (e.g., raw materials) per unit of product i released into production. (r_m is similar.)

Variables

$x_{i,t}$ = quantity of raw material for the manufacture of product i to be released during period t. $x_{i,0}, x_{i,-1}, x_{i,-2}, ..., x_{i,Li+1}$ must be given.

$y_{n,m,t}$ = quantity of bin n allocated to release of product m in period t.

$z_{m,t}$ = quantity of raw material for the manufacture of product m to be released during period t. $z_{m,0}, z_{m,-1}, z_{m,-2}, ..., z_{m,-Lm+1}$ must be given.

$I_{n,t}$ = inventory of bin n at time t. $I_{n,0}$ must be given.

$I_{m,t}$ = inventory of completed product m at time t. $I_{m,0}$ must be given.

$BO_{m,t}$ = shortage of completed product m at time t. $BO_{m,0}$ must be given.

Formulation

$$\text{Minimize} \sum_{i=1}^{N_I} \sum_{t=1}^{T} r_i x_{i,t} + \sum_{m=1}^{N_M} \sum_{t=1}^{T} (r_m z_{m,t} + h_m I_{m,t} + s_m BO_{m,t})$$

Subject to

$$\sum_{\tau=1}^{t} b_m z_{m,\tau - L_i} - I_{m,t} + BO_{m,t} = \sum_{\tau=1}^{t} d_{m,\tau} - I_{m,0} + B_{m,0}, \quad m = 1, ..., N_M, \quad t = 1, ..., T.$$

$$\sum_{\tau=1}^{t} f_{ni} b_i x_{i,\tau - L_i} + I_{n,0} - \sum_{m \in n} \sum_{\tau=1}^{t} y_{n,m,\tau} = I_{n,t}, \quad n = 1, ..., N_B, \quad t = 1, ..., T.$$

$$\sum_{n \in m} y_{n,m,t} = z_{m,t}, \quad m = 1, ..., N_M, \quad t = 1, ..., T.$$

$$\sum_{i=1}^{N_I} \sum_{j=1}^{N_i} a_{kij} b_{ij} x_{i,t - L_{ij}} + \sum_{m=1}^{N_M} \sum_{j=1}^{N_m} a_{kmj} b_{mj} z_{m,t - L_{mj}} \leqslant c_{kt}, \quad k = 1, ..., K, \quad t = 1, ..., T.$$

$$x_{i,t} \geqslant 0, \; z_{m,t} \geqslant 0, \; y_{n,m,t} \geqslant 0, \; I_{n,t} \geqslant 0, \; I_{m,t} \geqslant 0, \; BO_{m,t} \geqslant 0,$$

$$i = 1, ..., N_I, \quad n = 1, ..., N_B, \quad m = 1, ..., N_m, \quad t = 1, ..., T.$$

The modifications introduced in the sections on pages 4–7 may also be incorporated into a multistage formulation modeling binning and substitution such as the above.

Scheduling Initial WIP

The reentrant nature of semiconductor manufacturing flows means that new releases must compete for capacity with work in process (WIP). The basic formulation of Section 2.2 captures this competition by making use of parameters specifying past production releases $x_{i,0}$, $x_{i,-1}$, $x_{i,-2}$, and so on. However, the formulation implicitly assumes the following:

1. the initial WIP is distributed through the manufacturing process precisely according to the given time lags (i.e., flow times) and the given history of product releases,
2. the initial WIP must keep moving through the process according to the given flow times, and
3. the initial WIP has priority over new releases for allocation of scarce capacity.

In stochastic environments, assumption (1) may be unrealistic. Because of a severe process disruption or because of a downgrading of priority, it is possible that the initial WIP of a product seriously lags the target distribution calculated based on its release dates and the planned flow times. In such a case, the initial WIP will consume more resource capacity to exit the factory than the amount implied by assumption (1), and hence the planning model may overestimate the amount of new releases that may be made without extending flow times.

To circumvent this problem, one could precalculate the capacity required to flush the initial WIP according to the given flow times from each step to product completion. To do this, one can compute a product completion schedule equivalent to the initial WIP (calculated in terms of the given yields and flow times), and then apply a backward version of the capacity consumption formula introduced in Section 2.2. If we let

$$\bar{x}_i(t)$$

denote the output rate at time t calculated from the initial WIP, then the consumption of resource k capacity by initial WIP in the planning period $(t-1,\ t]$ is computed as

$$\sum_{i=1}^{N_I} \sum_{j=1}^{N_i} \int_{t-1}^{t} a_{kij} \bar{x}_i(\tau + L_i - L_{ij}) d\tau.$$

Instead of incorporating capacity consumption estimated from the history of past releases, one could subtract the capacity consumption precalculated as above from the right-hand sides c_{kt} of capacity constraints placed on new releases. This was the approach taken in the Berkeley Planning System (Leachman et al., 1996), which served as the planning engine of the IMPReSS planning system at Harris Corporation.

This precalculation approach overcomes problems arising from assumption (1), but it continues to make assumptions (2) and (3). Because of changed product forecasts, it may be desired in some applications to give higher priority to new product releases of one product over the initial WIP of another.

To handle this consideration requires an expansion of the planning model to schedule not only releases of raw material at the first manufacturing step, but to also schedule releases of initial WIP from manufacturing steps throughout the process. When process flows consist of upward of six hundred process steps, a model formulated with release variables defined for every step is likely to be computationally impractical. A more practical strategy is to formulate WIP release variables only for major process steps, that is, steps utilizing scarce resources included in the capacity analysis. In typical applications in the author's experience, there are in the order of fifty or less major steps per process flow. At the other minor steps, initial WIP may be assumed to flow according to planned time lags up to the next major step, where its release downstream will be scheduled by the model. We shall therefore assume in the following that the index j applies only to major process steps.

This strategy is implemented by introducing new parameters, new variables, and both new and modified constraints into the formulation of Section 2.2 as follows:

Parameters

w_{ijt} = initial WIP arriving at major step j during period t.

$L_{i\alpha j}$ = estimated flow time from initiation of step α to initiation of step j on product i, equal to $L_{ij} - L_{i\alpha}$.

$b_{i\alpha j}$ = estimated yield from initiation of step α to initiation of step j on product i, equal to $b_{ij}/b_{i\alpha}$.

b_{ij}^{d} = estimated yield from initiation of step j on product i to completion of product i, equal to b_i/b_{ij}.

Variables

$x_{i,j,t}$ = planned release of initial WIP at major step j of the process flow for product i during period t

Constraints

New and modified constraints of the formulation are as follows:

$$\sum_{\tau=1}^{t} x_{i,j,\tau} \leqslant \sum_{\tau=1}^{t} w_{ij\tau}, \quad \text{all } i, \ \text{all } j, \ \text{all } t$$

$$\sum_{j=1}^{N_i} \sum_{\tau=L_i - L_{ij}+1}^{t} b_{ij}^{d} x_{i,j,\tau - L+L_{ij_i}} + \sum_{\tau=1}^{t} b_i x_{i,\tau - L_i} - I_{it} + BO_{it} = \sum_{\tau=1}^{t} d_{i,\tau} - I_{i,0} + B_{i,0},$$

$$i = 1, ..., N_I, \quad t = 1, ..., T.$$

$$\sum_{i=1}^{N_I} \sum_{j=1}^{N_i} a_{kij}(b_{ij} x_{i,t - L_{ij}} + x_{i,j,t} + \sum_{\alpha=1}^{j-1} b_{i\alpha j} x_{i,\alpha,t - L_{i\alpha j}}) \leqslant c_{kt}, \quad k = 1, ..., K, \quad t = 1, ..., T.$$

$$x_{i,j,t} \geqslant 0, \ x_{i,t} 0, \ I_{i,t} \geqslant 0, \ BO_{i,t} \geqslant 0, \quad j = 1, ..., N_i, \quad i = 1, ..., N_I, \quad t = 1, ..., T.$$

We remark that, in order to implement the steady-state horizon condition proposed in the section on Controlling Horizon Effects, one cannot allow independent WIP release variables to be formulated for periods intersecting the frozen interval. It is proposed to define WIP release variables only for time periods preceding the frozen interval, that is, all initial WIP that is to be released must be released in time to be realized as completed product before the planning horizon. We also remark that enforcing exact inventory balance of finished goods is likely to be impractical when there are many major process steps with noninteger time lags. A practical approach is to formulate the balance constraints over a sufficiently fine time grid.

Accommodating Machine Arrangement Constraints

Nonhomogeneous sets of machines that perform a basic kind of semiconductor fabrication step such as photolithography exposure are increasingly prevalent in semiconductor factories. To accurately represent capacity relationships in the planning model, one must expand the model. Depending on the nature of the alternatives, there are various formulation strategies that minimize the model complexity necessary to accurately model capacity.

The simplest case is when process times are independent of the resource alternative selected, and the alternative machine types are nested. For example, suppose there are two types of exposure machines, type A and type B. Type B is a newer model that can perform any exposure step; type A is an older model that can perform only noncritical exposure steps. For this type of case, Leachman and Carmon (1992) show that ordinary capacity constraints defined for appropriate groups of machine types constitute an exact capacity model. For this particular example, two capacity constraints per period constitute an exact model. One constraint limits the total workload of critical steps by the available processing time of machine type B, and the other limits the total workload of all exposure steps to the sum of available processing times of both machine types. This approach also provides an exact model when process times of the nested machine types are proportional, since available processing times of alternative machine types may be appropriately scaled according to the process times of one machine type chosen as a standard.

A more difficult case arises when machine usage patterns are not nested. For example, suppose now there are three exposure machine types. Suppose some process steps must be performed on either machine types *A* or *B*, other process steps must be performed on either machines types *B* or *C*, and still others must be performed on machine types *A* or *C*. These more general patterns of allowed allocation of machines arise when engineering effort is expended to qualify machines one by one for critical process steps, and certain machines are found to perform better than others. The restrictions placed on machine allocation are thus an avenue for securing better process control and higher yields, albeit at the potential expense of reduced capacity and longer flow times.

When alternative machine types exhibit this more general pattern of allowed assignments to process steps, Leachman and Carmon (1992) show that the most compact exact model requires introduction into the model of new variables that allocate step workloads to the resource types. Fordyce and Sullivan (1995) formulated a planning model with static demand rates incorporating this approach. Allocation variables can also be incorporated into a dynamic planning model as we shall now illustrate.

We shall term the allowed pattern of assignments of machine types to process steps the machine arrangement table. For the case we consider, machine arrangements at one step are independent of arrangements made at other steps; we term this case one of static machine arrangement constraints. We introduce the following new notation and variables, and we modify the constraints of the formulation of Section on Scheduling Initial WIP as follows.

Notation

$k \epsilon (i, j)$ denotes that machine type k is one that is qualified to perform step j on product i.

$(i, j) \epsilon k$ denotes that step j on product i is one of the steps that machine type k is qualified to perform.

Variables

$A_{i,j,k,t}$ = workload of step j on product i in period t allocated to machine type k.

Constraints

Modified capacity constraints are as follows. Inventory balance constraints are unchanged.

$$b_{ij}x_{i,t-L_{ij}} + x_{i,j,t} + \sum_{\alpha=1}^{j-1} b_{i\alpha j}x_{i,\alpha,t-L_{i\alpha j}} = \sum_{k\in(i,j)} A_{i,j,k,t}, \quad j = 1, ..., N_i, \quad i = 1, ..., N_I, \quad t = 1, ..., T.$$

$$\sum_{(i,j)\in k} a_{kij}A_{i,j,k,t} \leqslant c_{kt}, \quad k = 1, ..., K, \quad t = 1, ..., T.$$

$$A_{i,j,k,t} \geqslant 0, \; x_{i,j,t} \geqslant 0, \; x_{i,t} \geqslant 0, \; I_{i,t} \geqslant 0, \; BO_{i,t} \geqslant 0,$$
$$k = 1, ..., K, \quad j = 1, ..., N_i, \quad i = 1, ..., N_I, \quad t = 1, ..., T.$$

Dynamic Machine Arrangement Constraints

The most difficult case of machine arrangement involves dynamic machine arrangement constraints, whereby the set of qualified process machines at one step is dependent on the machine type assigned at some previous process step. Efforts to achieve process control on the most advanced digital process technologies in the industry sometimes include dynamic machine allocation constraints between critical photolithography exposure steps, or between the lithography exposure step and the following etching step. (In such cases, most of the machine types are individual machines.) To illustrate dynamic arrangement constraints, suppose the qualified machines for the first critical exposure step are machines *A*, *B*, and *C*. If machine *A* is selected, then the qualified machines for the second critical exposure step are machines *A* or *C*. If machine *B* is selected at the first step, then the qualified machines for the second critical exposure step are machines *B* or *D*. Thus the qualified machines for performing the second critical step vary according to which machine was utilized at the first critical step.

To properly model capacity constraints in this case, the allocation of workloads to machines at different steps must be constrained so as to be consistent, if planned flow times are to be observed. To illustrate, we assume the case where one general type of machine, such as the photolithography exposure machines, is subject to dynamic machine arrangement constraints. We shall designate process steps whose machine assignment influences or is influenced by the machine assignment at other steps as critical steps. We assume the case where the machine assignment decision at the first critical step constrains the assignment at the second critical step, the assignment at the second critical step constrains the assignment at the third critical step, and so on.

Lin (1999) shows that this case is most efficiently modeled using routing variables that schedule the release of WIP for movement through particular machine types at downstream critical steps. In this approach, machines utilized at each critical step are categorized into machine groups. A machine group may consist of a single machine or several machines. Machine groups at the first critical step are defined with the property that product processed by any machine in the group has identical machine assignment possibilities at the next downstream critical step. Machine groups at subsequent critical steps are defined with the property that product processed by any machine in the group has identical machine assignment possibilities at both the immediately preceding upstream and the next downstream critical step.

Initial WIP at each process step is categorized into WIP groups. Each WIP group is distinguished based on which machine group was utilized at the preceding critical step. (If there are no preceding critical steps, then all WIP belongs to the same group. Also, if two WIP groups have identical machine group possibilities at the next downstream critical step, they may be consolidated into a single WIP group.) At the next critical step downstream, each WIP group will have one or more machine groups qualified to process that WIP.

In Lin's approach, routing variables are defined that route the initial WIP in each WIP group into all possible machine groups at all downstream critical steps. Constraints insuring consistency of these assignments are incorporated into the formulation.

New notation, parameters, variables, and revised constraints to revise the model of Section on Scheduling Initial WIP to precisely model dynamic arrangement constraints are as follows:

Notation

j = major process step; $j \in C_i$ denotes that major process step j is a critical process step for product i.

j_u = the first critical process step upstream from major process step j.

j_d = the first critical process step downstream from major process step j.

j_i = the first critical process step for product i.

g = machine group; $g \in k$ denotes that machine group g is one that includes machine type k.

a = WIP group; $a \in j$ denotes that WIP group a is one that applies to WIP at major process step j.

$(a,\ j')$ = WIP group a at process step j'; $(g, j)\ \varepsilon\ a$ denotes that machine group g at downstream critical step j is suitable for processing WIP in WIP group a.

Parameters

$w_{i,(a,j),t}$ = initial WIP in the WIP group that is a resident at steps between major steps $j-1$ and j arriving at major step j during period t.

Variables

$x_{i,t}^{(g,j)}$ = quantity of raw material for the manufacture of product i to be released during t and to be processed by maching group g at critical step j.

$x_{i,(a,j),t}^{(g,j')}$ = release of initial WIP from WIP group a at major step j to be processed by machine group g at downstream critical step j'.

$A_{i,k,t}^{(g,j)}$ = workload for machine group g at critical step j on product i allocated to machine type k in period t.

$A_{i,k,t}^{j}$ = workload at noncritical major step j allocated to machine type k in period t.

Constraints

$$x_{i,t} = \sum_{(g,j_i)} x_{i,t}^{(g,j_i)}, \quad t = 1, ..., T, \;\; i = 1, ..., N_I.$$

$$\sum_{(g',j_u)} x_{i,t}^{(g',j_u)} = \sum_{(g,j)} x_{i,t}^{(g,j)}, \quad \forall j \in C_i, \;\; t = 1, ..., T, \;\; i = 1, ..., N_I.$$

$$x_{i,j,t} = \sum_{a \in j} \sum_{(g,j_d) \in a} x_{i,(a,j),t}^{(g,j_d)}, \quad j = 1, ..., N_i, \;\; t = 1, ..., T, \;\; i = 1, ..., N_I.$$

$$\sum_{(g',j_u) \in a} x_{i,(a,j'),t}^{(g',j_u)} = \sum_{(g,j) \in a} x_{i,(a,j'),t}^{(g,j)} \quad \forall j \in C_i \;\text{ such that }\; j > j', \quad t = 1, ..., T, \;\; i = 1, ..., N_I.$$

$$\sum_{\tau=1}^{t} \sum_{(g,j_d) \in a} x_{i,(a,j),\tau}^{(g,j_d)} \leqslant \sum_{\tau=1}^{t} w_{i,(a,j),\tau}, \quad j = 1, ..., N_i, \;\; i = 1, ..., N_I, \;\; t = 1, ..., T.$$

$$A_{i,k,t}^{(g,j)} = b_{ij} x_{i,t-L_{ij}}^{(g,j)} + \sum_{j'=1}^{j} \sum_{\{a \in j' | (g,j) \in a\}} b_{ij'j} x_{i,(a,j'),t-L_{i,j',j}}^{(g,j)}, \quad j \in C_i, \;\; i = 1, ..., N_I, \;\; t = 1, ..., T.$$

$$A_{i,k,t}^{j} = b_{ij} x_{i,t-L_{ij}} + \sum_{j'=1}^{j} \sum_{a \in j'} \sum_{(g,j'_d) \in a} b_{ij'j} x_{i,(a,j'),t-L_{i,j',j}}^{(g,j)}, \quad j \notin C_i, \;\; i = 1, ..., N_I, \;\; t = 1, ..., T.$$

$$\sum_{i=1}^{N_I} \sum_{j \in C_i} \sum_{g \in k} a_{kij} A_{ikt}^{(g,j)} + \sum_{i=1}^{N_I} \sum_{j \notin C_i} a_{kij} A_{ikt}^{j} \leqslant c_{kt}, \quad k = 1, ..., K, \;\; t = 1, ..., T.$$

All variables are constrained to be nonnegative. Inventory balance constraints are unchanged.

Modeling Alternative Resource Combinations

There are instances of scarce tooling in semiconductor manufacturing, whereby capacity may be a function not only of the available machines, but also of the available tools, such as photomasks, burn-in boards, test handlers, and fixtures, and so on. An added dimension of complexity arises when alternative machine types for a process step each require distinct tooling. That is, there are alternative combinations of resources that can be assigned to carry out a process step. Modeling the capacities of machines and tools independently may lead to production plans impossible to carry out, because the plans may implicitly require joint application of incompatible tools and machines. The model must explicitly represent the assignment of a valid combination of resources to carry out the process step.

For example, in semiconductor test operations, several different types of tools are required, in addition to the testing CPU. These may include handlers, interface boards, software programs, jigs, and fixtures. There may be several alternative tools for any one of these tool types, for example, handlers capable of only room temperature tests, handlers capable of both room temperature and high temperature tests, and handlers capable of cold temperature, room temperature and high temperature tests. A similar phenomenon occurs in burn-in operations, where the set of resources to be applied includes the burn-in oven chambers, the burn-in board, and the load/unload equipment. There can be two or more alternative sizes of burn-in boards that may be assigned to burn in a given device, each compatible only with similar-sized oven chambers and load/unload machines. Each instance of a valid set of resources to carry out the process is what we term a valid resource combination.

We may extend the model of Section on Accommodating Machine Arrangement Constraints to handle such cases by introducing variables that assign resource combinations to process steps. New notation, parameters and variables, as well as revised capacity constraints, are introduced as follows.

Notation

c denotes a resource combination, that is, a set of resource types required to perform one or more process steps. $k \in c$ denotes that resource type k is a member of resource combination c.

$c \in (i, j)$ denotes that resource combination c is one that is qualified to perform step j on product i.

$(i, j, c) \in k$ denotes that step j on product i is one of the steps that machine type k performs as part of resource combination c.

Variables

$A_{i,j,c,t}$ = workload of step j on product i in period t allocated to resource combination c.

Constraints

Modified capacity constraints are as follows. Inventory balance constraints are unchanged.

$$b_{ij}x_{i,t-L_{ij}} + x_{i,j,t} + \sum_{\alpha=1}^{j-1} b_{i\alpha j}x_{i,\alpha,t-L_{i\alpha j}} = \sum_{c\in(i,j)} A_{i,j,c,t}, \quad j = 1, ..., N_i, \quad i = 1, ..., N_I, \quad t = 1, ..., T.$$

$$\sum_{(i,j,c)\in k} a_{kij}A_{i,j,c,t} \leqslant c_{kt}, \quad k = 1, ..., K, \quad t = 1, ..., T.$$

$$A_{i,j,c,t} \geqslant 0, \ x_{i,j,t} \geqslant 0, \ x_{i,t} \geqslant 0, \ I_{i,t} \geqslant 0, \ BO_{i,t} \geqslant 0,$$

$$k = 1, ..., K, \quad j = 1, ..., N_i, \quad i = 1, ..., N_I, \quad t = 1, ..., T.$$

2.4. Handling Integrality of Production Lots

Work in process in wafer fabrication plants is typically released into production and moves from step to step in lots of twentyfive or fifty wafers, less yield losses. In device assembly and testing areas production lots typically consist of thousands of packaged chips, with specific upper limits based on the capacity of lot carriers. The foregoing linear programming formulations make no allowance for lot integrality when calculating release schedules. It is computationally unattractive to enforce integer constraints on release schedules, so heuristic techniques are typically used to alter input data and/or solution results into a workable schedule.

For example, cumulative demands for each product may be rounded up into lot-sized step functions, where the lot size is chosen according to the desired starting lot size and the expected process yield. If the total demand over the horizon is less than one lot, the input demand can be rounded up to one lot, or alternatively, the model may be allowed to schedule the release of a reduced-size lot, if overproduction is not desired. If the demand is satisfied on time in the LP solution, the resulting release schedule will be lot-sized as desired. If the release schedule is fractional, it nevertheless can be interpreted as a target rate (remember the variables were defined as production rates) to which the actual discrete lot releases should adhere as closely as possible. Of course, the integrality problem is most acute for low-volume products and becomes insignificant for high-volume products.

2.5. Multiobjective Programming

It is sometimes difficult practically or institutionally to capture all the desirable qualities of a production plan in a single economic objective such as formulated in Section 2.2. Successive LP runs, possibly with different objectives for different priority classes of demands, were suggested in Section 2.4 as a means of preventing the LP from making trade-offs between servicing prior customer commitments and responses to other market opportunities. But there may be other objectives of interest as well.

For example, it may be desired to control the distribution of bin inventories remaining at the end of the planning horizon, encouraging the planning model to use the lowest-quality bins and the least capacity-consuming test flows necessary to meet the demands. If costs on inventories and costs on test flows were incorporated into the same objective function that minimizes shortage costs, it may admit solutions that trade-off product shortage costs with production and inventory costs, yet no such trade-off is desired. Handling bin inventory costs in a second objective function, to be optimized subject to the optimal value of the first, is a more desirable approach.

It is straightforward to add follow-on, multiple objectives to the basic formulation discussed herein.

3. Brief History of Optimization Applications to Semiconductor Production Planning

The first regularly-used optimization-based planning system in the industry known to the author was the A-Plus system at Intel Corporation, based on the Berkeley Planning System software, installed in 1986 to plan across the fabrication lines in that company. BPS was implemented to plan three fabrication plants at Harris Corporation in 1987, and beginning in 1989 an effort was undertaken to enhance and expand BPS for fully automated, company-wide production planning. This culminated in the IMPReSS planning system, fully implemented in 1992. The Franz Edelman Award from the Institute for Operations Research and the Management Sciences (INFORMS) was awarded to Harris in 1995 in recognition of the achievements of IMPReSS. Subsequently, enhanced versions of BPS were implemented at Advanced Micro Devices and at Samsung Electronics, Co., Ltd.

Commercial supply chain management and production planning software products incorporating linear programming calculations began to be marketed to the semiconductor industry following the success of IMPReSS. Major products currently available include MIMI from Chesapeake Decision Sciences (later Aspen Technologies), Rhythm from i2 Technologies, and the Supply Chain Planner from Paragon Management Systems (later Adexa Management Systems). Leachman and Associates LLC offers customized optimization-based planning systems to the industry.

At the start of the millennium, a substantial and growing fraction of the global semiconductor manufacturing industry employ optimization calculations to assist in the generation of production plans. This remarkable growth has been paced not only by the development of modeling software as above, but also by the rapid development of computer hardware capabilities and of optimization software. At the time of implementation of the A-Plus system at Intel, the leading commercial LP software could solve problems with up to

10,000 rows in a few hours using the simplex method operating on a top-of-the-line IBM mainframe computer. Six years later, in 1992, the IMPReSS system at Harris routinely optimized formulations with upward of 130,000 rows in a few hours using interior point methods operating on a UNIX workstation computer.

Despite this progress, it must be acknowledged that optimization still plays a limited role in the generation of production plans within most commercial planning systems. For example, Rhythm uses optimization calculations only to generate guidelines such as routings through the product structure and allocations of capacity to products. Heuristics and artificial intelligence rules subsequently act on the guidelines to actually compute schedules. With the exception of products from Leachman and Associates LLC, the commercial products utilize optimization models that incorporate only a subset of the capabilities described in Section 2.3. For example, formulations in most commercial systems require time lags to be integers. Some offer only crude capabilities for capacity analysis, yet others model alternative resource combinations. Some do not accommodate binning and substitution product structures, yet others readily handle such complex product structures.

While optimization modeling has made considerable progress over the last fifteen years in serving the semiconductor industry, it is the author's opinion that the opportunity remains for much more progress to be made.

References

Fordyce, K. and G. Sullivan. 1995. "A dynamically generated rapid response fast capacity planning model for semiconductor fabrication facilities." In *The Impact of Emerging Technologies on Computer Science and Operations Research*, edited by S. Nash and A. Sofer. Boston: Kluwer Academic Publishers.

Fordyce, K., D. Dalton, B. Gerard, R. Jesse, and G. Sullivan. 1992. "Daily output planning: integrating operations research, artificial intelligence, and real-time decision support with APL2." *Expert Sys. Appl.* **5**: 245–256.

Hackman, S. T. and R. C. Leachman. 1989. "A general framework for modeling production." *Mgmnt Sci.* **35(4)**: 478–495.

Hung, Y.-F. 1991. *Corporate-level production planning with simulation feedback of parameters*. Ph.D. diss. Berkeley: College of Engineering, University of California at Berkeley.

Hung, Y.-F. and R. C. Leachman. 1996. "A production planning methodology for semiconductor manufacturing based on iterative simulation and linear programming calculations." *IEEE Trans. Semiconductor Manufacturing* **9(2)**: 257–269.

Kang, J. 1996. *A method for target scheduling of semiconductor wafer fabrication based on event-based optimization modeling and discrete event simulation*. Ph.D. diss. Berkeley: College of Engineering, University of California at Berkeley.

Leachman, R. C. 1986. Preliminary design and development of a corporate-level production planning system for the semiconductor industry. ORC Report 86-11. Berkeley: Operations Research Center, University of California at Berkeley.

Leachman, R. C. 1993. "Modeling techniques for automated production planning in the semiconductor industry." In *Optimization in Industry*, edited by T. A. Ciriani and R. C. Leachman, pp. 1–30. Chichester, England: John Wiley and Sons, Ltd.

Leachman, R. C. and T. F. Carmon. 1992. "On capacity modeling for production planning with alternative machine types." *IIE Trans.* **24(4)**: 62–72.

Leachman, R. C. and D. J. Raar. 1994. "Optimized production planning and delivery quotation for the semiconductor industry." In *Optimization in Industry 2*, edited by T. A. Ciriani and R. C. Leachman, pp. 63–72. Chichester, England: John Wiley and Sons, Ltd.

Leachman, R. C., R. F. Benson, C. Liu, and D. J. Raar. 1996. "IMPReSS: an automated production-planning and delivery-quotation system at Harris Corporation—semiconductor sector." *Interfaces* **26(1)**: 6–37.

Lin, V. 1999. *Advanced Semiconductor Production Planning*. Ph.D. diss. Berkeley: College of Engineering, University of California at Berkeley.

Liu, T.-Y. 1998. *Equipment acquisition planning in the semiconductor industry considering learning effects in equipment efficiency*. PhD diss. Berkeley: College of Engineering, University of California at Berkeley.

Stray, J., J. W. Fowler, and W. M. Carlyle. 2000. "Enterprise wide strategic and logistics planning for semiconductor manufacturing." In *Proceedings of the International Conference on Modeling and Analysis of Semiconductor Manufacturing (MASM 2000)*, edited by W. Cochran, John W. Fowler, and Steven Brown, pp. 353–356. San Diego: The Society for Computer Simulation International.