Iterative Capacity Allocation and Production Flow Estimation for Scheduling Semiconductor Fabrication

Shi-Chung Chang, Loo-Hay Lee, Lee-Sing Pang
Department of Electrical Engineering
National Taiwan University
Taipei, Taiwan, R.O.C.
e-mail: scchang@ac.ee.ntu.edu.tw
Fax: 886-2-363-8247

Thomas W.-Y. Chen, Yi-Chen Weng,
Huei-Der Chiang, David W.-H. Dai
Taiwan Semiconductor Manufacturing Co.
Hsin-Chu, Taiwan, R.O.C.
Fax: 886-35-781-546

ABSTRACT

This paper presents an effective algorithm of determining daily production targets and the corresponding machine capacity allocation for semiconductor wafer fabrication. The algorithm adopts an iterative scheme and each iteration consists of two modules: the proportional Target Generation and Machine Allocation (TG&MA) and the Stage of Penetration Estimation Algorithm (SOPEA). In TG&MA, machine capacities are allocated to processing different types of products at various stages in proportion to their respective available workloads. With the capacity allocated to each product type, SOPEA then applies a recursive, deterministic queuing analysis to estimate the expected flow-in workload of a stage within a day. The flow-ins are fed into TG&MA for another iteration of capacity allocation. Field implementation of this algorithm has demonstrated significant effects on production move increase, cycle time reduction and line balancing.

1. Introduction

Semiconductor wafer fabrication involves one of the world's most complex manufacturing processes. There may be tens of product types in a wafer fabrication plant (fab). The fabrication process of each type of wafers may require more than 100 fabrication stages, each consisting of a few fabrication steps; the whole process involves tens of delicate and expensive machines. As a type of wafers have quite a few (10 - 30 or so) layers of fabrication and stages between two layers of wafer bear some basic similarity, the production flow of each type of wafers may reenter the similar sequence of machine groups from layer to layer in its fabrication process. Owing the reentrant nature, wafers of different types as well as wafers of the same type but at different layers of fabrication may compete for the finite capacity of a machine group. Complex and reentrant process flows and uncertainties of machine availability pose unique challenges to production scheduling of a fab for effective wafer-in-process (WIP) movement, machine utilization and on-time delivery.

There have been many results on scheduling and dispatching for wafer fabrication. Wein [Wei88] conducted simulation study and pointed out that short interval scheduling has a significant impact on fab performance. Bai et. al. [BSG90] and Conors et. al. [CFY92] adopted fluid network models for scheduling high volume fabs. Lu et al [LRK94] analyzed several distributed scheduling/dispatching policies and identified two best policies for minimizing both the mean and the variance of cycle time. Liao et. al. [LCK94] adopted a Lagrangian relaxation and network-based optimization approach for scheduling a pilot line. In [Lea94], Leachman provided a survey of scheduling practices across six companies. These schedulers are mostly cutomized designs, involving some commercially available modules, the Kanban Logic, cycle time tracking mechanism, rule-based system, deterministic simulation, etc.

This paper presents an iterative algorithm for determining daily production targets and the corresponding machine allocation by product type and by production stage. The algorithm consists of two modules: the proportional Target Generation and Machine Allocation (TG&MA) and the Stage of Penetration Estimation Algorithm (SOPEA). The daily production target generation problem is first described in Section 2. Sections 3 and 4 present the TG&MA and SOPEA algorithms respectively. Successes of their field application are then given in Section 5. Section 6 concludes the paper.

2. Daily Target Generation and Machine Allocation (TG&MA) Problem

Daily Target Generation and Machine Allocation is the production control function that determines for each product type the amount of wafers to be processed and the machine capacity allocated at each stage during a day. Under a given wafer release schedule, machine capacity, and initial WIP distribution, it aims at multiple operation objectives such as

1.) meeting the monthly target output volume,
2.) balancing the production line,
3.) reducing WIP and cycle time,
4.)maximizing bottleneck machine utilization, and
5.)meeting the due date and demanded volume of each production order.

Such a decision is difficult because of the complexity of IC fabrication.

To reduce the problem complexity to a comprehensible level and to focus on key issues, we make the following assumptions and/or simplifications.

0-7803-2037-9/95 $4.00 ©1995 IEEE

1. All product types are of the same priority.
2. The fabrication process of each part type is fixed.
3. Stage is adopted as the basic unit for describing process flows, where a stage of a product type is obtained by aggregating a few consecutive fabrication steps of the product type.
4. All the stages of various fabrication processes can be arranged into a global sequence of totally J stages in a way that if a stage k precedes a stage k' in one process, then stage k also precedes stage k' in the global stage sequence.5. The production unit is a wafer.
6. Each stage has a corresponding key machine group. As steps of a stage may require machines from different machine groups and different stages may also share the same machine groups, the machine group that has the highest ratio of processing time over number of machines among the steps in a stage is selected to be the key machine of the stage.
7. The aggregate capacity of a machine is defined by the average number of wafers it can process in a day; this value is obtained from empirical statistics without taking the factor of part mix into account.
8. Batching effects at the diffusion and photolithography stages and setup times at the implantation and photolithography stages are ignored.
9. Buffer space for WIP is large and can be considered as infinite.
10. A desirable WIP level, called the "Standard WIP", is given for each stage.

To perform daily target generation and machine capacity allocation, the input data set includes
- the daily target output;
- the process flow in terms of stages for each product type;
- current WIP level of each stage of a product;
- the desired WIP level (called standard WIP) of each stage;
- processing time of every product at each stage,
- the key machine group for each fabrication stage and the number of machines in it;
- the capacity of each machine in a day defined as wafers per day; and
- current machine status.

By using the above input data, the decision function is to decide daily production targets and machine allocation for each stage. The output targets and machine allocation are finalized in the daily production meeting, which are then given to the shop floor for execution. Operators carry out the actual dispatching of wafers so that the targets can be met.

3. The TG&MA Algorithm

TG&MA is an iterative algorithm that computes, for each wafer type at each fabrication stage, the target amount of wafers to be processed during a day. Given the desired outputs, WIP distribution and expected wafer flow-ins to each stage, each iteration of the TG&MA algorithm first ignores the factor of finite machine capacity and computes an upper bound demand by type and stage via a PUSH-PULL procedure. The PUSH procedure generates demands in a way that pushes the wafers at a stage to down stream stages except those needed for maintaining the standard WIP so that the unnecessary WIP is reduced and the throughput is maximized at the stage in a heuristic sense. The PULL procedure generates demands in a way that pulls for each stage the production flows from its up stream stages, attempting to maintain the standard WIP level of the stage and to meet the output demands of the line at the same time. The upper bound demand by stage and type is the larger of the PUSH and PULL demands. Note that the upper bound demands may not be satisfied because of insufficient machine capacity or insufficient wafers for processing.

The factor of finite machine capacity is then considered. Since different types of wafers and different fabrication stages of a type may compete for the same type of machines, the capacity of a machine group is allocated proportionally to the upper bound demands of stages that are competing for it and targets of individual types and stages are obtained. Finally, the target of each stage is further modified by considering initial WIP and how many wafers that may flow into the stage from its up-stream stages within one day.

To formalize the description of the algorithm, let us first define some notations.

Notations

I: total number of part type;
i: part type index, i =1,...I;
OUT_i:the desired output amount of type-i wafers for the day;
J: total number of stages in global sequence;
j: stage index, j =1,...J;
WIP_{ij}:the WIP level of type-i parts at stage j at the beginning of the day;
Std-WIP_{ij}: the standard WIP level of type-i parts atstage j;
$flow\_in_{ij}$: number of type-i wafers flowing to stagej from its up-stream stages during the day;
t_{ij}: average processing time of a type-i wafer at stage j;
S_j: the set of all the immediate up stream stages of stage j in various process flows;
R: amount of wafer start for the day;
M: the total number of machine groups;
m: the machine group index, m=l,...,M;
N_m: number of available machines in group m of the day;
Cm: capacity of a machine in group m in term of wafers per day;
m_j: index of machine group required by stage J;
k_{ij}: the immediately up stream stage index for stage j of type i part.

Decision Variables

n_j: number of machines allocated to process stage j for the day;

n_{ij}: number of machines allocated to process type-i parts at stage j -for the day;

$Target_{ij}$: number of type-i wafers leaving stage j to its down-stream stages.

ALGORITHM

Step 0: Initialization

Input all the necessary data.

Set $flow\_in_{i1} = R_i$ for all i, and $flow\_in_{ij} = 0$ for all i and j=2,...,J.

Step 1: PUSH

Do for j=1,..., J

Do for i=1,...,I

$$Push_{ij} = Push_j \times \frac{(WIP_{ij} + Flow\_in_{ij})t_{ij}}{\sum_i (WIP_{ij} + Flow\_in_{ij})t_{ij}}$$

enddo

enddo

Step 2: PULL

Set $Pull_{iJ} = OUT_i$, for i = 1,...,I.

Do for j=J-1, ..., I

$$Push_j = \max[0, \sum_i (WIP_{ij} + Flow\_in_{ij}) - std\_WIP_j]$$

Do for i=1,...,I

$$Pull_{ik_{ij}} = Pull_j \times \frac{(WIP_{ik_{ij}} + Flow\_in_{ik_{ij}})t_{ij}}{\sum_i (WIP_{ik_{ij}} + Flow\_in_{ik_{ij}})t_{ij}}.$$

enddo

enddo

Step 3: Upper Bound Demand

Set $UBD_{ij} \equiv \max(Push_{ij}, Pull_{ij})$

for i = 1,...,I and j = 1,...,J.

Step 4: Target Generation and Machine Allocation

Do for all i and j

$$Target_{ij} = \text{Min}\left(\frac{UBD_{ij} \times t_{ij}}{\sum_{j' \text{ with } m_{j'} = m_j} UBD_{ij'} \times t_{ij}} \times C_{m_j}, WIP_{ij} + Flow\_in_{ij}\right)$$

$$n_{ij} = \frac{Target_{ij}}{C_{m_j}} \times N_{m_j}$$

enddo

Step 5: Convergence Check

If the targets differ from the targets of the previous iteration by less than a preset small amount, then stop.

Step 6: Flow-in Update

Do for i=1,...,I

Update $Flow\_in_{ij}$ for j=1,...,J under the targets of this iteration by using the SOPEA algorithm.

enddo

Go to Step 1 for the next iteration.

4. Stages of Penetration Estimation Algorithm (SOPEA)

In this Section, an algorithm (SOPEA) is proposed to estimate how many stages that the initial WIP at each stage may go through after one day, which in turn is used to estimate the amount of flow-in wafers of each stage during a day. The key idea of SOPEA is that once the capacity allocation (n_{ij}) is obtained in TG&MA, individual production flows of different part types are essentially independent from each other. We therefore focus on analyzing a single type of wafer flow and develop a deterministic queuing analysis for it. In our analysis, we assume that the wafers at a stage are processed on a FIFO basis. The part type index i is omitted in the following derivations for simplicity of presentation.

Consider the production flow between stage j to stage k (k > j) as shown in Figure 1, where T_{jk} is the cycle time needed for the last piece of WIP_j to finish processing at stage k. SOPEA defines a recursive algorithm for computing T_{jk} by using $T_{j(k-1)}$ and $T_{(j+1)k}$ based on the following relationship.

Two Stage Case (k=j+1)

If $WIP_j t_j/n_j \le [(WIP_j - 1)+WIP_{j+1}]t_{j+1}/n_{j+1}$,
then $T_{j(j+1)} = (WIP_j+WIP_{j+1})t_{j+1}/n_{j+1}$;
else $T_{j(j+1)} = WIP_j t_j/n_j + t_{j+1}/n_{j+1}$.

General Case

If $T_{j(k-1)} \le T_{(j+1)k} + (WIP_j - 1)t_k/n_k$,
then $T_{jk} = T_{(j+1)k} + WIP_j t_k/n_k$;
else $T_{jk} = T_{j(k-1)} + t_k/n_k$.

Note that we can start with computing $T_{j(j+1)}$ for j = 1,..., J-1 by applying the two-stage case formula. Then we can compute $T_{j(j+d)}$ for j = 1,..., J-d, where d is increased from 2 to J-1, by applying the general case formula with $T_{j(j+d-1)}$ and $T_{(j+1)(j+d)}$ computed. Such a procedure generates all the T_{jk}'s for $1 \le j < k \le J$. Interested readers may refer to Section 4.2 of [Wan94] for more details.

The amount of wafers that may flow into a stage j during a day can be easily computed by adding up the WIPs of stage j's uptream stages that has $T_{j'j} \le 24$ hours, i.e.,

$$Flow\_in_{ij} = \sum_{j' \in A_{ij}} WIP_{ij'},$$

where $A_{ij} \equiv \{j' | j'$ a stage of type i process flow, $j' < j$ and $T_{j'j} \le 24$ hrs$\}$.

5. Field Implementation Results

A TG&MA module using an empirical rule for estimating flow-ins instead of SOPEA was first implemented in the field. Comparing the fab performance before and after the implementation, we observed significant improvements (Figure 2 and [Lea94]):

(1) overall fab WIP declined by 8%;
(2) daily fab total wafer moves increased by about 20%;
(3) average cycle time per layer fell from 3.25 to 2.96 days;
(4) the average +2 sigma fell from 4.63 to 3.68.

After integrating SOPEA with TG&MA in a later time, further improvements were immediately observed (Figure 3 and [Lea94]):

(5) daily fab total wafer moves increased by another 5%; and
(6) the number of stages that have more than 10% difference between the scheduled and the actual targets was reduced by about 10%.

VI. Concluding Remarks

The TG&MA and SOPEA algorithm presented in this paper combines production flow modeling, empirical rules, proportional resource allocation and deterministic queuing analysis into an effective target generation and machine capacity allocation tool for a semiconductor fab. Preliminary analyses of its convergence and line balancing properties [CCW95] are consistent with our observations from its field implementation. It has also been extended to weekly and monthly target generation [Wan94].

VII. Acknowledgements

The authors woould like to give special thanks to Mr. William Wang and Mr. Mu -Tao Chi for their vision in supporting this University-Industry cooperation project.

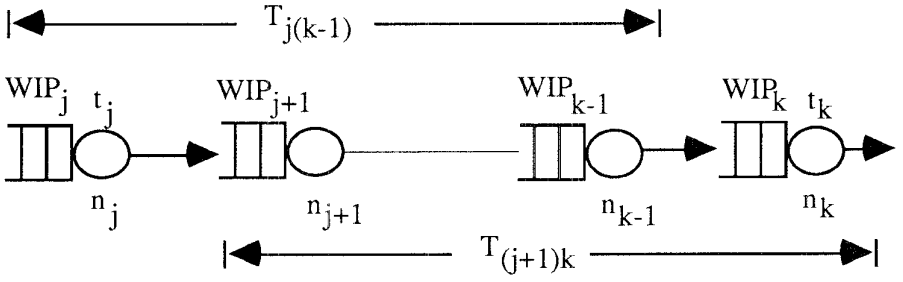

Figure 1: Partial Process Flow

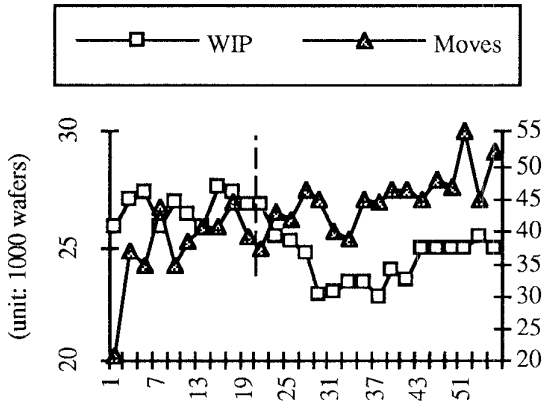

Figure 2: Performance before and after TG&MA

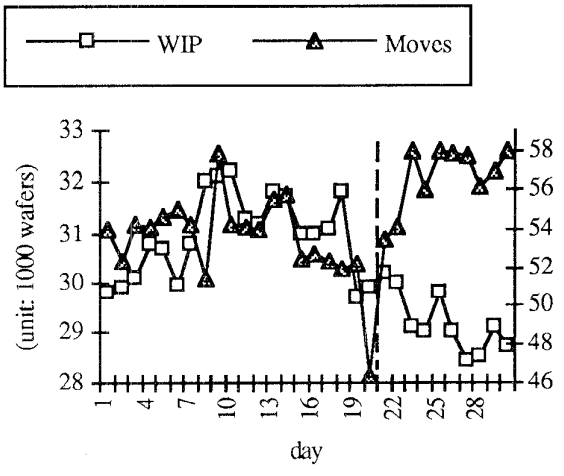

Figure 3: Performance before and after SOPEA

References:

[BSG90] S. Bai, N. Srivantsan, and S. Gershwin, "Hierarchical Real-Time Scheduling of a Semiconductor Fabrication facility,"*Proceedings of the 9th IEEE International Electronics Manufacturing Technology Symposium*, Washington D.C., October 1990.

[CCW95] S. Chang, W. Chan, T. Wang, C. Chang, "Proportional Machine Allocation and Line Balancing in a Re-entrant Line," *Proceedings of the Conference on Emerging Technology for Factory Automation*, Paris, France, Oct. 1995.

[CFY92] D. Connors, G. Feigin, and D. Yao, "Scheduling Semiconductor Lines Using Fluid Network Model," *Proceedings of the 3rd International Conference on Computer Integrated Manufacturing*, Troy, New York, May 1992, pp. 174 - 183.

[Lea94] R. Leachman, "Production Planning and Scheduling Practices Across the Semiconductor Industry,"*Technical Report*, ESRC 94-29/CSM-18, Engineering Systems Research Center, U. C. Berkeley, Berkeley, CA, Sept. 1994.

[LCK93] D. Liao, S. Chang, K. Pei, C. Chang, "Daily Scheduling for R&D Semiconductor Fabrication," submitted to *IEEE Transactions on Semiconductor Manufacturing*, Dec. 1993; under revision.

[LRK94] S. Lu, D. Ramaswamy, P. Kumar, "Efficient Scheduling Policies to Reduce Mean and Variance of Cycle-Time in Semiconductor Manufacturing Plants,"*IEEE Trans. on Semiconductor Manufacturing*, Vol. 7, 1994.

[Wan94] T.-H. Wang, "Design and Analysis of a Short Term Scheduling Method for Semiconductor Manufacturing," MS Thesis, Dept. of Mechanical Engineering, National Taiwan University, Taipei, June 1994.

[Wei88] L. Wein, "Scheduling Semiconductor Wafer Fabrication," *IEEE Trans. on Semiconductor Manufacturing*, Vol. 1, No. 3, Aug. 1988, pp. 115 - 130.