A Production Planning Model for the Semiconductor Industry with Processing Capabilities

Conference Paper · March 2017

1 author:

Cenk Çalışkan
Utah Valley University

A PRODUCTION PLANNING MODEL FOR THE SEMICONDUCTOR INDUSTRY WITH PROCESSING CAPABILITIES

Cenk Çalışkan, Woodbury School of Business, Utah Valley University, 800 W. University Pkwy, Orem, UT 84058, (801) 863-6487, cenk.caliskan@uvu.edu

ABSTRACT

Production planning in the semiconductor industry has challenges due to complicating factors such as binning and downgrading and the use of overlapping production capabilities instead of individual machines or machine types as resources. Binning is due to the inherent variability in semiconductor manufacturing processes. The output of manufacturing processes that produce integrated circuits (IC) is highly variable in terms of clock speed, memory and other characteristics. A process that is started with the objective of producing a specific IC with a desired speed may produce both faster and slower ones, as well as the desired ones. Faster ones are frequently downgraded to the desired speed level, in order to satisfy the demand at that level. Another complicating factor is the use of process capabilities. A process plan for producing a given product may require a specific type of equipment, or a specific processing capability that may exist on different types of machines. This creates an overlap between different types of machines, creating bundle constraints that complicates the problem further. We propose a mixed integer program that accounts for these complicating factors and creates an aggregate production plan, as the higher level of a two-level planning approach.

Keywords: production planning, semiconductor, mixed integer programming, process capabilities

1. INTRODUCTION

Production planning in the semiconductor industry is challenging due to complicating factors such as binning and downgrading, and the use of process capabilities as resources, instead of specific equipment units or equipment types. A silicon wafer consists of a number of integrated circuits (IC) put together in a circular disk. The equipment that is used in manufacturing silicon wafers is expensive. This necessitates using the equipment for a large number of processing steps. For this reason, process flows in semiconductor manufacturing are re-entrant, i.e. wafers visit the same type of machine many times during their manufacture. Binning is another challenge that complicates the problem by adding uncertainty to the output of the production process, not only in the form of variable output quantity, but also in the form of variable output product types. Downgrading is the practice of assigning higher quality (higher speed or memory) products to lesser quality demand, since additional demand for the higher quality product does not exist and/or the demand for the lesser quality ones must be urgently satisfied. Average bin fractions for a process plan are usually known. The usual practice is to plan for the average bin fractions rather than the intended output.

Another complicating factor is the use of process capabilities. A process plan for producing a given product may require a specific type of equipment, or a specific processing capability, which may exist on different types of machines. This creates an overlap between different types of machines, which complicates the problem further, by adding bundle constraints.

In this paper, we propose a mixed-integer programming model that accounts for these complexities that we mentioned and produces an executable aggregate production plan, which will be an input to the lower-level planning process, i.e. detailed scheduling.

2. LITERATURE REVIEW

Over the decades, many approaches were proposed to tackle production planning and scheduling problems in semiconductor manufacturing. Uzsoy et al. (1992, 1994) provide a detailed description of the semiconductor manufacturing process, including discussions of production planning, scheduling, control and performance evaluation. Mönch et al. (2013) provide a comprehensive summary of the literature on production planning models in semiconductor manufacturing. Gupta et al. (2006) discuss all aspects of planning and control in semiconductor manufacturing and describe the state of the art in the literature. Gallego et al. (2006) propose models to tackle the issue of binning and downgrading in an inventory management context and present two heuristics. Pfund et al. (2006) consider the practical aspects of the problems and surveys industrial planning and control practices and situations in the semiconductor industry.

Hung & Leachman (1996) suggest automated production planning based on iterative LP optimization and discrete-event simulation while Lee et al. (1997) present a model based on LP techniques considering variable cycle times. Chou & Hong (2000) propose a mixed integer programming model for product mix decisions in a semiconductor foundry and develop a bottleneck-based procedure, and Chen (2003) presents a fuzzy LP model for monthly production planning in a wafer fabrication facility. Hwang & Chang (2003) propose a two-level hierarchical planning and scheduling approach, which consists of medium term and long term models. Asmundsson et al. (2006) develop a nonlinear programming model for production planning in a wafer fabrication facility.

Venkateswaran & Son (2005) provides a review of the literature in hierarchical (multi-level) production planning approaches. Shanthikumar & Sargent (1983) classify these hierarchical approaches in which different levels use different modeling methodologies. Byrne & Bakir (1999) propose a hybrid method in which an LP model and a simulation model are used iteratively, while Kim & Kim (2001) use the workload and utilization of each machine for validating and updating parameters and data in the LP model. Byrne & Hossain (2005) modify the model of Kim & Kim (2001) by allowing split lots. Venkateswaran & Son (2005) propose a hierarchical production planning and scheduling method using a system dynamics model and a discrete-event simulation model in the two levels.

Jornsten & Leisten (1995) suggest an aggregation-disaggregation approach and present an iterative procedure for connecting the aggregate and disaggregate levels. Leisten (1998) discusses how a suboptimal or infeasible solution can be optimized in this iterative process. Vicens et al. (2001) propose an iterative aggregation-disaggregation scheme in which the processing times of aggregated product types can be adjusted at each iteration. Selçuk et al. (2006) show that excessive frequent updating of data in the aggregate model deteriorates the quality of production plans and schedules. Bang & Kim (2010) propose a two-level production planning method with LP-based aggregate planning and discrete-event simulation based scheduling in the two levels.

None of the aforementioned models incorporate the complexities added to the problem by overlapping processing capabilities. In this paper, we propose a mixed-integer programming based model that successfully incorporates binning, downgrading and processing capabilities. Our model generates an aggregate plan as part of a two level hierarchical modeling approach. In future research, we will add a simulation based scheduling model in the lower level.

3. PROBLEM DESCRIPTION

As it is common in aggregate planning, we assume a finite planning horizon with discrete planning intervals. The model is intended to be re-optimized on a rolling-horizon, typically on a daily basis. The following parameters represent the range of indices in our model:

$T =$ the number of intervals in the planning horizon,
$P =$ the number of process plans that are in use at the facility,
$K =$ the number of products manufactured at the facility,
$R =$ the number of resources (equipment types and processing capabilities),
$Q =$ the number of equipment units (individual machines) at the facility.

A process plan is a set of processing steps that is intended to produce a certain product, but it results in a set of products that differ by memory, clock speed, etc., due to binning. Note that each of the output items that result from a process plan that differ by at least one characteristic is a different product. The manufacturing resources that are used in the facility are represented by "equipment types" and "processing capabilities" in our model. One machine is capable of performing a number of different steps, each of which represents a "processing capability." Some process plan steps specify a certain equipment type, whereas others specify a certain capability. The following parameters link the equipment units to equipment types and processing capabilities:

$ET(q) =$ the type of equipment unit q,
$EK(q) =$ the set of processing capabilities that the equipment unit q has,
$EU(r) =$ the set of equipment units that this resource is defined on.

Binning and downgrading (downward substitution) are represented by the following parameters. We assume constant binning fractions that are calculated from historical production statistics. Any product can be (downward) substituted or downgraded to a lesser product, and this downgrading relationship is user-defined.

$O(p) =$ the set of output products for process plan p,
$G(k) =$ the set of products downgradeable to product k, including itself
$\phi_{kp} =$ binning fraction for product k, using process plan p,

We assume a constant deterministic demand for each product in each time interval. On a rolling horizon, the model is adjusted with changes in the demand and in the conditions of the machines and the progress of the existing production lots. The following are the parameters representing the demand and process plans:

$D_{kt} =$ the demand for product k in period t,
$\tau_p =$ the total cycle time for process plan p,
$\lambda_p =$ the lot size for process plan p,

Each processing step of each process plan consumes a specific resource, i.e. an equipment type or processing capability. The following parameters define this relationship between process plans and resources as well as the capacities of the resources. Capacities are defined on individual equipment units.

$\pi_{rp} =$ the amount of resource r required for process plan p,
$\theta_{rp} =$ the time resource r is required for process plan p from the start,
$U_{qt} =$ the capacity of equipment unit q in period t,

We consider the manufacturing costs, inventory holding costs, and backordering costs for the products. The manufacturing costs are defined for each execution of each process plan, and this includes the cost of input parts and raw materials, the cost of labor and the cost of resources consumed by the process plan execution, which really is a "lot start" on that process plan. These cost parameters are defined as follows:

$C_p^{\pi} =$ the cost of process plan p per lot,
$C_k^{\psi} =$ the cost of inventory holding per unit of product k,
$C_k^{\beta} =$ the cost of backordering per unit of product k.

The decision variables of the model are defined as follows.

$X_{tp} =$ the number of times process plan p is started in period t,
$I_{kt} =$ the number of units of product k carried in the inventory at the end of period t,
$B_{kt} =$ the number of units of product k backordered at the end of period k,
$Y_{rqt} =$ the amount of resource r that is used up on equipment unit q in period t

In our model, our objective is to determine a minimum cost aggregate plan that minimizes the manufacturing, inventory holding and backordering costs. Therefore, our objective function is as follows:

$$Z = \sum_{t=1}^{T} \sum_{p=1}^{P} C_p^{\pi} X_{tp} + \sum_{t=1}^{T} \sum_{k=1}^{K} C_k^{\psi} I_{kt} + \sum_{t=1}^{T} \sum_{k=1}^{K} C_k^{\beta} B_{kt} \tag{1}$$

For products $k = 1, 2, \ldots, K$, in periods $t = 1, 2, \ldots, T$, the initial inventory plus the production output in that period minus any inventory left at the end of that period should be equal to the demand in that period plus backordered amount at the beginning of that period minus any backordering left at the end of the period. This results in the following constraint:

$$I_{k(t-1)} + \sum_{p=1}^{P} \sum_{i \in G(k) \cap O(p)} \lambda_p \phi_{ip} X_{(t-\tau_p)p} - I_{kt} = B_{k(t-1)} + D_{kt} - B_{kt} \tag{2}$$

For each resource $r = 1, 2, \ldots, R$, in each time period $t = 1, 2, \ldots, T$, process plans consume a certain amount and they add up to the consumption of that resource in that period. This results in the following set of constraints which link process plans with resources:

$$\sum_{p=1}^{P} \pi_{rp} X_{(t-\theta_{rp})p} = \sum_{q \in EU(r)} Y_{rqt} \tag{3}$$

The total consumption of all resources (equipment type and capabilities) on each equipment unit cannot exceed its capacity for equipment units $q = 1, 2, \ldots, Q$, in time periods $t = 1, 2, \ldots, T$. This ensures that consumption of equipment types and processing capabilities will not be multiply allocated to individual equipment units. This results in the following set of constraints:

$$\sum_{r \in ET(q) \cup EK(q)} Y_{rqt} \leq U_{qt} \tag{4}$$

Finally, all decision variables must be nonnegative and production lot starts, or the number of times each process plan is executed (X variables) should be integer:

$$X_{tp} \geq 0 \tag{5}$$

$$\text{and integer}, \quad t = 1, 2, \ldots, T, \;\; p = 1, 2, \ldots, P \tag{6}$$

$$I_{kt}, B_{kt} \geq 0, \quad k = 1, 2, \ldots, K, \;\; t = 1, 2, \ldots, T \tag{7}$$

$$Y_{rt}, Y_{rqt} \geq 0, \quad r = 1, 2, \ldots, R, \;\; q = 1, 2, \ldots, Q, \;\; t = 1, 2, \ldots, T \tag{8}$$

4. SOLUTION METHODOLOGY

We applied the proposed mixed integer program on a real fab data from our industrial partner, and solved it using CPLEX. In our preliminary tests, we observed that the model is solvable in reasonable time (less than 20 minutes of CPU time) when the fab is lightly loaded, i.e. when the total demand only consumes up to 65% of the fab capacity. For loads heavier than 65%, the CPU time becomes excessive. For this reason, we will explore decomposition and heuristic approaches based on partitioning and relaxation of the capacity constraints. In this preliminary stage, we were able to optimize aggregate plans for up to 100 process plans, 50 products, 40 equipment types, and 150 process capabilities. In future research, we will explore ways to solve bigger, more realistic size problems, and we will also develop simulation based models, to be used for detailed scheduling of the resources, as part of a two-level hierarchical planning approach.

5. CONCLUSIONS

In this research, we develop a mixed integer programming model for the medium term production planning of a semiconductor manufacturing facility that incorporates the challenges inherent in the semiconductor production planning. Production planning in the semiconductor manufacturing industry is challenging due to complexities caused by variable output as a result of the phenomenon of binning, as well as re-entrant flows that is caused by expensive equipment that have to be used for a large number of processing steps, resulting in wafers visiting the same equipment many times during different stages of its manufacturing process. For this reason, resources are defined as "processing capabilities" that cause overlap of different equipment types and make capacity planning more difficult. We propose a model that successfully and realistically models binning and overlapping processing capabilities. The model is promising as it was successfully optimized for a real wafer fab with a 65% capacity utilization. Future research will address decomposition and heuristic approaches to optimize the mixed integer program for larger datasets and higher workloads.

References

Asmundsson, J., Rardin, R. L., & Uzsoy, R. (2006). Tractable nonlinear production planning models for semiconductor wafer fabrication facilities. *IEEE Transactions on Semiconductor Manufacturing*, 19, 95–111.

Bang, J. & Kim, Y. (2010). Hierarchical production planning for semiconductor wafer fabrication based on linear programming and discrete-event simulation. *IEEE Transactions on Automation Science and Engineering*, 7(2), 326–336.

Byrne, M. D. & Bakir, M. A. (1999). Production planning using a hybrid simulation-analytical approach. *International Journal of Production Economics*, 59, 305–311.

Byrne, M. D. & Hossain, M. M. (2005). Production planning: An improved hybrid approach. *International Journal of Production Economics*, 93, 225–229.

Chen, T. (2003). A fuzzy mid-term single-fab production planning model. *Journal of Intelligent Manufacturing*, 14, 273–285.

Chou, Y. C. & Hong, I. H. (2000). A methodology for product mix planning in semiconductor foundry manufacturing. *IEEE Transactions on Semiconductor Manufacturing*, 13, 278–285.

Gallego, G., Katirciglu, K., & Ramachandran, B. (2006). Semiconductor inventory management with multiple grade parts and downgrading. *Production Planning and Control*, 17(7), 689–700.

Gupta, J., Ruiz, R., Fowler, J., & Mason, S. (2006). Operational planning and control of semiconductor wafer production. *Production Planning and Control*, 17(7), 639–647.

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

Hwang, T.-K. & Chang, S.-C. (2003). Design of a lagrangian relaxation based hierarchical production scheduling environment for semiconductor wafer fabrication. *IEEE Transactions on Robotics and Automation*, 19, 566–578.

Jornsten, K. & Leisten, R. (1995). Linear programming aggregation: A heuristic for hierarchical production planning. *Asia-Pacific Journal of Operational Research*, 12, 161–177.

Kim, B. K. & Kim, S. Y. (2001). Extended model for a hybrid production planning approach. *International Journal of Production Economics*, 73, 165–173.

Lee, Y., Kim, S., Yea, S., & Kim, B. (1997). Production planning in semiconductor wafer fab considering variable cycle times. *Computers and Industrial Engineering*, 33, 713–716.

Leisten, R. (1998). An LP-aggregation view on aggregation in multi-level production planning. *Annals of Operations Research*, 82, 413–434.

Mönch, L., Fowler, J. W., & Mason, S. J. (2013). *Production Planning and Control for Semiconductor Wafer Fabrication Facilities*, volume 52 of *Operations Research/Computer Science Interfaces*. New York: Springer.

Pfund, M., Mason, S., & Fowler, J. (2006). Semiconductor manufacturing scheduling and dispatching. In J. W. Herrmann (Ed.), *Handbook of Production Scheduling* (pp. 213–242). New York: Springer.

Selçuk, B., Fransoo, J. C., & Kok, A. G. D. (2006). The effect of updating lead times on the performance of hierarchical planning systems. *International Journal of Production Economics*, 104, 427–440.

Shanthikumar, J. G. & Sargent, R. G. (1983). A unifying view of hybrid simulation/analytic models and modeling. *Operations Research*, 31, 10301052.

Uzsoy, R., Lee, C., & Martin-Vega, L. (1992). A review of production planning and scheduling models in the semiconductor industry part i: System characteristics, performance evaluation and production planning. *IEE Transactions*, 24(4), 47–60.

Uzsoy, R., Lee, C., & Martin-Vega, L. (1994). A review of production planning and scheduling models in the semiconductor industry part i: Shop-floor control. *IEE Transactions*, 26(5), 44–55.

Venkateswaran, J. & Son, Y.-J. (2005). Hybrid system dynamic-discrete event simulation-based architecture for hierarchical production planning. *International Journal of Production Research*, 43, 4397–4429.

Vicens, E., Alemany, M. E., Andres, C., & Guarch, J. J. (2001). A design and application methodology for hierarchical production planning decision support systems in an enterprise integration context. *International Journal of Production Economics*, 74, 5–20.