Computers & Operations Research 53 (2015) 353–363

Contents lists available at ScienceDirect

Computers & Operations Research

journal homepage: www.elsevier.com/locate/caor

Incorporating contractual arrangements in production planning

R. John Milne [a,*], Chi-Tai Wang [b], Brian T. Denton [c], Kenneth Fordyce [d]

[a] *Clarkson University School of Business, 107 B.H. Snell Hall, P.O. Box 5790, Potsdam, NY 13699, USA*
[b] *Institute of Industrial Management, National Central University, 300 Jhongda Road, Jhongli City, Taoyuan County 32001, Taiwan*
[c] *Department of Industrial and Operations Engineering, University of Michigan, 1205 Beal Avenue, Ann Arbor, MI 48109-2117, USA*
[d] *Arkieva, 5460 Fairmont Drive, Wilmington, DE 19808, USA*

ARTICLE INFO

Available online 18 April 2014

Keywords:
Linear programming
Foundry
Supply chain planning
Semiconductor manufacturing

ABSTRACT

The semiconductor supply chain is full of complexities outside of the traditional order, make/buy, and deliver process. One critical challenge occurs when part of a semiconductor fabricator's capacity is allocated to produce wafers designed by and provided to fabless companies. In this situation, linked customer requirements are expressed simultaneously at both the semiconductor level of the supply chain and the finished goods level. As a result of the complex contractual relationships between the foundry and the fabless company, a new solution model and method is needed to determine a production plan. In our approach, two linear programming (LP) models are solved sequentially where the results of a first LP are post-processed into input for a second LP. We describe the application of this approach for two different types of contracts where the goal is maintaining as much common modeling as possible while ensuring the unique features of each contract are covered. For one type of contract, the first LP model determines the minimum quantities of wafers required to be released into the fab to meet the contractual obligation; these required starts are added as a constraint for the second LP model. For the other type of contract, the first LP determines production at one level of the bills of materials and feeds these outputs into a second LP that determines production for later stages of manufacture.

1. Introduction

In many industries, manufacturers have traditionally both designed and produced the products they sell. However, companies have emerged that design products that are subsequently produced by contract manufacturers, referred to as *foundries* [1]. For manufacturers with high fixed costs, the economies of scale may be such that large manufacturers must conduct two modes of business to be profitable. In the first, they manufacture products of their own design. In the second, they act as a foundry, providing manufacturing services via contract to other firms. This hybrid production model has created challenges for supply chain managers who must allocate resources for these two competing purposes.

As a result of this hybrid business model, IBM and other firms with semiconductor fabrication facilities (often shortened to *fabs*) serve as foundries for fabless companies. In these relationships, the client designs the product and contracts its manufacturing to the foundry. With new fabs costing as much as five billion dollars each [2], foundry/fabless relationships are becoming increasingly common because the companies designing products do not want to incur the expense of building semiconductor fabs and foundry manufacturers must build a wide range of products to fully utilize their fabs. This mutual dependence motivates long term agreements so that semiconductor foundries can be assured of enough demand to fill their fabs while fabless companies can be assured of sufficient supply to fulfill their needs. As a result, fabless firms enter into contracts with IBM that stipulate minimum guaranteed production levels over an agreed upon time frame.

Contractual arrangements between foundries and fabless companies are influenced by the nature of the semiconductor manufacturing process. In semiconductor wafer circuit fabrication, four manufacturing steps are repeated dozens of times: deposition, photolithography, etching, and ion implantation. Through these steps, a set of three-dimensional, layered circuit structures are built on the two-dimensional surface of each wafer in lots of 4–25 wafers, where 25 is common. After these circuits have been built, they are connected through wiring (within the chip) in which the following four manufacturing steps are repeated numerous times: deposition, photolithography, etching, and metallization (wiring). Typical lead times range from 50 to 150 days to manufacture both the circuits and the wires which connect them. The time jobs spend waiting for equipment to become available comprises the largest component of semiconductor lead times.

Following wafer fabrication, finished circuits are tested, the good chips cut (diced) from the finished wafer and placed onto a

---

[*] Corresponding author. Tel.: +1 315 268 7919; fax: +1 315 268 4478.
E-mail addresses: jmilne@clarkson.edu (R.J. Milne), ctwang@mgt.ncu.edu.tw (C.-T. Wang), btdenton@umich.edu (B.T. Denton), kfordyce@arkieva.com (K. Fordyce).

http://dx.doi.org/10.1016/j.cor.2014.03.023
0305-0548/

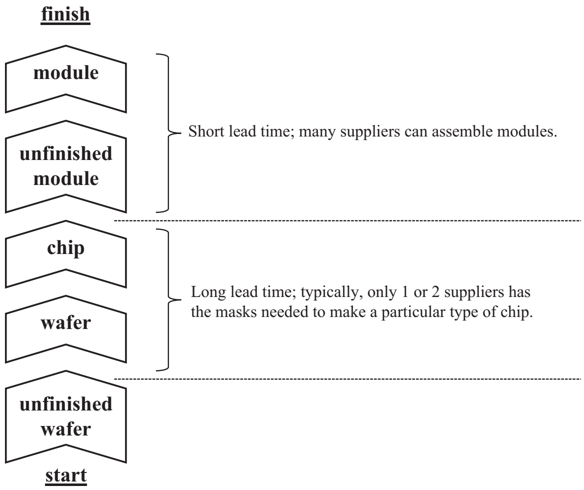

Fig. 1. Simplified bills of materials semiconductor manufacturing flow.

substrate and packaged to make modules. These modules are used in the assembly of a variety of products. Some IBM modules go into IBM servers such as supercomputers, mainframes, and workstations, but the chips and modules shipped to fabless companies often are assembled into consumer products such as cell phones, automobiles, global positioning systems, and game machines such as PlayStation, X-box, and Wii. Typical manufacturing lead times for module assembly and test are 10–20 days.

A set of masks (also referred to as reticles) are used in photolithography processes for etching circuit and wiring patterns onto a silicon wafer. Producing a set of masks for a single type of chip can cost several million dollars [3]. Because of the high cost of masks and other fixed costs involved in being able to manufacture a particular chip design, fabless firms will contract the manufacturing of a particular chip to a limited number of foundries—often only a single foundry. In contrast, module assembly operations are relatively few and less technologically complex resulting in short manufacturing lead times. Furthermore, as indicated in Fig. 1, the module assembly operations are more generic and can be performed economically by a large number of manufacturing contractors.

In contrast to a typical supply chain planning (SCP) process of order/make/deliver, the fabless/IBM partnerships involve the fabless company specifying their requirements simultaneously at both the start (wafer) or chip level and finished product (module) level in the supply chain (Fig. 1), and these requirements are linked. In addition to ordering finished modules, the fabless company places requirements on the production of the wafers/chips that are used to make the modules they order. This non-traditional SCP information flow (versus order/make/deliver) creates SCP challenges. While we discuss the modeling in the context of IBM's business, other semiconductor manufacturers can benefit from applying these concepts.

The remainder of this article is organized as follows. In Section 2, we review the related literature. In Section 3, we summarize the problem statement. In Section 4, we describe the core linear programming (LP) formulation used by IBM for production planning. In Section 5, we describe the usage of the core LP model and additional constraints required for minimum wafer starts contracts; these contracts require IBM to release a minimum number of wafer starts so long as there are sufficient orders placed by the fabless company to consume the output resulting from those wafer starts. In Section 6, we describe the usage of the core LP model and additional constraints required for complementary demand contracts under which customer requirements are simultaneously applied in a linked manner to both chips and modules. Section 7 provides numerical examples illustrating insights into the proposed methods and the advantages of these methods. Section 8 concludes with a summary of the insights from this article.

2. Literature review

Mathematical programming approaches have been applied to many contexts in which supply and demand must be matched subject to competing priorities and complex material flows such as those present in semiconductor manufacturing (e.g. product substitutions, alternative bills of materials, alternative manufacturing plant locations, and alternative capacities within a plant) within the traditional practice of order/make/deliver. For mathematical programming approaches to semiconductor SCP, see [4–22].

Kim et al. [23] propose algorithms for allocating (pegging) wafers lots to orders (demand). They peg the entire quantity of each lot to a set of orders. At IBM, it is possible for portions of lots to be pegged to portions of demands. As a result, although we use pegging in our method, the work of Kim et al. is not applicable for the IBM situation.

Hackman and Leachman [17] and Hung and Leachman [18] describe the modeling of lead times that are non-integral multiples of the LP time periods. In these papers, production starts made in one period may result in production output in multiple (typically two) periods and is often referred to as *fractional lead times*. The IBM team implemented this approach as an option in its original LP model. After extensive computational and usability exploration, IBM determined that best practices for detailed production planning models dictate that all starts made in one LP period should be modeled to arrive at stock in a single time period. This approach was preferred (but not perfect).

We illustrate the practical difficulty of the fractional lead time approach with a simple example. Suppose that production starts made in period 2 result in 40% of its production becoming available as output in period 4 and the remainder available in period 5. Further suppose there is only a single demand and it is for 100 pieces and is due in period 4. Typically, the objective function penalizes late deliveries more severely than anything else (case 1). Consequently, an LP model may recommend starting 250 pieces in period 2 so that 40% of them result in enough production output to meet the period 4 demand (assuming the yield rate is 100%); this would result in an excess inventory of 150 pieces. Conversely (case 2), if the ending inventory is penalized in the objective function to be more expensive than satisfying demand on time, then the LP may recommend starting 100 pieces in period 2 which will result in a backorder of 60% of the demand in period 4 that is not satisfied until period 5. In case 1, the planner sees 150 pieces produced and never consumed. In case 2, the planner sees the backordered demand even though capacity and components may be available. The result is unsatisfactory in both cases.

To account for demands of differing priorities, Leachman [19] and Leachman et al. [6] describe a goal programming type approach that invokes an LP run for each demand class priority in sequence of the most important demands first. During an LP run, the model is constrained to satisfy all demands more important than the current class of demands at least as well as the more important demands were satisfied during previous runs of the LP model. As a result, the on time delivery performance of satisfying the most important demands is as high as possible. For performance and “best practice” reasons, Denton et al. [4] extend the approach of Leachman et al. [6,19] so that multiple demand class priorities are accommodated within each run of an LP model. This approach involves running several demand class priorities within a single LP run and as in the Leachman et al. [6,19]