Computers & Operations Research 53 (2015) 338–352

Contents lists available at ScienceDirect

Computers & Operations Research

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

Product mix optimization for a semiconductor fab: Modeling approaches and decomposition techniques

Martin Romauch [a,*], Andreas Klemmt [b]

[a] University of Vienna, Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria
[b] Infineon Technologies, Konigsbrücker Straße 180, 01099 Dresden, Germany

ARTICLE INFO

Available online 14 May 2014

Keywords:
Semiconductor production planning
Product mix optimization
Master planning
Resource pooling
Decomposition
Linear programming

ABSTRACT

To optimize the product mix of a semiconductor fab, the production capabilities and capacities are matched with the demand in the most profitable way. In this paper we address a linear programming model of the product mix problem considering product dependent demand limits (e.g. obligations and demand forecast) and profits while respecting the capacity bounds of the production facility. Since the capacity consumption is highly dependent on choosing from different production alternatives we are implicitly solving a static capacity planning problem for each product mix. This kind of planning approach is supported by the fluid flow concept and complete resource pooling in high traffic. We propose a general model that considers a wide range of objectives, and we introduce a heuristic that is based on the decomposition of the static capacity planning problem. A computational study reports on the quality of the decomposition approaches, and examples from practice demonstrate the versatility of the model.

1. Introduction

In this paper we are considering a fab with given capabilities and capacities (technology and scale of the machinery) and we have to decide on the product mix to maximize profit. Typically, we encounter certain obligations (e.g. contracts, work in progress, customer service, development) and expectations on the sales (e.g. demand forecasts, down turns, market saturation). Consequently, we consider the product's demand limits and the marginal profits, while taking the capacity limits of the production facilities into account.

Certainly, the best thinkable model to answer the question for the optimal product mix integrates all necessary aspects; Simulation or Discrete Event Simulation [3,6,20] is a method that comes very close to these needs. Simulation is employed to analyze and forecast the dynamic behavior of complex and stochastic systems. In case of a semiconductor fab the complexity arises from the intricated product routes (re-entrant flow, hundreds of steps, alternative processing, etc.), the large number of equipments (usually hundreds), the variety of tool types (batch processing, pipeline tools, cluster tools, etc.) and the vast number of other factors (transport, dispatching, machine shutdowns, repair, yield, rework, operators, etc.) that can or should be considered [22]. Depending on the purpose of the simulation model (e.g. cycle time forecast) some of the selected factors will be stochastic in nature and they need to be modeled with the proper distributions, therefore several runs are usually necessary to achieve the desired level of significance. Furthermore, there are also cases that go beyond simple dispatching rules (e.g. local scheduling) and it gets necessary to integrate an optimizer in the simulation models (cf. [18] and [2]). We note that factors like that can make the simulation model complex and error-prone [24]. Therefore, dependent on the time span, the level of detail, and the number of runs, the task of developing a model that is suitable for optimizing certain parameters is usually difficult and time consuming. Models that consider various aspects, typically, have a considerably high development cost and simulation runtime. Therefore, only a small number of parameters can be optimized employing such a framework, and optimizing several interdependent parameters at the same time is computationally intractable. Besides simulation, there are several alternative approaches to model a semiconductor fab, for example Zisgen et al. [25] present a framework that enables to analyze steady and transient states for queuing networks to forecast the cycle time and utilization on tool basis (MES level), but there is no extension that covers product mix optimization. In [23] and [7] you can find approaches to optimize the product mix by solving a mixed integer linear program and in both cases the capacity is captured by considering a predefined set of bottlenecks and the process times are not tool dependent which

---

* Corresponding author.
E-mail addresses: martin.romauch@univie.ac.at (M. Romauch), andreas.klemmt@infineon.com (A. Klemmt).

http://dx.doi.org/10.1016/j.cor.2014.04.017
0305-0548/

prevents to model certain production flexibilities. Multiple periods are considered in both approaches, but the cycle time of the individual products is an input parameter since it is captured in a predefined shift of the capacity consumption.

In contrast to that, we employ a static capacity model that includes all tools and allows to model tool dependent processing times (cf. [14]). Furthermore we will also explain how to model multiple periods and how to link the periods to balance the load between periods.

We assume that the profit arising from each product is linear in the number of units produced, and we will formulate a linear program to model the optimal product mix problem for the steady state system, that is based on the idea of heavy traffic resource pooling [14].

The underlying capacity model balances the load and minimizes the capacity usage for bottlenecks. The capacity limits are determined in advance (in practice, notions like the operating curve [1] and other performance indicators can be used to estimate the limits) therefore cycle times are not addressed directly, but in practice the capacities are properly chosen and the proposed load distribution policy leaves sufficient freedom to absorb short term disruptions encountered in daily business. According to that, the approach combines the product mix decision and load optimization, i.e. we are maximizing the profit while optimizing the load within the given capacity limits.

In [19] you can find a decomposition method for solving the product mix problem (compare Algorithm 1 in Section 4) where the product mix problem is solved on equipment basis in the master problem. It turns out that this method leaves considerable optimization gaps for certain types of instances. The approach presented in this paper aims to improve these results by integrating the resource pools concept in the decomposition.

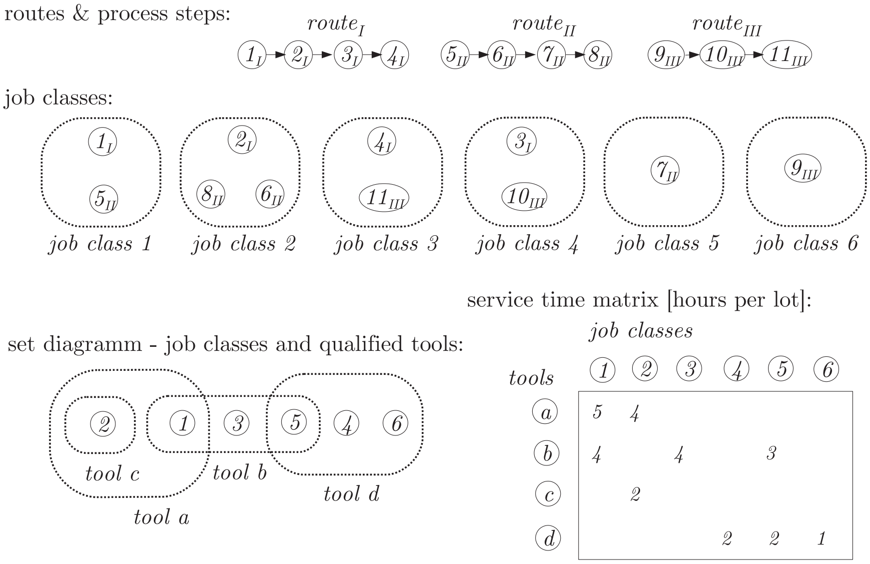
We note that uncertain demand will not be discussed extensively in this paper, but we indicate that stochastic programming and sensitivity analyses are apparent ways to deal with random demand. A demonstration of the sensitivity to changes in the product mix is contained in Section 7 which covers practical experiences.

For robust planning approaches we refer to [16] and [4] (objective: demand satisfaction and minimizing the number of tools) and in [12] you can find a minimum cost formulation that considers noisy demand that is related to [23].

The paper is organized as follows. In Section 2 we discuss the static capacity problem and we introduce the basic concepts of load balancing, connected components and resource pools. In Section 3 we define a (global) linear program for the optimal product mix problem. In addition, we discuss an extension of this model. In Section 4 we propose a decomposition approach for the (global) product mix problem which originates in the approach presented in [19]. Here we iteratively solve small-sized linear sub problems that decouple the product mix and the resource allocation problem. To evaluate both methods we are considering a set of randomly generated instances. For this purpose, we describe a benchmark scheme in Section 5 and we present the results of our computational experiments in Section 6. We conclude with a discussion of practical experiences from semiconductor industry in Section 7, including a demonstration of select use cases that give an outlook to closely related problem settings.

2. Load balancing, connected components and resource pools

To solve the product mix problem we need to check the expected utilization, which is covered by solving the corresponding capacity problem. In this section we will briefly introduce the notions of the static capacity problem and the concept of load balancing and resource pooling. For more details on capacity planning in semiconductor industry we refer to [5,9]. For the theoretic concept of resource pooling in high traffic we refer to [10,14,15]. We begin with the explanation of an example which illustrates the problem of capacity planning. Suppose, that we are considering a lithography work center with four steppers $\{a, b, c, d\}$ with different setups and processing speeds (different technological generations) while ignoring other work centers. Fig. 1 gives some details of the structure of three routes that consist of particular process steps. We suppose that we already identified the so-called ‘job classes’ to aggregate those processes that are ‘very similar’ – and ‘very similar’ means that the processes of the same job class need to be qualified on the same tools and the corresponding process time can be regarded as identical from a capacity planning point of view. We can also find the service time matrix on level job class and resource in Fig. 1. The units for service time are given in hours per lot and only positive entries can be interpreted as a qualified job class and tool combination. According to the set diagram the job classes are connected and the problem cannot be decomposed. We can observe that the



tools	1	2	3	4	5	6
a	5	4				
b	4		4		3	
c		2				
d				2	2	1

Fig. 1. Example: structure for three routes, job classes and qualified tools.