International Journal of Production Economics (xxxx) xxxx–xxxx

Contents lists available at ScienceDirect

Int. J. Production Economics

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

Capacity planning for cluster tools in the semiconductor industry

Martin Romauch[b,*], Richard F. Hartl[a]

[a] Department of Business Administration, University of Vienna, Oskar-Morgenstern-Platz 1, 1090 Wien, Austria
[b] University of Applied Sciences Wiener Neustadt, Johannes Gutenberg-Straße 3, 2700 Wiener Neustadt, Austria

ARTICLE INFO

Keywords:
Cluster tool
Capacity planning
Semiconductor industry
Linear programming
Duality

ABSTRACT

This paper proposes a new model for cluster tools with two load locks. Cluster tools are widely used to automate single wafer processing in semiconductor industry. The load locks are the entry points into the vacuum of the cluster tool's mainframe. Usually there are two of them available. Each lot being processed is dedicated to a single load-lock. Therefore at most two different lots (with possibly different processing times and qualification) can be processed simultaneously. This restriction is one of the major potential bottlenecks.

Capacity planning is one of the possible applications for the proposed model and the paper demonstrates the integration into a more general framework that considers different tool types and different operational modes.

The paper also generalizes an earlier model that is limited to three processing chambers. The proposed modeling approach is based on makespan reductions by parallel processing. It turns out that the performance of the new approach is similar, when compared to the generalized model for three chambers, but the new approach computationally outperforms the generalized model for four and more chambers.

1. Introduction

This paper considers cluster tools, a tool type that can be found in semiconductor fabrication front-end sites, also called wafer fabrication facilities. The wafer is the substrate used for the fabrication of integrated circuits for semiconductor devices. A wafer is a disc, usually a slice of mono-crystalline silicon with a diameter between 150 and 300 mm, and a thickness between 0.5 and 1 mm.

The input of front-end facilities are containers of wafers (also called lots or FOUPs, Front Opening Unified Pods). A lot usually consists of 20 or more wafers that require the same processing. The fabrication is characterized by a large number of loops or layers (compare Hutcheson (2000)). Each loop consists of one or more processing steps; one of them is a lithographic process that allows to cover parts of the wafer (photo resist) from the process that follows. Fig. 1 gives a schematic view on wafer fabrication.

The flow diagram in Fig. 1, is a representation of the fabrication process as a re-entrant flow shop. According to Thiesse and Fleisch (2008) the traversed path of a single lot can be several kilometers long.

In semiconductor industry most of the equipment is automated and characterized by load/unload operations, robot handling, testing, alignment, cooling and much more. Cluster Tools (compare Franssila (2010)) for instance, have one or more load locks and multiple processing chambers. Cluster tools are used to automate several process types. The internal software defines the behavior of the system. Cluster tools allow an automatic transfer of wafers between load ports and process chambers with possibly different processes inside a vacuum. The system can also be used for parallel processing to increase throughput and productivity. The wafers are transferred between chambers under a vacuum using a robotic arm to prevent exposure to air to avoid oxidation and contamination.

Since a cluster tool may operate on different lots at the same time, the combination of lots is important and the cluster tool may show different operation cycle times for the same type of lot. According to Mönch et al. (2011) several researchers studied the optimization of internal scheduling, but the corresponding models are not suitable to consider several cluster tools at the same time. One of the major bottlenecks of a cluster tool are the load locks (see Christopher (2008)). The load locks are the entry and exit point into the vacuum of the cluster tool.

In order to describe a cluster tool with respect to load locks and chambers, a simplified schedule for a cluster tool that is working in parallel mode is discussed in Example 1. The example is based on Fig. 2.

Example 1. On the left hand side of Fig. 2, a cluster tool that is working in parallel mode with two load locks LL_1 and LL_2 is illustrated. In this example, the chambers are supposed to be identical. Later in the example, one of the chambers will be less flexible than the others. In the beginning (t=0), each load lock is occupied by one lot (LL_1 with lot

* Corresponding author.
E-mail addresses: martin.romauch@univie.ac.at (M. Romauch), richard.hartl@univie.ac.at (R.F. Hartl).

http://dx.doi.org/10.1016/j.ijpe.2017.01.005
Received 29 April 2016; Received in revised form 3 January 2017; Accepted 4 January 2017

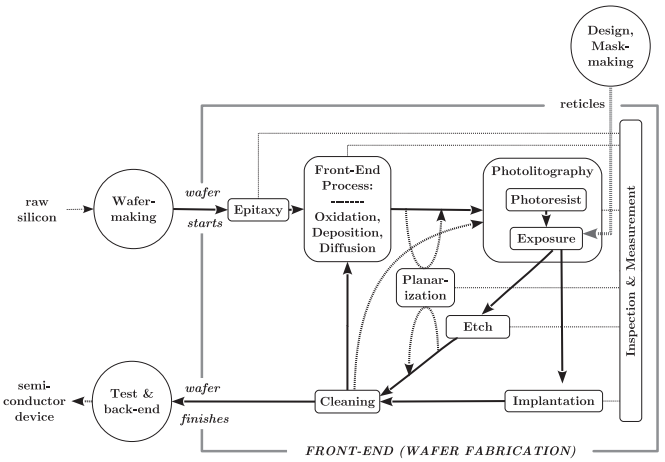

Fig. 1. Front end fabrication as a re-entrant flow shop (based on Sorenson (1999) and Mönch et al. (2011)).

L_1 and LL_2 with lot L_2), and each lot consists of three wafers. Each wafer needs to be processed in one of the chambers. The load lock LL_1 is occupied with the lot L_1 that needs the longer processing time. More precisely, each wafers from L_1 need a processing time of six time units. For L_2 the processing time is five time units per wafer. In this example the wafer handling time is neglected, therefore it is possible to start all processes in each chamber at the same time.

In the first step (i) with t=0, two wafers from lot L_1 are assigned to chambers A and B; and one wafer from L_2 is assigned to chamber C. At t=5 the wafer in chamber C is finished and it is passed back to L_2. Then, the second wafer from L_2 is assigned to chamber C. At t=6 the wafers in A and B are handed back to L_1 and chamber A will be occupied by the last wafer of L_1; chamber B will be occupied by the last wafer of L_2.

The corresponding Gantt-chart can be found on the right hand side of Fig. 2. The make-span is 12 with an idle time percentage of $\frac{1}{12}$ on average. If wafers from L_1 are excluded from chamber C then at least two wafers from L_1 need to be assigned to the same chamber (A or B), hence the schedule is optimal with respect to the make-span.

Now, additionally suppose that the problem is scaled by thirty, and each lot counts 90 wafers. Then the wafers can be distributed in such a way that the idle time percentage vanishes. If this is possible, then the total processing time is $90\cdot(6+5)=990$. Therefore it is sufficient to find a distribution where each chamber finishes after 330 time units. Note that 330 is divisible by five and six:

$$330 = \overbrace{55\cdot 6}^{A} = \overbrace{35\cdot 6 + 24\cdot 5}^{B} = \overbrace{66\cdot 5}^{C}$$

Therefore 55 wafers from L_1 can be assigned to chamber A and 66 wafers of L_2 can be assigned to chamber C. The remaining wafers can be assigned to chamber B.

This example shows that on a larger scale where larger quantities are considered, a continuous approximation - where it is allowed to computationally "split" wafers - leads to reasonable results.

LP based cluster tool models can be directly integrated in capacity models or in master planning like proposed in Ponsignon and Mönch (2012) and Romauch and Klemmt (2015). The combination of simulation and optimization (LP) as discussed in Almeder et al. (2009); Gansterer et al. (2014) and Juan et al. (2015) is also suitable for the integration of LP based cluster tool models.

Finally, real-time dispatching that integrates LP approaches can be found in Doleschal et al. (2013) and Ham et al. (2009) which can be extended to LP based cluster tool models. According to Duemmler and Wohlleben (2012), there are also various alignments necessary to achieve an effective WIP Flow Management, that assures that the planned capacity consumption is coherent with the dispatching reality. Therefore, improvements of LP based cluster tool models are important to several areas.

This paper concentrates on a specific static capacity planning problem, but serves as an example to demonstrate that the Cluster Tools model can be integrated into larger capacity planning models that consider various equipment types, including different types of Cluster tools. The considered objective function is linear, but the proposed Cluster Tool model is also suitable to be integrated in quadratic programming formulations that address batch server efficiency Gold (2004). Furthermore, robust capacity planning models like discussed in Barahona et al. (2005) where recourse decisions are covered in a two stage stochastic linear program are also a suitable framework. The considered static resource allocation problem also appears as a subproblem in Simulation (Almeder et al. (2009); Bang and Kim (2010); Romauch and Klemmt (2015)), master planning Romauch and Klemmt (2015) and also in dispatching and scheduling (Doleschal et al. (2012)).

In the context of capacity planning (see Hopp and Spearman (2011); Bermon and Hood (1999)) one of the major tasks is a concise prediction of the equipment utilization. This kind of calculation is meant to confirm that the capacity restrictions for a given demand (requested wafer starts for given product routes) is met and the identification of bottlenecks is one of the major outputs. Besides that, for volatile demands like in semiconductor industry, capacity expansion and down sizing are frequent measures that can be supported by

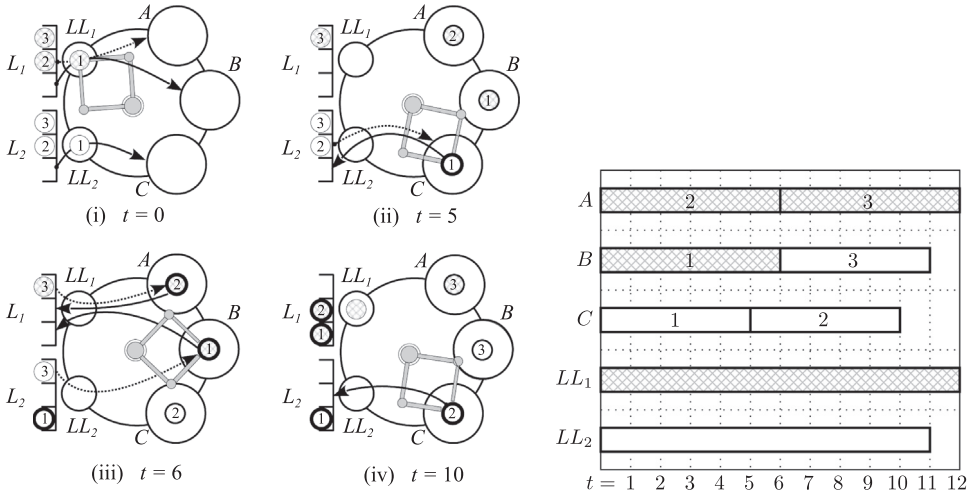

Fig. 2. A cluster tool with two load locks that is processing wafers in parallel. On left hand side: four snapshots of the cluster tool for t = 0, 5, 6, 10 can be found. The thickness of the outline of the wafers, indicates the processing progress. The corresponding Gantt-chart can be found on the right hand side of the Figure.