Computers & Industrial Engineering 65 (2013) 537–543

Contents lists available at SciVerse ScienceDirect

Computers & Industrial Engineering

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

On the complexity of short-term production planning and the near-optimality of a sequential assignment problem heuristic approach ☆

Adar Kalir [a,]*, Yonatan Zorea [b], Adir Pridor [c], Lev Bregman [c]

[a] Fab/Sort Manufacturing, Intel Corporation, Qiriat-Gat, Israel
[b] Manufacturing Systems Group, Micron, Qiriat-Gat, Israel
[c] Institute of Industrial Mathematics, Beer-Sheva, Israel

ARTICLE INFO

Article history:
Received 3 April 2012
Received in revised form 31 December 2012
Accepted 4 May 2013
Available online 11 May 2013

Keywords:
Assignment problem
WIP management
Optimization
Semiconductor manufacturing

ABSTRACT

In semiconductor manufacturing, the process of short-term production planning requires setting clear and yet challenging and doable goals to each operation and toolset in the process flow per each product type. We demonstrate the complexity of this problem using an experimental study performed with proficient workforce, and then show how the problem can be decomposed, aggregated, and solved using sequential recurrent linear programming assignment problems. We also refer to the improvements that the proposed algorithm has achieved in practice when applied to multiple semiconductor production facilities, and discuss its efficiency and uniqueness as a fast heuristic relative to other proposed methods.

1. Introduction

Today, in the semiconductor industry, more and more fabs run multiple processes with a diverse range of products (Bang, An, Kim, & Lim, 2005). Since a majority of the resources (equipment and head-count) are shared amongst the various products, the need for short-term production planning to efficiently meet on-time delivery and output goals becomes a complex continuous process of resource allocation planning for each toolset (or machine group) throughout all processes, operations, and products/lots (Zorea, Perez, Pridor, & Bregman, 2003).

In the late 1990s, drum-buffer-rope (DBR) based policies were attempted at the problem, which proved to work very well for regular production lines and systems. For example, Sivasubramanian, Selladurai, and Rajamramasamy (2000) report on a successful case study that has been taken up in a small-scale industry, and analysis has been carried out on the positive effect of the DBR approach on the performance of the system. However, even without the complexity emanating from the product mix environment, the highly re-entrant process in semiconductor manufacturing systems dictates that the same machines perform competing (different) operations on the lots as they are processed through the fab and consequently, implementing drum-buffer-rope is difficult and complex in re-entrant flows, because several bottleneck operations of each lot will appear on the drum in various locations (Wu & Yeh, 2006).

A number of specific dispatching rules have been widely applied to determine the best strategy of machine allocations to lots in process at the time of execution, such as 3–2–1, also known as Back-To-Front (Sohn & Kempf, 1996) and Critical Ratio (Subramaniam, Ming, & Mohan, 2005). A combination of these dispatching rules has also been evaluated (Dabbas & Fowler, 2003). However, attempting to solve an underlying short-term planning problem implicitly with dispatching rules, at factories with a range of several technologies, multiple products per technology, and hundreds of operations in the process flow – is almost set to failure, since these rules are by definition limited in the consideration of pertinent information for making the best decision.

In this paper, we illustrate the complexity of the short-term production planning decision problem, using a re-entrant line consisting of only two products and eight operations over a planning horizon of 1 week (fourteen shifts), where performance is measured on output and on-time delivery. We report the results of an experimental study that was performed with this line, with many proficient participants from across the semiconductor manufacturing industry that deal with this problem on a daily basis. The experimental study evaluates the goodness of solutions based on common practices such as 3–2–1, critical ratio, fast-box and others in obtaining near-optimal solution to the problem.

☆ This manuscript was processed by Area Editor Subhash C. Sarin.
* Corresponding author. Tel.: +972 54 7886932.
E-mail address: adar.kalir@intel.com (A. Kalir).

0360-8352/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cie.2013.05.005

We then compare these results with those obtained by a unique algorithm proposed for this problem. The algorithm is based on decomposition of the short-term production planning problem into a sequence of assignment sub-problems (of available machines to available lots), each solved to optimality by itself and, when aggregated together, form a comprehensive heuristic solution.

Similar to the approach by Qiu (2005), we present a practical solution to the problem while also addressing issues such as real-time performance, scalability, and re-configurability. Unlike in Qiu, which is based on distributed WIP control, the proposed algorithm is a fast and efficient heuristic that has easily been applied in real-time for large-scale semiconductor manufacturing facilities. The proposed algorithm has been demonstrated via multiple implementations in high volume production facilities, and we also reference the improvements that this algorithm has achieved in practice once it was implemented in these facilities.

2. Experimental study

In most factories, production line complexities have made it near impossible for humans to make optimal decisions regarding WIP management. This is a direct consequence of the fact that the magnitude of the problem is immense, in terms of the number of decision variables and constraints involved. In a typical high volume semiconductor manufacturing process, there exist hundreds of production steps (or operations), that compete over hundreds of different toolsets, shared across dozens of product types with different processing routes, process times, etc. Additionally, these resources may be subject to limitations such as restrictions on specific tools, dedications schemes, rework and special handling of specific lots in process and so forth.

As a result, without tackling the underlying problem explicitly and efficiently, these factories experience reduced performance on output and schedules. A more systematic approach is needed to comprehend the schedule requirements and variability to produce the best performance results possible.

2.1. Description of the experiment

To illustrate the complexity of the problem, as well as to demonstrate the principles that will improve output and schedule delivery performance, an interactive game has been developed. The target audience for this game includes manufacturing, planning, industrial engineering, and process engineering personnel from fab, sort, and assembly operations. The purpose in the game is to meet demand while maximizing output by minimizing misses against demand and maximizing line moves (=sum of output from all operations in the process flow).

The game is played as follows. A board, depicted in Fig. 1, with eight operations (or process steps) denoted from 1 to 8, represents the process flow for two product types, differentiated via chips of different color: purple–green (PG) and yellow–red (YR) (see Fig. 2). Machines for each operation are represented by cylinders with different heights, representing their capacities per shift, and different symbols, representing their operational capabilities (Fig. 3). There are a total of 5 machines: two machines with capacities of 3 and 4 (heart symbol) can perform operations 1, 3, and 7;

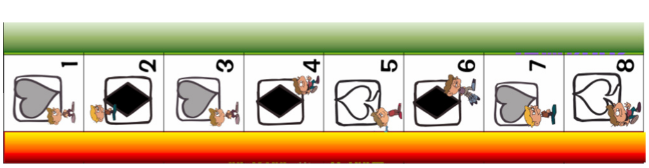

Fig. 1. Game board: simulation of 8-operations re-entrant production line.

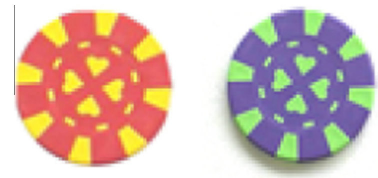

Fig. 2. Chips: simulating WIP of two products.

one additional machine with capacity of 8 (diamond symbol) can perform operations 2, 4, and 6; and two additional machines with capacities 2 and 3 (spade symbol) can perform operations 5 and 8. The rules of the game are:

- During each shift, a machine may be assigned to only 1 operation from 1 product type.
- Machines may only perform operations of their color and symbol.
- WIP movement to next operation is synchronized at end of shift, after all machines have been assigned to operations.
  - Starts are also introduced into operation 1 at end of shift.
- Misses are cumulative, i.e. a miss in shift *i* must be reported as a miss in all subsequent shifts, until closed.
- Machines are subject to downtime; they cannot be utilized in a shift if they are down.
- A machine cannot be split across two operations during the same shift, nor can it be split across two products at the same operation. However, two machines can be assigned to the same product at the same operation.

The game is totally deterministic, i.e. all the information is known and pre-determined in advance. Initial WIP position is setup prior to the start of the game (Fig. 4). The players are given a table before the game begins with all relevant data. For each shift and product type, the following is information provided (see Table 1):

- Starts (or raw materials) into operation 1, at end of shift.
- Demand of each product type at end of shift (against which misses are measured.)
- Tools down during the shift

The game is played for a total of 14 shifts. At the end of each shift, WIP is moved from each operation to the next per the machine assignments; then moves and misses are recorded on a scoring sheet (see Table 2). The players are requested to do their best on both moves and misses. The final performance score is calculated as the total number of moves over all shifts minus 10 times the total number of misses overall shifts (i.e. a weighted objective function).

2.2. Experiment results and conclusions

The game was played numerous times by homogeneous and heterogeneous populations from across the semiconductor manufacturing industry. Teams of sizes ranging between 5 and 8 participants, consisting of product planners, industrial and process engineers, shift managers have all attempted to do their best in this game. However, the average performance of these teams has consistently been in the range of 20–30% below the optimal solution, and the best performances were in the range of 15–20% below the optimal solution (Kalir et al., 2004) – a significant gap in terms of production system performance.

Various strategies were utilized by the teams in the game, but the most popular ones that were observed are the 3–2–1 (or back-to-front) strategy and the 'maximum output' strategy. In 3–2–1, priority is always given to WIP that is closer to end of line,