Computers & Industrial Engineering 90 (2015) 86–106

Contents lists available at ScienceDirect

Computers & Industrial Engineering

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

Improving performance of dispatch rules for daily scheduling of assembly and test operations

Shihui Jia[a], Jonathan F. Bard[a,*], Rodolfo Chacon[b], John Stuber[c]

[a] ORIE, University of Texas, Austin, TX 78712, United States
[b] Texas Instruments, WW TMG IT, 6550 Chase Oaks Blvd, Plano, TX 75023, United States
[c] Texas Instruments, DMOS6, 12500 TI Blvd, Dallas, TX 75243, United States

ARTICLE INFO

Article history:
Received 11 May 2015
Received in revised form 30 July 2015
Accepted 27 August 2015
Available online 7 September 2015

Keywords:
Machine setup
Reentrant flow
Semiconductor assembly and test
AutoSched
GRASP

ABSTRACT

In recent years, there has been an increasing effort to improve the performance of semiconductor assembly and test facilities given their critical role in achieving on-time delivery. Using the simulation package AutoSched AP (ASAP) as the analytic tool, the goal of this paper is to show how the logic of intelligent heuristics can be combined with discrete event simulation to evaluate various dispatch rules for machine setup and scheduling in such facilities. The problem addressed is defined by a set of resources that includes machines and tooling, process plans for each product, and four hierarchical objectives: minimize the weighted sum of key device shortages, maximize weighted throughput, minimize the number of machines used, and minimize makespan for a given set of lots in queue.

Three new dispatch rules are presented for configuring machines and assigning lots to them in assembly and test facilities. The first gives priority to hot lots containing key devices while using the setup frequency table obtained from our machine optimizer that takes the form of a greedy randomized adaptive search procedure (GRASP). The second embeds the more robust selection features of GRASP in the ASAP model through customization. This allows ASAP to explore a larger portion of the feasible region at each decision point by randomizing machine setups using adaptive probability distributions that are a function of solution quality. The third rule, which is a simplification of the second, always picks the setup for a particular machine that gives the greatest marginal improvement in the objective function among all candidates. The computational analysis showed that the three dispatch rules greatly improved ASAP performance with respect to the four objectives.

1. Introduction

Semiconductor devices are manufactured from wafers in a fabrication facility or fab in what are called *front-end* operations. After fabrication, the wafers are sent to an assembly and test (AT) facility where they are cut into chips, packaged, and tested in what are called *back-end* operations. During this process, a predefined sequence of steps is followed that involves several different machines and tooling pieces. In recent years, there has been an increasing effort to model and study back-end operations given their critical role in meeting customer demand and improving customer satisfaction.

When scheduling AT operations, the goals are to achieve low cycle times, high throughput and high utilization without violating agreed upon delivery dates. The first attempt to use optimization technology to achieve these goals was undertaken by Deng, Bard, Chacon, and Stuber (2010) who formulated the scheduling problem as a mixed-integer program (MIP) with the following four objectives given in order of priority: (1) minimize the shortage of critical devices, (2) maximize the weighted throughput of the remaining lots, (3) minimize the number of machines used, and (4) minimize the makespan. Solutions were obtained with a reactive greedy randomized adaptive search procedure (GRASP) designed to examine a diversity of machine-tooling combinations and lot assignments over many iterations [the literature on GRASP is extensive; e.g., see Festa and Resende (2009) for an annotated bibliography of algorithms, and Feo, Venkatraman, and Bard (1991) and Monkman, Morrice, and Bard (2008) for manufacturing applications]. In the original version of the model, three assumptions were made by Deng et al. that are rarely true in the real manufacturing environment: (1) machines could be set up only once over the planning horizon; (2) all the machines have

* Corresponding author.
E-mail addresses: sjia@utexas.edu (S. Jia), jbard@mail.utexas.edu (J.F. Bard), rchacon@ti.com (R. Chacon), stuber@ti.com (J. Stuber).

http://dx.doi.org/10.1016/j.cie.2015.08.016
0360-8352/

to start from idle; and (3) only the upcoming step in the *route* of the device needs to be considered.

In a follow-on paper, Bard, Gao, Chacon, and Stuber (2013) presented an enhanced methodology for dealing with changeovers, setups at time zero, and the multi-pass requirements associated with a route. To decide on the best machine-tooling configurations and how to assign lots to machines, a three-phase heuristic was implemented. As a real-time alternative tool to the GRASP, Bard, Jia, Chacon, and Stuber (2015) developed a deterministic discrete event simulation model using AutoSched AP (ASAP) that similarly schedules AT lots over a given planning horizon. The built-in rules in ASAP performed poorly compared to the enhanced GRASP, though, so three new dispatch rules were formulated. The first integrates the GRASP results with the ASAP logic, and the second and third give increased priority to "hot lots," which are defined as those lots containing critical or key devices associated with production targets.

The customization feature in ASAP was used to implement the new rules. The specific motivation for combining the best features of the two approaches was several fold. First, the standard rules in ASAP are inherently myopic in that set up and dispatch decisions reflect the best choice for each machine at the current point in time. In contrast, GRASP makes decisions in full view of system capacity and prioritized demand for the entire planning horizon. Second, ASAP provides one solution while GRASP makes repeated runs to explore a large portion of the feasible region. Third, ASAP handles the multiple-pass (reentrant flow) requirements of a lot easily and more efficiently because it updates the unassigned lot list when the first pass of a lot finishes. In contrast, the enhanced GRASP only starts to process subsequent passes when all the first passes of lots that require the same setup are completed.

With this in mind, the primary purpose of this paper is to (1) further customize ASAP rules by taking advantage of the type and frequency of machine setups recommended by GRASP results, (2) evaluate and compare the effectiveness of the various dispatch rules for machine setup and scheduling within ASAP, and more generally, (3) to demonstrate how to combine the logic of intelligent heuristics with discrete event simulation.

In the next section, we present the most recent literature on dispatch rules and reentrant flow models aimed at improving performance in the semiconductor industry. In Section 3, the enhanced GRASP and the discrete event simulation that were applied previously to schedule AT operations are reviewed. The three new scheduling rules are introduced in Section 4. Section 5 contains a comparative study, which was conducted with both real data sets obtained from the sponsoring company and randomly generated data sets based on the former. In all cases, the data were scaled and the machines renamed to avoid revealing true production capacity. Insights gained from the analysis are discussed in Section 6.

2. Literature Review

For an overview of dispatch rules typically applied in the semiconductor industry, see Atherton and Atherton (1995). Wu, Chiou, and Chen (2008) developed a dispatching algorithm that tries to balance the output rate of each product segment with the goal of improving on-time delivery for a make-to-order semiconductor wafer fab. The showed that the algorithm outperformed the scheduling procedures favored by the company on 10 test scenarios with respect to on-time delivery rates and cycle times. Saito (2007) proposed a pseudo periodical priority dispatching (P3D) rule for dynamic allocation of WIP in mixed products semiconductor manufacturing. The P3D rule evaluated both the amount of WIP and the arrival rate of lots for each quantum, where a quantum is defined as a period during which a single type of product is processed on a machine. Results comparing P3D with first-come, first served logic, and the shortest processing time rule for simulated data with Poisson arrivals showed that P3D uniformly outperformed the other rules in terms of adjustment rate, throughput, response time, and tardiness.

For scheduling semiconductor back-end operations, Chiang, Guo, and Pai (2008) introduced a fuzzy analytical hierarchy process to identify acceptable WIP deviation levels, which were then used to determine job priorities. The approach was shown to balance on-time delivery goals and WIP targets with the help of a simulation model that was calibrated with real data. Fu et al. (2011) presented a MIP model and a deterministic scheduling system (DSS) to minimize prioritized tardiness for the weekly production scheduling of a semiconductor back-end facility. Depending on customer orders, the DSS uses either a linear programming optimizer or a material-requirements-planning optimizer along with one of two scheduling rules: dynamic lot prioritization or dynamic machine prioritization, for finding schedules. The results were consistent and satisfactory from management's point of view, and required less solution time for randomly generated large problem instances than the MIP formulation. Related research in a job shop environment was undertaken by Sels, Gheysen, and Vanhoucke (2012) who compared 30 rules under two flow time-related and three tardiness-related objectives.

For a single product, Narahari and Khan (1996) proposed an approximation method for predicting the performance of heuristics for scheduling reentrant flows based on mean value analysis (MVA). They modeled reentrant lines with buffers as a nontraditional queuing network and were able to show that MVA was better than simulation-based methods with respect to accuracy and time complexity. One shortcoming of their approach was the need to treat each machine as a unique family, so they couldn't take advantage of situations in which some machines were identical. To address the more general case, Park, Kim, and Jun (2002) considered a facility that processed multiple products using multi-servers, where each server consisted of one or more identical machines. Choi, Kim, and Lee (2011) proposed a decision tree-based real-time scheduling mechanism for the reentrant hybrid flow shop scheduling problem. A decision tree was created using four attributes related to the jobs in the queue; the extremities of the tree contained the proposed dispatching rule of which one was identified as being the best through the roll-up logic. Testing showed that the approach led to higher throughput in less time when compared to discrete event simulation.

Freed, Doerr, and Chang (2007) developed a dispatcher within an Excel-VBA decision support system. The dispatcher takes current WIP data and sorts it based on due-date and processing requirements, and gathers feedback from managers to prioritize the use of resources before providing the operators with the final schedule. Testing showed that on-time delivery increased from 70% to 90% and lot lead time was reduced by 30% due to the dispatcher. Knutson, Kempf, Fowler, and Carlyle (1999) proposed a method for deciding the lot assignments on a given day with the overall goals of maximizing on-time delivery and minimizing excess product that had to be stored. The problem was formulated as nonlinear integer program with three objectives: maximize the number of die sent to the customers, minimize the number of die sent to the warehouse, and meet due date requirements for orders. A two-stage decomposition approach was used to find solutions. Stage 1 consisted of a knapsack problem whose objective was to maximize a combination of factory utilization and on-time-delivery, while Stage 2 was a modified bin covering problem in which the orders represented variable size bins. A first-fit-decreasing (FFD) heuristic with order sizes modified by their due date was used in Stage 1. The results were used to fill orders one