INTEGRATED SCHEDULING OF JOBS, RETICLES, MACHINES, AMHS AND ARHS IN A SEMICONDUCTOR MANUFACTURING

Andy Ham
School of Industrial Engineering
Purdue University
West Lafayette, IN 47907

Myoung-Ju Park
Industrial & Management Systems Engineering
Kyung Hee University
Global Campus
Yongin-si, South Korea

Ho-Jun Shin
Research and Development
CSPI
Suwon-si, South Korea

Si-Young Choi
Research and Development
CSPI
Suwon-si, South Korea

John W. Fowler
W. P. Carey School of Business
Arizona State University
Tempe, AZ 85287-5906

ABSTRACT

This paper studies simultaneous scheduling of production and material transfer in the semiconductor photolithography area. In particular, jobs are transferred by a material handling system that employees a fleet of vehicles. Reticles serving as an auxiliary resource are also transferred from one place to another by a different set of vehicles. The extremely complex scheduling problem that includes jobs, reticles, machines, and two different sets of vehicles is (apparently) studied for the first time. A novel constraint programming model is proposed.

1 INTRODUCTION

The photolithography (litho hereafter) operation imposes an additional operational complexity, mainly stemming from reticles that serve as an auxiliary resource to print a circuit pattern on a wafer. Namely, the litho operation requires both a production machine and a step-specific reticle to process jobs. Furthermore, jobs and reticles are transferred by an automatic material handling system (AMHS) and an automatic reticle handling system (ARHS), respectively, in a modern semiconductor fabrication line. Simultaneously orchestrating reticles, machines, AMHS, and ARHS for a given set of jobs has been a great challenge to practitioners and researchers. Industrial scheduling systems generate a Gantt-chart schedule of jobs, machines, and reticles, generally ignoring material handling systems, thus often causing one of the following conditions: idle-w/o-job or idle-w/o-reticle. The former refers to a situation where the machine experiences idle time waiting for the job to arrive and the later explains idle time waiting for the reticle to arrive. Both situations can be avoided if all resources are seamlessly coordinated together. Clearly, seeking an optimal solution for this problem under study is very challenging.
In this paper, we introduce two constraint programming (CP) models for automated planning and scheduling of the litho operation. The objective is to minimize the time (Cmax) required to process all jobs on machines and return jobs to a stocker.

Figure 1 depicts such a system. \(V_{job}\) (AMHS) navigates to stocker to pickup a job and drops it off to machine. Later, another (or the same) \(V_{job}\) visits the machine to pickup the completed job. Similarly, \(V_{reticle}\) (ARHS) drops off the appropriate reticle to the same machine. Later, another (or same) \(V_{reticle}\) visits the machine to pickup the reticle. Each transfer task requires two sub-tasks: pickup (approaching a source where a job is located) and dropoff (moving to a destination where a job will be staged).

![Figure 1: An illustration of transfer tasks in a photolithography area.](image)

Table 1 contains the selected articles related to design, dispatching, and integrated scheduling of the AMHS and the ARHS. Integrated scheduling (also known simultaneous scheduling) refers to the scheduling of vehicles in addition to scheduling the jobs on machines. These two decisions (vehicle-scheduling and machine-scheduling) are interrelated with each other and must be synchronized to operate manufacturing facilities efficiently. Integrated scheduling has been well studied in a job shop environment. In this integrated approach, vehicles perform a delivery task between two operations. The paper by Drießel and Mönch (2012) seems to be only one that addressed integrated scheduling in the context of semiconductor manufacturing. They proposed a shifting bottleneck heuristic. All the others listed on the table addressed integrated scheduling of a flexible job shop problem (FJSP) in the general manufacturing context.

To the best of our knowledge, there has been no attempt to exactly model concurrently vehicles, machines, and jobs in semiconductor manufacturing. With state-of-the-art modeling capability and computing power, we were able to successfully accomplish this goal. In this paper, we further study the integrated scheduling of jobs, reticles, machines, and two different sets of vehicles (AMHS and ARHS).

Table 1: Selected Articles Related to Design, Dispatching, and Scheduling of AMHS/ARHS.

<table>
<thead>
<tr>
<th>Design</th>
<th>Dispatching</th>
<th>Integrated Scheduling</th>
</tr>
</thead>
</table>
2 PROBLEM DESCRIPTION

The proposed unrelated parallel machine scheduling problem integrated with reticles, AMHS, and ARHS in litho operations can be stated as follows. A set of jobs \((j, J)\) that have to be processed by one of unrelated parallel machines \((m, M)\) with one of reticles \((r, R)\) with the goal of minimizing the makespan. Each job has a different processing time \((p_j)\). Not all reticles and machines are compatible with each job (commonly known as process dedication, track-in-prevent, process qualification in practice). All jobs, reticles, and vehicles are located at a stocker at time 0 (note we can easily set an initial location of each entity). An AMHS vehicle \((v, V_{job})\) performs a two-way transportation (pickup and dropoff) between a stocker and machine for loading and unloading. Similarly, an ARHS vehicle \((v, V_{reticle})\) performs a two-way transportation between a stocker and machine (or machine-to-machine). All vehicles perform two types of trips: a loaded trip and an empty trip (ET). A loaded trip is a delivery operation where the vehicle moves a job/reticle from the output buffer of a location to the input buffer of another location. In an empty trip, the vehicle moves from an idle position without carrying a job/reticle in order to pick up a job/reticle waiting to be transferred. Let \(t_{m,\hat{m}}\) represents the travel time between any two locations \(m\) and \(\hat{m}\). Transportation operations of both job and reticle are required before a given machine can proceed with actual processing.

A Gantt chart of the integrated scheduling of machines, reticles, \(V_{job}\), and \(V_{reticle}\), generated by the proposed CP model, is represented on Figure 2. For instance, J4 and R1 are transferred to M2 by \(V_{job}\) and \(V_{reticle}\) and arrive at time 2, respectively. \(V_{job}\) returns to stocker (empty trip: ET) to pick up J1 and transfers it to M1. Similarly, \(V_{reticle}\) returns to stocker to pick up R2 and transfers it to M1. After all jobs are processed, \(V_{job}\) transfers J5 from M2 to a stocker; completing at time 38. The parameters (traveling distances and route) used for this schedule are included in the same figure. The \(<j, m, r, pt>\) refers to <job, machine, reticle, processing time>.

![Figure 2: A sample of integrated scheduling of 5-job, 2-machine, 2-reticle, 1-job vehicle, and 1-reticle vehicle.](image)

We assume the same reticle cannot revisit the same machine multiple times during the planning horizon. Once a reticle is being used on a machine, the reticle may stay at the same machine or must be transported to another machine. However, the reticle cannot leave the machine and return back to the same machine. So far, we have been unable to adequately capture this scenario. This assumption seems to be acceptable to practitioners.

3 SOLUTION

The notation used in this paper is summarized in the following:

- Sets and parameters:
- \( M \): set of machines
- \( L \): set of locations (machines and stocker)
- \( J \): set of jobs
- \( O \): set of operations in a tuple structure \( \{i, m, r\} \)
- \( p_j \): processing time of job \( j \in J \)
- \( t_{i,l} \): vehicle travel time between \( l \) and \( l' \) \((l, l' \in L)\)
- \( T \): collection of \( t_{i,l} \) in a tuple structure \( \{l, l', t_{i,l}\} \)

**Decision variables:**
- \( J_{j \in J} \): interval variable representing job \( j \) with a size of \( p_j \)
- \( O_{o \in O} \): optional interval variable representing actual operation \( o \)
- \( J_{j \in J}^{\text{load}} \): interval variable representing loading on machine of job \( j \)
- \( Load_{j \in J} \): optional interval variable representing actual loading on machine of job \( j \)
- \( J_{j \in J}^{\text{unload}} \): interval variable representing unloading on machine of job \( j \)
- \( Unload_{j \in J} \): optional interval variable representing actual unloading on machine of job \( j \)
- \( RetPick_{j \in J} \): optional interval variable representing reticle pickup
- \( RetPick_{j \in J} \): optional interval variable representing reticle pickup by vehicle \( v \)
- \( RetDrop_{j \in J} \): optional interval variable representing reticle dropoff
- \( RetDrop_{j \in J} \): optional interval variable representing reticle dropoff by vehicle \( v \)
- \( Seq\text{Machine}_{m \in M} \): sequence variable representing all permutation sequences of the interval variables assigned to the sequence on each machine. It is necessary to add a tracer to keep track of the last reticle used. This is achieved by associating \( o.r \) as a type to this sequence variable.
- \( Seq\text{Reticle}_{r \in R} \): sequence variable representing all permutation sequences of the interval variables assigned to the sequence on each reticle. It is necessary to add a tracer to keep track of the last location where the reticle used. This is achieved by associating \( o.m \) as a type to this sequence variable.
- \( Seq\text{Job}_{v \in V} \): sequence variable representing all permutation sequences of the interval variables assigned to the sequence on each vehicle \( v \in V \).
- \( Seq\text{Reticle}_{v \in V} \): sequence variable representing all permutation sequences of the interval variables assigned to the sequence on each reticle \( v \in V \).
- \( Seq\text{Machine}_{m \in M} \): sequence variable representing all permutation sequences of the interval variables assigned to the sequence on each station.
- \( Seq\text{Vehicle}_{v \in V} \): sequence variable representing all permutation sequences of the interval variables assigned to the sequence on each transbot.

Then, the FISP\textsuperscript{transbots} can be formulated into CP as follows:

\[
\begin{align*}
\text{Min } & \max_{j \in J} \{ \text{endOf}(J_{j \in J}^{\text{unload}_j}) \} \\
\text{alternative } & \{ J_{j \in J}^{\text{load}_j}, [Load_{j \in J}^{\text{load}_j}]_{m \in M, r \in R, v \in V_{job}} \} & \forall j \in J \\
\text{alternative } & \{ J_{j \in J}^{\text{unload}_j}, [Unload_{j \in J}^{\text{unload}_j}]_{m \in M, r \in R, v \in V_{job}} \} & \forall j \in J \\
\text{PresenceOf}(O_{j \in J}) = & \sum_{v \in V_{job}} \text{PresenceOf}(Load_{j \in J}^{\text{load}_j}) & \forall j \in J, m \in M, r \in R \\
\text{PresenceOf}(O_{j \in J}) = & \sum_{v \in V_{job}} \text{PresenceOf}(Unload_{j \in J}^{\text{unload}_j}) & \forall j \in J, m \in M, r \in R
\end{align*}
\]
endAtStart(\text{Load}_{jmrv}, J_j)
endAtStart(J_j, \text{Unload}_{jmrv}, t_{m, stk})
presenceOf(O_{jmr}) \geq \text{presenceOf}(\text{RetPick}_{jmr})
presenceOf(O_{jmr}) \geq \text{presenceOf}(\text{RetDrop}_{jmr})
presenceOf(\text{RetPick}_{jmr}) = \text{presenceOf}(\text{RetDrop}_{jmr})
endAtStart(\text{RetPick}_{jmr}, \text{RetDrop}_{jmr})
presenceOf(\text{RetPick}_{jmr}) \rightarrow \text{sizeOf}(\text{RetPick}_{jmr}) =
= t_{\text{trans}} \text{typeOfPrev(seqRetVeh}_{p, \text{RetPickVeh}_{jmrv, stk, stk}), \text{typeOfPrev(seqRetVeh}_{p, O_{jmr}, stk, stk})}
presenceOf(\text{RetDrop}_{jmr}) \rightarrow \text{sizeOf}(\text{RetDrop}_{jmr}) =
= t_{\text{trans}} \text{typeOfPrev(seqRetVeh}_{p, O_{jmr}, stk, stk}) \cdot m
\forall j \in J, m \in M, r \in R, v \in V_{job} \quad (B7)
\forall j \in J, m \in M, r \in R, v \in V_{job} \quad (B8)
\forall j \in J, m \in M, r \in R \quad (B9)
\forall j \in J, m \in M, r \in R \quad (B10)
\forall j \in J, m \in M, r \in R \quad (B11)
\forall j \in J, m \in M, r \in R \quad (B12)
\forall j \in J, m \in M, r \in R \quad (B13)
\forall j \in J, m \in M, r \in R \quad (B14)
\forall j, j \in J, m \in M, r, r \in R \quad : j < > j \& r < > r \quad (B15)
\forall j \in J, m \in M, r \quad (B16)
\forall j \in J, m \in M, r \quad (B17)
\forall j \in J, m \in M, r \in R, v \in V_{reticle} \quad (B18)
\forall m \in M \quad (B19)
\forall r \in R \quad (B20)
\forall v \in V_{job} \quad (B21)
\forall v \in V_{reticle} \quad (B22)
\forall j \in J, m \in M, r \in R \quad (B23)
\forall j \in J, m \in M, r \in R, v \in V_{job} \quad (B24)
\forall j \in J, m \in M, r \in R, v \in V_{reticle} \quad (B25)
\forall j \in J, m \in M, r \in R, v \in V_{reticle} \quad (B26)
\forall m \in M \quad (B27)
\forall r \in R \quad (B28)
\forall v \in V_{job} \quad (B29)
\forall v \in V_{reticle} \quad (B30)

The objective (B1) is to minimize the time required to process all jobs and return jobs to the stocker. Constraints (B2)–(B6) ensure that each operation will have two vehicle tasks: loading and unloading. Constraints (B7)–(B8) ensure that a loading must complete before starting processing and an unloading must start after completing processing. Constraints (B9)–(B12) forbid any unnecessary reticle transfers and ensure a reticle pickup will complete prior to a drop off. Constraints (B13)–(B14) calculate a reticle pickup transfer-time \( t_{\text{vehicle, reticle}} \) and a reticle dropoff transfer-time \( t_{\text{reticle, machine}} \). Constraint (B15) ensures that a reticle pickup can start only after the same reticle completes at another machine. Constraints (B16)–(B18) ensure that a reticle drop off must complete if an operation needs the auxiliary resource. Constraints (B19)–(B22) prevent intervals in a sequence from overlapping. In particular, Constraint (B21) ensures a transfer-time of AMHS vehicle by using a transition matrix \( T \). Constraints (B23)–(B26) define interval variables of each task. Constraint (B29) associates job loading and unloading transfers to a sequence of each AMHS vehicle. Similarly, Constraint (B30) associates reticle pickup and dropoff transfers to a sequence of each ARHS vehicle.

We term this proposed model CP-1. Constraints (B3)–(B4) are redundant with Constraints (B5)–(B6). We remove Constraints (B3)–(B4) from the model and term it CP-2.
4 COMPUTATIONAL EXPERIMENTS

In this section, the effectiveness of the proposed model is examined. The CP and flow control models are all coded in IBM OPL 12.10.0 on a personal computer with an Intel® Core i7-4770 CPU with 16 GB of RAM. Since benchmark instances are not available in the literature, we generated random problem instances. The processing-times were drawn from a uniform distribution $[10, 25]$. The numbers of machines/reticles qualified to run each job were drawn from a uniform distribution of $[1, 2]$. Finally, to represent different layouts, two different sets of transportation times $(t_{m,m})$ were generated from a uniform distribution of $[2, 5]$ and $[2, 10]$, respectively.

Table 2 records the experimental results of the proposed models. Columns 1–5 identify the size of each test instance. Columns 6–9 $(10–13)$ report $c_{\text{max}}$ and optimality-gap of the models with layout A (B). When a model could not find any feasible solution within 60 seconds, the $c_{\text{max}}$ and gap have the symbol $\text{--}$. CP-2 showed a mild improvement over CP-1. However, both models failed to handle the medium-sized instances. All the test instances, CP logs, and detailed scheduling results (to build Gantt charts) are located at the following link:


5 CONCLUSION

This extremely complex scheduling problem that includes jobs, reticles, machines, and two different sets of vehicles is studied for the first time. A novel constraint programming model is proposed. The proposed models were able to prove optimality of the small-sized instances. But, they failed to even finding feasible solutions for the medium-sized instances.

A couple of areas can be foreseen for future research. For problems of this enormous complexity, we can relax some of constraints to find an initial solution and feed the solution to the original model as a partial solution. The CP solver propagates constraints out of the partial solution, infers the rest of decision variables, and uses them as a starting point. The questions are what constraints to relax and how to ensure the feasibility of the original model when we design a relaxed model. It will be also interesting to develop heuristic models. Since the proposed CP models can calculate the optimality gap, we can evaluate the performance of heuristic models.
ACKNOWLEDGMENTS

This research has been partially supported by CSPI Corporation.

REFERENCES


Im, K., K. Kim, T. Park, and S. Lee. 2009. “Effective vehicle dispatching method minimising the blocking and delivery times in automatic material handling systems of 300 mm semiconductor fabrication”. International journal of production research 47(14): 3997-4011.


Ham, Park, Choi, and Fowler


Sun, D. S., N. S. Park, Y. J. Lee, Y. J. C. Ahn, & T. E. Lee. 2005. “Integration of lot dispatching and AMHS control in a 300mm wafer fab”. In IEEE/SEMI Conference and Workshop on Advanced Semiconductor Manufacturing, April 11th-12th, Munich, Germany, 270-274.


AUTHOR BIOGRAPHIES

ANDY HAM received a Ph.D. in Industrial Engineering from Arizona State University in 2009 and an M.S. in OR/IE from University of Texas at Austin in 2000. He is currently working as an associate professor of Applied Engineering Technology, North Carolina A&T State University, while he serves as a technical advisor for CSPI, semiconductor manufacturers, and robotic warehouse companies. Prior to his current position, he worked for Samsung Electronics, Samsung Austin Semiconductor, GlobalFoundries, AMD, and IBM/ILOG. His research is currently focusing on real-time scheduling of vehicles, drones and robots in smart factories and smart warehouses in the logistics industry. His email address is <aham@ncat.edu>.

MIYOUNG-JU PARK received a Ph.D. in Industrial Engineering from Seoul National University in 2012. He is currently working as an associate Professor in Industrial and Management Systems Engineering, Kyung Hee University, Yongin, Korea. His research is currently focusing on combinatorial optimization, approximation algorithms, and scheduling. His email address is <pmj0684@khu.ac.kr>.

HO-JUN SHIN received a Ph.D. in Software Engineering from Daegu Catholic University in 2003. He works as a director of research and development in CSPI. He has been serving Samsung Semiconductor Fab/Backend, Samsung Display, Samsung Electro-Mechanics, SK Hynix Fab/Backend, SK Siltron, BOE, Hydis, etc. His email address is <hjshin@cspi.co.kr>.

SI-YOUNG CHOI is an executive director of research and development in CSPI. After working for Samsung Electronics in the semiconductor division over 10 years, he joined CSPI as a founding member and created the real-time dispatching/scheduling/simulation solution called ezDFS, which has been selected by Samsung Semiconductor Fab/Backend, Samsung Display, Samsung Electro-Mechanics, SK Hynix Fab/Backend, SK Siltron, BOE, Hydis, etc. His email address is <fasyc@cspi.co.kr>.

JOHN W. FOWLER is the Motorola Professor of Supply Chain Management in the W.P. Carey School of Business at Arizona State University. His research interests include discrete event simulation, deterministic scheduling, multi-criteria decision making, and applied operations research with applications in semiconductor manufacturing and healthcare. He has published over 130 journal articles and over 100 conference papers. He was the Program Chair for the 2002 and 2008 Industrial Engineering Research Conferences, Program Chair for the 2008 Winter Simulation Conference (WSC), and Program Co-Chair for the 2012 INFORMS National Meeting. He was the founding Editor-in-Chief of IE Transactions on Healthcare Systems Engineering and currently serves as a Healthcare Operations Management Departmental Editor. He is also an Editor of the Journal of Simulation and Associate Editor of IEEE Transactions on Semiconductor Manufacturing and the Journal of Scheduling. He is a Fellow of the Institute of Industrial and Systems Engineers (IISE) and recently served as the IIE Vice President for Continuing Education, is a former INFORMS Vice President, and served on the WSC Board of Directors. His email address is <john.fowler@asu.edu>.

1973