A PRODUCTION PLANNING AND SCHEDULING MODEL FOR SEMICONDUCTOR WAFER MANUFACTURING PLANTS

K. Rezaei¹, H. Eivazy¹,², A. Rezazadeh³ and S. Nazari-Shirkouhi¹

¹Department of Industrial Engineering, College of Engineering, University of Tehran, Tehran, Iran, email: snnazari@ut.ac.ir and krezaie@ut.ac.ir
²Department of Civil and Environmental Engineering, Faculty of Engineering, University of Alberta, Edmonton, Alberta, Canada, email: eivazy@ualberta.ca
³Department of Technical Engineering, Ahar Branch, Islamic Azad University, Ahar, Iran, email: shahab40757@yahoo.com

ABSTRACT

In this paper, a comprehensive production planning and scheduling model is developed for a typical semiconductor plant (fab). The proposed model includes two decision making policies on production planning and scheduling of the semiconductor manufacturing systems, release and dispatching policies. The proposed release and dispatching policies decide on flow of wafer lots in the fab in order to achieve production management goals. The proposed model integrates two successful approaches, theory of constraint (TOC) (Goldratt, 2006) and workload control (WLC) (Land and Gaalman, 1996) to improve the performance of semiconductor manufacturing systems and also uses the concept of program evaluation and review technique (PERT) (Nicholas 2004) to estimate lots’ cycle time. A number of simulation experiments are conducted to prove the applicability and superiority of the proposed model on two excellent models in the literature in association with some performance indicators: on-time delivery, cycle time, and total lead-time.

Keywords: Production Planning; Scheduling; Semiconductor Manufacturing; Release; Dispatching

1. INTRODUCTION

From its advent, semiconductor wafer fabrication industry has attracted many notices regarding the developing a production planning and scheduling model to achieve production goals. Existence of various complexities such as reentrancy characteristic, capital-technology-intensive equipment, long process route of orders, order variety, changing product-mix, uncertainty (e.g. machine breakdown), constant device miniaturization, and changing technologies in the wafer manufacturing systems are main factors making the production planning and scheduling of these systems very difficult (Glassey and Resende 1988a,b). The fundamental goal of each production planning and scheduling model is to control and facilitate the production to achieve the highest performance in term of a number of measures such as utilization and on-time delivery of orders. To do so, there are several models proposed for such systems in the literature which primarily, present plans for release of orders’ lots into the semiconductor wafer fabrication shop (fab) or dispatching (scheduling) them to guide the flow of orders’ lots inside fab. The release models are mainly classified in two classes of “open-loop” and “closed-loop” policies. Open-loop release methods send a new lot into fab ignoring the dynamic workload state of fab. The release models are Constant WIP (CONWIP) (Spearman et al. 1990), starvation avoidance (SA) (Glassey and Resende 1988a), load-oriented manufacturing control (LOMC) (Bechte 1994) and workload regulating (WR) (Wein 1988) are some of these policies. Generally, closed-loop methods outperform open loop rules in reaching the production objectives (Eivazy, 2009). On the other hand, there are numerous methods for dispatching of lots inside fabs. Some of the prominent dispatching policies are SA+ presented by Glassey and Resende (1988b) and other methods presented by Lu et al. (1994), Dabbas and Fowler (2003). Besides mentioned dispatching methods, Chen (2010) presents an optimized nonlinear fluctuation smoothing rule for scheduling of semiconductor plants. Also, Lee et al. (2011) uses Petri-nets for
improving the production planning and scheduling of semiconductor manufacturing processes. A more detailed literature review on release and dispatching policies could be found in Eivazy (2009).

The paper is structured as follows: in section 2, the proposed model including planning the order release and dispatching is presented. In section 3, the proposed model is evaluated by simulation modeling of a typical fab. Finally, the conclusion and future research directions are presented in section 4.

2. THE PROPOSED PRODUCTION PLANNING AND SCHEDULING MODEL

In this section, a WLC-TOC based production planning and scheduling model for a semiconductor wafer fab is presented. The main merits of WLC are buffering the shop from external dynamics (by job pool) and creating short and stable queue of workstations (Land and Gaalman 1996). Figure 1 shows the schematic view of the job pool and fab. The accepted lots are placed in the job pool until they can be released. WLC balances fab’s workload by setting an upper bound limit (LL) on the total workload each workstation can have (Eivazy et al. 2009). The total workload (TWL) for a typical workstation \( y \) is the summation of three workloads: 1) direct workload as the summation of processing times for all present lots in workstation \( y \), 2) upstream workload as the summation of processing time in workstation \( y \) of lots that are to be processed at workstation \( y \) in the next steps in their process route, and 3) releasing workload this is related to the workload of lots which are going to be released and will be processed at workstation \( y \) (Eivazy et al. 2009).

\[
TWL_y = \sum_{i \in S_y} PT_{iy} \cdot P_{iy} \quad (1)
\]

\[
P_{iy} = \prod_{m \in \text{upstream workstations of } y} \frac{PC_m}{TWL_m} \quad (2)
\]

In equation (1), \( S_y, PT_{iy}, P_{iy} \) stand for the set of lots currently placed in the fab and job pool that will pass from workstation \( y \), the processing time of lot \( i \) in workstation \( y \), the probability that lot \( i \), currently located in upstream workstation \( u \) passes through workstation \( y \) in the set planning period (PP), respectively. If lot \( i \) in workstation \( u \) passes \( m \) workstations before entering workstation \( y \), \( P_{iyu} \) is calculated from equation (2) in which \( PC_m \) and \( TWL_m \) are the planned capacity of workstation \( m \) in the planning period and the total workload of workstation \( m \) at a moment, respectively. The value of \( PC_m \) is constant in the planning period while \( TWL_m \) changes dynamically. For detailed elaboration on workload calculations and definitions, please refer to Eivazy (2009).

2.1 Order Release

Concisely, the release steps are “sequencing of lots in the job pool” and “examining the release opportunities and possibilities”. These steps are done continuously in the course of time. Figure 2 shows the framework of lot release policy in which the bold box represents the algorithm of examining lots’ release possibility.

2.1.1 Sequencing lots in job pool

The priority and ranking of lot \( i \) waiting in the job pool are calculated by the release priority (RP) reflected in equations (3) and (4). The more priority a lot has, the less rank it has.

\[
RP_i = \begin{cases} 
\sqrt{RWL_i,B} & \text{If release opportunity } \#1 \text{ is active} \\
CR_i & \text{Otherwise} 
\end{cases} \quad (3) \\
CR_i = \frac{d_i - T_{now}}{CT_i} \quad (4)
\]

\( RP_i, RWL_{i,B} \), and \( CR_i \) represent the release priority of lot \( i \), the releasing workload of lot \( i \) which imposes on the bottleneck workstation (calculated by equations (1) and (2)), and the critical ratio rule (Eivazy et al. 2009), respectively. It should be noted that release opportunities will be defined in section 2.1.2. \( CR_i \) is calculated from equation (4) in which \( d_i, T_{now} \), and \( CT_i \) stand for the due date of lot \( i \), the present time,
and the cycle time of lot \( i \) (the time-lag between the lot release time to fab and exit time from fab), respectively. Lots are sequenced in reverse order of \( \text{RP} \). In most of previous research, the values of lots’ cycle time are estimated by historical data or “actual to theoretical ratio” or successively running the pre-simulations. Here, similar to PERT, a weighted average of three times is proposed to predict the cycle time value of a typical lot \( i \). Three times are: 1) **Optimistic time** as the total processing time of lot \( i \) (\( \text{TPT}_i \)) in its process route which indicates lot \( i \) is processed in all workstations with no waste of time in workstations’ queues, 2) **The most probable time** as the time last a lot passes from bottleneck workstation, on average. According to TOC philosophy, all main production management goals such as cycle time are dominated by the bottleneck workstation (Goldratt, 2006). \( \text{B} \) presents probability that a lot passes from bottleneck workstation \( \text{B} \) during the planning period \( \text{PP} \). Thus, \( \text{PP} / (\text{B} \text{PC LL}) \) indicates, on average, the time a lot consumes to pass from the bottleneck workstation and 3) **Pessimistic time** as the possible longest time which is related to when all workstations in the process route of lot \( i \) have always their maximum load \( \text{LL} \). \( \prod \text{PC}_x / \text{LL}_x \) \( (x \text{ is a workstation in the process route of lot } \text{i}) \) is the minimum probability that \( \text{i} \) passes the all workstations in its process route in the planning period. Thus, \( \text{PP}/(\prod \text{PC}_x / \text{LL}_x) \) represents the maximum value of \( \text{CT}_i \). Weighted average of three mentioned time values with coefficients 1/6, 4/6 and 1/6 is used for estimation of a lot’s cycle time.

### 2.1.2 Releasing opportunities

This step defines *release opportunities* in which “release time” and that “what lots are released” are clarified. Before defining release opportunities, it should be noted that in each release opportunity, a lot can be released only when none of the workstations are expected to get blocked with release of that lot. The release opportunities Eivazy et al. (2009) are defined as: 1) **If total workload of potential bottleneck workstation lowers than a critical level (TWL<CL)**: This release opportunity exploits the benefits of prominent TOC approach in way that throughput and utilization of the manufacturing system are governed by monitoring the bottleneck’s utilization. Traditionally, the photolithography workstation is considered as the potential bottleneck in the semiconductor wafer fabs, 2) **If there is an urgent lot in the job pool**: All urgent lots (lots with \( \text{CR}<1 \)) are included in the release feasible set to be released into the fab as soon as possible to be delivered on-time, 3) **If all lots in the buffer of one workstation are non-urgent (non-urgent workstation)**: this pull-type release opportunity is set to prevent from the starvation/idleness of the head workstations, i.e. the workstations related to first step of lots. This release opportunity leads to smooth fab’s workload. Note that, *non-urgent and urgent lots in a workstation* will be defined in the dispatching section, 4) **If a workstation is starved**: here, the set of lots in the job pool that their first step in their process route is in that starving workstation forms the release feasible set. This pull-type release opportunity results in higher utilization, reduction in total lead-time, and the balanced fab’s workload (Eivazy et al. 2009).

After forming the corresponding release feasible set in each above-mentioned release opportunities, the release of lots in this set should be examined sequentially. If with release of the first ranked lot in the release feasible set, none of workstations gets blocked, that lot is released. Otherwise, the release of the next ranked lot is examined. Once a lot is released, total workload of all workstations, the sequence of lots in the job pool, and release feasible set are updated. This release procedure repeats for lots in the updated feasible set until all lots in this set are examined and none of them can be released.

### 2.2 Dispatch Model

The flow of lots in the fab is ruled by dispatching policy depicted in Figure 3. When, a machine gets idle, the first ranked lot is dispatched on that idle machine. Equation (5) is proposed for prioritizing and ranking of lots in each non-bottleneck workstation \( w \): The more \( \text{DP} \) value of a lot, the less rank it has.

\[
\text{DP}_{i,w} = \begin{cases} 
\frac{1}{\text{FWL}_{i,w} B} & \text{release opportunity #1 is active} \\
\text{cr}_{i,w} & \text{otherwise}
\end{cases}
\] (5)
Here, $DP_{i,w}$, $FWL_{i,w,B}$, and $cr_{i,w}$ represent dispatching priority of lot $i$, upstream workload of bottleneck originated from lot $i$, and critical ratio. To calculate $FWL_{i,w,B}$ and $cr_{i,w}$, equations (6) and (7) are used:

$$FWL_{i,w,B} = PT_{i,B} \cdot P_{i,B}$$  

$$cr_{i,w} = \frac{d_i - (d_i - r_i) \cdot (RPT_{i,w} / CT_i)}{PT_{i,w}}$$

In equation (6) and (7), $PT_{i,B}$, $P_{i,B}$, $RPT_{i,w}$, $CT_i$, $PT_{i,w}$, $r_i$, and $d_i$ represent the processing time of lot $i$ in the bottleneck workstation, the probability that if lot $i$ currently located in upstream workstation $w$ passes from bottleneck workstation in the planning period (can be calculated from equation (2)), the remaining processing time of lot $i$ located in the buffer of workstation $w$ (summation of processing times in next
workstations), the cycle time of lot $i$, processing time of lot $i$ in workstation $w$, the release time of lot $i$, and due date of lot $i$, respectively. The numerator of equation (7) calculates the rough virtual due date in that lot $i$ should exit from workstation $w$. For bottleneck workstation, lots are dispatched only according to equation (7).

### 3. SIMULATION EXPERIMENTS

Through a number of simulation experiments, the proposed production planning and scheduling model is validated. A small virtual fab with 10 workstations (workstations #7 and 3 are the photolithography and batch processes, respectively.) is simulated by Arena 13.0 simulation software. For batch workstation #3, the minimum batch size (MBS) of 6 is considered for dispatching. Six different types for orders are considered (A-F) with Poisson arrival rate of 5 lots per day. The simulation horizon is 350 days with the first 100 days as the warm up and each simulation replicates 30 times to test the statistical hypotheses. Three indicators, cycle time, total lead-time, and on-time delivery are considered which are calculated at each 10-days planning period. The total lead-time is the interval time between the arrival of a lot into job pool and its departure from the fab. The total processing time for each order type is shown in Table 1. A lot is delivered on-time if it departures the fab in a time-lag before or after its due date that are calculated by equations (8) and (9), respectively.

<table>
<thead>
<tr>
<th>Order type</th>
<th>TPT (hour)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>162</td>
</tr>
<tr>
<td>B</td>
<td>111</td>
</tr>
<tr>
<td>C</td>
<td>82</td>
</tr>
<tr>
<td>D</td>
<td>86</td>
</tr>
<tr>
<td>E</td>
<td>140</td>
</tr>
<tr>
<td>F</td>
<td>153</td>
</tr>
</tbody>
</table>

Where $d_{i,k}$, $A_{i,k}$, $CT_i$, $U(a,b)$, $AL$ represent the due date of lot $k$ of order type $i$, the arrival time of lot $k$ of order type $i$ into the job pool, cycle time of order type $i$, the uniform random distribution number between $a$ and $b$ (set to 1 and 3, respectively), and time-lag allowance (set to 0.05), respectively.

Moreover, values of $PP$, $CL$, $LL$, $AL$ are set to ten 24-hour days, 99 hours, and 330 hours, respectively. For validation purpose, results of applying two excellent production planning and scheduling models in the literature, SA-SA and SA-CR, are compared with the proposed model’s results. These benchmark models often have outperformed other production planning and scheduling models regarding the considered performance measures (Eivazy et al., 2009). The Duncan’s multiple range test is applied to grade these models into 4 levels, namely A, B, C and D in the descending order. Table 2 shows the statistical analysis of simulation results that indicates the proposed model works better than alternative excellent methods in term of considered performance indicators. In this table, SD means the standard deviation. Figure 4 shows the trend of on-time delivery in 25 10-day periods for the three models. This figure shows the higher and smoother trend of on-time delivery for the proposed model than benchmarking models. Better performance related to on-time delivery indicates that the proposed model can hand in orders to customers more on-time and with more reliability which leads to higher customers satisfaction and long-term profitability.

### 3. CONCLUSION / FUTURE RESEARCH DIRECTIONS

In this paper, a production planning and scheduling model, including two policies for planning of orders release and dispatching, is proposed for a typical semiconductor wafer manufacturing plant. Simulation experiments approve the superiority of the proposed model on two well-known benchmarking models in terms of on-time delivery, cycle time, and total lead-time. Besides superiority of the proposed model on benchmarks, it has some merits which briefly are as: 1) **Integration of two superior production planning and scheduling approaches, TOC and WLC**: the proposed model applies WLC to control, balance, and smooth the workload of fab which and ensures avoidance of bottleneck shifting phenomenon. Only with preventing bottleneck shifting, TOC can be implemented effectively. The proposed model applies WLC by workload balancing. Also, it applies TOC by monitoring of bottleneck workload and setting some release and dispatch rules sensitive to bottleneck’s workload. 2) **Applying...**
PERT to estimate the values of orders’ cycle time, even when there are no historical data, by current dynamic workload of fab: Previous proposed models in the literature mainly determine orders’ cycle time by either on-hand historical data on cycle time or a deterministic ratio called “actual to theoretical ratio”. When a new product is going to be manufactured, there is no historical data on its cycle time value. Also, a deterministic ‘actual to theoretical ratio’ cannot reflect the dynamic of fab workload. The proposed model estimates cycle time by fab workload and order characteristics while keeping away from mentioned previous methods’ shortages.

Table 2: Statistical analysis of models

<table>
<thead>
<tr>
<th>Performance Measures</th>
<th>Models</th>
<th>Mean</th>
<th>SD</th>
<th>Duncan test (Mean/SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cycle time (hour)</td>
<td>Proposed model</td>
<td>214.7</td>
<td>11.7</td>
<td>A/A</td>
</tr>
<tr>
<td></td>
<td>SA-SA⁺</td>
<td>275.2</td>
<td>31.1</td>
<td>BC/D</td>
</tr>
<tr>
<td></td>
<td>SA-CR</td>
<td>301.6</td>
<td>37.0</td>
<td>C/D</td>
</tr>
<tr>
<td>Total lead-time (hour)</td>
<td>Proposed model</td>
<td>425.5</td>
<td>20.4</td>
<td>A/A</td>
</tr>
<tr>
<td></td>
<td>SA-SA⁺</td>
<td>462.5</td>
<td>45.1</td>
<td>BC/BC</td>
</tr>
<tr>
<td></td>
<td>SA-CR</td>
<td>531.2</td>
<td>57.8</td>
<td>D/D</td>
</tr>
<tr>
<td>On-time delivery (%)</td>
<td>Proposed model</td>
<td>91.0</td>
<td>4.9</td>
<td>A/A</td>
</tr>
<tr>
<td></td>
<td>SA-SA⁺</td>
<td>77.4</td>
<td>10.2</td>
<td>B/B</td>
</tr>
<tr>
<td></td>
<td>SA-CR</td>
<td>82.4</td>
<td>11.3</td>
<td>AB/B</td>
</tr>
</tbody>
</table>

REFERENCES


