*Proceedings of the 2007 Winter Simulation Conference*
*S. G. Henderson, B. Biller, M.-H. Hsieh, J. Shortle, J. D. Tew, and R. R. Barton, eds.*

# SIMULATION EXPERIMENTAL INVESTIGATION ON JOB RELEASE CONTROL IN SEMICONDUCTOR WAFER FABRICATION

Chao Qi

Singapore-MIT Alliance (SMA)
Manufacturing Systems and Technology (MST) Programme
N3.2-01-36, 65 Nanyang Drive, 637460, SINGAPORE

Appa Iyer Sivakumar

School of Mechanical and Aerospace Engineering
Nanyang Technological University
50 Nanyang Avenue, 639798, SINGAPORE

Stanley B. Gershwin

Department of Mechanical Engineering
Massachusetts Institute of Technology
77 Massachusetts Avenue, Cambridge, MA 02139-4307, USA

## ABSTRACT

This paper presents a new job release methodology, WIPLOAD Control, especially in semiconductor wafer fabrication environment. The performance of the proposed methodology is evaluated in a simulation study on a simplified wafer fabrication model, in comparison with other existing release control methodologies. A case study is also conducted by simulating a real-life wafer fabrication facility. Based on the experimental results, it appears that WIPLOAD Control is a reliable job release methodology, which can efficiently reduce average cycle time and standard deviation of cycle time for a given throughput level, especially with the increase of system congestion level and system variability caused by stochastic events such as machine unreliability or processing time variability.

## 1 INTRODUCTION

Job release control is an essential part of scheduling issue, which determines the type, amount and time point of release of new jobs into a manufacturing facility. Although it can be applied in a more general manufacturing environment, job release control is emerging as an important research topic in semiconductor manufacturing, given the complexity and cost of modern wafer fabrications. The overall semiconductor manufacturing flow can be generally divided into four stages: wafer fabrication, wafer probe, assembly or packaging, and final test. This research focuses on wafer fabrication as it is the most technologically complex and capital-intensive phase among all the processes. The facility where wafer fabrication takes place is commonly referred to as a wafer fab. The significant impact of job release control on the wafer fab performance has been demonstrated by Wein (1988) and Glassey and Resende (1988).

Since 1970s, a number of release control methodologies have been developed and investigated. These methodologies can be generally classified into two categories: open-loop and closed-loop release methodologies. Open-loop release methodologies make release decisions regardless of any current system information; release is usually scheduled based on exogenous information such as prediction and demand; and the release time is not modified according to what is happening in the production process. For example, a widely used open-loop release methodology is to start a certain amount of new jobs into the facility after a certain time interval, which is referred to as uniform release (UNIF) in this paper. In contrast, closed-loop release control methodologies take into account the dynamic shop floor information according to their specific objectives. The majority of the existing closed-loop release methodologies adopt the idea of "WIP cap" (Hopp and Spearman 2000) to control the start of new jobs by limiting the workload. There are generally three methods to set the workload limitation relating to three levels of aggregation. The first method is to limit the load of the overall shop floor. The representative of this kind of release methodologies is CONWIP proposed by Spearman, Woodruff, and Hopp (1990), which is to maintain a constant WIP level by starting new jobs whenever the WIP level has fallen below a specific level. The second way is to only consider the workload of bottleneck workstation based on the Theory of Constraints (TOC). Workload Regulating (WR) proposed by Wein (1988) belongs to this category, and is widely discussed and compared for semiconductor manufacturing environment. The third type of load limited release methodology is to limit the workload for each workstation in the

**1-4244-1306-0/07/$25.00 ©2007 IEEE**

production line. A typical way is to set an upper bound workload threshold for each workstation; a job is released only if no machine on the job's path will be loaded over its threshold (Philipoom, Malhotra, and Jensen 1993); this method is referred to as StnLoad in this paper.

This paper presents a new closed-loop release methodology, WIPLOAD Control (WIPLCtrl), which is a shop load limited release methodology. We conduct a simulation experimental investigation to evaluate the performance of WIPLCtrl in comparison with that of CONWIP, WR, StnLoad, and UNIF, in terms of throughput (TH), average cycle time (CTAVG), and standard deviation of cycle time (CTSTD).

In section 2, WIPLCtrl is introduced. It is followed by a simulation study on a simplified wafer fab in section 3. In section 4, a case study considering a real-life wafer fab is presented. The conclusions are discussed in section 5.

## 2 WIPLOAD CONTROL

The notations used to describe WIPLCtrl are given as follows:

| | |
|---|---|
| $i$ | : workstation $i$ $\quad i = 1,...,k$ |
| $m$ | : part type $m$ $\quad m = 1,...,M$ |
| $J_m$ | : the total number of operations for part $m$ |
| $j, j'$ | : an operation step $\quad j, j' = 1,...,J_m$ |
| $m(j)$ | : the $j^{th}$ step in the route of part $m$ |
| $t$ | : time |
| $P_{m(j')}$ | : processing time of operation step $m(j')$ |
| $R_{m(j)}$ | : remaining processing time for job undergoing operation $m(j)$ |
| $L$ | : reference WIPLOAD level |
| $W_{m(j)}(t)$ | : the number of jobs undergoing $m(j)$ at time $t$ |
| $L(t)$ | : system WIPLOAD level at time $t$ |
| $e(t)$ | : difference between $L$ and $L(t)$ at time $t$ |

We define a new measure: system WIPLOAD, to measure the overall workload on shop floor, as the sum of the remaining processing times of all the jobs on the shop floor (see Equation 1). WIPLOAD is dynamic workload measure changing with the processing status of the jobs that are already in process.

$$L(t) = \sum_{m=1}^{M}\sum_{j=1}^{J_m} W_{m(j)}(t) \bullet R_{m(j)} \qquad (1)$$

where

$$R_{m(j)} = \sum_{j'=j}^{J_m} P_{m(j')} \qquad (2)$$

To smooth out the fluctuation of the workload on shop floor, one simeple way to control the release process based on WIPLOAD is to maintain system WIPLOAD at a prescribed level. This methodology is referred to as WIPLOAD Control (WIPLCtrl), which is depicted by the framework shown in Figure 1.

As the controllable variable in this framework, $L(t)$ is updated in real time by information system when a new job is released or when an operation step is completed on a workstation. Given a reference WIPLOAD level, $L$, the feedback calculation is performed to compute the difference between $L$ and $L(t)$. The release decision is determined based on this difference to maintain system WIPLOAD at $L$. According to this principle, a job release controller is designed and described in the form of flow chart (see Figure 2).

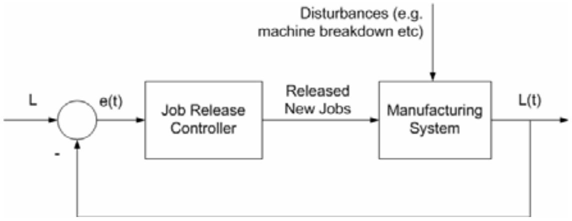


Figure 1. Framework of WIPLCtrl

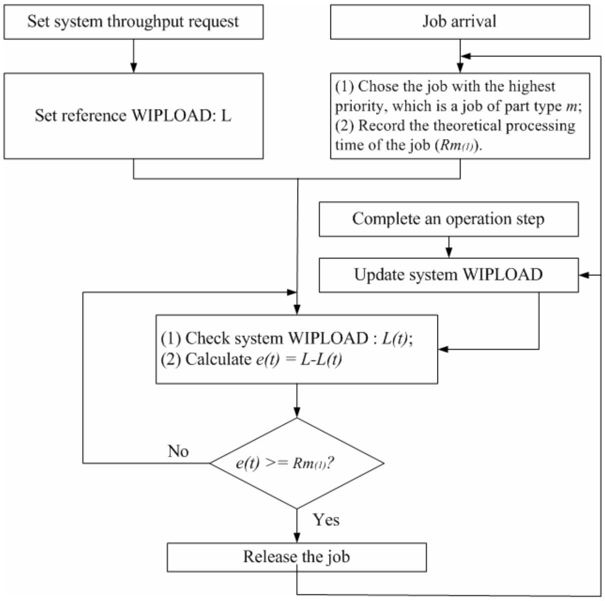

Set system throughput request
Job arrival
Set reference WIPLOAD: L
(1) Chose the job with the highest priority, which is a job of part type m; (2) Record the theoretical processing time of the job (Rm(1)).
Complete an operation step
Update system WIPLOAD
(1) Check system WIPLOAD : L(t); (2) Calculate e(t) = L-L(t)
No
e(t) >= Rm(1)?
Yes
Release the job

Figure 2. Release decision making process of WIPLCtrl

## 3 SIMULATION EXPERIMENTAL STUDY

A simulation study is conducted using AutoSched$^{TM}$ AP to evaluate the performance of WIPLCtrl. Below we discuss the results of this simulation study that illustrate some of the advantages of WIPLCtrl in comparison with CONWIP, WR, StnLoad and UNIF.

### 3.1 Model Description

There are 7 kinds of key operations in a typical wafer fab as depicted by Figure 3, which can each include multiple sub-steps that take place on different machines. Production of a particular type of circuit requires a specific sequence of processing steps, with unique processing times at each step for the product type.

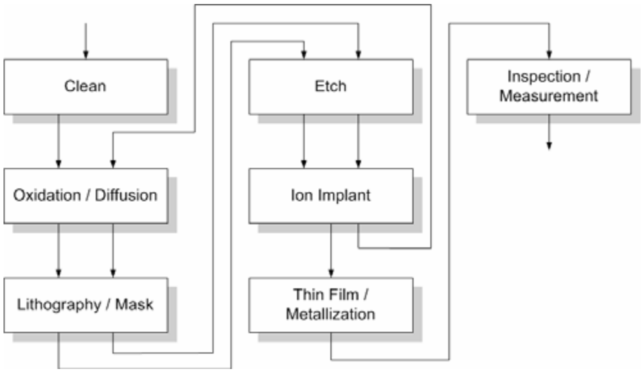


Figure 3: Basic operations in Wafer Fab

Since wafers of pure silicon are imprinted with tens or even hundreds of integrated circuits in dozens of layers, the sequence of processing steps in a wafer fab requires individual lots to revisit certain workstations numerous times at different steps. In between such visits, a number of other workstations may be visited. This kind of process is known as reentrant flow. For example, a wafer may have to visit the lithography workstation more than ten times to have all layers of circuitry fabricated. Reentrant processing is one of the distinguishing characteristics that make wafer fab different from traditional manufacturing.

A simulation model of a simplified wafer fab is studied using AutoSchedTM AP. The model consists of 7 single machine workstations, which correspond to the 7 basic operations involved in a typical wafer fab. The system can process 5 part types. Table 1 summarizes partial characteristics of the model including the function of the workstation, the number of reentries, and the processing time per step on the workstation for each part type. Table 2 shows the routes of the part types produced.

Table 1: Simplified Wafer Fab model

| STN | Part 1 | | Part 2 | | Part 3 | | Part 4 | | Part 5 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | R | P | R | P | R | P | R | P | R | P |
| CL | 4 | 6.5 | 6 | 11.1 | 5 | 8.4 | 6 | 6.0 | 8 | 4.5 |
| DF | 3 | 6.9 | 4 | 11.5 | 4 | 15.0 | 3 | 12.0 | 6 | 9.0 |
| LT | 4 | 10.0 | 9 | 7.4 | 5 | 14.4 | 5 | 6.0 | 5 | 7.2 |
| ET | 4 | 8.8 | 3 | 13.1 | 4 | 13.5 | 4 | 12.0 | 8 | 6.0 |
| IM | 1 | 18.8 | 4 | 6.8 | 2 | 18.0 | 4 | 7.5 | 2 | 15.0 |
| FL | 1 | 35.3 | 2 | 30.0 | 2 | 27.0 | 3 | 10.0 | 3 | 16.0 |
| ME | 6 | 2.6 | 14 | 1.3 | 6 | 5.0 | 12 | 2.5 | 10 | 4.2 |

R: Number of reentrant processes on the workstation.
P: Processing time (min) for each operations.

Table 2: Processing routes

| Part | Route |
|---|---|
| 1 | CL→DF→LT→ME→ET→CL→DF→IM→ET→DF→LT→ME→CL→FL→LT→ME→ET→CL→ME→LT→ME→ET→ME |
| 2 | DF→LT→ME→CL→LT→ME→IM→CL→ME→DF→LT→ME→IM→CL→LT→ME→DF→CL→ME→DF→LT→ME→IM→CL→ME→LT→ME→IM→ET→CL→ME→FL→LT→ME→ET→LT→ME→FL→LT→ME→ET→ME |
| 3 | CL→DF→LT→ME→ET→CL→DF→LT→IM→ET→DF→LT→ME→IM→CL→DF→LT→CL→FL→LT→ME→ET→CL→ME→FL→ME→ET→ME |
| 4 | CL→LT→ME→IM→CL→ME→DF→LT→ME→IM→CL→ET→LT→ME→DF→CL→ME→DF→LT→ME→IM→CL→ME→FL→LT→ME→IM→ET→CL→ME→FL→ME→ET→ME→FL→ET→ME |
| 5 | CL→DF→LT→ME→ET→CL→DF→LT→ME→ET→CL→DF→IM→ET→DF→LT→ME→CL→DF→IM→ET→DF→LT→ME→CL→FL→LT→ME→ET→CL→ME→FL→ME→ET→CL→ME→ET→CL→ME→FL→ET→ME |

The assumptions of this simplified wafer fab model include:

- All the workstations are single-machine workstations.
- For each part type, the mean processing time on a specific workstation is identical for different reentrant steps.
- Each machine can process only one lot at a time.
- Each machine is not reliable and subject to failures. Only unscheduled machine failures are considered. The time to failure and time to repair for each machine are assumed to be exponentially distributed. The mean-time-to-failure (MTTF) is assumed to be 500 minutes. The value of mean-time-to-repair (MTTR) is determined based on the machine availability level considered.
- The machine availability levels of each machine is identical, which are computed according to $A = MTTF/(MTTF + MTTR)$.
- Setup times are included in the processing times.
- Transportation time between workstations is not considered.
- Workpieces are not destroyed or rejected at any workstation in the line. Rework is not considered.
- There are no human errors made during the processing and issues with regard to operators are not considered.
- All of the processing steps are performed on a lot of wafers.
- Dispatching rule used for all the machines is First In First Out (FIFO).
- The initial level of WIP inventory in the system is set equal to zero. The output data during the warm-up pe-

riod are discarded as the data in the transient state. Only the data obtained after the system gets steady are used for performance analysis.

### 3.2 Model Verification and Justification

The computer program is checked and debugged in steps to verify the simulation model. In addition, the technique of "trace" is employed, which is one of the most powerful techniques that can be used to debug a discrete-event simulation program (Law and Kelton 2000). For instance, considering the single part type case, 6 lots of part type 1 are simultaneously released into the system at the beginning of a simulation run, then the operation events on these lots are traced. The event tracing record is compared to the expected event schedule. Table 3 shows partial trace data of the released lots. The model verification is confirmed since the achieved event schedule is coincident with the intended one.

Table 3: Partial trace data for model verification

| Entity | Event Start | Event End | Time (min) | State | Step | STN |
|---|---|---|---|---|---|---|
| LOT-1-1 | 01/01/03 00:00:00 | 01/01/03 00:06:30 | 6.50 | Proc | 1 | CL |
| LOT-1-1 | 01/01/03 00:06:30 | 01/01/03 00:13:24 | 6.90 | Proc | 2 | DF |
| LOT-1-1 | 01/01/03 00:13:24 | 01/01/03 00:23:24 | 10.0 | Proc | 3 | LT |
| … | … | … | … | … | … | … |
| LOT-1-2 | 01/01/03 00:00:00 | 01/01/03 00:06:30 | 6.50 | Wait | 1 | CL |
| LOT-1-2 | 01/01/03 00:06:30 | 01/01/03 00:13:00 | 6.50 | Proc | 1 | CL |
| LOT-1-2 | 01/01/03 00:13:00 | 01/01/03 00:13:24 | 0.42 | Wait | 2 | DF |
| … | … | … | … | … | … | … |

Although the constructed simulation models are not models of specific existing systems, they are built based on the knowledge of real-life wafer fabs. To increase the model validity and credibility, several issues are considered in the process of building the simulation model.

- Most data used to describe the workstations and the processing steps of the products are simplified based on the real-life processes in wafer fabs. The purpose of simplification is to avoid complicating the issue by the interactive factors of the system.
- The relevant simulation studies presented in the literature are referred to. For example, when the machine unreliability is considered, exponential distribution is adopted to describe the time to failure and time to repair since it is considered to be reasonable to assume the random machine failures to be exponentially distributed in the relevant simulation studies carried out using wafer fab models (Wein 1988; Glassey and Resende 1988; Kim, Leachman, and Suh 1996).
- In the course of creating the simulation model, the industrial engineers of Chartered Semiconductor Manufacturing (Chartered) have contributed their suggestions and comments on how to take account of practical issues in the simulation study.

### 3.3 Warm-Up Period

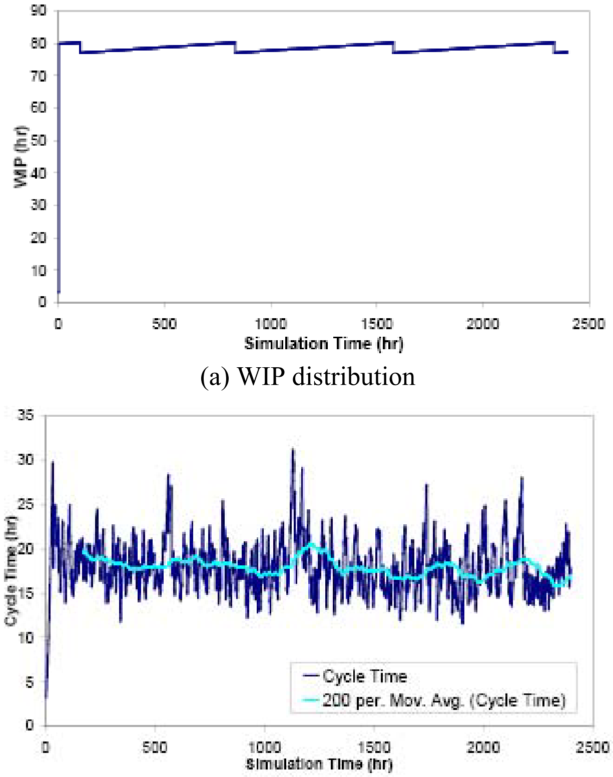
Only the steady-state output data are used for performance analysis in this simulation. To determine the length of warm-up period, Welch's procedure (Welch 1983) is utilized, which is the simplest and most general technique for determining warm-up period (Law and Kelton 2000). Welch's procedure is based on making a certain number of replications of a simulation; after calculating the average process of these replications, the moving average of the average process will be computed and plotted; then warm-up period is chosen beyond which the moving average appears to have converged.

The observed system performances are the distribution of WIP (in terms of hours) and the distribution of cycle time in different cases. Here a single part type (Part 1) case is considered as an illustration, in which the machine availability level is 90%. The adopted release methodology is CONWIP. The reference WIP level of CONWIP is 77 hours, under which the system is operating at a relatively high throughput level.

Figure 4 depicts the observed history of WIP and cycle time. The moving average for cycle time with a period of 200 hours is plotted to smooth out the oscillations in the cycle time history. Note that it is better to choose the length of warm-up period too large rather than too small (Law and Kelton 2000). Therefore the first 400 hours are set as the warm-up period. It is observed that 400 hours is also a long enough warm-up period in all the other cases. Only the data obtained after the warm-up period are used for performance analysis.

### 3.4 System Factors Considered

A manufacturing system is subject to continuous change due to various stochastic factors. One of the major purposes of improving production control policies is to enhance the capability to make effective response to these stochastic factors so that the system can operate in a desired state. Therefore, to evaluate the effects of release control methodologies, three sources of system variability contained in most manufacturing systems are considered in this simulation study including product mix, machine unreliability, and processing time variability. Considering the impact of system congestion level, 5 throughput levels are considered for each case.



(a) WIP distribution

(b) Cycle time and moving average

Figure 4: Determination of warm-up period

### 3.4.1 Job Release Control Methodologies

The evaluated job release methodologies include:

- UNIF: New jobs are released into the system with a constant rate, one every X minutes. The value of X is determined according to the expected throughout rate.
- CONWIP: Maintain system WIP level at a constant level. New jobs cannot begin on the line until the WIP has fallen below the specified level.
- WR: Control the workload level of the bottleneck workstation. A new job is released into the system whenever the total amount of remaining work in the system for the bottleneck workstation falls below a prescribed level.
- StnLoad: Set an upper workload bound for each workstation. At the beginning of every shift, the jobs waiting to be released are checked according to their priority. A new job is released if the target workload level of any workstation will not be exceeded.
- WIPLCtrl: Prescribe a reference WIPLOAD level. A new job is released when this reference WIPLOAD is not exceeded.

### 3.4.2 Product Mix

Product mix is one of the most significant issues that cause difficulties to manufacturing systems control. The steps to produce a wafer of different technologies and different wafer types could be very different. As a result, the machine utilization levels are constantly changing with the change of product mix. In particular, the system bottleneck machine may shift with the change of product mix, especially for the perspective of short-term production control. In order to assess the impact of product mix on the relative effect of release control methodologies, three product mix scenarios are considered in this simulation study including single part type (Part 1), a 50-50 mix of two part types (Part 1 and Part 2), and a five-part type case (Parts 1--5) with an equal proportion for each part. Note that for a long time period, the bottleneck workstation of this simplified wafer fab model is relatively deterministic at lithography workstation (LT) in these three scenarios.

### 3.4.3 Machine Unreliability

Wafer fab requires many highly sophisticated machines. Besides the periodic maintenance, these machines may also jam, work improperly, or cease working altogether and have to be serviced, and this is called unscheduled maintenance or a breakdown. In contrast to other types of manufacturing, a significant amount of time is spent in scheduled and unscheduled maintenance of machines in wafer fab (Hogg and Fowler 1991). In this simulation study, only the unpredictable machine breakdown is considered. The time to failure and time to repair for each machine are assumed to be exponentially distributed. The MTTF is fixed to 500 minutes. Two levels of machine availabilities are tested, which are 90% and 80%. Note that machine breakdown with a longer MTTR can bring more variability into the system than the one with a shorter MTTR.

### 3.4.4 Processing Time Variability

The processing time variability is also an important source of randomness in manufacturing systems. Two levels of processing time variability are considered in this simulation, i.e. the processing time is deterministic or is uniformly distributed with an offset of 10% of the mean value on both sides.

### 3.4.5 System Congestion Level

In order to observe the impact of system congestion level on the relative effect of the tested release methodologies, for each experimental case, the average cycle time and the standard deviation of cycle time are collected under five throughput levels. The motivation to consider different throughput levels is that the relative effect of release methodologies always depends upon the congestion level of the system.

### 3.5 Evaluation Methodology

To evaluate the effects of the considered factors, the performance criteria employed in this study include the average cycle time (CTAVG) and the standard deviation of cycle time (CTSTD) under different throughput levels (TH). Cycle time performance emerges as the number one performance metric in semiconductor industry (Fowler and Robinson 1995). The major benefits of reducing average cycle time include reducing the overall response time to customers; and carrying less work in process (WIP) level. The standard deviation of cycle time is a measure of how spread out a cycle time distribution is. Reducing the standard deviation of cycle time also has intrinsic benefits. It can imply smaller WIP and finished goods inventory for a given cycle time level; it also improves the predictability and service level of the system.

### 3.6 Simulation Experiments

The simulation experiments are organized into 5 cases as described in Table 4.

Table 4: Simulation cases

| Case | Part Type | Availability | Proc Time |
|---|---|---|---|
| 1 | Part 1 | 90% | Deterministic |
| 2 | Part 1,2 | 90% | Deterministic |
| 3 | Part 1-5 | 90% | Deterministic |
| 4 | Part 1-5 | 80% | Deterministic |
| 5 | Part 1-5 | 80% | Uniform |

For each case, the relative effect of the considered release control methodologies is tested under five throughput levels. Cases 1--3 consider the issue of product mix. Case 4 tests the impact of machine availability. The issue of processing time variability is addressed in Case 5.

The length for each simulation run is 2400 hours, in which the beginning 400 hours are considered as the warm-up period. The average values of 10 replications are presented as the results. The statistical analysis is performed using the paired student's t-test with a 95% confidence level.

### 3.7 Simulation Results and Analysis

Table 5 lists the percentage improvements of the evaluated closed-loop release methodologies over UNIF in this simulation study.

Relative to an open-loop release methodology such as UNIF, the appropriate choice of a closed-loop release control methodology can significantly improve the system performance in terms of both the average cycle time and the standard deviation of cycle time simultaneously for a certain throughput level. This is mainly because a closed-loop release methodology is able to adjust the release decision responding to the stochastic events such as machine failures. As stated by Gilland (2002), although closed-loop methodologies introduce variability into the release process, this variability is correlated with the production variability in a beneficial manner. It is also observed that the improvement by controlled release process becomes more significant with the increase of system congestion level (i.e. higher throughput level). In other words control of release process plays a more significant role when the manufacturing system is operating on a high congestion level. This is important because the throughput level of interest in a real-life wafer fab are usually high for the purpose of adequately utilizing the system capacity. However, care must be taken on the choice of release control methodology. We can observe from the experimental results that the performance of StnLoad is worse than that of UNIF in case 1 when a single part type situation is considered.

Among the evaluated closed-loop release methodologies, WIPLCtrl performs the best in most of the tested scenarios. We can observe that manufacturing system environmental conditions influence the relative performance of release control methodologies. For example, in case 1, WIPLCtrl significantly outperforms CONWIP, WR and StnLoad in terms of both mean and variance of cycle time for all the throughput levels. When the product mix is introduced in cases 2 and 3, the system bottleneck (LT) becomes more critical relative to case 1. In these settings, the improvements of WIPLCtrl over WR become less significant especially for average cycle time. The reason is probably that for a system with an explicit bottleneck, the major part of system WIPLOAD is created by the jobs queuing in front of the bottleneck machine. In other words the workload of the bottleneck machine is close to the value of system WIPLOAD. However, by efficiently compensating more system disturbances, WIPLCtrl significantly reduces the standard deviation of cycle time for all the tested cases. Based on the system configuration of case 3, a higher level of variability caused by machine failure is considered in case 4, and the processing time variability is further introduced in case 5. By observing the results of these 3 cases, we should notice that WIPLCtrl shows consistent improvements on both the mean and the standard deviation of cycle time with the increase of system variability.

WR works better than CONWIP and StnLoad especially on average cycle time because the long-term bottleneck of the system is relatively deterministic in all the tested cases. CONWIP leads to satisfactory improvements on standard deviation of cycle time. The benefit of StnLoad can be observed when the system is operating at a relatively high throughput level.

.

Table 5: Percentage improvements of closed-loop release methodologies over UNIF

| Case | Factors Considered | TH (%) | CONWIP | | WR | | StnLoad | | WIPLCtrl | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | CTAVG (%) | CTSTD (%) | CTAVG (%) | CTSTD (%) | CTAVG (%) | CTSTD (%) | CTAVG (%) | CTSTD (%) |
| 1 | Congestion level | 85 | 11 | 46 | -3 | 4 | -108 | -71 | 20 | 52 |
| | | 90 | 8 | 50 | -1 | 9 | -96 | -65 | 18 | 56 |
| | | 95 | 4 | 52 | 1* | 13 | -67 | -56 | 18 | 58 |
| | | 98 | -4 | 48 | 2* | 15 | -49 | -52 | 18 | 57 |
| | | 100 | 33 | 66 | 44 | 50 | 26 | 19 | 45 | 73 |
| 2 | Product mix; Congestion level | 85 | 33 | 73 | 42 | 73 | -21 | 45 | 39 | 79 |
| | | 90 | 49 | 83 | 54 | 84 | 15 | 66 | 52 | 87 |
| | | 95 | 48 | 74 | 56 | 82 | 42 | 68 | 62 | 85 |
| | | 98 | 36 | 67 | 58 | 81 | 48 | 73 | 66 | 87 |
| | | 100 | 35 | 64 | 60 | 80 | 45 | 72 | 64 | 84 |
| 3 | Product mix; Congestion level | 85 | 20 | 62 | 33 | 54 | -27 | 28 | 35 | 70 |
| | | 90 | 29 | 64 | 37 | 56 | 1* | 51 | 41 | 77 |
| | | 95 | 10 | 54 | 28 | 49 | 4 | 56 | 32 | 72 |
| | | 98 | 7 | 47 | 20 | 47 | 3 | 46 | 19 | 63 |
| | | 100 | 28 | 55 | 46 | 63 | 34 | 63 | 45 | 72 |
| 4 | Unreliability; Product mix; Congestion level | 85 | -12 | 16 | 6 | 15 | -8 | 14 | 5 | 33 |
| | | 90 | 0 | 23 | 14 | 23 | 2* | 27 | 19 | 50 |
| | | 95 | 14 | 36 | 22 | 29 | 16 | 38 | 28 | 50 |
| | | 98 | 31 | 51 | 31 | 51 | 34 | 56 | 38 | 64 |
| | | 100 | 38 | 58 | 44 | 59 | 48 | 65 | 51 | 70 |
| 5 | Proc time variability; Unreliability; Product mix; Congestion | 85 | -16 | 18 | 0 | 8 | -10 | 28 | -2* | 37 |
| | | 90 | -7 | 24 | -11 | 10 | -14 | 23 | -4 | 34 |
| | | 95 | -4 | 17 | 9 | 25 | -4 | 20 | 17 | 35 |
| | | 98 | 22 | 24 | 29 | 26 | 18 | 31 | 35 | 43 |
| | | 100 | 31 | 28 | 39 | 36 | 29 | 37 | 50 | 53 |

*: Not statistically significant;
TH: Normalized Throughput = (Achieved Throughput / Best Expected Throughput) * 100%
CTAVG: Average Cycle Time; CTSTD: Standard Deviation of Cycle Time
Percentage Improvement over UNIF = (1− Achieved Performance/ UNIF Performance)*100%

The advantage of WIPLCtrl should attribute to the characteristic of WIPLOAD, which can achieve more efficient response to system stochastic events by taking into considering the remaining processing times of the jobs in the system, so that WIPLCtrl can compensate for more system disturbances and reduce the unexpected WIP accumulation to some extent. Therefore, WIPLCtrl can be considered as an efficient job release methodology for a manufacturing system with high output and variability levels. Meanwhile, WIPLCtrl is not restricted by the issue of identification of the bottleneck machine of a production line

## 4 CASE STUDY

To further evaluate the performance of WIPLCtrl, the assumptions of the simplified wafer fab model presented above are relaxed by simulating a real-life wafer fab of Chartered Semiconductor Manufacturing (Chartered) using AutoSched$^{TM}$ AP. The performance of WIPLCtrl is compared with that of CONWIP and UNIF. Chartered, founded in 1987 in Singapore, is one of the world's top three pure-play silicon foundries, providing advanced technology wafer manufacturing services for the global semiconductor industry.

### 4.1 Model Description

The manufacturing system studied in this case study possesses all the characteristics and complexities of a typical wafer fab. There are 511 machines involved. A total of 37 products are produced, belong to 7 product categories. Each prodcut has an individual processing route. The operation step of each part type strictly follows its real-life process, which usually consists of 200 to 300 steps. The raw processing times range from 200 hrs to 400 hrs. Diverse equipment characteristics are considered. For example, batch processing machines are simulated, where a number of lots are processed simultaneously as a batch. The time interval that a multi-capacity machine wait before processing a subsequent batch and the time interval that a machine must wait before inducing a new piece for processing are modeled. The workstations are located at 23 areas. The transportation times between these areas are taken into account. Both preventive maintenance (PM) and unscheduled breakdowns are simulated. There are five levels of scheduled maintenance for the workstations including weekly, monthly, quarterly, half-yearly and yearly PM. The unpredictable breakdowns are assumed to be exponentially distributed with the mean values estimated according to the historical data. However, in this model, the issue of operator availability is not considered, and job rework is not modeled

### 4.2 Model Verification and Validation

To verify the simulation model, the program is checked and debugged carefully. The technique of "trace" is also used. The "correlated inspection approach" (Law and Kelton 2000) is utilized to validate the model. Historical data from the actual fab are collected. By comparing the outputs of the actual fab and the simulation model in terms of fab outputs and cycle time performance, the model is considered to be accurate enough.

### 4.3 Experimental Results

In this case study, two output levels are considered, which are referred to as low and high output levels respectively. The simulation results are the average values of ten independent replications. The simulation length for each run is three years (25,920 hrs), in which the beginning half a year (4320 hrs) is considered as warm-up period. Note that the wafer start volumes and the outputs of the simulation model presented here are not the real numbers used in Chartered due to the confidentiality of data.

Table 6 shows the simulation results with standard errors, which indicate that the system performance of the wafer fab can be significantly improved by an appropriate choice of release control methodology. WIPLCtrl outperforms UNIF and CONWIP in terms of both the mean and the variance of cycle time for a given output level. The improvement of WIPLCtrl becomes more significant when the system is operating at a relatively high output level. From the industrial practice point of view, this is an important meritorious characteristic of WIPLCtrl since wafer fabs are usually expected to be operating at a high output level so that the system capacity can be fully utilized.

Table 7 lists the percentage improvements of WIPLCtrl over UNIF and CONWIP on average cycle time and standard deviation of cycle time. Paired student's t-test is used to do the statistical analysis with a 95% confidence level. These improvements can also be understood as the improvements on throughput for a given cycle time level. The advantage of WIPLCtrl could potentially lead to a considerable amount of increased benefits due to the reduced costs and the increased revenue, given the large capital investments and sales revenue of semiconductor manufacturing.

Table 6: Case study simulation results

| Release | Output Level | Output (lots) | CTAVG (hrs) | CTSTD (hrs) |
|---|---|---|---|---|
| UNIF | Low | 43282.7 | 431.72 (±7.06) | 122.95 (±2.75) |
| | High | 45530.1 | 687.40 (±34.37) | 211.09 (±12.67) |
| CONWIP | Low | 43294.6 | 426.63 (±3.80) | 105.84 (±1.02) |
| | High | 45622.2 | 590.54 (±1.28) | 138.52 (±1.17) |
| WIPLCtrl | Low | 43296.9 | 408.60 (±1.02) | 98.49 (±0.64) |
| | High | 45625.4 | 543.18 (±1.80) | 123.66 (±1.06) |

Table 7: Percentage improvements of WIPLCtrl

| Output Level | Over UNIF (%) | | Over CONWIP (%) | |
|---|---|---|---|---|
| | CTAVG | CTSTD | CTAVG | CTSTD |
| Low | 5 | 20 | 4 | 7 |
| High | 21 | 41 | 8 | 11 |

## 5 CONCLUSIONS

Job release control has a significant impact on wafer fab performance. In this paper, we proposed a new job release methodology, WIPLCtrl. The main idea is to maintain "system WIPLOAD" at a specified level by controlling the release of new jobs into the system. "System WIPLOAD" is proposed to measure the workload of the overall shop floor, taking into account the distribution of jobs along the production line with the involvement of their remaining processing times. The performance of WIPLCtrl is evaluated in a stimualtion study on a simplifed wafer fab model. A case study is also conducted by simulating a real-life wafer fab. The results of the simulation experiments indicate that WIPLCtrl could efficiently improve the cycle time performance for a given throughput level in comparison with WR, CONWIP, StnLoad, and UNIF. This benefit can also be understood as the improvement on system productivity for a given cycle time level. Moreover, WIPLCtrl is observed to bring more performance improvement when the system is operating on a relatively high throughput level. The performance of WIPLCtrl is reliable with the increase of variability in a manufacturing system.

WIPLCtrl is looked upon as an effective and reliable release control methodology for semiconductor wafer fabs. First, WIPLCtrl can be easily implemented in a wafer fab environment; the data needed can be collected by the computer-integrated-manufacturing system. By setting the reference WIPLOAD level, the manufacturing manager is able to control the workload level of the system, and consequently to obtain an expected system output level. Secondly, it was observed that the performance of WIPLCtrl is reliable with the increase in the variability in the manufacturing system it is applied. A typical semiconductor manufacturing is extremely complex and highly dynamic with the involvement of multiple product types with different processes and hundreds of unreliable machines. Under this circumstance, a release control methodology possessing reliable responsiveness and robustness to system disturbances such as WIPLCtrl is preferable. Moreover, the product types present in a wafer fab greatly vary with the oscillations of customer demand. Equipment is often out of service for scheduled maintenance or unpredictable breakdowns. Therefore system bottleneck machine can appear at different plances at different times, especially from the shor-term production control perspective. This practical issue constrains the implementation of the bottleneck-based release methodologies such as WR because any mistake in identifying the bottleneck equipment will significantly deteriorate the system performance.

Our ongoing work is to study a multi-stage WIPLCtrl problem. The manufacturing process can be divided into multiple stages, for each stage WIPLCtrl is used to control the workload level.

## ACKNOWLEDGMENTS

This research work is supported by Singapore-MIT Alliance (SMA). The authors would like to acknowledge the members of Industrial Engineering Manufacturing Systems Group of Chartered Semiconductor Manufacturing Limited (Chartered), especially to Mr. Tan Hiap Keong and Mr. Lieu Sip Khean, for helping to build the simulation model of Chartered fab.

## REFERENCES

Fowler, J. and J. Robinson. 1995. Measurement and Improvement of Manufacturing Capacity (MIMAC) Designed Experiment Report. Technology Transfer 95062860A-TR. SEMATECH.

Gilland, W. G. 2002. A Simulation Study Comparing Performance of CONWIP and Bottleneck-Based Release Rules. *Production Planning and Control*. 13: 211-219.

Glassey, C. R., and M. G. C. Resende. 1988. Closed-Loop Job Release Control for VLSI Circuit Manufacturing. *IEEE Transactions on Semiconductor Manufacturing*. 1: 36-46.

Hogg, G. and J. Fowler. 1991. Flow Control in Semiconductor Manufacturing: A Survey and Projection of Needs. Non-Confidential Document 911110757A-GEN. SEMATECH.

Hopp, W. J., and M. L. Spearman. 2000. *Factory Physics*. Irwin/McGraw-Hill.

Kim, J., R. C. Leachman, and B. Suh. 1996. Dynamic Release Control Policy for the Semiconductor Wafer Fabrication Lines. *Journal of the Operational Research Society*. 47: 1516-1525.

Law, A. M. and W. D. Kelton. 2000. *Simulation Modeling and Analysis*. Singapore: McGraw-Hill.

Philipoom, P. R., M. K. Malhotra, and J. B. Jensen. An Evaluation of Capacity Sensitive Order Review and Release Procedures in Job Shops. *Decision Sciences.* 24: 1109-1133.

Spearman, M. L., D. L. Woodruff, and W. J. Hopp. 1990. CONWIP: A Pull Alternative to Kanban. *International Journal of Production Research*. 28: 879-894.

Wein, L. W. 1988. Scheduling Semiconductor Wafer Fabrication. *IEEE Transactions on Semiconductor Manufacturing* 1:115-130.

Welch, P. D. 1983. *The Computer Performance Modeling Handbook*. New York: Academic Press.

## AUTHOR BIOGRAPHIES

**CHAO QI** is a research fellow of Singapore-Massachusetts Institute of Technology (MIT) Alliance (SMA), Manufacturing Systems and Technology (MST) Programme. She received the Bachelor degree from Huazhong University of Science and Technology, China, in 1998, and the Ph.D. degree in systems and engineering management from Nanyang Technological University, Singapore, in 2006. Her research interests include production planning and control, modeling and simulation for manufacturing systems.

**APPA IYER SIVAKUMAR** is an Associate Professor in the School of Mechanical and Aerospace Engineering (MAE) at the Nanyang Technological University, Singapore and a Faculty Fellow of Singapore - Massachusetts Institute of Technology (MIT) Alliance (SMA-MST programme) He was at Gintic Institute of Manufacturing Technology, Singapore prior to this appointment. His research interests are in the area of OR, advanced Manufacturing Systems engineering, discrete event Simulation, Scheduling, Logistics, Supply chain design, and Research Methodology. He has many years of industrial and research experience in the UK prior to post in Singapore in 1993. He held various senior positions including technical manager and project manager for a number of years at Lucas Systems and Engineering and Lucas Automotive, UK. During this period he was responsible for manufacturing re-engineering projects and involved in the introduction of IT and Scheduling systems. He received a Bachelors of Engineering in Manufacturing Systems Engineering and a PhD in Manufacturing Systems Engineering from University of Bradford, UK. He has made many contributions at international conferences and journals. He has trained and supervised a number of PhD students and many Master's level students. He has been the technical chair and co-edited the proceedings of the 3rd and 4th International Conference on Computer Integrated Manufacturing (ICCIM '95 and ICCIM'97), Singapore.

**STANLEY B. GERSHWIN** is a Senior Research Scientist at the MIT Department of Mechanical Engineering. He received the B.S. degree from Columbia University in 1966; and the M.A. and Ph.D. degrees in Applied Mathematics from Harvard University in 1967 and 1971. In 1970-71, he was employed by the Bell Telephone Laboratories, where he studied telephone system capacity. At the C. S. Draper Laboratory in Cambridge, Massachusetts, from 1971-75, he investigated problems in manufacturing and in transportation. He worked in the MIT Laboratory for Information and Decision Systems during 1975-1987. He was Professor of Manufacturing Engineering at the Boston University College of Engineering in 1986-1987. Dr. Gershwin teaches MIT courses in Manufacturing Systems He is a member of the MIT Laboratory for Manufacturing and Productivity and he is also affiliated with MIT's Leaders for Manufacturing Program and Operations Research Center. Dr. Gershwin is the author of "Manufacturing Systems Engineering" (Prentice-Hall, 1994) and numerous papers in international journals. He is a co-editor of "Analysis and Modeling of Manufacturing Systems"(Kluwer, 2002). One of his papers was awarded both the Best Paper Award for the IIE Transactions focus issues on Design and Manufacturing for 2000, and the Outstanding IIE Publication Award for 2000-2001. He is a co-author of a paper that was awarded both the Best Paper Award for the IIE Transactions focus issues on Design and Manufacturing for 2006. His research interests include real-time scheduling and planning in manufacturing systems; hierarchical control; dynamic programming in hybrid (discrete and continuous state) systems; and decomposition methods for large scale systems. His major research goal is the development of an engineering theory of manufacturing systems. He and his students have performed research projects and consulted for such companies as Boeing, General Motors, Hewlett Packard, Johnson & Johnson, Peugeot, Polaroid, United Technologies, and others. Dr. Gershwin is a member of the IEEE Control Systems Society, the IEEE Robotics and Automation Society, the Operations Research Society of America, the Institute of Industrial Engineers, and the Society of Manufacturing Engineers. He has been an Associate Editor of several international journals, including International Journal of Production Research, Operations Research, and IEEE Transactions on Automatic Control. He has been a member of the Scientific committee of the 1997-2007 series of Conferences on Analysis of Manufacturing Systems. Dr. Gershwin is a Fellow of the IEEE.