Prediction of Incoming Work-In-Progress in Semiconductor Fabrication Foundry using Long Short-Term Memory Recurrent Neural Network

Tin Tze Chiang

Master of Science

2018

Prediction of Incoming Work-In-Progress Prediction in Semiconductor Fabrication Foundry using Long Short-Term Memory Recurrent Neural Network

Tin Tze Chiang

A thesis submitted

In fulfilment of the requirements for the degree of

Master of Science

Faculty of Computer Science and Information Technology

UNIVERSITI MALAYSIA SARAWAK

2018

UNIVERSITI MALAYSIA SARAWAK

Grade: _____

Please tick (√)

Final Year Project Report ☐

Masters ☑

PhD ☐

DECLARATION OF ORIGINAL WORK

This declaration is made on the 10 day of October 2018.

Student's Declaration:

I, TIN TZE CHIANG, 15020052, FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHONLOGY, hereby declare that the work entitled, Prediction of Incoming Work-In-Progress in Semiconductor Fabrication Foundry using Long Short-Term Memory Recurrent Neural Network, is my original work. I have not copied from any other students' work or from any other sources except where due reference or acknowledgement is made explicitly in the text, nor has any part been written for me by another person.

10/10/2018
Date submitted

TIN TZE CHIANG (15020052)
Name of the student (Matric No.)

Supervisor's Declaration:

I, CHIEW KANG LENG, hereby certifies that the work entitled, Prediction of Incoming Work-In-Progress in Semiconductor Fabrication Foundry using Long Short-Term Memory Recurrent Neural Network, was prepared by the above named student, and was submitted to the "FACULTY" as a * partial/full fulfillment for the conferment of MASTER OF SCIENCE (COMPUTER SCIENCE), and the aforementioned work, to the best of my knowledge, is the said student's work

Received for examination by: _____ Date: 10 OCT 2018

(Name of the supervisor)
CHIEW KANG LENG

I declare this Project/Thesis is classified as (Please tick (√)):

☐ **CONFIDENTIAL** (Contains confidential information under the Official Secret Act 1972)*
☐ **RESTRICTED** (Contains restricted information as specified by the organisation where research was done)*
☑ **OPEN ACCESS**

Validation of Project/Thesis

I therefore duly affirmed with free consent and willingness declared that this said Project/Thesis shall be placed officially in the Centre for Academic Information Services with the abide interest and rights as follows:

- This Project/Thesis is the sole legal property of Universiti Malaysia Sarawak (UNIMAS).
- The Centre for Academic Information Services has the lawful right to make copies for the purpose of academic and research only and not for other purpose.
- The Centre for Academic Information Services has the lawful right to digitise the content to for the Local Content Database.
- The Centre for Academic Information Services has the lawful right to make copies of the Project/Thesis for academic exchange between Higher Learning Institute.
- No dispute or any claim shall arise from the student itself neither third party on this Project/Thesis once it becomes sole property of UNIMAS.
- This Project/Thesis or any material, data and information related to it shall not be distributed, published or disclosed to any party by the student except with UNIMAS permission.

Student's signature _____ (Date) 10 Oct 2018

Supervisor's signature: _____ (Date) 10 Oct 2018

Current Address:
No. 139, Taman Sunny Hill, Jalan Rock, 93250 Kuching, Sarawak.

Notes: * If the Project/Thesis is **CONFIDENTIAL** or **RESTRICTED**, please attach together as annexure a letter from the organisation with the period and reasons of confidentiality and restriction.

The instrument was duly prepared by The Centre for Academic Information Services]

DECLARATION

I declare that the work in this thesis was carried out in accordance with the regulations of Universiti Malaysia Sarawak. It is original and is the result of my work, unless otherwise indicated or acknowledged as referenced work. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Name : Tin Tze Chiang

Metric No. : 15020052

Faculty : Faculty of Computer Science and Information Technology
Universiti Malaysia Sarawak

Dated : April, 2018

ACKNOWLEDGEMENT

The funding for this project is made possible through the research grant obtained from UNIMAS under the Special FRGS 2016 Cycle [Grant No: F08/SpFRGS/1533/2017].

I would like to express my sincere gratitude to my supervisor Dr. Chiew Kang Leng for our weekly meetings and his great supervision. His continuous guidance, support and encouragement have helped me to overcome the challenges encountered during the research process. His advices were crucial in completing this thesis.

I would like to extend my appreciation X-FAB Sarawak Sdn. Bhd. for their support in this research by providing the environment and resources to extract and compute for the relevant data for this project.

Finally, I would like to thank my parents, who has taught me the value of learning and not to give up on tough times and always showed me unwavering love. This work is dedicated to them.

ABSTRACT

Semiconductor process is one of the most complex processes in the manufacturing industries. In order to meet the high demands of semiconductor components, cycle time reduction in semiconductor manufacturing is crucial. One of the factors that can increase cycle time is the equipment preventive maintenance activity. Preventive maintenance activity is crucial to maintain or increase the operational efficiency and reliability of the equipment and minimises unanticipated failures due to faulty parts. Despite its criticality, preventive maintenance activity requires significantly long hours, thereby increasing cycle time. Therefore, this activity is usually performed when the equipment group is forecasted to have low incoming Work-in-Progress. However, current statistical forecasting approach has low accuracy because it lacks the ability to capture the time-dependent behaviour of the Work-in-Progress. In this work, we present an approach to Work-in-Progress arrival forecasting by using LSTM neural network. The LSTM forecasting model is trained using the historical incoming Work-in-Progress of an equipment group. The LSTM forecasting model is then required to forecast the incoming Work-in-Progress of three consecutive weeks to the same equipment group. The results of this study are compared with the results of the statistical method practised by the manufacturing production control engineers in the sponsoring company of this research. The comparison showed that the result proposed in this study is superior to the statistical method in the sponsoring company. The proposed model achieved value accuracy of 100%, 100%, 85.7% and trend accuracy of 66.7%, 100%, 50%, with respect to the three consecutive weeks of comparison.

Keyword: Semiconductor, cycle time, preventive maintenance activity, Work-In-Progress, forecasting, statistical method, LSTM

Peramalan Ketibaan Kerja-dalam-Kemajuan di FAB Semikonduktor dengan Menggunakan Pembelajaran Mesin

ABSTRAK

Proses semikonduktor adalah salah satu proses yang paling kompleks dalam industri pembuatan. Untuk memenuhi permintaan komponen semikonduktor yang tinggi, pengurangan kitaran masa (CT) dalam pembuatan semikonduktor adalah penting. Salah satu faktor yang boleh meningkatkan CT ialah aktiviti penyelenggaran pencegahan (PM). Aktiviti PM adalah penting untuk mengekalkan atau meningkatkan kecekapan operasi, kebolehpercayaan peralatan, dan meminimumkan kegagalan peralatan yang tidak dijangkakan. Aktiviti PM boleh meningkatkan CT sebab ia memerlukan masa yang panjang. Oleh itu, aktiviti PM biasanya dilaksanakan apabila ketibaan Kerja-Dalam-Proses (WIP) ke kumpulan peralatan diramalkan rendah. Tetapi, cara ramalan statistik semasa kurang tepat sebab ia tidak dapat memfaktorkan hubungan WIP terhadap masa. Dalam penyelidikan ini, satu cara peramalan ketibaan WIP menggunakan rangkaian saraf LSTM adalah dipersembahkan. Model peramalan LSTM dilatih menggunakan sejarah ketibaan WIP ke salah satu kumpulan peralatan. Kemudian, ia diuji dengan meramalkan ketibaan WIP tiga minggu yang berturutan ke kumpulan peralatan tersebut. Keputusan peramalan dibandingkan dengan keputusan peramalan kaedah syarikat penajaan penyelidikan ini. Perbandingan ini menunjukkan kaedah yang dicadangkan adalah lebih unggul berbandingkan cara syarikat penaja. Kaedah yang dicadangkan mencapai ketepatan nilai 100%, 100%, 85% dan ketepatan trend 66.7%, 100%, 50% untuk ketiga-tiga minggu perbandingan yang berkenaan.

Kata-kunci: *Semikonduktor, kitaran masa, aktiviti penyelenggaran pencegahan, Kerja-Dalam-Kemajuan, peramalan, cara statistik, LSTM*

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AC	Autocorrelation
AIC	Akaike Information Criterion
ANN	Artificial neural network
AR	Autoregressive
ARIMA	Autoregressive Integrated Moving Average
ASR	Automatic speech recognition
BIC	Bayesian Information Criterion
BPTT	Back-propagation-through-time
CMP	Chemical mechanical planarization
CT	Cycle time
DBN	Deep-belief network
DNN	Deep neural network
DPML	Day per Mask Layer
FNN	Feedforward Neural Network
GP	Gaussian Process (GP)
IWIP	Incoming Work-In-Progress
kNN	K-Neareset Neighbour
LSTM	Long Short-Term Memory
MA	Moving average
MAE	Mean absolute error
MES	Manufacturing execution system
MLE	Maximum Likelihood Estimation
MLP	Multilayer Perceptron
MRE	Mean relative error

MSE	Mean squared error
PAC	Partial Autocorrelation
PM	Preventive maintenance
RMSE	Root mean squared error
RNN	Recurrent neural network
RNN	Recurrent neural network
SAS	Simple Exponential Smoothing/Single Exponential Smoothing
SGD	Stochastic Gradient Descent
SRN	Simple recurrent network
SSE	Sum of squared errors
SVC	Support Vector Classification
SVM	Support Vector Machine
SVR	Support Vector Regression
TAT	Turn-around-time
WIP	Work-In-Progress

CHAPTER 1

INTRODUCTION

The semiconductor process is one of the most complex processes in the manufacturing industries with unique characteristics such as extremely long manufacturing process, complex series of hundreds of sequential processing steps, process steps dedication at designated tools, strict process window time frame, and dynamic product mix-run environment (Rozen & Byrne, 2016). These processes need to be completed almost flawlessly as defects in the manufactured items can bring about costly impacts in today's world of technologies where electronic devices play vital roles in daily living.

In general, semiconductor fabrication consists of seven major process steps. These process steps are photolithography, etching, chemical mechanical planarization (CMP), oxidation, ion implantation and diffusion. Briefly described, photolithography process is the process to transfer the pattern from a photomask to the surface of a wafer. By forming a layer of photoresist on the top of the wafer, the wafer is then sent to the etching process for the photoresist to be developed.

An etching mask is applied in this process to define the parts to remain and remove. At the deposition process, various layers of different materials is deposited on the wafer as part of the fabrication process. CMP process is a process used to plane the wafer surface using chemical slurry. This process is necessary because processes such as etching and deposition modify the wafer surface, leading to a non-planar surface. Oxidation is a process to convert silicon on the wafer into silicon dioxide, which is a layer to be used as masks for ion implantation process.

Ion implantation and diffusion process are used to introduce a controlled amount of impurities into the wafers to change their electrical properties. Ion implantation introduce

the dopant impurities into the crystalline silicon using electric fields and the diffusion process then moves the impurity atoms in the wafer at high temperatures. These steps are repeated until the required design is completed. The sequence of steps for the wafer to move through is defined by manufacturing execution system (MES). The defined steps are commonly called a process flow. Figure 1.1 further illustrates a non-exhaustive example of process flow of semiconductor fabrication process (Quirk & Serda, 2001).

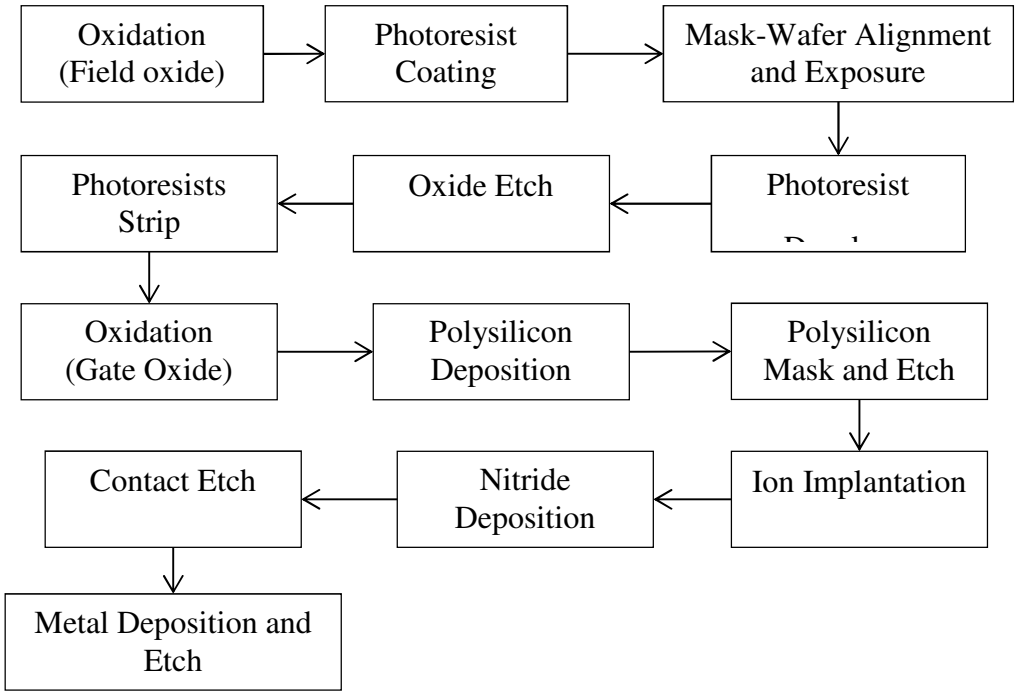

Figure 1.1: Example of semiconductor fabrication process flow

One of the main goals of semiconductor manufacturing plant, or commonly known as fab, is to reduce cycle time (CT) because reduced CT results in lower inventory costs, improved quality, and better response to demand changes and shorter time-to-market for new products (Rozen & Byrne, 2016). In the context of semiconductor manufacturing, CT is the total time it takes for a lot, which is a batch of wafers stored in a single cassette pod serve as the transportation means for the wafers production line, to be processed, plus the total time spent on waiting. Lots that are waiting to be processed in the production line are

commonly referred to as Work-in-Progress (WIP) in the manufacturing industries. The measurement unit for CT is Days per Mask Layer, (DPML), which is the number of days committed to process one photolithography layer.

One of the factors that can contribute to long waiting time, thereby increasing CT, is the equipment preventive maintenance activity. Preventive Maintenance (PM) is an activity that takes the entire machine off-line, or partially off-line, to carry out prescribed maintenance activity for maintaining or increasing the operational efficiency and reliability of the tool and minimizing unanticipated failures due to faulty parts (Ramirez-Hernandez et al., 2010). An unhealthy machine can cause various defects and chemical contaminations on the wafers. An affected wafer may need a rework by re-entering a sequence of chemical process steps to remove and fabricate the affected layers. Major defects or contamination could lead to scrapped wafer. Both of these activities incur losses to the company in both time and revenue. Most importantly, it affects the trust of the customers for potential future businesses.

In a fab environment, a set of machines is usually allocated for each process step to allow the lots to flow through in shorter time. The set of machines that perform the same wafer fabrication process step are commonly grouped logically and termed as equipment group. Figure 1.2 illustrates the logical grouping of six machines into an equipment group called Group A.

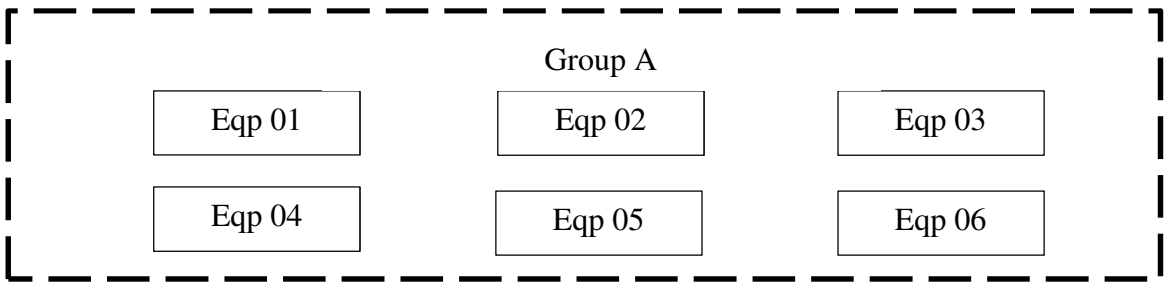

Figure 1.2: A logical group of equipment into an equipment group

Although PM activities are vital, PM downtime can be costly because it takes significantly long hours. As the complexity of the semiconductor manufacturing equipment increases, PM activities have also become more complex and would require longer time to complete. If one of the machines is taken offline for PM activities, the equipment group will experience insufficient capacity to process the incoming WIP. This will cause the production line to suffer WIP bottleneck that results in non-linear distribution of the WIP across the fab. Non-linear WIP distribution across the fab is highly undesirable because it will cause operational losses to the equipment at subsequent process steps as they become frequently idle due to the lack of WIP to process. In addition, WIP bottleneck also brings negative impact to the CT. This is because the CT duration will increase as the total waiting time of the impacted lots increases.

In the terminology of semiconductor manufacturing, the total wafers that each machine completed is called total moves of that machine, and the total wafers completed by the machines of an equipment group is called the total moves of that equipment group. Consider the example of logical group of six machines presented in

Figure 1.2, the total moves of the six machines in equipment group A on a particular day is the total moves of equipment group A for that day. The completed wafers from equipment group A are then moved to their designated process steps in the form of cassette pods. With this wafer movement, the completed wafers from equipment group A become the WIP of the next equipment groups. Figure 1.3 illustrates the typical relation of moves and WIP among equipment groups. In Figure 1.3 , S_1, S_2, S_3 and W represents four different equipment groups in a fab. S_1, S_2 and S_3 represents three equipment groups that have completed wafers designated to equipment group W. The wafers completed from S_1, S_2 and S_3 are the moves of S_1, S_2 and S_3, respectively. These completed wafers from S_1, S_2 and S_3 designated to be processed next at W will become the WIP of W.

It is also possible that the completed wafers from W becomes the WIP of W again.

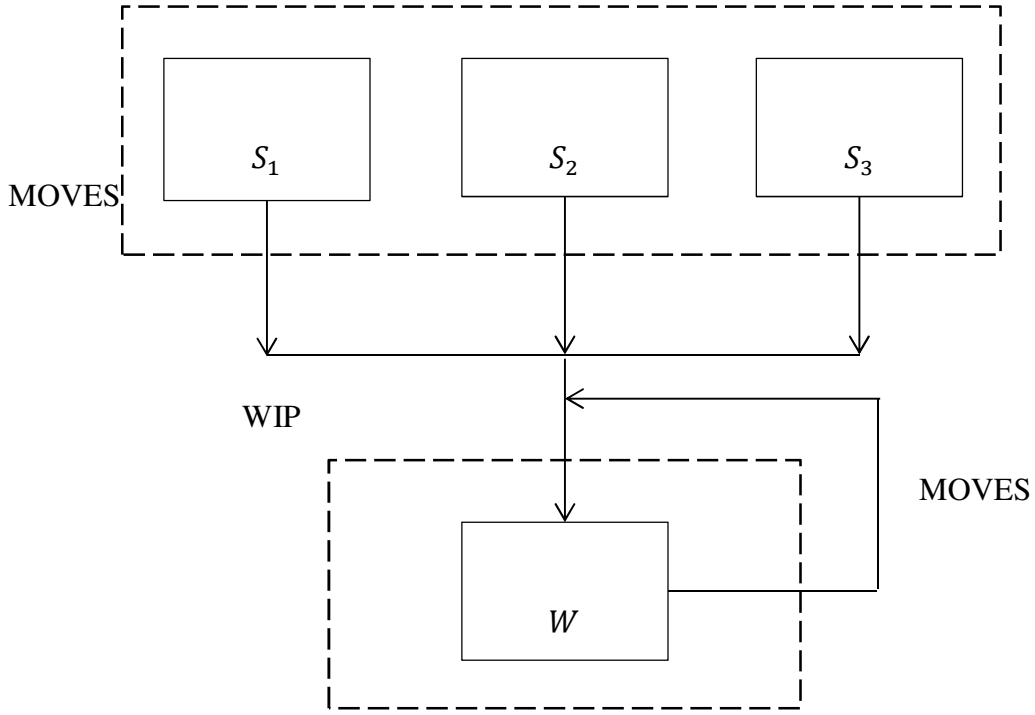

Figure 1.3: Typical relation of moves and WIP among fab equipment groups

In order to prevent long CT during PM activities is to perform the PM when the equipment group of the target machine is expected to have low incoming WIP. Referring to Figure 1.3, if one of the machines in W is scheduled for PM activity, it is desirable that the expected WIP arriving at W to be low. It is therefore essential to forecast the arrival of WIP to assist in PM activities planning for machines in W.

In semiconductor manufacturing, the WIP arrival forecasting commonly done by manufacturing production control engineers. A commonly to used approach to forecast the WIP arrival is through statistical aggregation based on the turn-around-time (TAT) of the wafers per operation step in the process flow. This is accomplished by first determine the product of interest to forecast, followed by retrieving the complete process flow of the product. For each of the process steps in the retrieved process flow, the TAT for an operation step is calculated, followed by calculating the rate of wafer flow in the

production. With both TAT and flow rate calculated, the CT for each operation step can be estimated. If a forecast of lot arrival for the next 24 hours is required, the engineer will first sum up the CT of each operation step of a lot of the product until it reaches 24 hours, and from there, examine the sum of estimated wafer to arrive at the step of interest.

1.1 Problem Statement

Although statistical forecasting method is commonly used to forecast the WIP arrival to an equipment group, the current statistical forecasting method contains the following limitations.

i. Statistical forecasting approach lacks the ability to capture the time-dependent behavior of the WIP in the production lines. This is because statistical approach assumed that an observed series of data is caused by random process and each occurrence is independent from each other. However, such assumption is not true for WIP behavior in the production lines which is time series in nature.

ii. Statistical forecasting approach is product-specific. Each product that is present in the production line needs to be modelled separately to create its forecasting model. The reason for this specificity is because each product has its own process steps sequence, and the statistical forecasting model is dependent on these process steps to perform forecasting. As a result of the separate modelling, statistical forecasting approach indirectly assumed that each product has dedicated resources such as manpower and equipment to cater to its production activity in the production environment. This assumption is not true as both manpower and equipment are shared among all the products in the production environment.

iii. The assumptions of non-dependency in the time series data and non-resource sharing production environment reduced the accuracy of statistical forecasting model. The forecast results of most of the statistical forecasting models are not within the accuracies level where they are sufficient for critical decision making in the production environment, but can only serve as general guidelines or references.

1.2 Motivation

Recently, a new approach to arrival forecasting was introduced by using a type of recurrent neural network (RNN), called the Long-Short Term Memory (LSTM) neural network. LSTM has been proven to achieve higher accuracy in traffic flow prediction over statistical approach by various research works (Vlahogianni et al., 2005; Lv et al., 2015; Ma et al., 2015; Tian & Pan, 2015; Duan, Lv, & Wang, 2016; Fu et al., 2016). Besides traffic flow prediction, LSTM is also used in research such as log-driven information technology system failure prediction to discover long-range structure in historical data (Zhang et al., 2016), gesture recognition (Zhu et al., 2017), voice conversion (Lai et al., 2016), and aircraft engines excess vibration events predictions (ElSaid et al., 2016). Motivated by the cited works, this research work aims to investigate the efficiency of LSTM in predicting the incoming WIP arrival for an equipment group of interest as a mean to assist in PM activities planning. With high prediction accuracy, PM activities planning could be improved to avoid major obstacle to WIP movement in the production line and thus avoid negative impact to the CT of the fab.

1.3 Research Objectives

The main objective of this research is to develop an incoming WIP arrival forecasting model that can minimize occurrence of long CT due to PM activities and