Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 36 (2009) 3421–3427
Expert Systems with Applications
www.elsevier.com/locate/eswa

# Using an artificial neural network prediction model to optimize work-in-process inventory level for wafer fabrication

Yu-Hsin Lin [a,*], Jie-Ren Shie [b], Chih-Hung Tsai [c]

[a] *Department of Industrial Engineering and Management, Ming-Hsin University of Science and Technology, No. 1, Hsin-Hsin Road, Hsin-Feng, 30401 Hsin-Chu County, Taiwan*
[b] *Department of Mechanical Engineering, Ming-Hsin University of Science and Technology, No. 1, Hsin-Hsin Road, Hsin-Feng, 30401 Hsin-Chu County, Taiwan*
[c] *Department of Industrial Engineering and Management, Ta-Hwa Institute of Technology, 1 Ta-Hwa Road, Chung-Lin, Hsin-Chu, Taiwan, ROC*

**Abstract**

A proper selection of a work-in-process (WIP) inventory level has great impact onto the productivity of wafer fabrication processes, which can be properly used to trigger the decision of when to release specific wafer lots. However, the selection of an optimal WIP is always a tradeoff amongst the throughput rate, the cycle time and the standard deviation of the cycle time. This study focused on finding an optimal WIP value of wafer fabrication processes by developing an algorithm integrating an artificial neural network (ANN) and the sequential quadratic programming (SQP) method. With this approach, it offered an effective and systematic way to identify an optimal WIP level. Hence, the efficiency of finding the optimal WIP level was greatly improved.


*Keywords:* Work-in-process level; Neural network; Sequential quadratic programming

## 1. Introduction

The semiconductor manufacturing is the most complex manufacturing process in the world (Leachman & Hodges, 1996). The maximization of the throughput rate, the minimization of the cycle time and the standard deviation of the cycle time are the primary performance measurements in wafer fabrication (Huang et al., 1999). However, the semiconductor business environments are challenged by the rapid changes in responsiveness (Qiu, 2005). In practice, production planners use the work-in-process (WIP) inventory level profile to control the material flow and simplify the production control. Controlling WIP is a proven approach in improving the responsiveness of manufacturing systems and services (Qiu, 2005). However, the WIP and the cycle time are convex increasing functions of the throughput (Lin & Lee, 2001). An infinite WIP level maximizes the throughput which cannot exceed the capacity of the bottleneck workstation (Buzacott, 1971). The inherent conflict in the determination of a proper WIP level is obvious when attempting to both maximize the throughput rate and minimize the cycle time. Since different WIP levels can influence the throughput rate (TP), the mean cycle time (CT) as well as the standard deviation of the cycle time (SD), a proper selection of the WIP is crucial to the productivity of wafer fabrication (Lin & Lee, 2001; Miller, 1990). Although there are many WIP-based release control policies which have been proven to be effective (Glassey & Resende, 1988a,b; Glassey and Weng, 1991; Graves, 1995; Qiu, 2005; Tsai, Chang, & Li, 1997; Wein, 1988) for semiconductor manufacturing, few methods have been proposed to find a suitable WIP level as a parameter for those policies. The objective of this study is to determine an optimal WIP level that meets the tradeoff of maximizing

* Corresponding author. Tel.: +886 968 623 090; fax: +886 3 5593142x3212.
*E-mail address:* brianlin@must.edu.tw (Y.-H. Lin).

0957-4174/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2008.02.009

the throughput rate, minimizing the mean cycle time and the standard deviation of the cycle time.

Simulation, regression and neural networks are the most widely used approaches to predict the production performance (Huang et al., 1999). Simulation software packages (e.g. Autoched, Mansim, and Adexa etc.) are common tools to determine the WIP levels in practice. However, there are few drawbacks for this method. First of all, a simulation model is applicable to a specific system only to the extent that the features contained adequate information representing that system, and it is very time-consuming to run a simulation model (Connors, Feigin, & Yao, 1996; Miller, 1990). Furthermore, it takes a long time to build and debug simulation models (Suri, Diehl, de Treville, & Tomsicek, 1995); and it tends to find a suitable rather than an optimal solution. Thus, it is important to recognize that this approach only works for a specific system. In addition, applying the model hinges on the definition of simulation objectives, the availability and accuracy of data and assumptions, the verification and validation of the model for the specific system under study, and the analysis and interpretation of simulation results (Miller, 1990).

Although regression equations are normally utilized to find the relationship between the inputs and the outputs, it is very difficulty to determine a proper order for regression equations. If the order of a selected equation is not sufficient, it results in low prediction fidelity of output values. On the other hand, if a higher than necessary order of equations is chosen, undesirable oscillations of the output values occur and even reduce the generalization (prediction) capability of equations.

Recently, neural networks become very popular in various applications and are successfully implemented in manufacturing processes (Udo, 1992; Huang & Zhang, 1995). In this study, a multi-output parameter design problem, which intends to identify an optimal WIP level that yields a maximum throughput as well as minimizes the cycle time and the standard deviation of the cycle time, is solved by integrating the neural network capability with the sequential quadratic programming (SQP) method (Fletcher, 1981). To resolve this type of problems, Das (1999) proposed to select one response variable as a primary variable, which is then optimized by adhering to the other constraints set by the criteria. Tong, Su, and Wang (1997) determined a multi-output signal-to-noise (MRSN) ratio through integrating the quality loss of each response. Liao (2004) incorporated an artificial neural network (ANN), data envelopment analysis (DEA) and an improved Taguchi method to optimize a multi-output problem. Egorov–Yegorov and Dulikravich (2005) adapted an advanced semi-stochastic evolutionary algorithm for constrained multi-objective optimization and determined optimum concentrations of alloying elements in heat-resistant austenitic stainless steel alloys and super alloys that will simultaneously maximize a number of alloy's mechanical properties. However, these methods are either incomplete or difficult to determine effective objective functions. Moreover, some methods can only identify an optimal setting restricted to discrete values.

Hence, in order to obtain an optimal WIP level that strikes the balance between maximizing the throughput rate and simultaneously minimizing the mean cycle time and the standard deviation of the cycle time, this study proposed to train an ANN to learn the relationship between the WIP levels vs. the throughput rate, the mean cycle time and the standard deviation of the cycle time. Then the SQP is implemented to resolve a multi-output constrained problem. The remainder of this article is organized as follows. Section 2 describes the optimization model of ANN. Section 3 illustrates the results and the confirmation tests of the proposed approach. Finally, conclusions of this study are contained in Section 4.

## 2. Optimization process

Fig. 1 depicts a flowchart of finding an optimal WIP level for the wafer fabrication. Details of each step are shown at the following subsections.

### 2.1. Identify the objective of the problem

The objective is to identify an optimal WIP level to maximize the throughput rate and simultaneously minimize the mean cycle time and the standard deviation of the cycle time. An algorithm integrating an ANN and the SQP method is implemented. An ANN is an effective modeling tool to map the relationship between the inputs and the

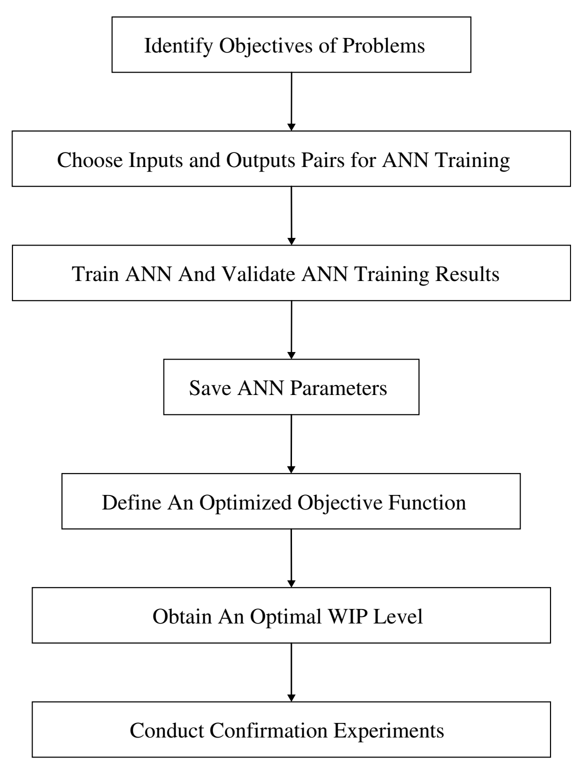


Fig. 1. Flowchart of finding an optimal WIP level.