Chapter 22
Bottleneck Detection Method Based on Production Line Information for Semiconductor Manufacturing System

Xiao-yu Yu, Fei Qiao and Yu-min Ma

Abstract Semiconductor wafer fabrication system is a typical complex manufacturing system, since it has large-scale, reentrant, multi-objective, uncertain and other characteristics. It's too difficult to achieve the capacity balance to lead the existence of the bottleneck. According to the theory of TOC, the accurate detection of bottleneck is the key to implement DBR thought. For the characteristics of semiconductor production line, this paper proposes a bottleneck detection method based on the starvation and blockage information of the production line. The method is verified on HP-24 model by simulation. Compared to the relative load method, the equipment utilization law and the queue length method; the experimental results show that this method makes performance better than them.

Keywords Bottleneck detection · DBR · Production line · Reentrant

22.1 Introduction

Drum-Buffer-Rope (DBR) theory is put forward by Doctor Goldratt based on the theory of constraints to solve the production scheduling problem (Rahman 1998). According to the theory, the whole system's output is decided by bottleneck's output and maximizing the utilization of the bottleneck's capacity is the key to improve system productivity and economic benefit. Bottleneck detection is a very important step for the operation and management of manufacturing enterprise, because it will not only affect the decision of the feeding strategy, but also influence the dispatching of the jobs on bottleneck equipment.

X. Yu (✉) · F. Qiao · Y. Ma
The School of Electronics and Information Engineering,
TongJi University, Shanghai, China
e-mail: yuxiaoyu_20002000@163.com

E. Qi et al. (eds.), *The 19th International Conference on Industrial Engineering and Engineering Management*, DOI: 10.1007/978-3-642-38391-5_22,

At present, much research effort has been devoted to bottleneck detection and can be divided into two categories. One is to detect the bottleneck before the start of production system. Literature (Zhang and Wu 2012) firstly establishes an optimization model by reducing some traditional constraints of a standard Job-shop, then calculates the bottleneck's characteristic value and chooses the excellent by the simulated annealing algorithm. Literature (Zhai et al. 2010) uses the orthogonal table and different assigned rules to construct a test program, with production system job target as measurement index to identify bottleneck. The other method to identify bottleneck is to conduct collection, imitation and simulation analysis of the data from the production system which has been online for a period of time. As in (Roser et al. 2002, 2003), system bottlenecks are divided into independent bottlenecks and shift bottlenecks, they are identified by calculating the maximum active time of the machines, which is similar to the common equipment utilization method. Literature (Li et al. 2007, 2009; Wang et al. 2008) is based on the data of the production line, the blockage and starvation information are made full use of to detect bottlenecks. Literature (Kasemset 2009; Kasemset and Kachitvichyanukul 2009, 2010) classifies the candidate bottleneck equipment according to the static and real-time data from simulation and testifies the correctness of the bottleneck equipment selection via confidence interval level.

Semiconductor wafer fabrication system is recognized as one of the most complex manufacturing systems, which normally contains as many as three or four hundreds processing steps. In addition, wafer manufacturing has reentrant characteristics, namely the same product will go through some processing center more than once, which is different from the traditional Job-shop and Flow-shop system (Wu et al. 2006). Most methods mentioned above are applicable to Flow-shop; some can be used for Job-shop as in (Zhai et al. 2010). It will ignore some stochastic disturbance such as equipment failure or seasonal variation of need though it is faster and more convenient. Based on literature (Li et al. 2007), this paper mainly studies the semiconductor wafer fabrication system by changing its constraints and proposing a concept of relatively blocking rate, which can record the detailed variation of the count of the jobs in the buffer throughout the whole production system operation period. The method can make full use of outline and online information of the production line, and it is not necessary to consider machine sets, type, product type, processing route and all kinds of factors such as random fluctuation in the process of identification. Therefore, it is a convenient and accurate identification method.

22.2 Common Bottleneck Detection Methods

Due to the high complexity of the semiconductor manufacturing system, the existing manufacturing system bottleneck identification methods are not suitable as in (Sengupta et al. 2008), which identifies bottlenecks by analysing the departure time among the equipments. However, it does not exist the logic

upstream and downstream equipments in semiconductor wafer fabrication system because of its serious reentrant characteristic. At present, common methods of bottleneck detection in semiconductor manufacturing system are as follows.

22.2.1 Analyze the Queue Length of the Equipment

In this approach, the queue length or waiting time of the equipment is measured and the one which has the longest queue length or waiting time is considered as manufacturing bottleneck, as is shown in formula 22.1

$$\text{Bottleneck} = \text{Machine}_j = \max_{1 \le j \le m} \left(\max_{0 < t \le T} (W_{tj}) \right) \tag{22.1}$$

where T stands for simulation period, m is the number of the processing center in a system or a model, W_{tj} is the number of jobs in the buffer for equipment j at a certain time t during the simulation period.

22.2.2 Measure the Utilization Rate of the Equipment

The equipment which has the highest utilization rate is system bottleneck as is shown in formula (22.2) (Zhou and Rose 2009).

$$\text{Bottleneck} = \text{Machine}_j = \max_{1 \le j \le m} \left(\frac{WT_j(T) + OT_j(T)}{T} \right) \tag{22.2}$$

where T is the time period considered, m is the number of the processing center in a system or a model, WTj(T) and OTj(T) are the processing time and off-line time of equipment j during time period T.

22.2.3 Calculate the Relative Load of the Equipment

On the basis of the order, the load of every processing center is calculated according to the job's craft, and the machine which has the maximal relative load is system bottleneck (Ding et al. 2008), as is shown in formula (22.3) and (22.4).

$$L_B = \text{Max}(L_h) \tag{22.3}$$

$$L_h = \sum_{i=1}^{x} q_i \sum_{j=1}^{y} \frac{\theta t_{ij}}{\mu} (h = 1, 2, \ldots m) \tag{22.4}$$

where L_B is the load of the system bottleneck, work center B is system bottleneck, i is the type of the job, x is the number of the type of the job, y is the step number of the job, θ is the related coefficient of the equipment (if some job is processed in the center, then θ is 1, otherwise θ is 0), t_{ij} is the processing time of step j of job i. The method can identify the system bottleneck through a simple calculation using some relevant technological parameters.

22.3 Bottleneck Detection Based on the Information of the Production Line

Using the blockage and starvation information in buffer to detect bottlenecks is a method based on data, the basic idea is that a bottleneck machine will often cause the upstream machines to be blocked and downstream machines to be starved, therefore, the bottleneck machine will often have a lower total blockage plus starvation time than its adjacent machines. LIN LI validates his theory according to the thought that the disturbance of the bottleneck equipment has the largest impact on the system. At last, he applies his idea to the production line including three and more machines (Li et al. 2007).

Semiconductor production line has a great difference with general industrial manufacturing line for its complex processing flow and serious reentrant characteristics. In the actual production line, bottleneck equipment is always the one which owns more reentry times and the jobs in the bottleneck buffer possibly come from multiple upstream machines, therefore, upstream or downstream machines can't be judged by physical location. In addition, the starvation rate and blockage rate can't be simply added together because there are lots of parallel and group equipments. This paper proposes a new method based on the thought of utilizing historical information in the production line aiming at the characteristics of semiconductor production line. At first, the capacity of buffer is set to be infinite. If the capacity is a fixed value, it will lead to buffer overflow when the production line is crowded and the block degree of different machines can't be distinguished. Secondly, the formula of the starvation rate is defined according to its basic concept. Finally, this method puts forward a concept of relative blocking rate and the formula is defined. The details are as follows.

(1) Starvation rate The idle time divided by processing period reflects when the machine is leisure in the production process. It can be expressed as $S_i = \frac{T_{idle}}{T}$ (i is equipment number, T_{idle} is leisure time, T is processing period);
(2) Relative blocking rate generally, each buffer will appear the phenomenon of accumulation, this paper proposes the concept of relative blocking rate to distinguish the congestion degree among different equipment at different times, namely the equipment is relative to other equipments in degree of obstruction. The computing method is: obtain the number of jobs in each

buffer at regular time called q_i (i is equipment number), calculate the total queue length of jobs of all the equipment buffers at the present moment called $\sum_{i=1}^{n} q_i$(n is the number of all the equipment), then the relative blocking rate of the equipment is $B_i = \frac{q_i}{\sum_{i=1}^{n} q_i}$. Assume that the collection period is Δt, the total simulation time is T, then the relative blocking rate is $\delta_{B_i} = \frac{\Delta t \sum \frac{q_i}{\sum_{i=1}^{n} q_i}}{T}$

In the real production line, bottleneck machine is the weak link of the whole system and its processing ability is the weakest. For that reason, the buffer before it often has a large accumulation of jobs waiting for processing. Compared to the other machines, blockage occurs more frequently in bottleneck machine and starvation is on the contrary. Thus, the starvation rate plus the opposite number of the relative blocking rate will be the least of all the equipments. To avoid confusion, the opposite number of the relative blocking rate is defined as non-blocking rate which is equal to $1 - \delta_{B_i}$. According to the above description, the method is expressed as:

$$\begin{aligned} \text{Bottleneck} = \text{Machine}_i &= \min_{0<i\le n} (1 - \delta_{B_i} + S_i) \\ &= \min_{0<i\le n} \left(1 + \frac{T_{idle} - \Delta t \sum \frac{q_i}{\sum_1^n q_i}}{T} \right) \end{aligned} \tag{22.5}$$

where i is equipment number and n is the total number of the equipments.

Because of the high complexity of semiconductor manufacturing system, it may exist multiple bottlenecks (Cao et al. 2010). When this method is used to detect bottleneck, the distribution of starvation rate and non-blocking rate of each equipment should be analyzed. For the equipment whose result of starvation rate plus non-blocking rate is similar to the system bottleneck can be considered as the second bottleneck. For the main aiming of this paper is single bottleneck detection, we don't make much analysis on multiple bottlenecks.

22.4 Example Validation

This paper chooses the model of HP-24 semiconductor production line as object of study and EM-PLANT as simulation platform. HP-24 Model comes from silicon wafer production technology center lab and most parameters are collected from real devices. There are 24 equipment groups in the model and most of them are single machine except the lithography (one group contains two machines, the other group contains three machines). As one kind of simplified model, HP-24 only

Table 22.1 Parameters for machines in Hp-24 model (Murphy and Dedera 1996) and simulation data

Machine group		Count	Reentrant times	Starvation rate (%)	Nonblocking rate (%)
ID	Name				
1	CLEAN	1	19	14.73	97.72
2	TMGOX	1	5	23.97	99.31
3	TMNOX	1	5	20.89	99.50
4	TMFOX	1	3	58.23	99.90
5	TU11	1	1	83.30	100.00
6	TU43	1	2	56.30	99.98
7	TU72	1	1	81.39	100.00
8	TU73	1	3	59.87	99.93
9	TU74	1	2	71.97	99.98
10	PLMSL	1	3	65.82	99.96
11	PLMSO	1	1	78.44	100.00
12	SPUT	1	2	65.85	99.97
13	PHPPS	2	13	14.22	99.05
14	PHGCA	3	12	5.00	65.33
15	PHHB	1	15	63.23	99.94
16	PHBI	1	11	6.00	64.45
17	PHFI	1	10	53.88	99.93
18	PHJPS	1	4	57.36	99.91
19	PLM6	1	2	22.84	99.80
20	PLM7	1	2	67.12	99.96
21	PLM8	1	4	13.29	99.39
22	PHWET	1	21	37.82	99.44
23	PHPLO	1	23	26.00	99.36
24	IMPP	1	8	9.05	100.00

processes one type of products which has 172 processing steps (Ding et al. 2008). During the simulation period, the buffer information is collected every half an hour and the feeding method is subject to uniform distribution. The simulation period is set to be 1 year and the data collected is shown in Table 22.1. The starvation rate and non-blocking rate of each equipment are calculated by the formulas introduced in the third section.

As is shown in Table 22.1 and Fig. 22.1, the starvation rate of machine 14 is 0.05, the non-blocking rate is 0.64, so the starvation rate plus the non-blocking is 0.69, which is the smallest of all machines. Thus, Machine 14 can be considered as the bottleneck machine according to the thought of the method based on production line information. As machine 14 is a parallel processing machine, it can be regarded as the bottleneck processing center.

Bottleneck machine is the short and fat son of the system, so the system output depends on the processing speed of bottleneck machine. The following is comparisons among different bottleneck detection methods.

(1) To prove the effectiveness of the machine group 14, we feed jobs into the production line based on the processing speed of machine 14. Through

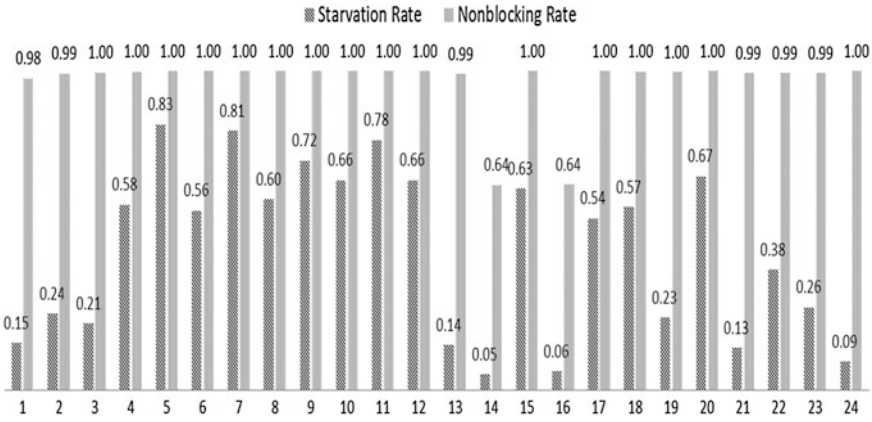

Fig. 22.1 Bar graph of starvation rate and nonblocking rate

calculation, the processing speed of machine 14 is $L_{14} = 31.28$h, namely machine 14 needs 31.28 h to process one lot of product.

(2) When we use the relative load method to identify bottlenecks, machine 16 holds the largest load according to the calculation formula introduced in Sect. 22.2. Thus, machine 16 is the bottleneck machine and its processing speed is $L_{16} = 32.56$h.

(3) Under the condition that the feeding speed obeys the uniform distribution, the equipment utilization of machine 24 is the highest. Thus, machine 24 is the bottleneck machine and its processing speed is $L_{24} = 30.88$h by the method.

(4) By analyzing the queue length of every machine in the whole process, machine 1 has the longest queue length and it is the system bottleneck with processing speed $L_1 = 29.45$h.

The feeding tables of all kinds of methods are drawn up according to the analysis above and they are validated on HP-24 model by simulation. The dispatching rule of bottleneck machine and non-bottleneck machines is FIFO (first in first out). The strengths and weaknesses of different feeding methods and scheduling strategies need to be evaluated by performance and the common performance indexes including:

(1) Average processing period. In the reentrant manufacturing system, the time from when a raw job is put into the production line to the time the job leaves the system is processing period, which can be expressed as $CT = T_{out} - T_{in}$ · T_{out} is the time the job leaves the processing system as finished product, T_{in} is the time the job enters the system. The scheduling goal is to make the average processing period minimum.

(2) Productivity. Productivity refers to the number of finished products per unit time, the formula is $PR = \frac{Q}{T}$. Q is the number of the finished lots during the processing period, T is processing period. Productivity is inverse to processing

Table 22.2 Performance of different bottleneck detection methods

	Average processing period	Processing period variance	Productivity	Wip	Bottleneck machine utilization
Production information	60024.49	6961.08	0.655	29.5	0.95
Relative load	63869.54	7097.10	0.643	30.57	0.93
Equipment utilization	78717.38	14856.57	0.634	39.96	0.921
Queue length	83229.23	20667.08	0.623	43.37	0.869

period. The shorter the processing period is, the higher the productivity will be. Productivity determines the cost of final product, processing period, customer satisfaction and so on.

(3) WIP (work in process) is the number of products in process online every day and the scheduling goal is to make the index minimum.

(4) Utilization rate of bottleneck machine. The formula is $U_B = \frac{T_{work}}{T_{open}}$, T_{work} is the time that the machine is in the state of process, T_{open} is the uptime of the machine. An overview of the comparison of different performance of each method is shown in Table 22.2

From Table 22.2 we can see that when we use the information of the production line to detect the bottlenecks, for average processing period, there is a 6.0 % reduction compared to the relative load method, a 23.7 % reduction compared to the equipment utilization, a 27.9 % reduction compared to the queue length. For processing period variance, there is a 1.9 % reduction compared to the relative load method, a 53.1 % reduction compared to the equipment utilization, a 66.3 % reduction compared to the queue length. For productivity, there is a 1.9 % increase compared to the relative load method, a 3.3 % increase compared to the equipment utilization, a 5.1 % increase compared to the queue length. For average day wip, there is a 3.5 % reduction compared to the relative load method, a 26.2 % reduction compared to the equipment utilization, a 32.0 % reduction compared to the queue length. For bottleneck machine utilization, there is a 2.2 % increase compared to the relative load method, a 3.1 % increase compared to the equipment utilization, a 9.3 % increase compared to the queue length. The relevant results are displayed in Table 22.3.

Table 22.3 Performance analysis of the other three bottleneck detection methods compared to the proposed method

	Average processing period (%)	Processing period variance (%)	Productivity (%)	WIP (%)	Bottleneck machine utilization (%)
Relative load	−6.0	−1.9	+1.9	−3.5	+2.2
Equipment utilization	−23.7	−53.1	+3.3	−26.2	+3.1
Queue length	−27.9	−66.3	+5.1	−32.0	+9.3

Data analysis presents that the proposed method in this paper has different range of ascension than other methods and thus proves its practicality and effectiveness.

22.5 Conclusion

Detecting bottleneck accurately is the first step to implement the DBR thought. Common bottleneck detection method has some limitations in the complex semiconductor manufacturing system. This paper utilizes the production line information to detect bottleneck which is based on data mining and obtains good effect. Historical data underlies the process information of manufacturing system and it can be regarded as a knowledge base which should be made full use of to detect bottlenecks. Future work: 1. There are many uncertain factors in a real production line, a single bottleneck feeding strategy may not be achieved good effect all the time, it should be combined with the bottleneck scheduling strategy. 2. How to further mine the underlying experience, knowledge and rules of the historical and online data to optimize the production line needs further study.

Acknowledgments This research was supported by Chinese National Natural Science Foundation (61034004), Science and Technology Commission of Shanghai (10DZ1120100, 11ZR1440400), Program for New Century Excellent Talents in University (NCET-07-0622) and Shanghai Leading Academic Discipline Project (B004).

References

Cao Z, Peng Y, Wu Q (2010) DBR-based scheduling for re-entrant manufacturing system (in Chinese). In: 2010 Proceedings of contemporary integrated manufacturing system, vol 2010, pp 566–572

Ding X, Qiao F, Li L (2008) Research of DBR scheduling method for semiconductor manufacturing (in Chinese). Integr Mech Electr 2008:29–31

Kasemset C (2009) TOC based Procedure for Job-shop Scheduling. Doctoral Dissertation, Industrial Engineering and Management, School of Engineering and Technology, Asian Institute of Technology, Bangkok, Thailand

Kasemset C, Kachitvichyanukul V (2009) Simulation tool for TOC implementation. In: Proceedings of ASIMMOD 2009, ASIMMOD 2009, Bangkok, Thailand, pp 86–97

Kasemset C, Kachitvichyanukuf V (2010) Effect of confidence interval on bottleneck identification via simulation. In: Proceedings of the 2010 IEEE IEEM. IEEE, DC, USA, pp 1592–1595

Li L, Chang Q, Ni J (2007) Bottleneck detection of manufacturing systems using data driven method. In: Proceedings of IEEE international conference on symposium on assembly and manufacturing. IEEE, Washington, DC, USA, pp 76–81

Li L, Chang Q, Ni J (2009) Data-driven bottleneck detection of manufacturing systems. Int J Prod Res 47(18):5019–5036

Murphy RE Jr, Dedera CR (1996) Holistic toc for maximum profitability. In: 1996 IEEE/SEMI advanced semiconductor manufacturing conference

Rahman S (1998) Theory of constraints: a review of the philosophy and its applications. Int J Oper Prod Manag 18(4):336–355

Roser C, Nakano M, Tanaka M (2002) Shifting bottleneck detection. In: Proceedings of the 34th conference on winter simulation. IEEE, Washington, DC, USA, pp 1079–1086

Roser C, Nakano M, Tanaka M (2003) Comparison of bottleneck detection methods for AGV systems. In: Proceedings of the 2003 winter simulation conference. IEEE, Washington, DC, USA, pp 1192–1198

Sengupta S, Das K, VanTil RP (2008) A new method for bottleneck detection. In: Proceedings of the 2008 winter simulation conference. IEEE, pp 1741–1745

Wang Z, Chen J, Wu Q (2008) A new method of dynamic bottleneck detection for semiconductor manufacturing line. In: Proceedings of the 17th world congress the international federation of automatic control, vol 17(1). Elsevier, Seoul, Korea, pp 14840–14845

Wu Q, Qiao F, Li L (2006) Semiconductor manufacturing system scheduling. Publishing House of Electronics Industry, Beijing

Zhai Y, Sun S, Wang J, Wang M (2010) Bottleneck detection method based on orthogonal experiment for job shop (in Chinese). Comput Integr Manuf Syst 16(9):1945–1952

Zhang R, Wu C (2012) Bottleneck machine identification method based on constraint transformation for job shop scheduling with genetic algorithm. Inf Sci 188:236–252

Zhou Z, Rose O (2009) A bottleneck detection and dynamic dispatching strategy for semiconductor wafer fabrication facilities. In: Proceedings of the 2009 winter simulation conference. Institute of Electrical and Electronics EngineersInc, Austin, TX, USA, pp 1646–1656