ELSEVIER

European Journal of Operational Research 131 (2001) 78–94

EUROPEAN JOURNAL OF OPERATIONAL RESEARCH

www.elsevier.com/locate/dsw

Theory and Methodology

A total standard WIP estimation method for wafer fabrication

Yu-Hsin Lin [a,b,*], Ching-En Lee [b]

[a] *Department of Industrial Engineering and Management, Ta Hwa Institute of Technology, Hsin Chu, Taiwan, ROC*
[b] *Department of Industrial Engineering and Management, National Chiao Tung University, Hsin Chu, Taiwan, ROC*

Received 15 October 1996; accepted 17 October 1999

Abstract

The standard work-in-process (WIP) level in a wafer fabrication factory is an important parameter which can be properly used to trigger the decision of when to release specific wafer lots. There are many WIP-based release control policies which have been proven to be effective for wafer fabrication manufacturing, few methods have been proposed to find the suitable WIP-level as a parameter for these release policies. This paper proposes a queueing network-based algorithm to determine the total standard WIP level so that the Fixed-WIP release algorithm to determine the total standard WIP level so that the Fixed-WIP release control policy can apply. A numerical example is provided to elaborate the algorithm. A simulation model of a real-world wafer fabrication factory in Taiwan is built and analyzed. Results of simulation experiment indicate that under the Fixed-WIP control policy, the total standard WIP level estimated from this study achieves a target throughput rate while keeping the corresponding cycle time relatively low. Results also demonstrate that the queueing network-based algorithm is a very useful method to determine the standard WIP level efficiently.

Keywords: Wafer release; Fixed-WIP; Total standard WIP level; Queueing network theory

1. Introduction

Semiconductor manufacturing is probably the most complex manufacturing process in the world. The semiconductor manufacturing process is a multistage process which transfers silicon in the form of thin, polished disk into integrated circuits. The entire process basically includes four main steps: raw wafer manufacturing, wafer fabrication, probe and die, package and test (Miller, 1990). The wafer fabrication is most time-consuming and complicated one, consisting primarily of at least six major types of phases: Diffusion, Lithography, CVD, Thin Film, Etching and Ion Implantation. The wafer passes through these six major phases numerous times. The flow time (cycle time) of a wafer piece is typically 30–45 days. The entire fabrication process involves hundreds of operation steps performed on a variety of machines. Wafers are grouped in lots and transferred in a standard cassette. A wafer must visit some machine groups

* Corresponding author. Tel.: +886-35-72673; fax: +886-35-722392.
E-mail address: bghlin@msl9.hinet.net (Y.-H. Lin).

0377-2217/01/$ - see front matter
PII: S0377-2217(99)00446-4

more than once. This is known as a reentrant process. For instance, a wafer may have to visit the photolithography machine 9–25 times for all layers of circuitry to be fabricated. Owing to the complexity of the wafer fabrication process, wafer release and dispatching decisions are extremely difficult to achieve. In addition, the cycle time is difficult to make.

The maximization of critical resource utilization as well as throughput rate and the minimization of cycle time are the primary goals of the release and dispatching policy in wafer fabrication because of its capital-intensive nature and the need to attain a competitive advantage. Goldratt and Cox (1986) recognized that the workstation with the lowest capacity (bottleneck) often governs the production rate of the entire manufacturing line. Thus, it is crucially important in wafer fabrication to maximize the bottleneck workstation's utilization to achieve high throughput rate while maintaining reasonable cycle time. However, the WIP level in the system affects both the cycle time and the throughput (Miller, 1990). WIP and cycle time are convex increasing functions of throughput. Infinite WIP level maximizes throughput which cannot exceed capacity of the bottleneck workstation (Buzacott, 1971). Queueing theory (Gross and Harris, 1974) has shown that minimal WIP levels produce the minimal cycle time in the steady state. The inherent conflict in the determination of a proper WIP level is obvious when attempting to both maximize throughput and minimize cycle time. The proper management of WIP level is critical to the success of manufacturing operations (Enns, 1995). It can regulate the material flow, make full use of bottleneck capacity, and act as a valuable parameter for wafer release control.

Simulation studies indicate that the wafer release mechanism has a stronger impact on system performance than the dispatching rule (Glassey and Resende, 1988a,b; Wein, 1988; Miller, 1990). Therefore, most researchers and practitioners in wafer fabrication focus primarily on wafer release control strategies. The wafer release control policy can be classified as closed-loop and open-loop. The closed-loop control policy is generally better than the open-loop one (Miller, 1990). The idea of the closed-loop control policy is to keep an optimal WIP level (i.e. workload) in the factory. The wafer release decisions will then be made according to the discrepancy between the actual and projected WIP levels (Graves et al., 1995). Starvation avoidance (SA) (Glassey and Resende, 1988a,b), Workload Regulating (WR) (Wein, 1988), Two-Boundary (TB) (Lou, 1989; Lou and Kager, 1989; Yan et al., 1996), CONWIP (Spearman et al., 1989, Spearman and Zazanis, 1992), Fixed-WIP (Burman et al., 1986; Glassey and Resende, 1988a,b; Lozinski and Glassey, 1988; Wein, 1988; Roderick et al., 1992), and Load-Oriented order release (Bechte, 1988a,b, 1994; Wiendahl et al., 1992; Wiendahl, 1995) are several well-known closed-loop control policies. The WIP level in a closed-loop control policy can be defined as either the total WIP in the system or the WIP level between two specific operation steps (Huang, 1995). In this paper, the WIP level is defined as the former one. Although the WIP level is crucial to wafer fabrication releasing decisions, few methods have been proposed to find the suitable WIP amount.

Miller (1990) used simulation to determine the number of lots in a fabrication line under Fixed-WIP control policy. He pointed out that a simulation model was applicable to a specific system only to the extent that the features it contained adequately represented that system, and it was time-consuming to run a simulation model. Furthermore, simulation models can take a long time to build and debug (Suri et al., 1995). It is important to recognize that simulation model construction is only just in case for a specific system. 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).

Queueing network model has been proven to be useful to analyze the performance of complex systems (Whitt, 1983). Burman et al. (1986) developed a queueing network model for IC manufacturing processes. He found that the performance measures (WIP and throughput time) of the queueing network model deviated by 7–20% from those of the simulation model, but the run

times were only one-tenth of that of the simulations. Chen et al. (1988) also constructed a simple queueing network model to predict specific key system performance measures. His modeling approach is the use of a mixed network model, composed of a serial of M/M/1 queues and closed with respect to non-monitor lots, in which the Fixed-WIP level is viewed as a design parameter and the throughput rate is viewed as a performance characteristic. The values predicted by his model were found to be within about 10% of those actually observed. He concluded that queueing network models could provide a useful quantitative guidance to designers in wafer fabrication factories. In fact, the queueing network-based method is efficient to estimate a wider range of manufacturing parameters, such as WIP level, and it can answer questions accurately and quickly under some conditions and ignore certain details of the manufacturing system (Burman et al., 1986; Askin and Standrige, 1993; Suri, 1995; Connors et al., 1996). According to studies by Suri et al. (1995), faculty members of US and European business schools agreed that the queueing-based approach was more efficient that simulation. Therefore, this paper proposes a queueing network-based algorithm to determine the total standard WIP level so that the Fixed-WIP release control policy can apply.

The production planner and the manufacturing engineer can monitor and control the WIP level and trigger the wafer release decision at the right time by using the estimated WIP level information. If the total WIP level falls below the defined standard WIP level, wafers must be released into the factory to avoid the bottleneck from starvation and keep the target throughput rate to be satisfied. The WIP level estimated by the proposed method herein and the corresponding wafer release control mechanism can achieve target throughput rate with a reasonable cycle time. To validate the result, a simulation model with real factory data (actual average processing times and routings) is experimented.

The remainder of this paper is organized as follows. Section 2 describes assumptions of the proposed method. Section 3 illustrates the standard WIP level estimating model. A numerical example is then given. In Section 4, a simulation model is built and results of simulation experiment are presented. Conclusions of this study are contained in Section 5.

2. Model assumptions

In this paper, a queueing network model to estimate the total standard WIP level in the wafer fabrication factory is developed. Some transformations and assumptions are made to suit the application to the queueing network. In the next two subsections, the corresponding transformations and assumptions will be discussed, respectively.

2.1. Processing time estimation

Wafers generally move through machines in lots in a wafer fabrication factory. The operational batch sizes range from a single wafer to several lots. Processing times depend mainly on the operation types. For lot splitting operations, such as photolithography, the processing time depends on the number of wafers in a lot. For batch operations, such as diffusion, no matter how many lots are loaded at a time (the number of lots batched must be smaller than the maximal batch size which is usually 6), the processing time generally does not vary with the number of lots batched.

The batch operation plays an important role in semiconductor manufacturing processes and have attracted many researches (Glassey and Weng, 1991; Fowler et al., 1992; Gurnani et al., 1992; Weng and Leachman, 1993; Robinson et al., 1995). A lot of researches (Neuts, 1967; Medhi, 1975) on batch operations use exponential services for computational tractability although the service time of batch operation may not be really distributed by an exponential. A single lot processing time is assumed in our queueing network model. For simplicity, the average processing time of a single lot at batch-type operation is estimated by dividing the processing time of a batch operation by its average batch sizes and is assumed exponential distributed.

The information of PM (preventive maintenance) duration, PM interval, machine breakdown

duration, and machine breakdown interval have a significant influence in the cycle time estimation (Uzsoy et al., 1992). To more adequately estimate the processing time of a single lot, these interruptions have to be included. To do so, we assume that the service time at a machine is the summation of its actual processing time and the average duration of breakdowns and PMs that occur within the specific service. We defined it as the "effective service time", the same name as defined by Chen et al. (1988). The effective service time will be applied to the queueing network model as the service time of each machine.

The effective service time of workstation j (EST_j) is defined as

$$m_j(1 + D_d/D_n + P_d/P_m), \tag{1}$$

in which m_j is the average processing time of a single lot in workstation j which does not include machine breakdowns and PMs; D_n the average time interval between machine breakdowns (mean time between failure, MTBF); D_d the average duration of machine breakdowns (mean time to repair, MTTR); P_m the average time interval between machine PMs (mean time between PM, MTBPM); P_d is the average duration of machine PMs (mean time to finish a PM, MTTPM).

The EST_j represents the long-run average service time of workstation j. The EST_j may not necessarily be consistent with an assumption of exponentiality, one still gets a good approximation (Chen et al., 1988). Such a system, the effective service time is applied, can be revised as an equivalent network without service interruptions (Vinod and Altiok, 1986; Chen et al., 1988). Other relevant assumptions will be described in the next section.

2.2. General assumptions

We assume that each product type has its own distinct fixed route. Each route comprises several workstations. Each workstation may contain one to several machines. The average processing time of each machine in a specific workstation is assumed the same and is exponentially distributed. The first-in-first-out (FIFO) discipline is assumed at each workstation.

In building a queueing network model for the wafer fabrication factory, each wafer lot is defined as an arriving customer and each workstation is viewed as a single or multi-server in the model. Because the studied case is a new factory, the service time variability at each workstation is relatively large. Therefore, each workstation is thought of as an M/M/S model. The entire factory therefore comprises a network of M/M/S queues. The bottleneck workstation, which will be defined in the next section, dominates the activities of other workstations. M/M/S queues of non-bottleneck workstations are independent of each other. Sufficient capacity is available to complete the arriving wafers at each workstation. That is, it is assumed that the service rate is greater than the arrival rate. Finite queues exist at the steady state.

3. Standard WIP level estimating model

One of the most important features of this model is to apply the concept of "the bottleneck workstation controls the throughput rate" (Goldratt and Cox, 1986). We treat the throughput rate, α, as a design parameter. In fact, the throughput rate (wafer out rate) is a very important production target which needs to be set and reviewed periodically by production control department in any wafer fabrication factories according to their rated capacity. After setting the target throughput rate, the arrival rate of the bottleneck workstation can be determined by means of multiplying the throughput rate of each individual product with the corresponding number of times that the bottleneck workstation is visited. The entire wafer fabrication route can be viewed as a virtual route flow illustrated in Fig. 1.

As shown in Fig. 1, the wafer fabrication route is divided into loops based on the bottleneck workstation. The bottleneck workstation is viewed as the starting point in each loop. The wafers flow through the bottleneck workstation and enter the next routing loop and the next until

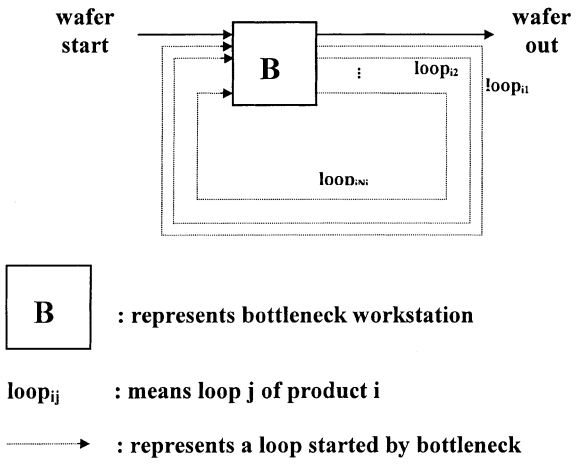

Fig. 1. The virtual process loop of wafer fabrication.

they finish their operations. That is, the entire wafer fabrication routing is composed of a series of virtual loops. The number of loops that a product requires depends on the number of times that the bottleneck workstation is visited by that specific product. Every loop of a product is unique. There are some operation steps between the wafer start and the first visiting of the bottleneck workstation. The ending operation step of the last loop of each product is wafer out. Since the wafer flow is controlled by the bottleneck workstation, the bottleneck workstation acts like a wafer distribution center. A wafer lot is distributed into a specific process loop through the bottleneck workstation.

Although there are many WIP-based release control policies which have been proven to be effective (Bertrand and Wortmann, 1981; Glassey and Resende, 1988a,b; Wein, 1988; Spearman et al., 1989; Lou, 1989; Lou and Kager, 1989; Graves et al., 1995; Yan et al., 1996) for wafer fabrication, few methods have been proposed to find the suitable WIP-level as a parameter for those policies. Among those lot release control policies, Fixed-WIP is a simple but very effective control mechanism in reducing cycle time and increasing mean throughput rate (Wein, 1988; Miller, 1990; Roderick et al., 1992). Simulation software packages (e.g. Autoched, Mansim, and Pacemaker etc.) are common tools to determine the WIP levels in practice. However, it is time-consuming and case specific as discussed in Section 1. There are many queueing network-based software packages like CAN-Q (Suri et al., 1995), QNA (Whitt, 1983), and PANACEA (Ramakrishnan et al., 1982) etc. in a number of applications. However, their applications in semiconductor manufacturing are very few. It is the objective of this research to propose a queueing network-based algorithm to determine the total standard WIP level so that the Fixed-WIP release control policy can apply. The wafer release decisions are made in accordance with the total WIP level. Wafer lots are released into the factory when the actual WIP level drops below the estimated standard WIP level. Because the number of operations from the wafer start to the first visiting of the bottleneck workstation are relatively few, these operation steps are not considered in our WIP estimation model.

3.1. Heuristic algorithm

The proposed algorithm is throughput-rate and bottleneck oriented. The total standard WIP level is calculated by summarizing each individual workstation's queue. As mentioned in Section 2.2, each individual workstation can be treated as an M/M/S model. According to Gross and Harris (1974) to calculate the WIP level of each machine the information of arrival rates, service rates, and the number of workstations is needed. Because the performance of a bottleneck workstation is much more important than that of others and the entire throughput rate is determined by bottleneck workstation, the WIP in front of the bottleneck workstation is determined first. The WIPs of non-bottleneck workstations will be derived according to the equivalence property and Jackson's network (Jackson, 1963).

In order to clearly describe the proposed algorithm, the following indexes are defined first. The bottleneck workstation is indexed by B, the non-bottleneck workstations by K, the product types by i, and the loops by l. The standard WIP level of the entire factory is represented by std_WIP and the corresponding definitions are given below.

α overall average throughput rate of the factory. It is expressed in lots per hours
α_i average throughput rate at which lots of product type i are processed through the factory
D_i fraction of product type i processed through the factory
λ_j average number of lots visit workstation j per hour
μ_j effective service rate per machine at workstation j, that is the average number of lots completed per hour
S_j number of machine at workstation j
W_j average WIP level in from of workstation j, including the processing wafer lot
N_{iB} number of visits to bottleneck workstation B by product type i
N_i number of loops for product type i; $N_i = N_{iB}$
P_{ijk} number of visits to non-bottleneck workstation K in loop l of product type i
R'_{il} probability of entering the loop l of product type i when finishing the operation at the bottleneck workstation
RO_{il} output rate of the bottleneck workstation to loop l of product type i when finishing the operation at the bottleneck workstation

The heuristic algorithm is divided into eight steps. First of all, the bottleneck workstation have to be identified. APICS (1992) indicates that the bottleneck workstation can be defined as "a same function of the machine group whose capacity is equal to or less than the demand placed on it". In this paper, the bottleneck workstation is defined as the one with the average greatest loading of an aggregate lot place on it. The flow chart of the heuristic method is illustrated in Fig. 2 and its detail is elaborated thereafter.

Step 1: Calculating the average throughput rate of each product type i by multiplying the average overall throughput rate α and the product mix ratios D_i:

$$\alpha_i = \alpha \times D_i. \tag{2}$$

Step 2: Computing the expected average arrival rate, λ_{iB}, of product type i visiting the bottleneck workstation based on the projected average throughput rate of product type i:

$$\lambda_{iB} = \alpha_i \times N_{iB}. \tag{3}$$

Summarizing the average arrival rates of all product type i visiting the bottleneck workstation to get the average arrival rate, λ_B, of all products visiting the bottleneck workstation:

$$\lambda_B = \sum_i \lambda_{iB}. \tag{4}$$

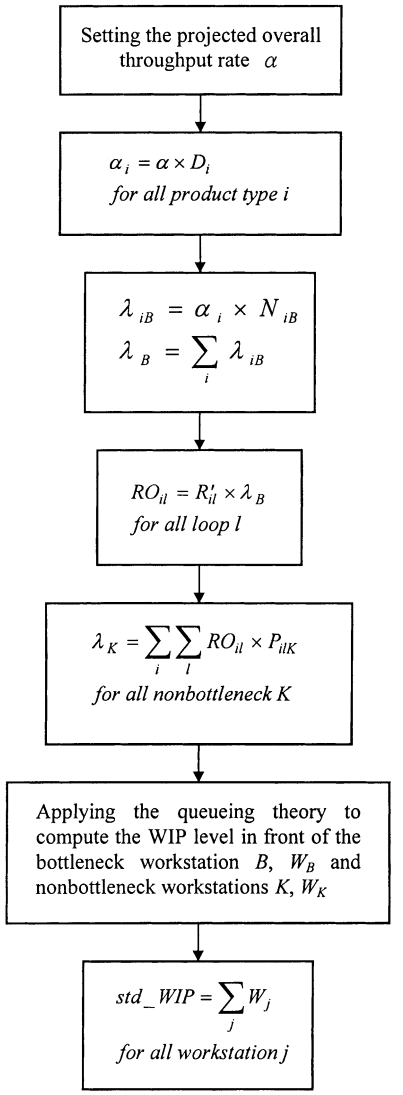

Fig. 2. The flow of std_WIP estimation algorithm.

Step 3: Applying the queueing theory (Gross and Harris, 1974) to compute the WIP level in front of the bottleneck workstation, W_B:

$$W_B = L_B + \frac{\lambda_B}{\mu_B} = \frac{P_0(\lambda_B/\mu_B)^{S_B}[\lambda_B/(S_B \times \mu_B)]}{S_B![1-\lambda_B/(S_B \times \mu_B)^2]} + \frac{\lambda_B}{\mu_B}, \quad (5)$$

where

$$P_0 = \frac{1}{\sum_{n=0}^{S_B-1}(\lambda_B/\mu_B)^n/n! + [(\lambda_B/\mu_B)^{S_B}/S_B!] \times [1-\lambda_B/(S_B \times \mu_B)]^{-1}}.$$

Step 4: Computing the probability R'_{il}; For a specific product type i, it is assumed that the probability of entering each loop l from the bottleneck is the same and the fraction is $1/N_i$. The probability R'_{il}s for loops l of a product type i are independent of each other and can be represented as follows:

$$R'_{il} = \frac{D_i \times N_i}{\sum_j (D_j \times N_j)} \times \frac{1}{N_j} = \frac{D_i}{\sum_j (D_j \times N_j)}. \quad (6)$$

Step 5: Calculating the product type i's output rate at the bottleneck workstation to loop l according to the probability R'_{il} and λ_B:

$$RO_{il} = R'_{il} \times \lambda_B. \quad (7)$$

Step 6: Calculating the anticipated arrival rate, λ_K, of non-bottleneck workstation K. For a specific product type i, the output rate at the bottleneck workstation to loop l (RO_{il}) is equal to the arrival rate of loop l. In loop l of product type i, the non-bottleneck workstation K may be visited several times (P_{ilk}). Thus, the average arrival rate at non-bottleneck workstation K in the loop l of product type i is derived from multiplying the average arrival rate at non-bottleneck workstation K by the corresponding number of times that it is visited. By summarizing the arrival rates of all products at non-bottleneck workstation K in each loop, the anticipated arrival rate, λ_K, at non-bottleneck workstation K can be derived:

$$\lambda_K = \sum_i \sum_l RO_{il} \times P_{ilK}. \quad (8)$$

Step 7: Applying Eq. (5) in step 3 to compute the WIP levels in front of all non-bottleneck workstations K; W_K.

Step 8: Summarizing the WIP levels in front of all workstations j to get the total standard WIP level of the entire factory

$$\text{std\_WIP} = \sum_j W_j. \quad (9)$$

3.2. Numerical case

In this section, a wafer fabrication factory of a global company is analyzed. This is a new factory located in the Science Based Industrial Park in Taiwan. In the beginning stage of their manufacturing, five major types of products: A, B, C, D, and E are focused on to be analyzed. The product mixes are 15.71%, 16.67%, 4.76%, 34.76%, and 28.01%, respectively. Each product type has its own distinct route. Each route is composed of 218–315 operation steps. Moreover, these operation steps are performed on 60 kinds of workstations. Tables 1 and 2 give the average processing time of a single lot, PM information, and breakdown information on each type of workstation. The average processing times, PM, and breakdown information are gathered from the real factory. The effective service time of workstation is derived from Eq. (1) and is given in Table 1.

The recently established wafer fabrication factory sets an initial production target (throughput rate) of 318 lots per month. Each lot contains 25 pieces of wafers. The daily mean throughput rate is equal to 10.6 lots (318 lots/month ÷ 30 days/month = 10.6 lots/day). According to the proposed algorithm, the bottleneck workstation has to be identified first. The rough-cut capacity estimation method is used to compute the average loads of an aggregate wafer lot place on workstations, which do not consider the timing of lot releasing. The workstation number 25 is identified as the bottleneck (see Table 3).

Table 1
The average and effective processing time of a single lot[a]

Workstation j	1	2	3	*4	5	6	*7	*8	9	*10
APT_j (min/lot)	10.00	10.00	10.00	60.00	35.00	46.96	6.25	38.92	60.00	12.50
EST_j (min/lot)	10.99	10.87	10.10	60.61	58.33	47.43	6.94	40.97	81.08	13.16
Workstation j	11	12	13	14	15	16	17	18	19	20
APT_j (min/lot)	25.34	31.00	27.64	60.00	15.00	10.00	65.02	20.67	30.00	62.98
EST_j (min/lot)	25.60	31.31	30.38	60.61	15.15	10.10	75.61	30.40	30.30	71.57
Workstation j	*21	*22	23	24	25	26	27	28	29	30
APT_j (min/lot)	86.86	10.67	10.00	10.00	40.00	3.00	57.35	74.06	90.00	53.47
EST_j (min/lot)	105.92	14.22	10.10	10.10	41.67	3.03	63.72	77.96	93.75	62.18
Workstation j	*31	*32	*33	*34	*35	*36	37	38	39	40
APT_j (min/lot)	10.02	63.63	22.01	13.25	10.90	12.65	8.77	10.00	84.41	25.00
EST_j (min/lot)	11.51	69.93	22.23	17.21	13.80	13.04	10.20	10.10	88.85	26.32
Workstation j	41	42	43	44	*45	46	*47	*48	49	*50
APT_j (min/lot)	25.00	25.00	10.00	10.00	66.75	42.22	49.51	17.05	73.76	100.00
EST_j (min/lot)	27.17	25.25	10.10	10.53	87.83	46.40	50.01	17.86	99.68	109.89
Workstation j	*51	52	53	54	55	56	*57	58	*59	*60
APT_j (min/lot)	57.11	67.79	10.00	55.00	55.00	120.00	200.00	15.00	40.91	57.19
EST_j (min/lot)	62.08	68.47	10.10	58.51	67.90	121.21	235.29	15.15	54.55	61.50

[a] APT_j represents the average processing time of a single lot at workstation j which does not include machine breakdowns and PMs; EST_j represents the effective service time of a single lot at workstation j; * represent batching process workstations. The average batch size of workstation number 4, 7, 21, 22, 45, 59, and 60 is 2 lots; the average batch size of workstation number 8, 10, 32, 47, 50, 51, and 57 is 3 lots; the average batch size of workstation number 31, 34, 35 and 36 is 4 lots.

According to the algorithm presented in Section 3.1, the average throughput rate of each product type i, α_i, is calculated first. The throughput rates of five product types are 1.665, 1.767, 0.505, 3.685, and 2.979 lots per day corresponding to products A, B, C, D, and E, respectively. The expected average arrival rates, λ_{iB}, of each product type i visiting the bottleneck workstation can be derived. Because the number of times that product types A–E visit the bottleneck workstation are 11, 8, 7, 10, and 9, the average arrival rates of products A–E visit the bottleneck workstation are 18.315 ($= 1.665 \times 11$), 14.136 ($= 1.767 \times 8$), 3.535 ($= 0.505 \times 7$), 36.850 ($= 3.685 \times 10$), and 26.811 ($= 2.979 \times 9$) lots per day. Thus, the average arrival rate of all products to visit the bottleneck workstation is 99.647 lots per day. This expected value shows that the production target (10.6 lots per day) will be reached if the average arrival rate at the bottleneck workstation is 4.152 lots per hour (Table 5).

In step 4, the probability R'_{il}s are determined and shown in Table 4. Through steps 5 and 6, the anticipated arrival rate at each workstation is derived and is given in Table 5. By summarizing the WIP levels in front of all workstations, presented in Table 6, the total standard WIP (145 lots) is estimated.

In the next section, a simulation model applying Fixed-WIP lot release policy and simulation experiment is designed to demonstrate how the derived standard WIP level performs well.

4. Simulation experiment

The most common modelling technique in semiconductor manufacturing is simulation (Dayhoff and Atherton, 1984; Atherton and Dayhoff, 1986; Burman et al., 1986; Miller, 1990; Glassey and Resende, 1988a,b; Wein, 1988). Thus, a simulation model is adopted to analyze the effectiveness of the standard WIP level determined through the proposed algorithm under the Fixed-WIP lot release policy.

Table 2
The PM and breakdown information of workstations

Workstations	1	2	3	4	5	6	7	8	9	10
P_d (h)	158	17,520	17,520	17,520	10	17,520	17,520	90	14	17,520
P_m (h)	0.30	1.00	1.00	1.00	3.45	1.00	1.00	3.30	2.75	1.00
D_d (h)	55	325	500	300	10	185	98	156	20	40
D_n (h)	5.33	28.35	5.05	10.00	3.22	1.83	10.83	2.48	3.10	2.10
Workstations	11	12	13	14	15	16	17	18	19	20
P_d (h)	17,520	16,520	75	17,520	200	17,520	62	12	75	183
P_m (h)	0.15	0.50	1.00	1.00	1.00	1.00	10.90	3.10	0.50	3.00
D_d (h)	140	190	70	300	200	500	590	15	75	25
D_n (h)	1.42	1.90	6.00	3.03	1.00	5.05	1.40	3.20	0.25	3.00
Workstations	21	22	23	24	25	26	27	28	29	30
P_d (h)	55	6	17,520	17,520	163	420	79	114	130	23
P_m (h)	6.86	1.51	1.00	1.00	1.50	2.00	5.00	3.00	2.57	1.00
D_d (h)	37	30	500	500	46	420	69	190	104	36
D_n (h)	3.50	2.43	5.05	5.05	1.50	2.00	3.30	5.00	2.28	3.60
Workstations	31	32	33	34	35	36	37	38	39	40
P_d (h)	20	64	12,520	21	10	600	16	17,520	34	115
P_m (h)	1.33	4.86	1.50	1.50	1.21	1.50	2.00	1.00	0.90	0.30
D_d (h)	30	156	150	104	35	95	60	500	312	100
D_n (h)	2.50	3.60	1.50	23.67	5.00	2.70	2.27	1.00	6.60	5.00
Workstations	41	42	43	44	45	46	47	48	49	50
P_d (h)	17	290	17,520	40	11	150	1250	95	15	42
P_m (h)	1.00	1.00	1.00	1.10	2.80	1.10	1.00	1.00	2.54	3.00
D_d (h)	69	300	500	40	65	120	320	100	69	109
D_n (h)	1.80	2.00	1.00	1.00	4.00	11.00	3.00	1.00	3.45	3.00
Workstations	51	52	53	54	55	56	57	58	59	60
P_d (h)	61	600	17,520	14	25	240	39	85	12	90
P_m (h)	5.00	3.00	1.00	0.85	2.60	1.20	6.10	0.50	3.00	2.00
D_d (h)	195	600	500	650	50	235	70	85	24	48
D_n (h)	1.00	3.00	1.00	25.00	6.46	1.20	1.40	0.33	1.90	2.55

4.1. Simulation model descriptions

Because the studied wafer fab is newly installed, the service time variability at each workstation is relatively large. Therefore, the exponential distribution is also assumed. In the simulation model, real average processing time as an exponential distribution's mean at each workstation for a specific product is applied. As indicated in Section 1, the WIP level in a system affects both cycle time and throughput rate (Miller, 1990). There is inherent conflict in WIP level decisions when attempting to both maximize throughput and minimize cycle time. Therefore, simulation experiments are designed to analyze throughput rate and cycle time performance under different WIP levels, represented by times ($\times$) of total standard WIP level estimated from our proposed algorithm. Ten experiments for different WIP levels

$$(0.5 \times \text{std\_WIP},\ 0.6 \times \text{std\_WIP},\ 0.7 \times \text{std\_WIP},\ 0.8 \times \text{std\_WIP},\ 0.9 \times \text{std\_WIP},\ 1.0 \times \text{std\_WIP},\ 1.1 \times \text{std\_WIP},\ 1.2 \times \text{std\_WIP},\ 1.3 \times \text{std\_WIP},\ 1.4 \times \text{std\_WIP})$$

are experimented. The lot release mechanism is Fixed-WIP and the lot dispatching rule is FIFO. The wafer lot to be released is randomly generated according to the product mix ratio. The batching

Table 3
The average loading of each workstation[a]

Workstations	1	2	3	4	5	6	7	8	9	10
Number of machines	1	3	4	1	2	3	4	1	1	2
T_load (min)	82.99	177.28	174.84	60.61	175.00	300.39	12.00	104.36	81.08	165.79
A_load (min)	82.99	59.09	58.28	60.61	87.50	100.13	6.00	104.36	81.08	82.89
Workstations	11	12	13	14	15	16	17	18	19	20
Number of machines	3	1	3	2	1	1	2	2	2	5
T_load (min)	218.12	71.73	259.49	181.82	90.77	50.51	226.82	163.71	80.95	556.06
A_load (min)	72.71	71.73	86.50	90.91	90.77	50.51	113.41	81.82	40.48	111.21
Workstations	21	22	23	24	*25*	26	27	28	29	30
Number of machines	2	2	3	2	*3*	1	2	3	2	2
T_load (min)	204.54	130.11	10.10	172.68	*355.95*	42.56	180.69	122.20	234.84	170.09
A_load (min)	102.27	65.05	3.37	86.34	*118.65*	42.56	90.35	40.73	117.42	85.04
Workstations	31	32	33	34	35	36	37	38	39	40
Number of machines	1	2	2	2	1	1	1	1	2	1
T_load (min)	31.31	147.31	197.98	54.24	21.35	61.27	46.55	17.94	106.54	99.76
A_load (min)	31.31	73.66	98.99	27.14	21.35	61.27	46.55	17.94	53.27	99.76
Workstations	41	42	43	44	45	46	47	48	49	50
Number of machines	2	1	2	2	1	2	2	1	3	2
T_load (min)	69.36	108.94	55.17	79.50	87.83	90.37	97.42	66.24	276.74	61.75
A_load (min)	34.68	108.94	27.59	39.75	87.83	45.19	48.71	66.24	92.25	30.87
Workstations	51	52	53	54	55	56	57	58	59	60
Number of machines	2	2	1	1	1	1	2	1	1	2
T_load (min)	230.88	196.29	40.88	42.07	3.23	5.77	235.29	30.30	39.22	130.60
A_load (min)	115.44	98.15	40.88	42.07	3.23	5.77	117.65	30.30	39.22	65.30

[a] T_load: represents the average loading of a single lot placed at a workstation. A_load: represents the average loading of each machine. Workstation 25 is the identified bottleneck at those cells with values given in italics.

Table 4
The probability R'_{il} of product A, B, C, D and E

Loops	1	2	3	4	5	6	7	8	9	10	11
Product A	0.0167	0.0167	0.0167	0.0167	0.0167	0.0167	0.0167	0.0167	0.0167	0.0167	0.0167
Product B	0.0177	0.0177	0.0177	0.0177	0.0177	0.0177	0.0177	0.0177			
Product C	0.0051	0.0051	0.0051	0.0051	0.0051	0.0051	0.0051				
Product D	0.0370	0.0370	0.0370	0.0370	0.0370	0.0370	0.0370	0.0370	0.0370	0.0370	
Product E	0.0299	0.0299	0.0299	0.0299	0.0299	0.0299	0.0299	0.0299	0.0299		

operations in diffusion, etching, thin film, and ion implant areas are also implemented in this simulation model, although a single lot operation is assumed in our queueing network model, to approach the real wafer fabrication process.

The performance measures, average throughput rate and cycle time, are compared under these 10 simulation experiments. Ten simulation runs are made for each experimental WIP level. Common random number streams, one of the variance-

Table 5
The arrival rates of each workstation

Workstation j	1	2	3	4	5	6	7	8	9	10
λ_j (lots/h)	3.040	7.409	7.850	0.441	1.323	2.996	0.637	1.107	0.441	5.924
Workstation j	11	12	13	14	15	16	17	18	19	20
λ_j (lots/h)	3.706	0.983	4.147	1.323	2.693	1.323	0.882	2.445	1.233	3.262
Workstation j	21	22	23	24	25	26	27	28	29	30
λ_j (lots/h)	0.779	4.241	0.441	7.281	4.152	5.613	1.250	0.857	0.901	1.094
Workstation j	31	32	33	34	35	36	37	38	39	40
λ_j (lots/h)	1.302	1.134	4.002	1.323	0.788	2.059	2.111	0.706	0.584	1.531
Workstation j	41	42	43	44	45	46	47	48	49	50
λ_j (lots/h)	0.970	1.397	2.187	3.040	0.441	0.723	0.723	1.485	1.147	0.347
Workstation j	51	52	53	54	55	56	57	58	59	60
λ_j (lots/h)	1.744	1.118	1.834	0.420	0.069	0.069	0.441	0.441	0.420	1.134

Table 6
The WIP levels in front of each workstation

Workstation j	1	2	3	4	5	6	7	8	9	10
W_j (lots)	1.256	1.491	1.461	0.804	2.195	4.555	0.074	3.098	1.476	2.248
Workstation j	11	12	13	14	15	16	17	18	19	20
W_j (lots)	1.878	1.054	3.248	2.417	9.716	0.287	1.609	2.010	0.690	5.692
Workstation j	21	22	23	24	25	26	27	28	29	30
W_j (lots)	2.612	1.345	0.074	1.963	25.089	0.396	2.373	1.183	10.471	1.672
Workstation j	31	32	33	34	35	36	37	38	39	40
W_j (lots)	0.333	2.348	3.292	0.394	0.221	0.810	0.560	0.135	1.064	2.047
Workstation j	41	42	43	44	45	46	47	48	49	50
W_j (lots)	0.462	1.427	0.381	0.574	1.823	0.606	0.662	0.792	2.604	0.706
Workstation j	51	52	53	54	55	56	57	58	59	60
W_j (lots)	9.683	2.149	0.447	0.694	0.085	0.163	12.647	0.125	1.216	1.756

reduction techniques (Law and Kelton, 1991), are used. In each simulation run, 365 days were simulated and the throughput rate and cycle time were collected. Startup statistics were discarded for the first 265-day warm-up period to minimize the potential startup bias. Data was then collected for the rest of 100 additional days.

4.2. Experimental design

As described in the previous section, simulation experiments are designed to analyze the line performance under different standard WIP levels. A total of 10 distinct WIP level cases are tested, varying the number of WIP lots in the factory from 50% of standard WIP level to 140% of standard WIP level. Each WIP level is simulated to observe the average throughput rate and cycle time of the factory. It takes 13–23 hours in general, depending on the amount of WIP level in the system, to run a simulation experiment on a Pentium-133 PC. Hypothesis tests are conducted for the differences on average throughput rates and cycle times among these 10 WIP levels. 45 $(10 \times 9 \div 2)$ paired t-tests are conducted for each of the two performance measures. The results are presented in the next section.

4.3. Analysis results

4.3.1. Analysis of mean throughput rate, cycle time and standard deviation of cycle time

Results ($p < 0.0001$) of ANOVA indicate that significant distinction exist in the performance of WIP levels with respect to mean throughout rate, mean and standard deviation cycle time at $\alpha = 0.05$. A Dunce's multiple test (Kirk, 1982) is then conducted on the WIP levels and results are presented in Tables 7–9 for mean throughput rate, mean cycle time, and standard deviation of cycle time at $\alpha = 0.05$, respectively. According to Table 7, no significant differences arise in mean throughput rates at WIP levels of $1.1 \times$ std_WIP, $1.2 \times$ std_WIP, $1.3 \times$ std_WIP and $1.4 \times$ std_WIP. Although the mean throughput rate at the $1.0 \times$ std_WIP falls in Duncan Group B, the highest percentage improved over that of $1.0 \times$ std_WIP is 1.69%. However, when the WIP levels are below $1.0 \times$ std_WIP, the mean throughput rates are significantly smaller than the mean throughput rate at $1.0 \times$ std_WIP and above. Table 7 and Fig. 3 reveal that as the WIP level drops further from $1.0 \times$ std_WIP, the percentage lost in mean throughput rate becomes faster (from 3.01% decrease at $0.9 \times$ std_WIP to 40.54% decrease at $0.5 \times$ std_WIP).

When WIP levels are at and below the standard WIP level, their corresponding mean cycle times are significantly smaller than those of the above (see Table 8 and Fig. 3). The higher the WIP level is, the longer the mean cycle time presents. As the WIP level exceeds $1.0 \times$ std_WIP, the percentage increased in mean cycle time over $1.0 \times$ std_WIP ranges from 8.85% to 37.80%. Little's results

Table 7
Duncan's multiple range test results for daily mean throughput rate

WIP levels	*N* (replications)	Mean (days)	Duncan grouping	% Improvement in mean TP (over $1.0 \times$ std_WIP)
$1.4 \times$ std_WIP	10	11.068	A	1.69
$1.3 \times$ std_WIP	10	11.048	A	1.51
$1.2 \times$ std_WIP	10	11.028	A	1.32
$1.1 \times$ std_WIP	10	11.036	A	1.40
$1.0 \times$ std_WIP	10	10.884	B	–
$0.9 \times$ std_WIP	10	10.556	C	−3.01
$0.8 \times$ std_WIP	10	9.912	D	−8.93
$0.7 \times$ std_WIP	10	8.952	E	−17.75
$0.6 \times$ std_WIP	10	7.840	F	−27.97
$0.5 \times$ std_WIP	10	6.500	G	−40.54

Table 8
Duncan's multiple range test results for daily mean cycle time

WIP levels	*N* (replications)	Mean (lots/day)	Duncan grouping	% improvement in mean CT (over $1.0 \times$ std_WIP)
$1.4 \times$ std_WIP	10	18.226	A	−37.80
$1.3 \times$ std_WIP	10	16.964	B	−28.26
$1.2 \times$ std_WIP	10	15.646	C	−18.30
$1.1 \times$ std_WIP	10	14.396	D	−8.85
$1.0 \times$ std_WIP	10	13.226	E	–
$0.9 \times$ std_WIP	10	12.386	F	6.35
$0.8 \times$ std_WIP	10	11.656	G	11.87
$0.7 \times$ std_WIP	10	11.342	H	14.24
$0.6 \times$ std_WIP	10	11.084	I	16.20
$0.5 \times$ std_WIP	10	11.262	H	14.85

Table 9
Duncan's multiple range test results for daily mean cycle time

WIP levels	N (replications)	Mean (h)	Duncan grouping	% improvement in std of mean CT (over 1.0 × std_WIP)
1.4 × std_WIP	10	54.698	A	−26.93
1.3 × std_WIP	10	51.870	B	−20.36
1.2 × std_WIP	10	49.596	C	−15.09
1.1 × std_WIP	10	45.920	D	−6.56
1.0 × std_WIP	10	43.094	E	–
0.9 × std_WIP	10	40.662	F	5.64
0.8 × std_WIP	10	38.448	G	10.78
0.7 × std_WIP	10	37.484	G	13.02
0.6 × std_WIP	10	37.154	G	13.78
0.5 × std_WIP	10	38.414	G	10.86

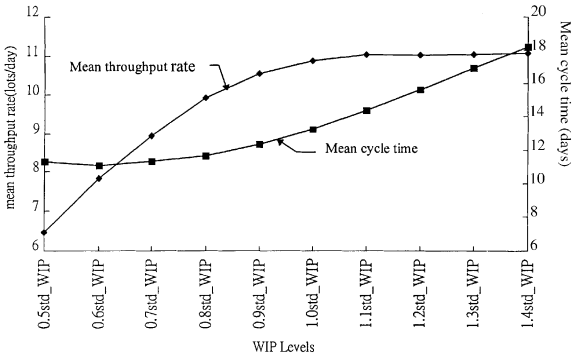

Fig. 3. Throughput rate and cycle time vs. WIP level.

(Gross and Harris, 1974) suggest that the linear model should exist between mean cycle time and WIP for the region of WIP $\geq$ std_WIP. A F-test (Montgomery and Peck, 1992) is made. It is found that the F-ratio ($=7821.139$) is greater than $F_{0.05,1,5}(=6061)$. Therefore, the result of this study confirms to Little's results.

For WIP levels exceeding 1.0 × std_WIP, as indicated above, the highest percentage improved in mean throughput rate over that at 1.0 × std_WIP is only 1.69%. Meanwhile, it may pay up to 37.80% increase in mean cycle time, which is not worth it. Moreover, Table 9 demonstrates that the standard deviation of cycle time at the standard WIP level is significantly smaller than those at 1.1 × std_WIP and above. This finding suggests that those the derived standard WIP level (1.0 × std_WIP) is more robust in cycle time performance than higher WIP-level situations which may therefore result in a better order promising performance. For WIP levels below 1.0 × std_WIP, their reductions on mean cycle times cannot compensate for the loss of mean throughput rate and the loss of expensive capacity at bottleneck workstation (as shown in Table 10, the utilization of bottleneck workstation drops from 95.12% at 1.0 × std_WIP to 56.24% at 0.5 × std_WIP). Therefore, the obtained standard WIP level, although it cannot guarantee the optimal, is a safe and reliable amount to achieve a high mean throughput rate with an acceptable short mean cycle time.

Both large and small batch sizes produce long flow times. That is, the batch size and the flow time form a U-shaped curve relationship (Karmarkar

Table 10
Utilization of bottleneck workstation under different WIP levels

WIP levels	0.5 × std_WIP	0.6 × std_WIP	0.7 × std_WIP	0.8 × std_WIP	0.9 × std_WIP
Utilization	0.5624	0.6806	0.7810	0.8627	0.9173
WIP levels	1.0 × std_WIP	01.1 × std_WIP	1.2 × std_WIP	1.3 × std_WIP	1.4 × std_WIP
Utilization	0.9512	0.9658	0.9661	0.9670	0.9674

et al., 1985; Jacobs and Bragg, 1988; Glassey and Weng, 1991). Small lot sizes produce longer flow times due to excessive setup requirements. Experimental results point toward a similar U-shaped phenomenon. When WIP levels are below $0.6 \times$ std_WIP, their mean cycle times increase owing primarily to the batching effect. A low total WIP level requires more time to accumulate a sufficient number of wafer lots to trigger a batching operation, thereby leading to a longer cycle time. The phenomenon that a lower WIP level does not guarantee a shorter average cycle time is also confirmed in this research.

Previous studies indicated that queueing network models had an accuracy within 5–10% on throughput rate (Solberg, 1977; Lazowaska et al., 1984; Chen et al., 1988). Solberg (1977) found that the discrepancy of throughput rate prediction between a queueing network model and a detailed simulation model was within 2.2%. In this study, the initial production target is at 10.6 lots daily. According to simulation results, the mean throughput rate is 10.884 lots daily under the estimated std_WIP level. The discrepancy is only 2.70%.

The actual throughput rate of the studied new wafer fab is 325 lots a month (10.833 lots per day). Its actual average WIP level is about 225 lots, i.e. nearly 1.55 times of the estimated standard WIP level. The actual average cycle time is around 21.2 days. The actual average throughput rate is very close (2.45% difference) to the simulated result after extrapolation by setting the WIP level to be $1.55 \times$ std_WIP. However, when the WIP level is $1.55 \times$ std_WIP, the actual cycle time (21.2 days) exceeds the extrapolated one (18.857 days). The cycle time discrepancy is attributed primarily to the dispatching decisions made almost arbitrarily by shop operators and some other unpredictable shop disturbances. Therefore, a WIP reduction plan (including a wafer releasing and dispatching policy and the corresponding shop management procedures) is currently taken.

In summary, it can be concluded that the standard WIP level estimated from this proposed algorithm can achieve a high throughput rate with a reasonable cycle time.

4.3.2. Theoretical and observed value comparisons on utilization of bottleneck workstation, arrival rates, R'_{il}s, and individual WIP levels at each workstation

The simulated result of bottleneck utilization (Table 10), 0.9512, is fairly close (1.02% difference) to the theoretical value $(\lambda_B/(S_B \times \mu_B))$, 0.9610. The arrival rate at each workstation estimated in Section 3.1 is called theoretical arrival rate. The observed arrival rate is collected from each simulation run. Table 11 presents both theoretical and observed arrival rates at each workstation. The average difference between theoretical and observed arrival rates is only 6.44%. For a specific product type i, the theoretical probability of entering loop l, R'_{il}, is derived from Eq. (2). The simulated result of probability, R'_{il}, and the theoretical one are given and compared in Table 12. The average discrepancy between theoretical and observed R'_{il} s is 0.8% which is very small.

The comparison of WIP level at each workstation between the simulated and the queueing network-based predictions are presented in Table 13. Larger differences of WIP levels at workstations between the results of simulation and queueing model occur mostly at heavier loading workstations. Although the discrepancy of WIP level at each workstation is not very close (32% items' discrepancies of individual WIP values between the queueing network-based model and the simulation model are within 10%), the purpose of this research is to determine the total standard WIP level so that the Fixed-WIP release policy and achieves a high throughput rate with a reasonable short cycle time. If one only concerns with such an aggregate WIP characteristic, the queueing network-based model is quite adequate (the total discrepancy on total WIP estimation is only 2.90%).

5. Conclusions

Although there are many WIP-based release control policies which have been proven to be effective for wafer fabrication, few methods, have been proposed to find the suitable WIP-level as a parameter for those release policies. This paper

Table 11
Comparisons of observed and theoretical arrival rates (lots/h)

Workstation j	1	2	3	4	5	6	7	8	9	10
Observed	3.589	8.101	8.556	0.455	1.368	3.059	0.656	1.149	0.457	6.572
Theoretical	3.040	7.409	7.850	0. 441	1.323	2.966	0.637	1.107	0.441	5.924
Workstation j	11	12	13	14	15	16	17	18	19	20
Observed	4.279	1.014	4.281	1.366	2.776	2.275	1.365	2.570	1.229	3.821
Theoretical	3.706	0.983	4.147	1.323	2.693	1.323	0.882	2.445	1.233	3.262
Workstation j	21	22	23	24	25	26	27	28	29	30
Observed	0.803	4.453	0.455	7.515	4.281	6.250	1.293	0.886	0.927	1.128
Theoretical	0.779	4.241	0.441	7.281	4.152	5.613	1.250	0.857	0.901	1.094
Workstation j	31	32	33	34	35	36	37	38	39	40
Observed	1.346	1.170	4.586	1.441	0.813	2.131	2.179	0.728	0.604	1.582
Theoretical	1.302	1.134	4.002	1.323	0.788	2.059	2.111	0.706	0.584	1.531
Workstation j	41	42	43	44	45	46	47	48	49	50
Observed	1.000	1.971	2.712	3.590	0.455	0.746	0.745	1.529	1.185	0.358
Theoretical	0.970	1.397	2.187	3.040	0.441	0.723	0.723	1.485	1.147	0.347
Workstation j	51	52	53	54	55	56	57	58	59	60
Observed	1.798	1.150	1.895	0.433	0.071	0.071	0.456	0.455	0.433	1.171
Theoretical	1.744	1.118	1.834	0.420	0.069	0.069	0.441	0.441	0.420	1.134

Table 12
Comparisons of observed and theoretical $R'_{il}s$

Product	Loops	1	2	3	4	5	6	7	8	9	10	11
A	Theoretical	0.0167	0.0167	0.0167	0.0167	0.0167	0.0167	0.0167	0.0167	0.0167	0.0167	0.0167
	Observed	0.0170	0.0168	0.0168	0.0168	0.0167	0.0167	0.0166	0.0166	0.0165	0.0164	0.0162
B	Theoretical	0.0177	0.0177	0.0177	0.0177	0.0177	0.0177	0.0177	0.0177			
	Observed	0.0179	0.0178	0.0178	0.0177	0.0177	0.0176	0.0176	0.0175			
C	Theoretical	0.0051	0.0051	0.0051	0.0051	0.0051	0.0051	0.0051				
	Observed	0.0052	0.0052	0.0051	0.0051	0.0051	0.0051	0.0050				
D	Theoretical	0.0370	0.0370	0.0370	0.0370	0.0370	0.0370	0.0370	0.0370	0.0370	0.0370	
	Observed	0.0376	0.0374	0.0374	0.0373	0.0371	0.0370	0.0370	0.0368	0.0367	0.0362	
E	Theoretical	0.0299	0.0299	0.0299	0.0299	0.0299	0.0299	0.0299	0.0299	0.0299		
	Observed	0.0303	0.0302	0.0301	0.0299	0.0299	0.0298	0.0297	0.0296	0.0295		

proposes a queueing network-based algorithm to determine the total standard WIP level in a wafer fabrication factory so that a effective WIP-based release control policy, Fixed-WIP release control policy, can apply

A simulation model with real numerical data of a newly installed wafer fabrication factory in the Science Based Industrial Park in Taiwan is tested. The experimental result shows that the std_WIP level derived from the heuristic algorithm performs well. It serves as a suitable parameter for the Fixed-WIP lot control policy and achieves a high throughput rate with a reasonable short cycle time under that control policy. Results also demonstrate that the queueing network-based model can be applied to the total standard WIP estimation problem and the proposed algorithm is a very efficient method to determine the standard WIP level for wafer fabrication.

Table 13
The WIP level at each workstation estimated by simulated and queueing network-based model

Workstation j	1	2	3	4	5	6	7	8	9	10
W_j (lots)	1.256	1.491	1.461	0.804	2.195	4.555	0.074	3.098	1.476	2.248
Simulated WIP	3.413	2.163	1.734	0.588	2.897	3.869	0.790	2.084	1.600	2.432
Workstation j	11	12	13	14	15	16	17	18	19	20
W_j (lots)	1.878	1.054	3.248	2.417	9.716	0.287	1.609	2.010	0.690	5.692
Simulated WIP	2.270	0.959	3.076	2.349	7.495	0.890	4.524	3.828	0.698	8.826
Workstation j	21	22	23	24	25	26	27	28	29	30
W_j (lots)	2.612	1.345	0.074	1.963	25.089	0.396	2.373	1.183	10.471	1.672
Simulated WIP	2.527	2.598	0.085	1.359	16.783	0.557	2.339	1.192	4.390	1.820
Workstation j	31	32	33	34	35	36	37	38	39	40
W_j(lots)	0.333	2.348	3.292	0.394	0.221	0.810	1.560	0.135	1.064	2.047
Simulated WIP	1.970	2.573	3.468	4.460	2.046	2.077	2.160	0.121	1.093	2.270
Workstation j	41	42	43	44	45	46	47	48	49	50
W_j (lots)	0.462	1.427	0.381	0.574	1.823	0.606	0.662	0.792	2.604	0.706
Simulated WIP	0.492	3.305	0.473	0.692	1.913	0.805	1.575	1.361	3.396	1.689
Workstation j	51	52	53	54	55	56	57	58	59	60
W_j(lots)	9.683	2.149	0.447	0.694	0.085	0.163	12.647	0.125	1.216	1.756
Simulated WIP	3.683	2.045	0.589	1.026	0.113	0.154	3.287	0.015	1.163	1.966

Acknowledgements

The authors acknowledge department manager Richard Huang for the support of production planning department of UMC Fab 3 and are grateful to B.M. Sonug and Fransia Hsu for the discussion on the model building. The work was supported in part by the National Science Council, Taiwan, ROC, under project number NSC 88-2213-E-233-003.

References

APICS Dictionary, seventh ed. American Production and Inventory Control Society, USA.

Askin, R.G., Standrige, C.R., 1993. Modeling and Analysis of Manufacturing Systems. Wiley, Singapore.

Atherton, R.W., Dayhoff, J.E., 1986. Signature analysis simulation of inventory cycle time and throughput trade-offs in wafer fabrication. IEEE Transactions on Components Hybrids Manufacturing Technology 9, 498–507.

Bechte, W., 1988a. Theory and practice of load-oriented manufacturing control. International Journal of Production Research 26 (3), 375–395.

Bechte, W., 1988b. Load oriented manufacturing control. In: 23rd Annual APICS Conference Proceeding. Falls Church, VA, USA, pp. 148–152.

Bechte, W., 1994. Load-oriented manufacturing control just-in-time production for job shops. Production Planning & Control 5 (3), 292–307.

Bertrand, J.W.M., Wortmann, J.C., 1981. Production Control and Information Systems for Component-Manufacturing Shops. Elsevier, Amsterdam.

Burman, D.Y., Gurrola-Gal, F.J., Nozari, A., Sathaye, S., Sitarik, J.P., 1986. Performance analysis techniques for IC manufacturing lines. AT & T Technical Journal 65 (4), 46–57.

Buzacott, J.A., 1971. The role of banks in flow-line production systems. International Journal of Production Research 9 (4), 425–436.

Chen, H., Harrison, J.M., Mandelbaum, A., Ackere, A.V., Wein, L.M., 1988. Empirical evaluation of a queueing network model for semiconductor wafer fabrication. Operations Research 36 (2), 202–215.

Connors, D.P., Feigin, G.E., Yao, D.D., 1996. A queueing network model for semiconductor manufacturing. IEEE Transactions on Semiconductor Manufacturing 9 (2), 412–427.

Dayhoff, J.E., Atherton, R.W., 1984. Simulation of VLSI manufacturing areas. VLSI Design, December, pp. 84–92.

Enns, S.T., 1995. An integrated system for controlling shop loading and work flow. International Journal of Production Research 33 (10), 2801–2820.

Fowler, J.W., Phillips, D.T., Hogg, G.L., 1992. Real-time control of multiproduct bulk-service semiconductor manufacturing processes. IEEE Transactions on Semiconductor Manufacturing 5 (2), 158–163.

Glassey, C.R., Resende, M.G.C., 1988a. Closed-loop job release control for VLSI circuit manufacturing. IEEE

Transactions on Semiconductor Manufacturing 1 (1), 36–46.

Glassey, C.R., Resende, M.G.C., 1988b. A scheduling rule for job shop release in semiconductor fabrication. Operations Research Letters 7 (5), 213–217.

Glassey, C.R., Weng, W.W., 1991. Dynamic batching heuristic for simultaneous processing. IEEE Transactions on Semiconductor Manufacturing 4 (2), 77–82.

Goldratt, E., Cox, J., 1986. The Goal: A Process of Ongoing Improvement. North River Press, New York.

Graves, R.J., Konopka, J.M., Milne, R.J., 1995. Literature review of material flow control mechanisms. Production Planning & Control 6 (5), 395–403.

Gross, D., Harris, C.M., 1974. Fundamentals of Queueing Theory. Wiley, New York.

Gurnani, H., Anupindi, R., Akella, R., 1992. Control of batch processing systems in semiconductor wafer fabrication facilities. IEEE Transactions on Semiconductor Manufacturing 5 (4), 319–327.

Huang, H.W., 1995. The design of production activity control policy for wafer fabrication factories. Master Thesis, National Chiao-Tung University, Taiwan, ROC.

Jackson, J.R., 1963. Jobshop-like queueing systems. Management Science 10 (1), 131–142.

Jacobs, F.R., Bragg, D.J., 1988. Repetitive lots: Flow-time reductions through sequencing and dynamic batch sizing. Decision Sciences 19, 281–294.

Karmarkar, U.S., Kekre, S., Kekre, S., 1985. Lotsizing in multi-item multi-machine job shops. IIE Transactions 17 (3), 290–297.

Kirk, R.E., 1982. Experimental Design: Procedure for the Behavioral Sciences. Brooks/Cole, USA.

Law, A.M., Kelton, W.D., 1991. Simulation Modeling & Analysis. McGraw-Hill, Singapore.

Lazowaska, E.D., Zahorjan, J., Graham, G.S., Sevcik, K.C., 1984. Quantitative System Performance. Prentice-Hall, Englewood Cliffs, NJ.

Lou, S.X.C., 1989. Optimal control rules for scheduling job shops. Annals of Operations Research 17, 233–248.

Lou, S.X.C., Kager, P.W., 1989. A robust production control policy for VLSI wafer fabrication. IEEE Transactions on Semiconductor Manufacturing 2 (4), 159–164.

Lozinski, C., Glassey, C.R., 1988. Bottleneck starvation indicators for shop floor control. IEEE Transactions on Semiconductor Manufacturing 1 (4), 147–153.

Medhi, J., 1975. Waiting time distribution in a Poisson queue with a general bulk service rule. Management Science 21 (7), 777–782.

Miller, D.J., 1990. Simulation of a semiconductor manufacturing line. Communications of the ACM 33 (10), 99–108.

Neuts, M.F., 1967. A general class of bulk queues with Poisson input. Annals of Mathematical Statistics 38 (3), 759–770.

Montgomery, D.C., Peck, E.A., 1992. Introduction to Linear Regression Analysis, second ed. Wiley, New York.

Robinson, J.K., Flower, J.W., Bard, J.F., 1995. The use of upstream and downstream information in scheduling semiconductor batch operations. International Journal of Production Research 33 (7), 1849–1869.

Roderick, L.M., Philips, D.T., Hogg, G.L., 1992. A comparison of order release strategies in production control systems. International Journal of Production Research 30 (3), 611–626.

Solberg, J.J., 1977. A mathematical model of computerized manufacturing systems. In: The Fourth International Conference on Production Research, Tokyo, pp. 22–30.

Spearman, M.L., Woodruff, D.L., Hopp, W.J., 1989. CONWIP: A pull alternative to kanban. International Journal of Production Research 28 (5), 879–894.

Spearman, M.L., Zazanis, M.A., 1992. Push and pull production systems: Issues and comparisons. Operational Research 40 (3), 521–532.

Suri, R., Diehl, G.W.W., Treville, S., Tomsicek, M.J., 1995. From CAN-Q to MPX: Evolution of queuing software for manufacturing. Interfaces 25 (5), 128–150.

Uzsoy, R., Lee, C.Y., Martin-Vega, L.A., 1992. A review of production planning and scheduling models in the semiconductor industry. Part I: System characteristics, performance evalution and production planning. IIE Transactions 24 (4), 47–60.

Vinod, B., Altiok, T., 1986. Approximating unreliable queuing networks under the assumption of exponentially. Journal of Operational Research Society 37, 309–316.

Wein, L.M., 1988. Scheduling semiconductor wafer fabrication. IEEE Transactions on Semiconductor Manufacturing 1 (3), 115–130.

Weng, W.W., Leachman, R.C., 1993. An improved methodology for real-time production decisions at batch-process work stations. IEEE Transactions on Semiconductor Manufacturing 6 (3), 219–225.

Whitt, W., 1983. The queueing network analyzer. The Bell System Technical Journal 62 (9), 2779–2815.

Wiendahl, H.P., Glassner, J., Petermann, D., 1992. Applications of load-oriented manufacturing control in industry. Production Planning & Control 3 (2), 118–129.

Wiendahl, H.P., 1995. Load-Oriented Manufacturing Control. Springer, Berlin.

Yan, H., Lou, S., Sethi, S., Gardel, A., Deosthali, P., 1996. Testing the robustness of two-boundary control policies in semiconductor manufacturing. IEEE Transactions on Semiconductor Manufacturing 9 (2), 285–288.