International Journal of Production Research

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tprs20

Job order releasing and throughput planning for multi-priority orders in wafer fabs

S. H. Chung [a], W. L. Pearn [a], A. H. I. Lee [a] & W. T. Ke [a]

[a] Department of Industrial Engineering & Management, National Chiao Tung University, 1001 Ta Hsueh Road, Hsin Chu, Taiwan, 30050, ROC

Published online: 14 Nov 2010.

To cite this article: S. H. Chung, W. L. Pearn, A. H. I. Lee & W. T. Ke (2003) Job order releasing and throughput planning for multi-priority orders in wafer fabs, International Journal of Production Research, 41:8, 1765-1784, DOI: 10.1080/0020754021000049862

To link to this article: http://dx.doi.org/10.1080/0020754021000049862

INT. J. PROD. RES., 2003, VOL. 41, NO. 8, 1765–1784

Taylor & Francis
Taylor & Francis Group

Job order releasing and throughput planning for multi-priority orders in wafer fabs

S. H. CHUNG†*, W. L. PEARN†, A. H. I. LEE† and W. T. KE†

To meet the production target of multi-level (multiple priority rank) orders in wafer fabs, this paper uses a hierarchical framework based on a mathematical model, and without the assistance of any simulation tool, to build a production scheduling system to plan wafer lot releasing sequence and time. This system first applies capacity loading analysis to set up the batch policy for each level (rank) of orders. Next, the production cycle time of each product level is estimated with considerations of batching and loading factor. The cycle time is then used to derive system control parameters such as the most appropriate level of work in process (WIP) and the number of daily operations on the bottleneck workstation. Lastly, a Constant WIP mechanism is applied to establish a wafer release sequence table and a throughput timetable. The due date designation for each specific order can hence be confirmed. With the comparison with the result of simulation, it shows that under the designed system the performance and planning measures in the master production schedule can be drawn up quickly and accurately, and the system throughput target and due date satisfaction can be achieved. Overall, the proposed production scheduling system is both effective and practicable, and the planning results are supportive for good target planning and production activity control.

1. Introduction

At the same time as upgrading manufacturing technology, wafer manufacturers raise more capital to increase capacity. This has made the supply of products increase tremendously and the competition become even fiercer. The different profit rate of products and the varied importance level of clients result in different levels of orders in a fab, and this makes production scheduling deal with many types of products and different levels of orders. The existence of rush orders, because of their priority for processing, oppresses normal orders, and as a result the average production cycle time of normal orders and its variance are increased, and the estimation of production cycle time and schedule planning becomes more difficult. In the past, scholars researching production planning and scheduling for wafer fabs usually assumed a uni-level product type, and very few considered multi-level products. In order to increase a company's competition edge and profitability, an effective schedule planning system, with the ability of setting wafer release sequence and time as well as throughput, must be built so as to achieve the system throughput target under the given combination of product type and priority rank.

Revision received September 2002.

† Department of Industrial Engineering & Management, National Chiao Tung University, 1001 Ta Hsueh Road, Hsin Chu, Taiwan 30050, ROC.

* To whom correspondence should be addressed. e-mail: t7533@cc.nctu.edu.tw

International Journal of Production Research ISSN 0020–7543 print/ISSN 1366–588X online
http://www.tandf.co.uk/journals
DOI: 10.1080/0020754021000049862

This paper covers master production scheduling and material release planning. Based on the throughput plan of a wafer fab, the most appropriate WIP level, production cycle time, wafer release sequence table and completion time table for each rank of product orders are prepared. Such a planning result can be valuable in setting the due date and for decision-making in production activity control system.

2. Literature review

Production cycle time is the time spent by a wafer batch from material release to completion (Kramer 1989). It can be further separated into theoretical cycle time and waiting time. Theoretical cycle time includes process time, set-up time, air pumping in and out time, and wafer carrying time, etc. (Winston 1991). Waiting time includes the time waiting for process and the time waiting for move (Raos 1992). Usually, theoretical cycle time has a smaller variance and often is treated as a constant, while waiting time has a high degree of uncertainty.

For cycle time estimation, Chung and Huang (1999), with the application of queueing theory and the observation of the characteristics of material flow, developed a production cycle time estimating formulation, the Block-Based Cycle Time (BBCT) estimation algorithm, which has distinguishable performance. Conway *et al.* (1967) adopted Laplace transforms to estimate the job cycle time on a single machine, and the cycle time and WIP level can both be considered at the same time.

For master production scheduling, the planning method can be separated into three categories: simulation, mathematical programming and the combination of the two. The advantage of using a simulation tool to plan schedules is that it is easy to construct the system that matches with the production activity control stage (Liu *et al.* 1995). The simulation can offer what-if analysis and control well the WIP level in a certain range, but it takes time. Linear programming based on capacity constraints can quickly derive the best production plan, but there are too many assumptions in describing the real phenomenon (Chu 1995, Hackman and Leachman 1989). For this reason, some researches combined the simulation and mathematic algorithm to describe better the environment than by separately adopting either of the two methods (Burman *et al.* 1986, Hung and Leachman 1996, Thompson and Davis 1990).

Miller (1990) has tried to make the system WIP level constant and decide on batch size in fab by simulation tool. Neuts (1975) presented the idea of minimal batch size (MBS), which means that even if a machine is idle, the jobs must be held until a certain level of batch size has been gathered before they can be processed. To minimize the waiting time in a queue line, Glassey and Weng (1991) considered the future arrival rate and applied Dynamic Batching Heuristic (DBH) to set up the batch size for a single product and single furnace machine.

For the performance of throughput, Chung *et al.* (1997) focused on layers completed in a fab to measure the production performance for shop-floor activity. Leachman and Hodges (1996) designed an integrated production ability index to measure the performance in different fabs by quality metrics, productivity metrics, and production speed and output target. Chung and Huang (1999) investigated 12 material release rules for wafer fabs. Among them, Constant WIP (CONWIP) is a type of material release control between Just-in-Time (JIT) and a push system. It considers the use of capacity-constrained resource (CCR), controls WIP in number of units instead of time and dynamically modifies the material-release time when machines break down. Therefore, CONWIP is a relatively good production activity

control rule and thus will be used in this research. The present paper will estimate cycle time based on BBCT and will plan MPS, which includes material release time and order, daily bottleneck throughput and throughput time. Little's law will be applied, which reveals the relation of WIP level (L), releasing rate (λ) and production cycle time (W), and the relation can be shown as $L = \lambda W$ (Little 1961).

3. Model construction

3.1. *Master production planning*

Generally speaking, the priority levels for orders in a wafer fab are classified into several different levels. In this study, the levels of orders are categorized into three: hot lots, rush lots and normal lots. Hot lots have the highest processing priority and are not restricted to the batch policy. In other words, a hot lot can be processed even if it consists of only a single lot. Rush lots have the second processing priority, and the batch policy is determined by the proposed scheduling system under rough-cut capacity planning. Normal lots have the lowest processing priority and are constrained by the full-loaded batch policy. To simplify the complexity of the problem, the assumptions in this study are made as follows.

- Monthly throughput target, product type and rank mix are predetermined.
- No carrying cost is considered.
- No human resource and material supply shortage problem.
- Processing steps and the corresponding processing times are different only in the product type, not in the processing priority.
- Size of each customer order is a multiple of release batch size.

Normally, the planning horizon for master production schedule (MPS) is 12 weeks, while the planning period (T) is 4 weeks. The framework for MPS planning is shown in figure 1.

First, an MPS planning module evaluates machine loading and determines batch policy according to system throughput target, product type and priority mix. Second, the 'block-based cycle time estimation algorithm for multiple-priority orders' (BBCT-MP), developed by Chung *et al.* (2001), is used to calculate the cycle time for each product type in each rank. The information about cycle time and throughput is used to derive the suitable WIP level and to calculate the number of daily bottleneck operations. A wafer lot releasing sequence and a timetable are then established. The corresponding order completion time and order due date are determined in consequence.

3.2. *Capacity loading evaluation and batch policy determination*

Adopting a partial loading batch policy for a hot lot and a rush lot will shorten their production cycle time. However, the machine utilization rate for each batch machine will be raised, and this will lead to a bottleneck or critical machine wandering, which in turn will increase the variation in cycle times.

To stabilize utilization rates among machine types, the MPS planning will analyse the machine loading levels for each type of batch policy to meet the throughput target under a predetermined product type and rank mix. The batch policy is set to be one lot for hot lots and full-loaded for normal lots. Since the maximum batch size for many batch machines is six lots, and these machines such as furnace usually have quite long processing times, we present the batch policy evaluation process for machines with a maximum batch size of six in figure 2. In this case, the batch

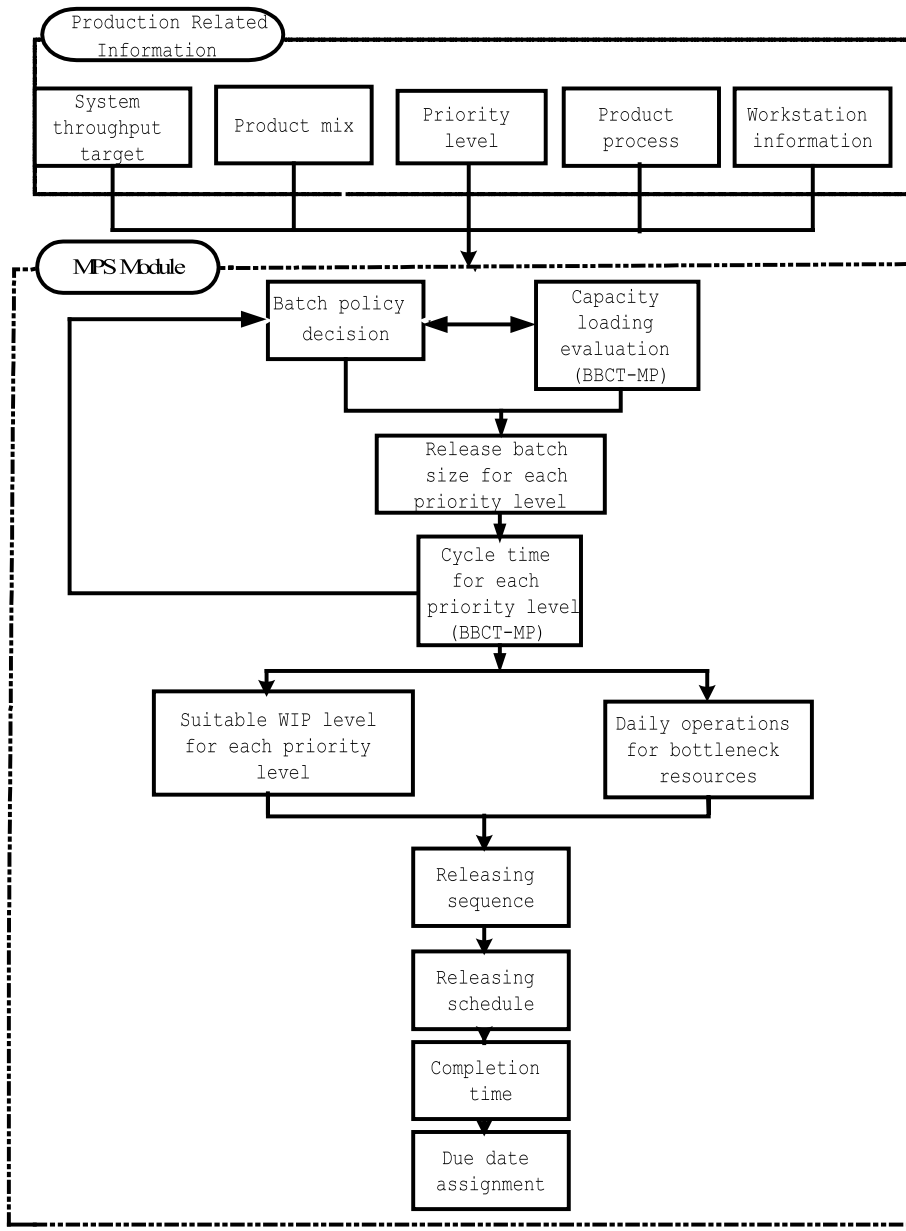

Figure 1. Framework for MPS planning.

policy is set to be six lots, full-loaded for normal lots. While for rush lots we will evaluate whether the batch size of one lot will make the machine overloading happen. If there is not sufficient capacity, we need to increase the batch size of rush lots. Even if there is sufficient capacity, we still need to check the utilization rate of CCR. The utilization of CCR should be at least 10% less than that of bottleneck to avoid bottleneck wandering. The batch size of rush lots is determined when there is little chance of bottleneck wandering.

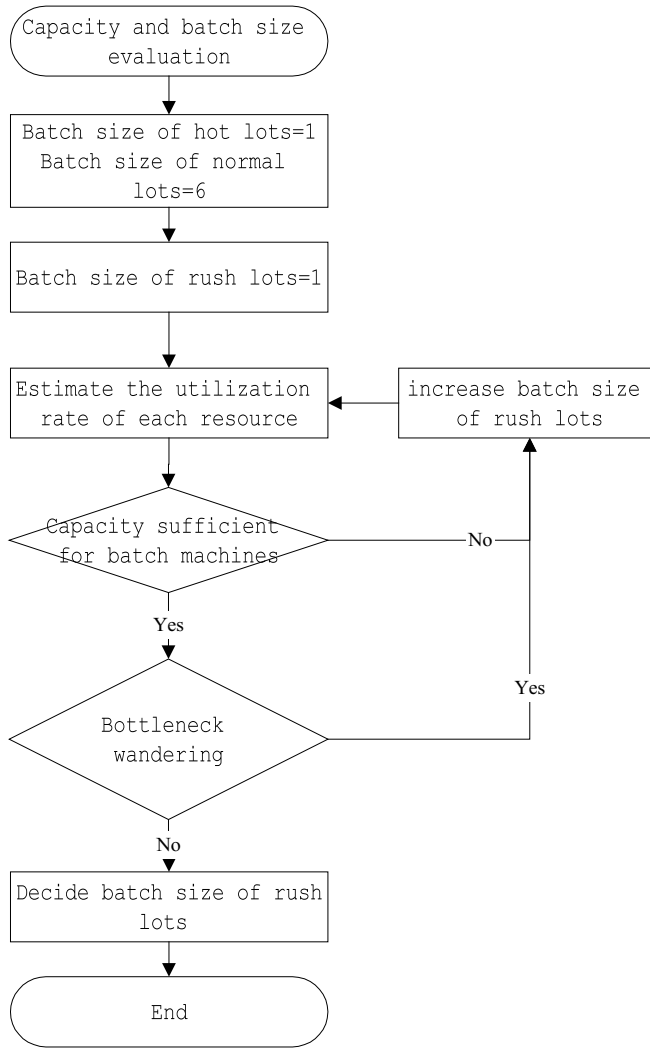

Figure 2. Batch policy evaluation process.

A queuing model is often adopted to estimate the utilization rate of a resource. Owing to expensiveness and essentials in making a photo layer, the photolithography stepper is treated as the bottleneck resource in the real world. In addition, the number of process steps and the total process time for each photo layer are significantly different. Such characteristics make a large variation in time for re-entry to one specific workstation. We thus assume that each workstation is an M/M/c queuing system, and workstations are independent in loading evaluation.

In order to derive the arrival rate for each level of product orders coming to a workstation, we let that quantity for system input equal the planned throughput of the planning period based on CONWIP material release control policy. The procedures for deriving utilization rate are as follows.

Step 1. Calculate the average system hourly arrival (throughput) rate, α_{hr}, based on throughput target $\Re$ in a planning period (T):

$$\alpha_{\mathrm{hr}} = \frac{\Re}{Tz \times 24}. \tag{1}$$

Step 2. Estimate the mean arrival rate λ_k^{pri} for each rank of product orders arriving workstation k per hour. It was derived by the average system hourly arrival (throughput) rate, mix ratio for product type i ($i = 1, \ldots, I$) to all product types, π_i, proportion of throughput with rank *pri* to throughput of all product type i (π_i^{pri}), number of process i visiting workstation k, f_{ik}, and average rework rate for workstation k, γ_k:

$$\lambda_k^{pri} = \alpha_{\mathrm{hr}} \times \sum_i [\pi_i \times \pi_i^{pri} \times f_{ik}(1 + \gamma_k)], \text{ for each } k \text{ and } pri. \tag{2}$$

Step 3. Estimate the equivalent available units for machine type k, c_k, which equals the total available units (n_k) deducting the downtime ratio and maintenance time ratio:

$$c_k = \sum_{m=1}^{n_k} \left(1 - \frac{\mathrm{MTTR}_{k_m}}{\mathrm{MTBF}_{k_m} + \mathrm{MTTR}_{k_m}} - \frac{\mathrm{MTTPM}_{k_m}}{\mathrm{MTBPM}_{k_m} + \mathrm{MTTPM}_{k_m}}\right) \tag{3}$$

for each k, where MTTR_{k_m} is the mean time to repair (MTTR) for the mth machine for machine type k, MTBF_{k_m} is the mean time between failure (MTBF) for the mth machine for machine type k, MTTPM_{k_m} is the mean time to preventive maintenance (MTTPM) for the mth machine for machine type k, and MTBPM_{k_m} is the mean time between preventive maintenance (MTBPM) for the mth machine for machine type k.

Step 4. Estimate the number of machines used for producing all kinds of hot lots ($pri = h$), c_k^h. Because the minimum batch size for hot lot flowing through machine type k, $B_k^{\min,h}$, is set to be 1, the batch size consideration is omitted from this formula:

$$c_k^h = \alpha_{\mathrm{hr}} \times \sum_{i=1}^{\mathrm{I}} \left[\pi_i \times \left(\pi_i^h \times \sum_{p \in \{M(i,p)=k\}} \mathrm{PT}_{ip}\right)\right], \text{ for each } k, \tag{4}$$

where PT_{ip} is the process time for product i at pth step, and $M(i,p)$ is the workstation type used for process type i at pth step.

Step 5. Estimate the number of machines used for producing all kinds of rush lots ($pri = r$), c_k^r, with the temporary batch size set at $B_k^{\min,r}$:

$$c_k^r = \alpha_{\mathrm{hr}} \times \sum_{i=1}^{\mathrm{I}} \left[\pi_i \times \left(\pi_i^i \times \sum_{p \in \{M(i,p)=k\}} \frac{\mathrm{PT}_{ip}}{B_k^{\min,r}}\right)\right], \text{ for each } k. \tag{5}$$

Step 6. Estimate the number of machine type k available for normal lots, c_k^n. It is the remaining machine units after providing all the capacity needs to hot lots and rush lots:

$$c_k^n = c_k - c_k^h - c_k^r, \text{ for each } k. \tag{6}$$

Step 7. Estimate the mean process time for all kinds of processes passing through machine type k, $\overline{\mathrm{PT}_k}$ (hr):

$$\overline{\mathrm{PT}_k} = \sum_{i=1}^{I} \left[\pi_i \times \sum_{pri} \left(\pi_i^{pri} \times \sum_{p \in \{M(i,p)=k\}} \frac{\mathrm{PT}_{ip}}{B_k^{\min,pri} \times f_{ik}} \right) \right], \text{ for each } k. \tag{7}$$

Step 8. Estimate the mean output rate for wafer lots with rank *pri* passing through workstation k, O_k^{pri}:

$$O_k^{pri} = \frac{c_k^{pri} \times B_k^{\min,pri}}{\overline{\mathrm{PT}_k}}, \text{ for each } k \text{ and } pri. \tag{8}$$

Step 9. Estimate the mean output rate for each workstation k, O_k:

$$O_k = \sum_i \left[\pi_i \times f_{ik} \times \sum_{pri} (\pi_i^{pri} \times O_k^{pri}) \right], \text{ for each } k. \tag{9}$$

Step 10. Estimate the mean service rate in a planning period for each workstation, μ_k. It is the product of efficiency ratio of workstation k (e_k, the fraction of time that the workstation is in a condition to perform its intended function to the total production available time) and mean output rate for each workstation k:

$$\mu_k = e_k \times O_k, \text{ for each } k. \tag{10}$$

Step 11. Estimate the mean utilization rate of the resources k, ρ_k. It is calculated by dividing the mean arrival rate for all product orders arriving a workstation to the mean service rate for that workstation:

$$\rho_k = \frac{\sum_{pri} \lambda_k^{pri}}{\mu_k}, \text{ for each } k. \tag{11}$$

3.3. *Estimation of cycle time for each priority level*

Once a batch policy is determined, the release batch size for each priority rank of orders is set the same as the largest batch size of all workstations set for that priority rank to utilize workstations effectively and to make flow smoothly. Let n^{pri} be the number of lots contained in a release batch with rank *pri*, where *h*, *r*, and *n* stand for hot, rush and normal, respectively. After the releasing batch size for each rank of orders is defined, an accepted order will be cut into several suborders, based on the corresponding releasing batch size and for the ease of production control. Next, after the confirmation of product mix, we can apply the block-based cycle time estimation algorithm for multiple-priority (BBCT-MP), developed by Chung *et al.* (2001), to estimate cycle time for each product type in every priority rank. Input data include monthly throughput target, product type and rank mix, process plan, workstation-related information, and batch policies.

Because batch forming before a batch workstation is one of the major reasons for congestion of material flows, BBCT-MP treats each batch workstation as a cutting point and divides a manufacturing process into blocks. In each block, as depicted in figure 3, all steps are serial machine steps except the first and last process steps, which are batch-type process steps. The difference in batch sizes and in output velocity between preceding and succeeding machines induces different waiting times before machines. We thus recognize the two reasons for incurring waiting as batching and loading factors for every block.

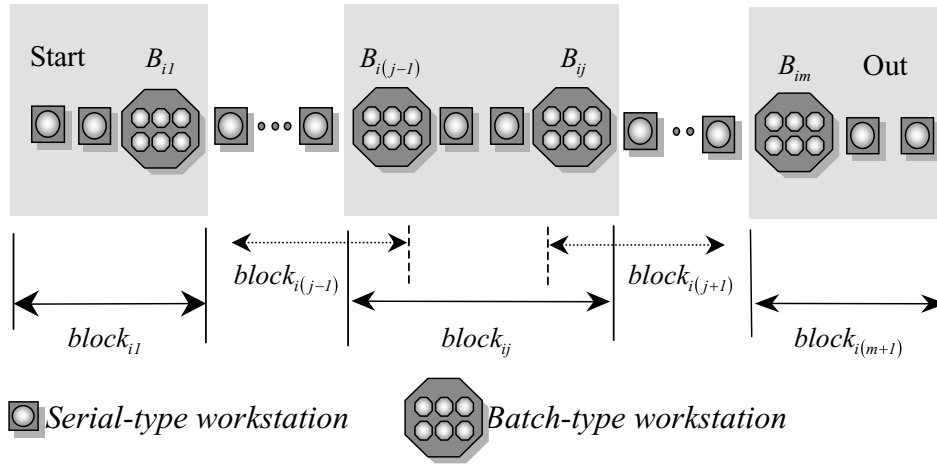

Figure 3. Block j for product type i.

The waiting time caused by batching factor consists of two parts. One is due to gathering the required batch size before batch processing. The other is due to the releasing of all the lots contained in a batch after the batch operation. This causes the increase of transient loading for a downstream serial workstation. The batch factoring flow time (BFFT) algorithm is used to estimate such a waiting time (Chung *et al.* 2001). On the other hand, the waiting time caused by the loading factor is due to the average load raised on each workstation so as to achieve the system throughput target. The waiting time for each rank of orders that incurred by the loading factor is estimated by a non-preemptive priority queuing model (Chung *et al.* 2001). Adding two kinds of waiting time and the total process time of all blocks of each process derives the product cycle time for a specific priority rank.

The BBCT-MP algorithm has a remarkable performance in cycle time estimation because it captures the exact interaction between two batch workstations and all serial workstations in a block.

3.4. *System WIP level estimation*

For production smoothing, i.e. minimizing cycle time variation, a constant WIP control is adopted here for setting the wafer release plan, and the suitable system WIP level for each priority rank of wafer lots, L_i^{pri}, needs to be derived. Little's Law in queuing theory (Little 1961), $L_i^{pri} = \lambda_i^{pri} \times \mathrm{CT}_i^{pri}$, is applied, where λ_i^{pri} is the arrival rate and CT_i^{pri} is the corresponding cycle time for the specific product type and rank. By adopting a constant WIP control policy, wafer lot(s) can be released to the shop floor only when the same quantity of wafers are finished and transferred out. The arrival rate for product type i with rank pri, λ_i^{pri}, is thus equivalent to the corresponding throughput rate.

The estimated system WIP level, L, is computed by summing up all the estimating values of L_i^{pri}. Under a constant WIP control policy, once a system WIP level is lower than L, a fixed number of consecutive lots as defined in the releasing sequence will be released into the shop floor. Note that if a system WIP level is set too low, the amount of throughput will be affected directly. The procedures for deriving system WIP level are as follows.

Step 1. Calculate the average hourly arrival rate for each product type in each priority rank, λ_i^{pri}. It is derived by multiplying α_{hr} with the product mix ratio of product i to all system throughput, π_i, and the ratio of rank *pri* orders among all product i outputs, π_i^{pri}:

$$\lambda_i^{pri} = \alpha_{\text{hr}} \times \pi_i \times \pi_i^{pri}. \tag{12}$$

Step 2. Estimate the appropriate WIP level for product i in rank *pri*, L_i^{pri}, by Little's law:

$$L_i^{pri} = \lambda_i^{pri} \times \text{CT}_i^{pri}. \tag{13}$$

Step 3. Estimate the system WIP level, L, by summing up all L_i^{pri} values:

$$L = \sum_i \sum_{pri} L_i^{pri}. \tag{14}$$

3.5. *Calculating the number of daily moves for a bottleneck resource*

As mentioned above, two major characteristics of the process of wafer fabrication are a long cycle time and numerous re-entry operations. In order to achieve the throughput target and to implement PAC efficiently, the number of daily bottleneck operations needs to be derived. A photolithography stepper, an extremely expensive resource and one in charge of the most critical operation for building every photo layer, is treated as a bottleneck resource. A high utilization rate and long queue line are expected for a photolithography stepper. Completion of a bottleneck operation implies that the most critical operation of a layer is finished and therefore is the completion of a 'system move'.

The concept of production smoothing is proposed here, i.e. the system throughput for each day is close to the average daily throughput target. Also, the output combination must fit the predetermined product type and rank mix ratio. To meet these conditions, the number of moves at each layer must be the same as the daily system output. Meanwhile, the combination of each layer output must be the same as the combination of the system output. The procedures for daily moves planning are shown below.

Step 1. Calculate planned daily output for product type i with rank *pri*, α_{id}^{pri}:

$$\alpha_{id}^{pri} = \lambda_i^{pri} \times 24. \tag{15}$$

Step 2. Estimate the number of daily moves for product type i with rank *pri*, M_{id}^{pri}, which is the product of α_{id}^{pri} and the number of photo layers for product i, N_i:

$$M_{id}^{pri} = \alpha_{id}^{pri} \times N_i. \tag{16}$$

Step 3. Number of the system daily moves, M_d is the summation of all M_{id}^{pri} values:

$$M_d = \sum_i \sum_{pri} M_{id}^{pri}. \tag{17}$$

3.6. *Building a cyclic wafer release sequence*

For products with the same priority rank, the same release batch size is defined for every product type. In this situation, the number of release times for each product type with the same rank depends on its product mix ratio to all product types. On the

other hand, for products with different priority ranks, different release batch sizes may be applied. The number of release times is proportionally reverse to their release batch sizes for product types with same product mix ratio. Thus, to make the combination of daily throughput satisfy a predetermined product type and rank mix, a table showing a cyclic release sequence must be built considering the product type mix, rank mix and the release batch size designed for each rank.

To set up a table of the releasing sequence and time, the throughput interval for each product type with each rank is calculated first. Then, the release interval can be set to match with the predetermined release batch size. For a hot run, the releasing interval is the same as its average throughput interval since its release batch size is one lot. For production orders with normal rank, their releasing interval is six times the average throughput interval since their release batch size is six lots. A similar rule is applied to rush orders once the release batch sizes are determined.

For each product type with each rank, once the release interval is derived, one can calculate the release time for every release batch lot so as to make the distribution of all WIP in the shop floor satisfy the product type and rank mix ratio. We then sort the release times of all product types and all ranks, from small to large, to establish the release timetable and release sequence. Finally, the release cycle table is determined after all products reach their expected throughput ratio. The release plan will be made based on this table. The algorithm for setting a cyclic release table is as follows.

Step 1. Calculate the product/rank mix for product i with rank *pri*, pro_i^{pri}, by dividing α_{id}^{pri} by the maximum common divider for all α_{id}^{pri}:

$$pro_i^{pri} = \frac{\alpha_{id}^{pri}}{\{\text{maximum common divider for all } \alpha_{id}^{pri}\}}, \quad \text{for each } i \text{ and each } pri, \text{ and } \alpha_{id}^{pri} \neq 0. \tag{18}$$

Step 2. Find the minimum common multiplier for all kinds of release batch sizes, i.e. $[n^h, n^r, n^n]$, and calculate the relative frequency for job order with rank *pri* by dividing $[n^h, n^r, n^n]$ by the release batch specified for rank *pri*, n^{pri}. The release number for product i with rank *pri*, q_i^{pri}, is the multiplication of pro_i^{pri} with the corresponding relative release frequency for rank *pri*:

$$q_i^{pri} = pro_i^{pri} \times \frac{[n^h, n^r, n^n]}{n^{pri}}. \tag{19}$$

Step 3. Compute the maximum sequencing number in the release table, $q_{\max}$ by summing up all q_i^{pri} digits of all product types with all ranks:

$$q_{\max} = \sum_i \sum_{pri} q_i^{pri}. \tag{20}$$

Step 4. Calculate the daily average interval between the throughput of product type i with rank *pri*, $I_{\text{out},i}^{pri}$. It is the reciprocal of α_{id}^{pri}:

$$I_{\text{out},i}^{pri} = \frac{1}{\alpha_{id}^{pri}}, \text{ for each } i, \text{ each } pri. \tag{21}$$

Step 5. Estimate the average release interval for product type i with rank pri, $I^{pri}_{\text{arr},i}$, by multiplying the average interval between throughput $I^{pri}_{\text{out},i}$ with its corresponding release batch size n^{pri}:

$$I^{pri}_{\text{arr},i} = I^{pri}_{\text{out},i} \times n^{pri}, \text{ for each } i, \text{ each } pri. \tag{22}$$

Step 6. For every product type i with rank pri, set release numbering $q = 1$.

Step 7. Let the first release time for every product type i with rank pri, $t^{pri}_{\text{arr},i}$, equal its average release interval $I^{pri}_{\text{arr},i}$:

$$t^{pri}_{\text{arr},i} = I^{pri}_{\text{arr},i}, \text{ for each } i, \text{ each } pri. \tag{23}$$

Step 8. Choose the smallest release time among all product types and all ranks. Identify its product type i^*, rank pri^*, and put it on the release table, with sequence numbering q. If there are more than two product types or ranks that have the same smallest release time, choose the one with the smaller number of times being chosen to break the tie. However, if the tie still exists, arrange the sequence randomly. Record the $\text{seq}_q = (i^*, pri^*)$ information in the cyclic release table.

$$t^{pri^*}_{\text{arr},i^*} = \min\{t^{pri}_{\text{arr},i}, \text{ for all } i \text{ and all } pri\} \tag{24}$$

$$\text{seq}_q = (i^*, pri^*) \tag{25}$$

$$q = q + 1. \tag{26}$$

Step 9. Renew the next release time for product type i^* with rank pri^* only. The others remain the same:

$$t^{pri}_{\text{arr},i} = t^{pri^*}_{\text{arr},i^*} + I^{pri}_{\text{arr},i}, \text{ for } i = i^*, pri = pri^* \text{ only.} \tag{27}$$

Step 10. Repeat Steps 8 and 9 until $q = q_{\max}$. Then, the cyclic release table is built.

For example, assume that planned daily throughput is 20 lots and daily outputs α^{pri}_{id} for products A, B and C are 2, 6 and 12 lots respectively. Further assume that product A ranks hot and the release batch size is one lot, while products B and C are normal and the release batch size is six lots. The product/rank mix calculated using Step 1 is as in table 1.

Product type_rank	α^{pri}_{id}	$pro^{pri}_i = \dfrac{\alpha^{pri}_{id}}{\{\text{maximum common divider for all } \alpha^{pri}_{id}\}}$
A_H	2	$\dfrac{2}{\{\text{maximum common divider for 2, 6, 12}\}} = 1$
B_N	6	$\dfrac{6}{\{\text{maximum common divider for 2, 6, 12}\}} = 3$
C_N	12	$\dfrac{12}{\{\text{maximum common divider for 2, 6, 12}\}} = 6$

Table 1. Calculations of product/rank mix.

Product type_rank	pro_i^{pri}	$pro_i^{pri} \times \frac{[n^h, n^r, n^n]}{n^{pri}}$
A_H	1	$1 \times \frac{[1,6,6]}{1} = 1 \times \frac{6}{1} = 6$
B_N	3	$3 \times \frac{[1,6,6]}{6} = 3 \times \frac{6}{6} = 3$
C_N	6	$6 \times \frac{[1,6,6]}{6} = 6 \times \frac{6}{6} = 6$

Table 2. Calculations of maximum sequencing number.

The pro_i^{pri} ratio is thus 1:3:6 for products A_H, B_N and C_N, respectively. The maximum sequencing number in the release table, $q_{\max}$, can be derived following Steps 2 and 3 as in table 2.

From table 2, we can calculate that $q_{\max} = 6 + 3 + 6 = 15$.

The corresponding average throughput intervals are 1/2, 1/6 and 1/12 day, and the average release intervals between throughput are thus 1/2, 1 and 1/2 day, respectively, for products A, B and C. As shown in figure 4, the release sequence is A-C-B-A-C—A-C-B-A-C—A-C-B-A-C. Note that an alternative option is acceptable by replacing all the A-C sequence with C-A.

3.7. *Releasing and delivery schedule setting*

Once the cyclic release sequence is known, the release schedule of the planning period can be easily derived. First record the last release time in the previous planning period for each product type with each rank, $t_{\mathrm{arr},i}^{pri}$. For each job order with distinguishable product type and rank, we can derive its planned release time by using the corresponding information of throughput interval, release batch size

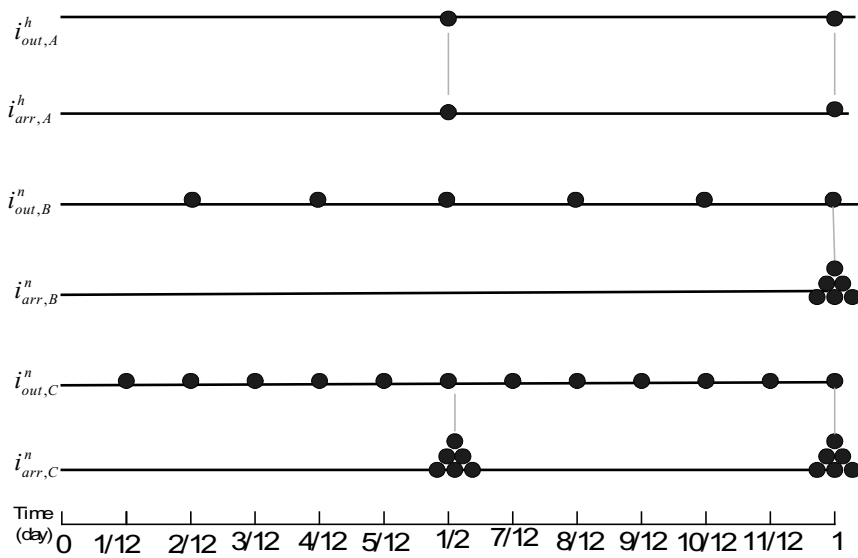

Figure 4. Setting of the wafer release sequence and timing for the example.

and the sequencing specified in the release sequence table. Adding the corresponding cycle time to the planned release time derives the delivery schedule for the specified job order. To set a due date, a cycle time-based allowance for each job order is needed to absorb the impact of the disturbance. Because orders with higher ranking have the preferential right for processing, this will cause higher variance in cycle time for job orders with lower ranks. Lower priority job orders thus need a larger allowance to increase the ratio of on-time delivery and reduce the risk of lateness. Procedures for setting releasing and delivery schedule are as follows.

Step 1. Set the initial job lot numbering as $v = 1$ for the current planning period. Lot numbering will be assigned for lots with a release time already being scheduled.

Step 2. Trace the final release time in the previous planning period for every product type i with rank pri, $t^{pri}_{\mathrm{arr},i}$. Also, trace the final release serial number q in the previous planning period.

Step 3. According to the release serial number q, identify product type i^*, its priority pri^*, and the $\mathrm{seq}_q = (i^*, pri^*)$ information from the cyclic release table. Set its job lot number as v and record its release time, product type and rank:

$$\mathrm{Re}\, l_v = t^{pri}_{\mathrm{arr},i} + (I^{pri}_{\mathrm{out},i} \times n^{pri}), \text{ only for } i = i^* \text{ and } pri = pri^* \tag{28}$$

$$\text{Reset } t^{pri}_{\mathrm{arr},i} = \mathrm{Re}\, l_v, \text{ only for } i = i^* \text{ and } pri = pri^*. \tag{29}$$

Step 4. Estimate the completion time of job lot v, Comp_v, by summing up the release time and cycle time of the corresponding product type and rank, $\mathrm{CT}^{pri(v)}_{i(v)}$:

$$\mathrm{Comp}_v = \mathrm{Re}\, l_v + \mathrm{CT}^{pri(v)}_{i(v)}. \tag{30}$$

Step 5. Calculate due date of lot v, Due_v, by summing up the release time, cycle time of the corresponding product type and rank, and a predetermined fractional allowance of cycle time with fraction σ^{pri}:

$$\mathrm{Due}_v = \mathrm{Re}\, l_v + (1 + \sigma^{pri}) \times \mathrm{CT}^{pri(v)}_{i(v)}. \tag{31}$$

Step 6. If $n^{pri^*} > 1$, copy information of $\mathrm{Re}\, l_v$, $i(v)$ $pri(v)$, Comp_v, Due_v for job lot numbers v, $v+1, \ldots, v+n^{pri^*}-1$. If $\mathrm{Re}\, 2_v$ is greater than the closing time of the planning period, end the planning. Otherwise, set $v = v+1$ and $q = q+1$. In the case $q > q_{\max}$, reset $q = 1$. Go back to Step 3.

4. Verification with simulation

To verify the effect of this planning system, data from a wafer fabrication facility in Hsinchu, Taiwan, are used.

- Production information: there are five different product types, A–E, to be fabricated in this system. The products are categorized into two product families, logic I.C. and memory I.C. All product types have different process routes, and every product type has only one route. The required workstations and process times used in each route are known.
- Order priority information: there are three priority ranks in the system.
  - Hot: highest priority. The job orders are not limited by batching policy and they can be released into the shop floor and be loaded onto any batch machine with only a single lot.

  - Rush: secondary priority. Batching policy is defined according to the model proposed in this paper, and release policy is determined by the result of batching policy.
  - Normal: lowest priority. Full loading policy is required, and release batch is six lots.
- Workstation information: there are 83 different types of workstations (coded from W1 to 83), including serial and batch workstations. There are 15 six-lot workstations (W24–36), three four-lot workstations (W38–40) and 19 two-lot workstations (W07, W08, W13–15, W47, W48, W67–77, W81). The photo stepper, W46, is the bottleneck, and furnace, W24, is the CCR.
- Down time distribution of workstations: MTBF and MTTR are exponentially distributed and are traced for each workstation, while MTBPM and MTTPM are known constants.
- Machine set-up time: average set-up time is included in process time for most of workstation types.

4.1. *Assumptions for master production schedule planning*

- Product type and rank mix: the throughput target for each planning period is 640 lots. The product type mix for products A–E is 5, 7, 3, 4 and 1 respectively, i.e. π_A, π_B, π_C, π_D and π_E are 5/20, 7/20, 3/20, 4/20 and 1/20, respectively. The rank ratio $(\pi_i^h, \pi_i^r, \pi_i^n)$ for products A–E are (0, 2/5, 3/5), (0, 3/7, 4/7), (0, 1/3, 2/3), (0, 1/4, 3/4) and (1, 0, 0), respectively. Based on the above, the throughput target for each product type and rank is listed in table 3.
- Allowance for setting due date: assume that the due date allowances for hot, rush, and normal rank of orders are 0.1, 0.12 and 0.15 of the estimated cycle time.
- Dispatching rule: for jobs waiting before a resource, the one with a higher rank has the preferential right. For jobs with same priority rank, first-in first-out is applied.

4.2. *System design for simulation*

To verify results of the system, a simulation model is built by SiMPLE++ software developed by AESOP Co. To match the length of the MPS planning horizon, a simulation horizon is set to 168 days. The first 84 days are a warm-up period; hence, only results belonging to the next 84 days are collected. The simulation model is run 15 times to get adequate statistical results.

4.3. *Process and results of the MPS module*

4.3.1. *Capacity loading and batching policy for rush-rank orders*

Since the capacity loading for the furnace (W24), a CCR, is next to the bottleneck workstation (W46), its batching policy is taken seriously. When evaluating the capacity loading for rush orders, W46 and W24 are thus the main focus. Table 4 shows the evaluation results.

The above results show that when the minimum loading batch size for rush orders is less than five lots, there is a big possibility of a bottleneck shifting or wandering. When the minimum loading batch size is set to five or six lots, there is a significant difference in utilization rate between W46 and W24. Note that the maximum loading batch sizes for batch machines can be two, four or six lots. Considering the existence of a multiplication factor between different batch machines

	Total	A	B	C	D	E
Total	640	160	224	96	128	32
H	32	0	0	0	0	32
R	224	64	96	32	32	0
N	384	96	128	64	96	0

Table 3. Throughput target for each product type and rank in a planning period (unit: lot).

Minimum batch size for rush orders (lots)	1	2	3	4	5	6
Utilization of W46	0.86	0.86	0.86	0.86	0.86	0.86
Utilization of W24	1.70	1.13	0.95	0.85	0.79	0.76

Table 4. Result of the capacity loading under different batching policy.

will benefit the material flow, we set six lots as the batch policy for W24 and adopt a full batch policy for all kinds of batch machines. Furthermore, since the release batch size is better to match with the batch policy for critical resources, rush orders wafers are released in six continuous lots.

4.3.2. *Estimation of production cycle time by product type and rank*

The BBCT-MP algorithm (Chung *et al.* 2001) is applied to estimate the production cycle time. The difference in cycle times between the estimated digit and simulation result are shown in table 5. The accuracy of the cycle time estimation is above 96.25%, and the maximum difference is only 11.57 h. In general, the estimation method is good for predicting production cycle time.

4.3.3. *Estimation of WIP level*

After deriving the cycle time of each product type and rank, we estimate the system work in process level with Little's law. As shown in table 6, the estimated system WIP level is about 280 lots. To test and verify the theoretical WIP level properly, we set 280 lots as system WIP levels and ran the simulation model 15 times. With the constant WIP release control policy, the simulation system throughput was 639.42 lots, which is very near to our throughput target of 640 lots.

Product type_rank	A_R	A_N	B_R	B_N	C_R	C_N	D_R	D_N	E_H
(1) Estimated value	259.83	289.38	280.46	312.08	270.19	296.76	309.32	335.80	221.10
(2) Simulation result	264.21	297.63	274.66	304.57	271.58	308.33	312.28	345.88	224.97
Difference (3) = (1)–(12)	−4.38	−8.25	5.8	7.51	−1.39	−11.57	−2.96	−10.08	−3.87
Error rate (4) = (3)/(2)(%)	−1.66	−2.77	2.11	2.47	−0.51	−3.75	−0.95	−2.91	−1.72

Table 5. Comparison of production cycle time by product type and rank (unit: h).

	Total	A_R	A_N	B_R	B_N	C_R	C_N	D_R	D_N	E_H
Total	279.92	24.75	41.34	40.06	59.44	12.86	28.25	14.73	47.97	10.53

Table 6. Estimation of WIP level by product type and rank (unit: lot).

4.3.4. *Estimation of daily movement*

Using equations (15–17), the daily total movement for the bottleneck and the daily layer throughput are derived as shown in table 7.

4.3.5. *Building a cyclic release sequence*

Since the release batch size for each rank has been determined, we can build the cyclic wafer release table using equations (18–27) as shown in table 8.

4.3.6. *Release and delivery schedule*

Follow the procedures in Section 3.7, we can derive the release time, completion time and due date for each wafer lot. For space saving, only planned and simulation results of 160 wafer lots released during the 112th to 119th days (the fifth week after the warm-up period) are shown in figure 5. The analysis will be given below.

4.4. *Effect analysis of MPS planning*

4.4.1. *Analysis of the accuracy in releasing time prediction*

Figure 5 and table 9 show that the prediction of release time is quite close to the time obtained by simulation. Note that we adopt the constant WIP level policy for

Layer movement	Total	A_R	A_N	B_R	B_N	C_R	C_N	D_R	D_N	E_H
1	22.86	2.28	3.43	3.43	4.57	1.43	2.29	1.14	3.43	1.14
2	22.86	2.28	3.43	3.43	4.57	1.43	2.29	1.14	3.43	1.14
3	22.86	2.28	3.43	3.43	4.57	1.43	2.29	1.14	3.43	1.14
4	22.86	2.28	3.43	3.43	4.57	1.43	2.29	1.14	3.43	1.14
5	22.86	2.28	3.43	3.43	4.57	1.43	2.29	1.14	3.43	1.14
6	18.29	2.28	3.43	3.43	4.57			1.14	3.43	
7	22.86	2.28	3.43	3.43	4.57	1.43	2.29	1.14	3.43	1.14
8	22.86	2.28	3.43	3.43	4.57	1.43	2.29	1.14	3.43	1.14
9	22.86	2.28	3.43	3.43	4.57	1.43	2.29	1.14	3.43	1.14
10	19.43	2.28	3.43	3.43	4.57			1.14	3.43	1.14
11	19.43	2.28	3.43	3.43	4.57			1.14	3.43	1.14
12	19.43	2.28	3.43	3.43	4.57			1.14	3.43	1.14
13	22.86	2.28	3.43	3.43	4.57	1.43	2.29	1.14	3.43	1.14
14	22.86	2.28	3.43	3.43	4.57	1.43	2.29	1.14	3.43	1.14
15	22.86	2.28	3.43	3.43	4.57	1.43	2.29	1.14	3.43	1.14
16	17.14			3.43	4.57	1.43	2.29	1.14	3.43	1.14
17	17.14			3.43	4.57	1.43	2.29	1.14	3.43	1.14
18	22.86	2.28	3.43	3.43	4.57	1.43	2.29	1.14	3.43	1.14
19	9.14					1.43	2.29	1.14	3.43	1.14
20	22.86	2.28	3.43	3.43	4.57	1.43	2.29	1.14	3.43	1.14
Final	22.86	2.28	3.43	3.43	4.57	1.43	2.29	1.14	3.43	1.14
Total	440	41.14	61.72	68.57	91.43	19.43	38.86	24	72	22.86

Table 7. Estimation result of daily movement (unit: lot).

Serial no.	Product type	Priority	Batch release (lots)	Serial no.	Product type	Priority	Batch release (lots)
1	E	H	1	16	E	H	1
2	B	N	6	17	C	R	6
3	A	N	6	18	D	R	6
4	B	R	6	19	A	R	6
5	D	N	6	20	C	N	6
6	E	H	1	21	A	N	6
7	A	R	6	22	B	R	6
8	C	N	6	23	D	N	6
9	B	N	6	24	B	N	6
10	E	H	1	25	E	H	1
11	A	N	6				
12	B	R	6				
13	D	N	6				
14	E	H	1				
15	B	N	6				

Table 8. Cyclic wafer release sequence.

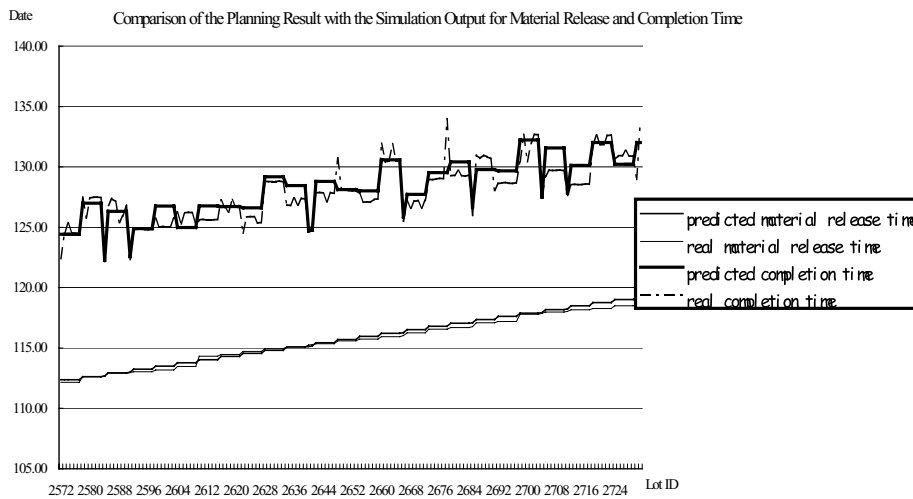

Figure 5. Comparison of the planning result with the simulation output for material release and completion time.

Absolute error in release prediction (day)	Number of job orders (lot)	Percentage to total job orders
0–0.2	60	37.50
0.2–0.4	80	50.00
0.4–0.6	18	11.25
> 0.6	2	1.25
Total	160	100

Table 9. Error analysis for release time estimation.

Type_rank	A_R	A_N	B_R	B_N	C_R	C_N	D_R	D_N	E_H
Throughput in simulation	64.20	95.80	96.07	127.13	32.00	64.80	32.33	96.47	32.27
Planned throughput	64	96	96	128	32	64	32	96	32
Difference	0.20	−0.20	0.07	−0.87	0.00	0.80	0.33	0.47	0.27
Error percentage	0.31	−0.21	0.07	−0.68	0.00	1.25	1.04	0.49	0.83

Table 10. Comparison of average throughput in simulation to the planned throughput of a planning period (unit: lot).

	A_R	A_N	B_R	B_N	C_R	C_N	D_R	D_N	E_H	Total
Average no. of tardy lots (lot)	4.00	4.67	0.87	1.27	2.07	2.60	1.07	4.87	0.47	21.87
Average throughput per period (lot)	64.20	95.80	96.07	127.13	32.00	64.80	32.33	96.47	32.27	641.07
Percentage of tardiness	6.23	4.87	0.90	1.00	6.46	4.01	3.30	5.04	1.45	3.41
Percentage of on-time delivery	93.77	95.13	99.10	99.00	93.54	95.99	96.70	94.96	98.55	96.59
Avg. Tardiness (day)	0.37	0.75	0.20	0.18	0.31	0.87	0.61	0.67	0.22	0.56

Table 11. Analysis of on-time delivery (unit: lot).

release time control. Only 1.25% of the total job orders have an error in release prediction of more than 0.6 days. This implies quite good control in throughput time and throughput quantity.

4.4.2. *Analysis of throughput target achievement*

Table 10 shows the results of throughput target achievement of the planning period (4 weeks). For every product type and rank, the average difference between the planned target and the simulation result is less than one lot. This means that the designed release schedule and fixed WIP level policy play a successful role in throughput mix and quantity control.

4.4.3. *Analysis of due date performance*

Table 11 shows that the percentage of on time delivery for each kind of product is quite high, ranging from 93.54 to 99%. The ratio of tardy lots is low, and the average tardiness is less than 1 day. Without using any due-date-oriented dispatching rule in the production activity control (PAC) phase, such a result reflects the value of the proposed MPS planning and releasing method.

5. Conclusion and future research

Multiple-priority orders and numerous product types are a universal phenomenon for current wafer fabrication. Quick response in due date setting and high satisfaction in on time delivery are critical issues for market competence. This paper has presented a planning system for multi-priority orders to set batch policy, estimate cycle time, determine a suitable system WIP level, set daily bottleneck operations, plan a release schedule and set due date for job orders. Without the need of simulation, this system responds quickly for any change in product type and

rank mix, or in machine units. The case study reveals the following effects of the proposed system.

- Accurate estimation of the production cycle time for each product and level is necessary. Since cycle time is the foundation for MPS in setting a proper system WIP level and in predicting the release schedule for each job order, the paper adopts a BBCT-MP module (Chung *et al.* 2001) for estimating cycle time and gets results near to those of simulation experiments.
- Release interval is determined by using the projected throughput rate and product mix. With a release interval, cycle time estimation and predetermined release batch size, we can determine the timing of each release batch. With release wafer lots based on the release time, the system can produce the right amount of products at the right time as predicted. The case study shows that the difference between expected and actual release time is quite small.

Consequently, this model has a relatively good result in processing stable product mix and multiple-priority orders. We can plan the MPS and detail schedule efficiently without the support of a simulation tool. The output of this planning system includes the information of production cycle time, proper WIP, daily move, batch policy, release cycle list, release time list, output time list, and due date setting of each job order. Such information is quite useful for making a decision in the production activity control system. Future research can be focussed on a production scheduling planning based on a fluctuating product mix or capacity use level. A target planning system that considers the issues such as financial measures, cycle time, throughput, market environment and competitiveness can also be researched on.

Acknowledgements

This work was partially supported by the National Science Council, Taiwan, ROC under Grant No. NSC 89-2622-E-009-002.

References

Burman, D. Y., Gurrola-Gal, F. J., Nozari, A., Sathaye, S. and Sitarik, J. P., 1986, Performance analysis techniques for IC manufacturing lines. *AT&T Technical Journal*, 46–56.

Chu, S. C. K., 1995, A mathematical programming approach towards optimised master production scheduling. *International Journal of Production Economics*, **38**, 269–270.

Chung, S. H. and Huang, H. W., 1999a, The block-based cycle time estimation algorithm for wafer fabrication factories. *International Journal of Industrial Engineering*, **6**, 307–316.

Chung, S. H. and Huang, H. W., 1999b, The design of production activity control policy for wafer fabrication factories. *Journal of the Chinese Institute of Industrial Engineers*, **16**, 93–113.

Chung, S. H., Pearn, W. L., Kang, H. Y., Chen, C. C. and Ke, W. T., 2001, Cycle time estimation model for wafer fabrication with multiple-priority orders. *Proceedings of the Advanced Simulation Technologies Conference*, Seattle, WA.

Chung, S. H., Yang, M. H. and Cheng, C. M., 1997, The design of due-date assignment model and the determination of flow time control parameter for the wafer fabrication factories. *IEEE Transaction on Components, Packaging, and Manufacturing Technology*, **20**, 278–287.

Conway, R., Maxwell, W. and Miller, L. W., 1967, *Theory of Scheduling* (Reading: Addison-Wesley).

Glassey, C. R. and Weng, W. W., 1991, Dynamic batching heuristics for simultaneous processing. *IEEE Transactions on Semiconductor Manufacturing*, **4**, 77–82.

HACKMAN, S. T. and LEACHMAN, R. C., 1989, A general framework for modeling production. *Management Science*, **35**(4), 478–496.

HUNG, Y. F. and LEACHMAN, R. C., 1996, A production planning methodology for semiconductor manufacturing based on interactive simulation and linear programming calculations. *IEEE Transaction on Semiconductor Manufacturing*, **9**, 257–269.

KRAMER, S. S., 1989, Total cycle time management by operational elements. *International Semiconductor Manufacturing Science Symposium*, pp. 17–20.

LEACHMAN, R. C. and HODGES D. A., 1996, Benchmarking semiconductor manufacturing. *IEEE Transaction on Semiconductor Manufacturing*, **9**, 158–169.

LITTLE, J. D. C., 1961, A proof for the queueing formula $L = \lambda W$. *Operations Research*, **9**, 383–387.

LIU, C., THONGMEE, S. and HEPBURN, R., 1995, A methodology for improving on-time delivery and load levelling starts. *Proceedings of the IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, pp. 95–100.

MILLER, D. J., 1990, Simulation of a semiconductor manufacturing line. *Communication of the ACM*, **33**, 98–108.

NEUTS, M. E., 1975, A general class of bulk queues with poisson input. *Annuals of Mathematical Statistics*, **21**, 777–782.

RAOS, S. S., 1992, The relationship of work-in-process inventories, manufacturing lead times and waiting line analysis. *International Journal of Production Economics*, **26**, 221–227.

THOMPSON, S. D. and DAVIS, W. J., 1990, An integrated approach for modeling uncertainty in aggregate production planning. *IEEE Transactions on Systems, Man, and Cybernetics*, **20**, 1000–1012.

WINSTON, W. L., 1991, *Operations Research: Applications and Algorithm* (Belmont: Duxbury).