Proceedings of the 1993 Winter Simulation Conference
G.W. Evans, M. Mollaghasemi, E.C. Russell, W.E. Biles (eds.)

PRECISE AND FLEXIBLE MODELING FOR SEMICONDUCTOR WAFER FABRICATION

Shinji Nakamura, Chisato Hashimoto

NTT LSI Laboratories
3-1 Morinosato Wakamiya, Atsugi-shi,
Kanagawa, 243-01 Japan

Osamu Mori

Information System Laboratories
1-2356 Take, Yokosuka-shi, Kanagawa,
238-03 Japan

ABSTRACT

For efficient planning and management of a semiconductor manufacturing line, such line performances as turnaround time and throughput have to be evaluated precisely. We have developed a simulation system, "SEMALIS", to handle complicated lot processing and equipment failures. This paper describes this line model, evaluates and analyzes its performance when using special lot processing, such as continuous processing, time-critical express lot processing, and some line operations for efficient lot processing.

1. INTRODUCTION

A semiconductor manufacturing line consists of several hundred kinds of processing equipment and processing steps. These can be very large and complicated depending on the production targets. For minimizing the investment in resources, optimizing process equipment and improving line performance by making line operations more efficient are of great concern.

In ASIC manufacturing lines, it is especially important to shorten turnaround time (TAT) so that line management can respond so quickly and flexibly to changes in marketing demand and to reduced productivity caused by such problems as equipment failures.

Simulation is the best method for evaluating and analyzing the complicated line performance [1]-[4], but there is a question of whether line model can be fit to actual line operations. We have developed a discrete-event simulation system, referred to as "SEMALIS" (SEmiconductor MAnufacturing LIne Simulator), that facilitates efficient planning and management of manufacturing lines. This system accurately reflects many details of actual lot processing and line operations by the use of precise and flexible line modeling.

This paper describes this line model, special lot processing applied to specific equipment and continuous process steps, and some model experimentation results for evaluation and analysis of its performance .

2. MODEL CONFIGURATION AND DEFINITION

The line model consists of four main blocks (Fig1.): lot release block (LRB), lot schedule block (LSB), lot processing block(LPB), definition and status information management block (DSMB).

The simulation model is written in the general-purpose simulation language SLAM-II and FORTRAN.

2.1 Lot release block (LRB)

The LRB releases wafer manufacturing lots according to a uniform release rate specified for each product. Attributes such as lot identities, product and process sequence code, lot-size (the number of wafers), and priority are individually assigned to each released lot . Lot release stops when the accumlated number of released lots exceeds the specified maximum number of released or work-in-process (WIP) lots.

2.2 Lot scheduling block (LSB)

The LSB increases the current WIP process step number after each step is completed. It assigns the equipment for the next process step by referring to the process sequence and puts it into the

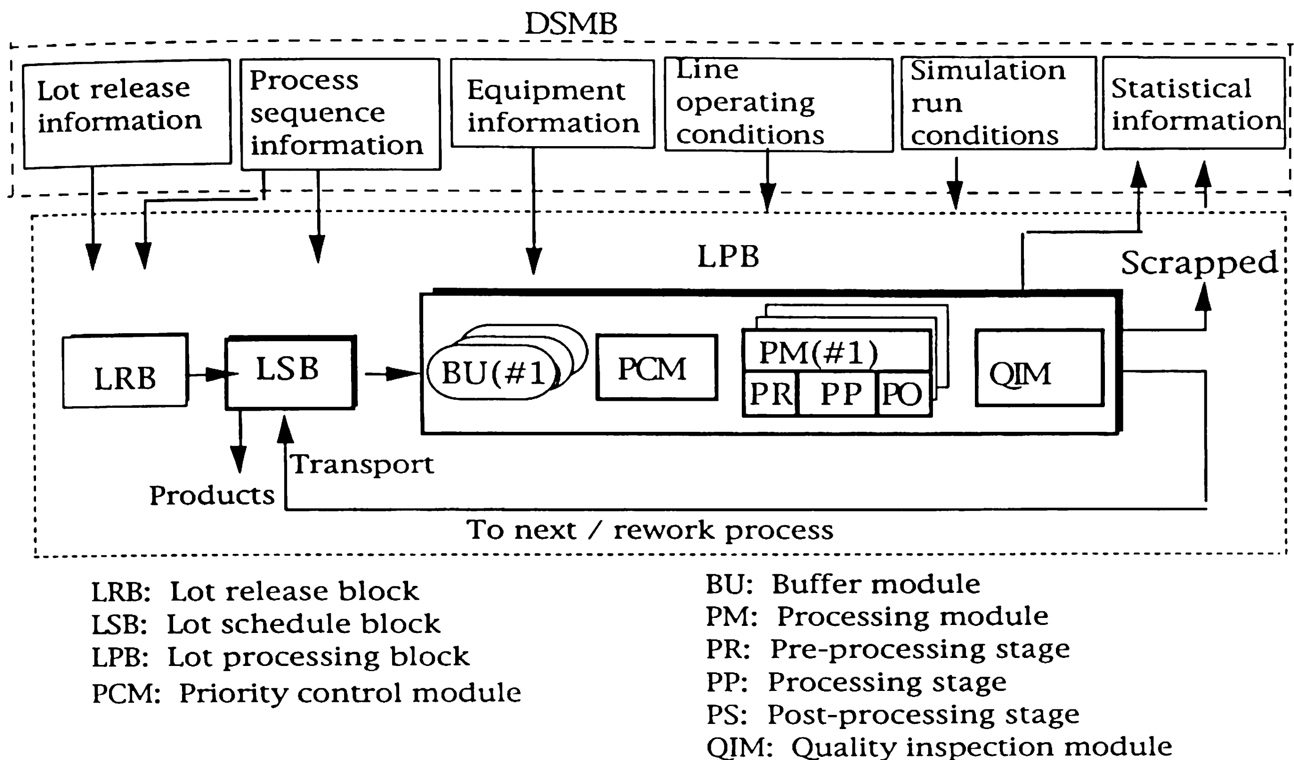


Figure 1: Configuration and definition of line model

buffer in order of the assigned priority. The assigned equipment has the shortest waiting time of all available equipment that is defined in the process step instructions. The lot waiting time is a simple estimate of the total processing time of all the waiting lots with the same or higher priority. If the number of the lot arrival current processing step is the specified step number, such as the terminating step of a process sequence, the end of the specific processing steps, e.g., forming a MOSFET on the silicon substrate (these processing steps are referred to as the substrate process sequence), or the end of the wiring process sequence (the ramaining processing steps in the substrate process sequence), then the manufacturing statistics are collected respectively.

2.3 Lot processing block (LPB)

The LPB consists of four parts. The buffer (BU), the lot priority control module (PCM), the processing module (PM), and the lot quality inspection module (QIM). The BU is the virtual storage for lots received from and prioritized by LSB. The PCM selects the next lot for processing out of the lots waiting in the buffer by referring to both the expected completion time based on equipment availability and lot priorities. Lot selection conditions are classified into five categories: specific products (C1), specifications for continuous processing steps (C2), required processing equipment (C3), processing step number (C4), and arrival time (first in first out) (C5).

A lot processing priority is assigned for each of these five conditions, with priority I being the hightest, and priority V the lowest. An example of priority assignation is shown in Figure 2. At the start of processing, those lots meeting the priority I condition are selected. If more than one lot are selected, selection process continues until finally only one lot can be selected based on C5 (FIFO). For selection condition C3, the lots that use the same recipe of the equipment as that specified are given top priority. The PM has three processing stages :

(i) During pre-processing, lots are formed into wafer batches or multiple lots to conform to the processing stage load-size defined in the recipe instruction.

(ii) During processing, each batch is processed by using single or multi-servers.

(iii) During post-processing, the batches and multiple lots are restored to their original lots.

Conditions for lot selection	Priority high ←				→ low
	I	II	III	IV	V
Specific product (C1)	○				
Process specification for continuous processing steps (C2)		○			
Required equipment lot processing (C3)			○		
Processing step number order (C4)					○
Arrival time order (C5)				○	

Figure 2: Example of priority assigned to the conditions for lot selection

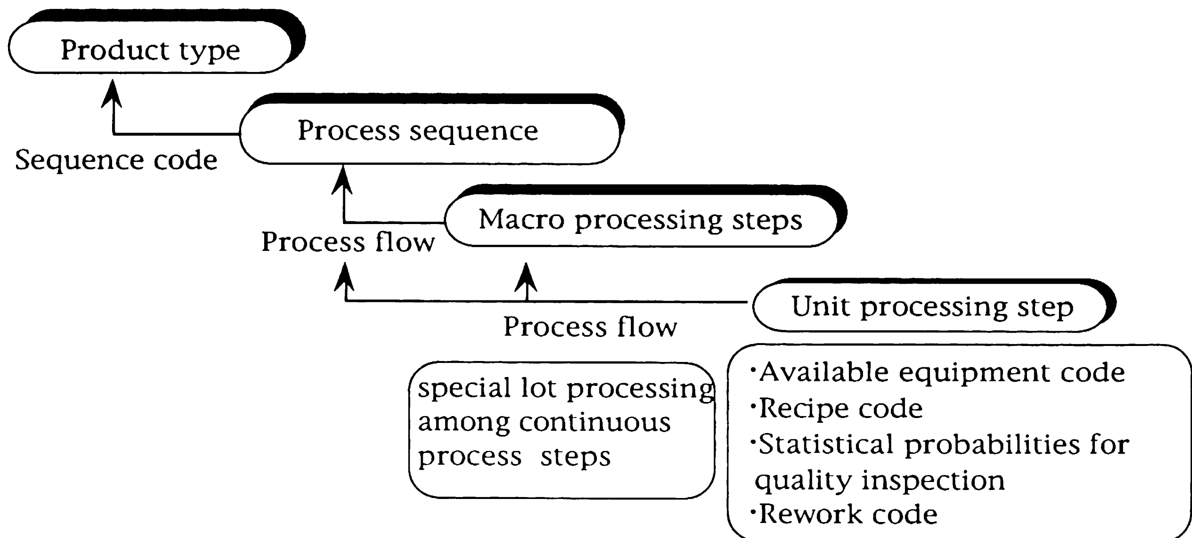


Figure 3: Process sequence definition with layered structure

The processing time for each processing stage is defined by the recipe data .

The QIM determines the next step for each lot: move to the next processing step, return for reworking, or be scrapped due to defects. This is done through quality inspection of each completed lot using the statistical probabilities assigned to the three alternatives in process step definition.

Lots progressing to the next process step are moved to the buffer for the next step by the LSB. For rework processing, a rework code is added to the lot attributes and the lot is moved to the rework processing steps defined in the macro process sequence described in section 2.4. During the rework, none of the lot wafers may be scrapped or cause another rework to be initiated. Upon rework completion, the lot is returned to the initiated processing step.

For scrapping, the scrap statistics for the entire lot are collected and the lot is terminated.

2.4 Model definition and status information management block (DSMB)

Five line model definition information classes are managed in this block. They are product, process, equipment, line

operating conditions, and simulation run conditions. The main definition items of the line model are shown in Table 1. The process sequence, which depends on each product are flexibly defined with a layered structure containing both unit and macro process steps. (Fig. 3) A macro process is defined as a processing group corresponding to certain process steps, such as the n or p channel forming process steps of a MOSFET, and contains some unit process steps. Unit process step is the minimum step in the process sequence. Special lot processing is assigned to the macro process steps and equipment definition described in section 3.

The time for the unscheduled maintenance that is carried out after specific number of wafers are processed, such as cleaning a sputtering chamber, can be assigned in addition to the time for regular maintenance.

Equipment reliability is defined as the MTBF (mean time between failures), the MTTR(mean time to repair), and the wafer handling in the failed equipment. Equipment fails based on the probability exponential distribution.

Transient behavior of line performance is evaluated dynamically by changing one of the line operating conditions. Statistical informations, such as turnaround time (TAT), throughput, equipment utilization and available time, lot waiting time and the number of lots waiting at each equipment, and the number of WIP lots in line, are collected at specified time interval .

2.5 Model limitations

The operators for equipment and resources for material handling are excluded as subjects of the modeling. This is because the operation to assign them to equipment or WIP lots with precise scheduling is so complicated that the long-range simulation time is too long.

3. SPECIAL LOT PROCESSING

Two main types of lot processing are simulated. The first is processing that depends on the load-size specifications of and the processing capacity of each equipment. The second is lot processing that depends on the specifications of continuous processing steps.

Type I

(a) Multiple lot processing: Each batch is equal to or less than the processing stage load-size.

(b) Pipeline processing among multi-servers: the maximum number of servers is five. The input interval time must be more than the maximum processing time of all servers .

(c) Setup match required processing: a maximum of recipe classes is three. The setup time is to change over another recipe class, e.g., gas purge time in ion implanter.

Type II

(a) Continuous processing: a second processing step to be performed within a critical time interval after completion of a first processing step.

(b) Trial processing: must be done just before the main lot processing to determine the detailed recipe for the same process steps as the main lot processing.

(c) Split lot processing: a lot may be subdivided so that some parts can be processed within the remaining available time for the equipment on that day.

(d)Pipeline processing between continuous processing steps: the load-size of two continuous equipment must be the same. Post-processing of the first and pre-processing of the second may be omitted.

(e) Time-critical express lot processing: express lots can seize equipment with the highest priority and be processed with raw processing time (RPT). In this processing, no failures or maintenance are assigned to any equipment.

Continuous processing and time-critical express lot processing are discussed here in detail.

3.1 Continuous processing

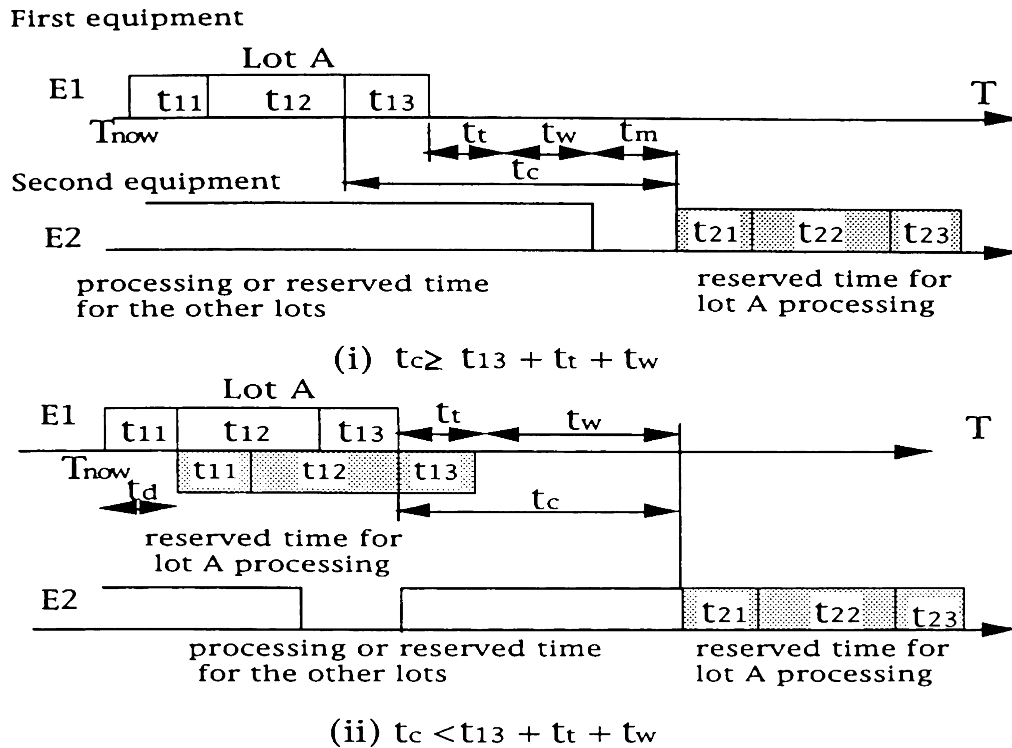
Continuous processing means time-constrained processing along continuous process steps. In this processing, the pre-processing of a lot on the second equipment has to be carried out within the previously specified time interval after completion of the processing stage of the first equipment.

If this fails, the lot is reworked or discarded. Therefore, the equipment for the second process step must be reserved prior to starting the first step.

The processing time reservation method

Table 1: Main line model definition items

Definition classes		Definition items
Product		Product code Process sequence code
Process sequence		Sequence code Macro processing steps Unit processing step (available equipment and the recipe code, statistical probability for next process, rework, and scrap, transport time, etc.)
Equipment		Equipment code Maximum number Recipe data (processing time and load-size of each processing stage, recipe change over time, etc.)
Line operations	Equipment operating conditions	Available time Maintainance conditions (scheduled and unscheduled maintainance time) Reliability conditions (MTBF, MTTR, and wafer handling for the failed equipment)
	Release lot conditions	Release rate Lot attributes (product code, number of wafers and priority in substrate and wiring process steps) priority for lot selection Time-critical express lot release time
	Dynamic modefication conditions	Release rate coefficient Additional product code Equipment code for increase or reduction End time of available equipment MTBF for specific equipment
Simulation run conditions		Simulation period Warm-up period Time interval for collecting of statistical information The maximum number of WIP and released lots



t_{11}-t_{13}, t_{21}-t_{23} : pre-processing, processing, and post-processing time of the equipment
t_t : lot transport time
t_w: lot waiting time until start of pre-processing on E2
t_c: constrained time interval between two continuous processing steps
t_d: delay time until start of pre-processing on E1 to satisfy eq.(1)
t_m: time of appropriation for processing or reservation for the other lots

Figure 4: Processing time reservations in continuous processing

for continuous processing is shown in Figure 4. In this model, the following relationship holds:

$$t_c \geq t_{13} + t_t + t_w \quad \text{----------------(1)}$$

where tc is the constrained time interval between the continuous processing steps , t13 is the processing time of lot A that is assigned to the first equipment E1 post-processing stage, tt is the transport time from E1 to the second equipment E2, and tw is the waiting time in the E2 buffer, equal to the sum of the processing time for the lots with a higher priority than that for lot A. In order to ensure proper continuous processing, the second equipment must be reserved so that the time relationship in equation (1) is guaranteed. The reserved processing start times for E1 and E2 (TE1 and TE2) are determined by the following equations.

$$TE1 = T_{now} \quad \text{-------------------(2)}$$

$$TE2 = T_{now} + t_{11} + t_{12} + t_c - t_a \quad \text{----(3)}$$

or

$$TE1 = T_{now} + t_a \quad \text{-----------------(4)}$$

$$TE2 = T_{now} + t_{11} + t_{12} + t_c + t_a \quad \text{----(5)}$$

where T_{now} is the present time, and t_a ($\geq$ 0) is the adjusting time for assuring that the reservation time for both equipments do not overlap the previously reserved times for the other lots, the regular maintenance time, and the end of service time on that day. Equations (2) and (3) show that only start time TE2 can be advanced by t_a ($\leq t_m$) and not overlap the other reservations and equations (4) and (5) show that both start times are delayed by t_a (=t_d) (here, the meaning of t_m, t_d is shown in Fig.4). The reservations for the continuous processing are canceled after the start of pre-processing on the second equipment.

3.2 Time-critical express lot processing

A time-critical express lot is one that

requires processing in the shortest possible TAT. As soon as the release time into the line has been determined, the processing start and end times for those equipments corresponding to a series of processing steps (e.g., N steps) are scheduled by using processing time ($t_{i1}+t_{i2}+t_{i3}$: i=1, n) and transport time (tt).

To ensure that time-critical express lot processing is performed, it is necessary to reserve the processing time. When pre-processing on the equipment at the i-th step is started, the reservations at the step for the express lot processing are canceled. After completion of post-processing, the equipment corresponding to the step number (i+N) is reserved. The reserved equipment may process other lots that can be completed before the start time of the express lots.

4. SIMULATION RESULTS

Figure 5 shows the changes in TAT and throughput when the maximum number of WIP lots is fixed and the available line time is changed from 9 hours/day operation to 24 hours/day operation, all without equipment failures and maintenance. It demonstrates that an appropriate number of WIP lots and available line time for achieving the required TAT or throughput of each product are obtained.

Figure 6 shows the effect of constrained time intervals in continuous processing on line performance. Continuous processing is assigned to all pairs of exposure and development processing steps in photo lithography, these steps sum to about 16% of all processing steps that are over three hundred. In this case, trial processing is also assigned to a portion of the above pairs. A maximum 10% reduction in TAT and throughput is observed within about sixty minutes, corresponding to the exposure processing time for a lot. This is because the utilization of steppers (first set of equipment) is limited, causing a bottleneck in the processing capacity of the developers (second set).

Figure 7 shows the effect of the number of WIP lots on line performance when continuous processing is applied to the process steps above. With less than eighty WIP, applying continuous processing improves line performance. It

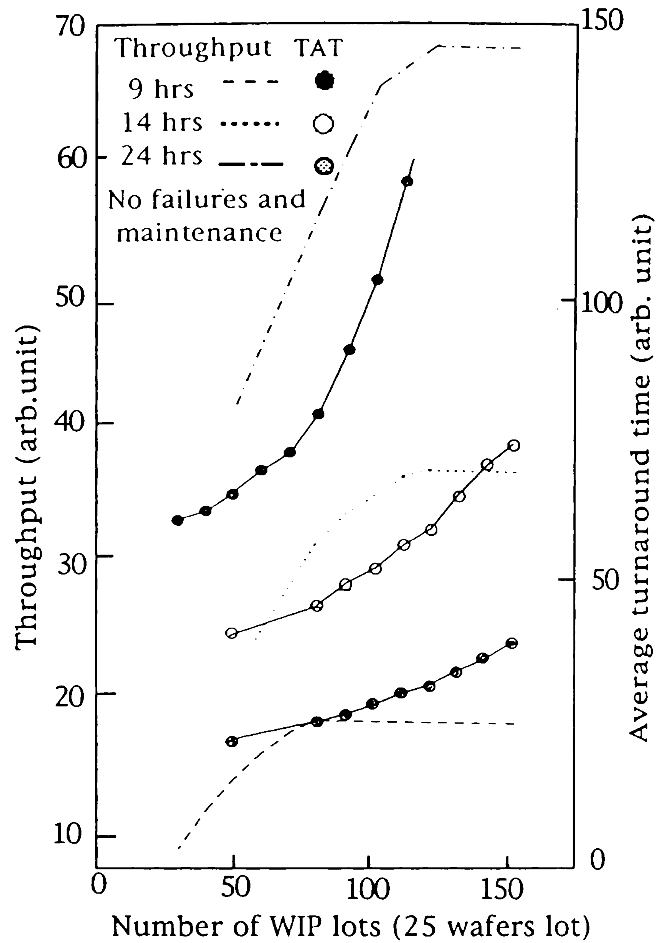


Figure 5: Effect of the number of WIP on line performance

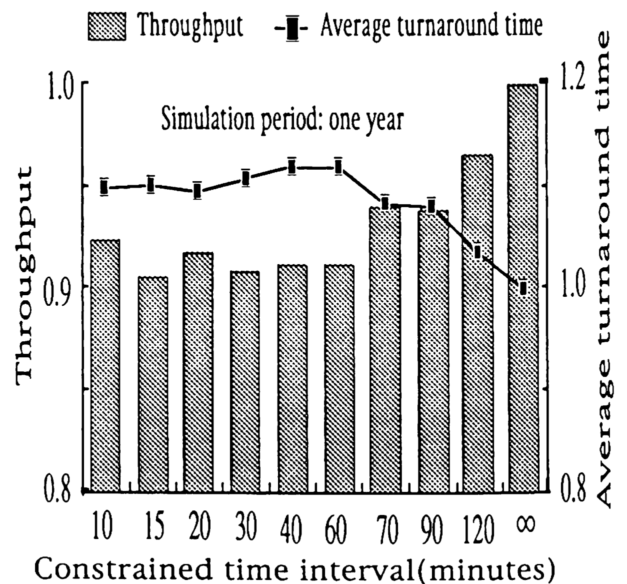


Figure 6: Effect of constrained time interval in continuous processing on line performance (normalized by the results for no continuous processing)

is assumed that as the number of the lots waiting in the stepper buffers increases, steppers utilization also increases, accelerating lot processing at each equipment after the exposure processing step since there are fewer waiting lots in their buffers. With more than eighty WIP, the utilization of bottlenecked equipment without continuous processing is more than that with continuous processing (maybe steppers) and as the WIP lots are dispersed to the various equipments, lot waiting time increases.

Figure 8 shows the effectiveness of various line operation methods for improving line performance when using continuous processing. Both TAT and throughput are improved about 5% by extending the available time of the second set of equipment (developers) and using split lot processing. Figure 8 also shows that pipeline processing and split lot processing are effective for reducing TAT and increasing throughput, respectively.

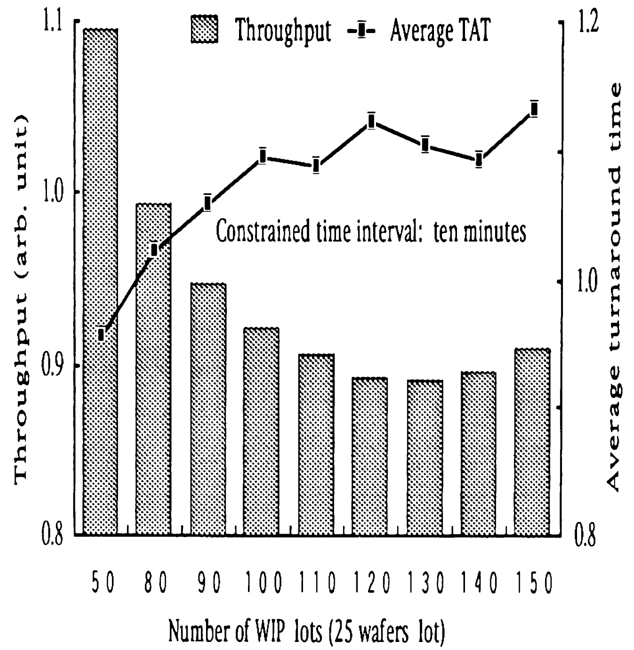


Figure 7: Effect of the number of WIP lots on line performance

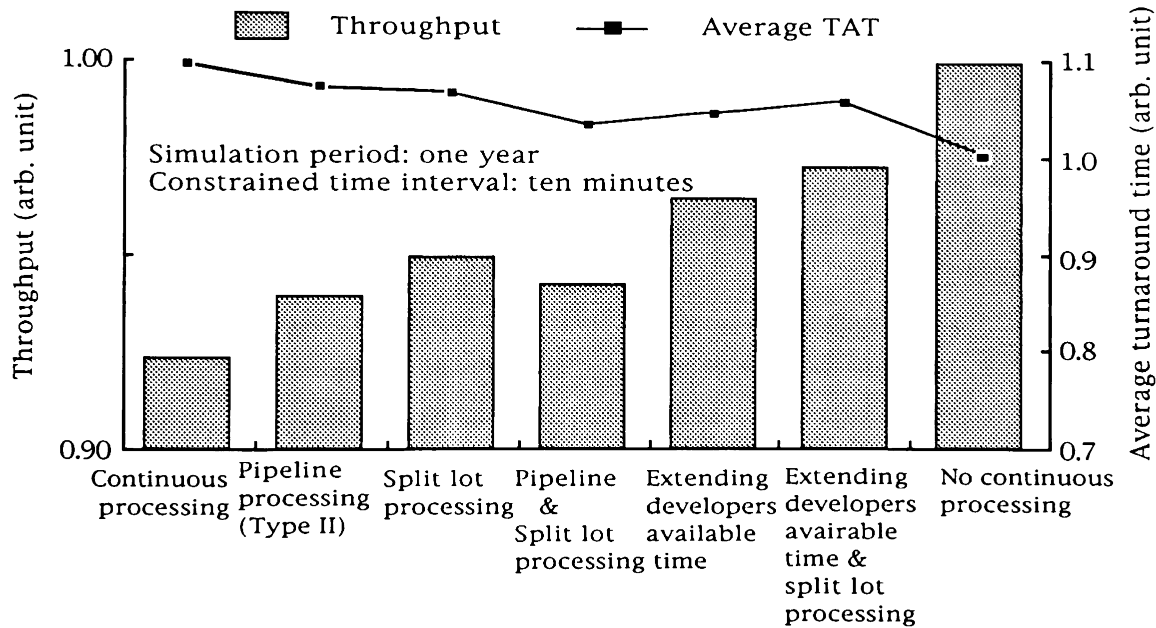


Figure 8: Line performance under different line operation methods (normalized by the results for no continuous processing)

Figure 9 shows an example of the time sequential change in line performance when only one time-critical express lot is released into a steady-state manufacturing line in which the number of WIP lots is constant. At thirty five and forty days after release, regular lot throughput decreases and then increases dramatically. This is because these regular lots were in the latter part of the substrate process sequence at the release time of the express lot and the express lot caught up with them during the early part of the wiring process sequence. Regular lot processing was thus interrupted until the express lot passed through and the lots were then produced together in a short term after the

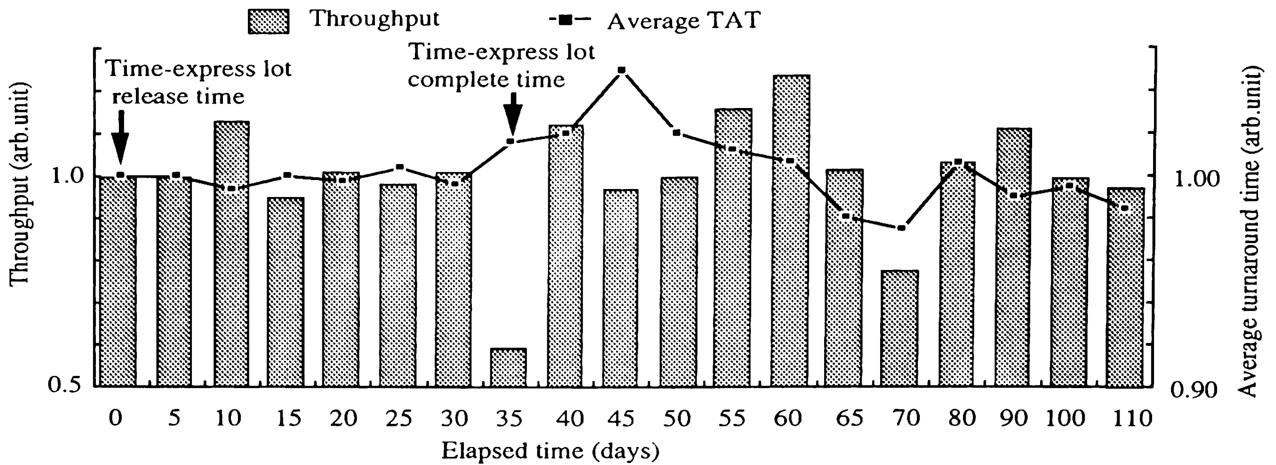


Figure 9: Effect of time-critical express lot release on line performance (normalized by the results for no time-critical express lot)

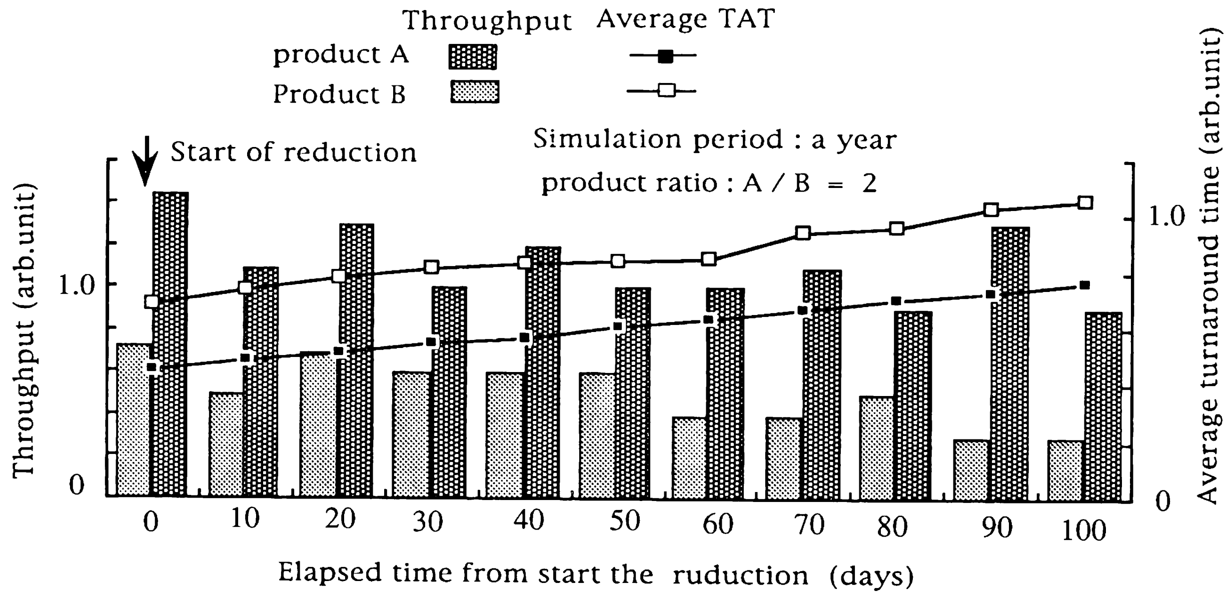


Figure 10: Influence of dynamic modification of model

completion of the express lot. Production of regular lots released at the same time as the express one also decreased due to the express interruption and other preceding regular lots. A maximum of 10% and 45% fluctuation in TAT and regular lot throughput was observed, respectively, for a while after the express lot completion. This is mainly because express lot processing interrupts regular lot processing, resulting in equipment bottlenecks or starvation dynamically.

Figure 10 shows transient behavior of line performance due to dynamic modification of line operating conditions on the way of simulation.

In this case, the maximum number of WIP lots is not limited. The number of steppers is reduced from four to three at a certain time in a steady state product mix manufacturing line. After modification, throughput for each product decreases gradually with periodic fluctuation for a while and TATs for both products increase uniformly along with the increase in elapsed time.

5. CONCLUSION

We proposed a precise and flexible simulation model which enables complicated lot processing and equipment failures in wafer fabrication to be handled accurately. The newly developed

simulator "SEMALIS" based on these proposals was verified as a useful tool for providing information on facility planning and the design of highly efficient line operating methods in ASIC manufacturing lines.

ACKNOWLEDGEMENTS

The authors would like to especially thank Dr. Atsusi Iwata, Dr. Eisuke Arai, Mr. Akira Shindo, and Mr. Junro Nose for their many helpful suggestions and guidance during the course of this work. We would also like to thank Dr. Tetsuma Sakurai for his many useful suggestions from the view point of a line manager and user.

REFERENCES

[1] D.J. Miller, Simulation of a semiconductor manufacturing line, Communications of the ACM, Vol. 33, No. 10, Oct.1990, pp. 99-108.

[2]B. Tullis et al., Successful modeling of a semiconductor R&D facility, ISMSS '90 Proc.

[3] M. Enomoto et al., Line productivity improvement using simulation system for VLSI manufacturing, ECS Proc., pp. 603, 1991.

[4] K.A. Pitts, Discrete-event simulation of wafer fabrication facility, Proc. of WSC, pp. 712, 1988.

AUTHOR BIOGRAPHIES

SHINJI NAKAMURA is a Senior Research Engineer at the Manufacturing Systems Tecnology Laboratory of NTT LSI Laboratories. He received the B.E and M.E degrees in precision engineering from Yamanashi National University, Japan, in 1973 and 1975, respectively. His research interests are focused on factory automation (FA), modeling and simulation for line analysis and evaluation of flexible manufacturing system (FMS).

CHISATO HASHIMOTO is a Senior Research Engineer at the Manufacturing Systems Tecnology Laboratory of NTT LSI Laboratories. He received the B.E and M.E degrees in electrical engineering from Yokohama National University, Japan, in 1972 and 1974, respectively. His research interests are focused on LSI manufacturing technology, particulary improving productivity through fabrication line optimization and process development. He is currently engaged in LSI yield analysis.

OSAMU MORI is a Senior Research Engineer at the Base Systems Architecture Laboratory of NTT Network Information Systems Laboratories. He received the B.E degree in electric communication engineering from Musashi Institute of Technology, Japan, in 1968. His research interests are focused on system modeling, evaluation, diagnostics, and CASE.