Dynamic Release Control Policy for the Semiconductor Wafer Fabrication Lines

ILHO LIM and JONGSOO KIM

Department of Industrial Engineering
Hanyang University

ABSTRACT

We propose a policy for controlling the release of raw wafers into the semiconductor wafer fabrication lines. The proposed policy exploits up-to-date factory floor information gathered by tracking systems used to calculate the time and amount of a new release to minimize mean flow times and mean tardiness while maintaining the maximum output rates of the system.

Extensive computer experiments show that the proposed policy results in significant improvements for the same output rates compared to existing release rules.

INTRODUCTION

An Integrated Circuit(IC), commonly referred to as a semiconductor chip, is a complex device that consists of miniaturized electronic components and their connections. The population of IC's is accomplished in a four-stage process, i.e., wafer fabrication(fab), wafer probe, assembly and test. Wafer fabrication is the most technologically complex and capital intensive of the four stages. It involves the creation of layers of circuitry on the wafer through a long sequence of processing steps involving many separate pieces of equipment. The lots of wafers are routed through the processing steps in the traditional job shop fashion.

Typically, cycle through the processing sequences include hundreds of steps. An important unique feature of these sequence is the re-entrant nature of the sequence. Because each layers of the wafer requires the exposure of a a photoresistive material through a mask(the process is known as photolithography), each batch of wafers makes many visits to the lithography workstation, interpersed with visits to other workstations. Machines in the wafer fab, such as the lithography machine, are highly specialized and very expensive. Nevertheless they have considerable downtimes including time to perform unscheduled repairs, scheduled maintenance, engineering tests, etc. Downtime for many types of machines is on the order of 50 percent.

Since the capital investment and sales revenue are extremely large, implementation of an improved scheduling policy results in a considerable amount of increased profits. But the scheduling of a wafer fab is challenging due to the long cycle time, ever-changing flux of products, re-entrant feature of the production sequence, and stochastic aspects of the wafer fab including machine failures. There are basically two types of scheduling decisions in the wafer fabrication stage. The first and most familiar one is dispatching. Each time a work center is ready to commence processing on another order and there is a queue of work-in-process(WIP) waiting to be processed at the work center, the dispatching decision selects the order or orders to start processing. The second types of shceduling decision, which is referred to as release control decision, is to decide the type, amount, and time point of release so as to maximize efficiency of the wafer fab. The simplest release policy is to release wafers without restraint on the acceptance and release of new work, i.e., any new work order that is received is immediately released to the wafer fab, and control of the wafer fab is reduced entirely to dispatching decisions. While such a policy would provide the opportunity for maximum capacity utilization, it has been shown that this policy leads to excessive amounts of WIP in the factory and long average flow times through the factory. [2,3,5] There are a number of negative consequences of long flow times particularly in the wafer fab. The most serious of these consequences is exposed to particles in the clean room and results in low production yield. In addition, a major crash of a machine can only be detected after a wafer that had been processed by the crashed machine completes all the fabrication steps and enters the following test phase. Thus in case of the major crash, longer flow times result in larger amount of completed but unacceptable wafers. Since the average flow times of wafers for the major products usually exceeds two months, it is not uncommon for the discarded wafers to

cost millions of dollars. Without technological change, reduction in the amount of work present in the factory. The trade-off is that if the workload is reduced sufficiently, the output of the factory will decrease. Thus, a release control is necessary to gate new work into the factory to maintain maximum output without creating any necessary work to maintain that output.

A number of papers have been published to introduce release control policies for semiconductor wafer fabrication. Glassey and Resende[2] proposed a continuous-review policy termed the 'starvation avoidance' policy for a single product, single bottleneck case. Solorzano[6] and Leachman et al.[4] proposed a release policy named 'queue management policy'. The policy measures the total workload on a work center to some finite time horizon and compares this load to a target load to determine if release is required to prevent loss of capacity. Wein[7,8], Chevalier and Wein[1] solved the problem of input control and and priority sequencing in a two-station and a multistation queueing metworks by approximating the problem as a Brownian control problem. The solution is interpreted in terms of the original queueing system to obtain an effective scheduling policy termed 'workload regulating input policy'. This policy was successfully applied to fictitious semiconductor wafer fabrication lines with from one to four bottleneck stations.

The main weakness of the above stated policies is that they are designed to use static information rather than real-time dynamic shop information. Tracking systems installed in most of the semiconductor fabrication shops as a part of the computer-integrated manufacturing systems generate global factory state information such as the current location of WIP, status of machines, and average flow times. This expanded factory information raises a question : How should the global and dynamic information be summarized and used for a release control to achieve the maximum improvement compared to decisions based on local and static information. We propose an efficient release control policy as an extension of the work by Leachman et al.[4] The policy is designed to minimize mean flow times and mean tardiness while maintaining the maximum achievable output rates of a wafer fab. The main contribution is that the proposed policy exploits dynamic factory floor information and demonstrates improved efficiency over the previously introduced policies. In this article, we introduce the proposed policy first and then present the results of our computational experiments. We conclude with some remarks and discussions.

DESCRIPTION OF THE PROPOSED HEURISTIC

In this section, we define notations and describe the key ideas of the proposed policy.

Indices

i = product type, $i = 1,...,I$

l = production step (operation)

(i,l) = lth step in a route for product type i

n = lot number, $n = 1,...,N$

k = work center, $k = 1,...,K$

t = time period

Parameters

$S(i)$ = the number of steps to complete product i

$w_{i,l}$ = units of WIP initially resident at step (i,l)

$d_{i,l}$ = standard flow time for step (i,l)

$a_{i,l,k}$ = standard loads (in unit times) on work center k to process one unit of wafer undergoing operation (i,l)

$C_k(t)$ = projected capacity (in unit times) of work center k in period t

$L_{i,k}$ = estimated time for product i to reach work center k for the first time

Ω_i = set of indices of work centers to be visited by product i

M_i = the smallest integer greater than the time for product i to reach each work center for the first time, i.e., $M_i = [\ \max_{k \in \omega_i} (L_{i,k})\ ]$

$D(n)$ = due date of lot number n

$p(n)$ = product type of lot number n

$v(n)$ = number of blank wafers in lot number n

R = review horizon; the length of time forecasting future workloads at each time period

τ = current time period

Variables

$X_k(t)$ = projected *cumulative* workload from WIP arriving at work center k by end of period t

$x_k(t)$ = projected workload from WIP arriving at work center k during period t

$Y_{i,k}(t)$ = projected *cumulative* workload from new workload from newly released work order i arriving at work center k by end of period t

$y_{i,k}(t)$ = projected workload from newly released work order i arriving at work center k during t

$q_k(t)$ = projected queue of work at work center k at the end of period t

The policy can be divided into three major parts : calculation of projected workloads, calculation of projected queues, and release decision parts.

Calculation of projected workloads

At the starting time point, future arrivals of WIP at each work center during the review horizon, R, are projected and converted into hours of workload. The review horizon is determined so that it is longer than the maximum of the throughput times of products. The current location of WIP and the previous data of flow times that are available from the tracking system are prepared. To obtain the amount and the time distribution of future arrivals of WIP at each work center, *cumulative* workloads associated with future arrivals of current WIP's are estimated and summarized by work center in discrete time periods in Equation (1). Then, as shown in Equation (2), the *cumulative* workload is differenced to obtain projected workloads at each time period.

$$X_k(t) = \sum_{i=1}^{I} \sum_{l=1}^{S(i)} \{ a_{i,l,k}\, w_{i,l} + \sum_{\Lambda} a_{i,\Lambda,k}\, w_{i,l} \},$$

$$\text{where } \Lambda = \left\{ l_0 \mid \sum_{\lambda=l+1}^{l_0} d_{i,\lambda} \le t \right\}, \quad k = 1, \ldots K, \quad t = \tau, \ldots, \tau + R \qquad (1)$$

$$x_k(t) = X_k(t) - X_k(t-1), \quad k = 1, \ldots, K, \quad t = \tau, \ldots, \tau + R \qquad (2)$$

Refer to Figures 1 and 2 for an illustration of the cumulative workloads and the projected workloads for a fictitious work center with $C_k(t) = 1$.

Calculation of projected queues

Projected workloads are then used to calculate the future(projected) inventory of uncompleted work(i.e., the queue size) at the end of each time period using standard discrete time balance equations in (3)

$$q_k(t) \;=\; Max\ \{\ 0, q_k(t-1) + x_k(t) - C_k(t)\ \}, \quad k = 1, \ldots, K, \quad t = \tau, \ldots, \tau + R \qquad (3)$$

Figure 3 shows the peojected queues for the fictitious work center described in Figures 1 and 2.

Release decision

The decision to release order is based upon examination of the projected queue sizes at each work center over the review horizon. Here we assume that new wafers to be released are devided into a lot of 24 or 48 blank wafers, which usually corresponds to an order of the same product type with the same due date.

The release decision starts with calculation of a priority value considering due date for each lot of raw wafers waiting to be released. $\Psi(n)$, the priority value of the lot number n, is determined using Equation (4).

$$\Psi(n) = D(n) - \sum_{\Omega_{p(n)}} d_{p(n),\Omega_{p(n)}} , \quad n=1,\ldots,N \tag{4}$$

Lots are then reordered in decreasing order of the priority value. From now on, the reordered sequence of the lots is referred to as a release sequence. The next step is

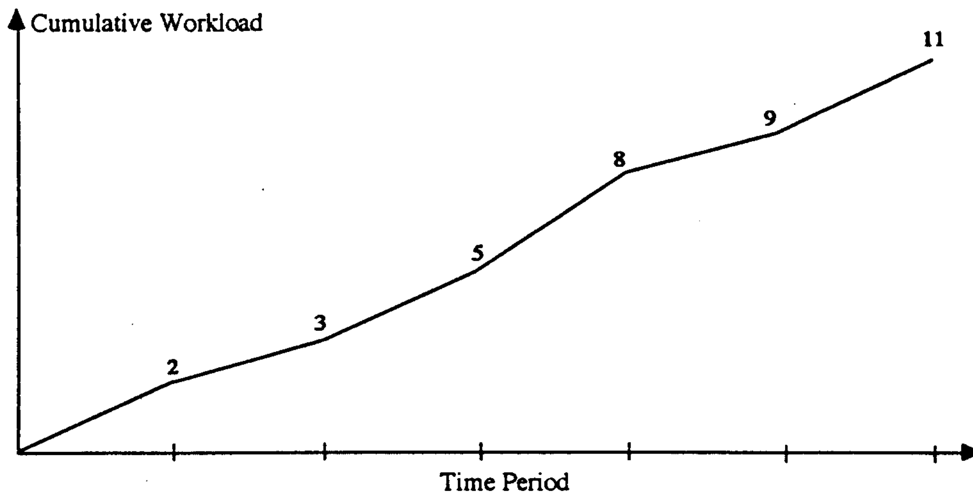

Figure 1. Cumulative workload $X_k(t)$

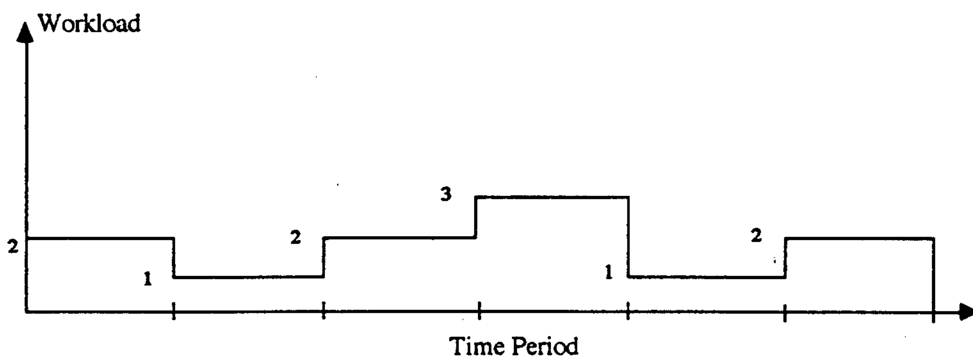

Figure 2. Workload $x_k(t)$

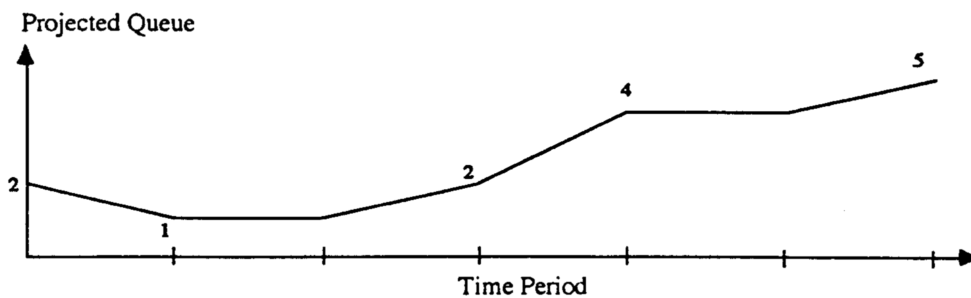

Figure 3. Projected queue $q_k(t)$

to determine whether or not to release each lot in the release sequence. Assume that the lot that is to be considered for release is lot number n in product type i. If the queue sizes at any work center to be visited by the lot are projected to be above the safety level over the time period $(\tau,\ \tau+M_i)$, the lot should not be released. Note that $\tau+M_i$ is the estimated earliest time when the order, if released, would visit all work centers that appears in the process route at least once.

On the other hand, if the queue sizes of *all* work centers to be visited by the lot are estimated to be below the safety levels within the time period $(\tau,\ \tau+M_i)$, the lot should be released now to avoid starvation, which otherwise will occur during the time period $(\tau,\ \tau+M_i)$. After the release of the lot is decided, the increased future queue sizes due to the *scheduled* release of the lot are calculated using Equations (5), (6), and (7).

$$Y_{i,k}(t)=\sum_{\Lambda} a_{i,\Lambda,k}\ v(n),\ \text{where}\ \Lambda=\left\{ l_0 \mid \sum_{\lambda=l+1}^{l_0} d_{i,\lambda} \le t \right\},\ k\in\Omega_i,\ \ t=\tau,\ldots,\tau+R \quad (5)$$

$$y_{i,k}(t)=Y_{i,k}(t)-Y_{i,k}(t-1),\quad k\in\Omega_i,\quad t=\tau,\ldots,\tau+R \quad (6)$$

$$x_k(t)=x_k(t)+y_{i,k}(t),\quad k\in\Omega_i,\quad t=\tau,\ldots,\tau+R \quad (7)$$

Using this updated queue sizes, release of the next lot in the release sequence is considered in9 the same manner. The production is continued until all lots waiting to be released are considered. When finished, we release orders as scheduled and wait until the start of the next time period.

At the start of the next time period, the average flow times of the previous time period, $\overline{d_{i,l}}$, gathered by the tracking systems, are used to recalculate the standard flow times using exponential smoothing as shown in Equation (8).

$$d_{i,l}=(1-\alpha)d_{i,l}+\alpha\overline{d_{i,l}}\ ,\qquad \text{for all } i \text{ and } l \quad (8)$$

We refer to α as a smoothing constant and set α to a number between 0 and 1. Thus, the standard flow times are updated dynamically at the start of each time period to reflect shop

conditions of the previous time period. This is the reason why we call the policy 'Dynamic Release Control Policy(DRCP)[1]. Now we may continue the same release decision procedure for the current time period.

COMPUTATIONAL EXPERIENCE

A simulation model, written in the SLAM simulation language and FORTRAN, is used to test the efficiency of the proposed release control policy, DRCP, in a simplified semiconductor wafer fabrication environment. The model consists of seven work centers and three defferent product types(lots). Each product type requires 20 processing steps to finish fab operations and visits work centers in a predetermined processing route shown in Table 1. Table 1's entries refer to the work center numbers. The processing routes are designed to imitate an actual wafer fab processing sequence and exhibit the re-entrant nature of the sequence, where each product may visit a station three or four times.

Table 1. Processing Routes

Product Number \ Step Number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	1	2	4	3	5	6	2	4	3	1	4	5	6	3	1	2	4	3	5	7
2	1	4	6	3	2	1	4	2	3	7	6	4	5	3	1	2	4	3	6	7
3	1	5	4	3	2	6	4	5	6	3	1	3	7	1	4	3	2	6	4	7

In our simulation study, it is assumed that all visits by each product type to a specific work center have the same processing time distribution. All processing time distributions are assumed to be exponentially distributed with different parameters for each work center. Machines at each work center break down according to a known probability distribution.

Exponential distributions are used to fit time-between-failure and time-to-repair distributions. Two different types of machine failure rates are simulated, and hereafter are referred to as low and high failure rates. We use exponential time-between-failure distributions with means 40 and 30 for the low and high failure systems, respectively. Time-to-repair distribution is same for the two types of systems and is set to an exponential distribution with a mean of four. Thus, machines of the high failure rate system are simulated to fail approximately once every 10 working days(3 shifts per day with the unit time equivalent to a shift) and the failed machine requires about 1 day to return to a working condition. The low and high failure type systems are tested with two different due date settings, tight and loose. The loose due date means that due date setting is done so that lots are processed normally and will be completed on time. Due dates with a 10 percent decrease are used for the tight due date setting to simulate the actual wafer fab lines for the higher demand items such as 1M DRAM's. Thus, four different configurations in terms of failure rate and due date tightness are simulated during the experiment.

Four types of release control rules combined with three different dispatching rules are evaluated for each type of configuration. The release control rules tested are the DRCP(the proposed policy), the poisson, the deterministic(times between releases are held constant), and the workload regulating input policies(Wein, 1993). The workload regulating input policy(WRIP) is arguably the most efficient rule so far introduced. Dispatching rules used are shortest processing time(SPT), first-come first-served(FCFS), and least work remaining(LWKR). For each set of input regulation and sequencing rules, 10 independent runs are made, and the average of the output is used as a result of the test. Simulation length of each test run is set to 1000 units times, which amounts to 1 working year. The output of the initial 100 unit times is discarded to remove a transient part. The performance measures of interest are the mean flow time, the mean waiting time, and the mean tardiness of completed jobs for the same output rates. The mean waiting time represents the portion of the mean flow time that can be improved by scheduling and thus, the mean waiting time is a more significant indicator for evaluating the performance of the rules tested than the mean flow time. The smoothing constant is set to a number from a uniform distribution [0.01, 0.3] throughout the computational experiments.

Table 2 illustrates the performance of the rules in the loose due date and low failure rate configuration. DRCP successfully decreases the mean waiting time by 19.8 percent compared

to WRIP. Poisson and deterministic rules deviate from the best performance by 122.5 percent and 59.0 percent respectively. The improvements from using different dispatching rules with the same release control mechanism are quite modest except for poisson release control cases. It is not clear if the improvement in mean tardiness is statistically significant. Table 3 shows the result for the tight due date and low failure rate systems. DRCP shows an improvement of 20.0 percent in mean waiting time and 49.6 percent in mean tardiness over WRIP. We also observe that the performance of the other two rules grows worse in all three performance measures. For the high failure rate system with loose due date, we observe in Table 4 that the 39.2 percent and 22.0 percent improvements in mean waiting time and mean tardiness are achieved by DRCP when compared to the WRIP. The improvement increases drastically in our final model, high failure rate system with tight due date cases. Table 5 shows that the DRCP is 50.0 percent and 37.8 percent more efficient than WRIP in mean waiting and tardiness. The other two rules, deterministic and poisson rules, report the worst performance deviating up to 173.3 percent and 107.9 percent from the best performance in mean waiting time and mean tardiness criteria respectively.

CONCLUSIONS

In summary, the proposed rule shows noticeable improvements over three rules that were compared in the computational experiments. It is also noted that the improvements increase as machine failure rates are increased and due dates are set tighter. This is a desired property for any release control rule to cope with highly stochastic features of real wafer fab lines. Based on the previous findings, the results also convince us that scheduling has a significant impact on wafer fab operations. The larger improvements come from discretionary release control rather than from dispatching.

Thus we may conclude that the proposed rule, DRCP, may be applied to real wafer fab lines and will improve performance of the shop significantly. Possilble extensions to our rule include employing a more complicated priority value determination rule such as an Artificial Intelligence(AI) based method and investingating sensitivity of the smoothing constant.

Table 2. Output of Loose Due Date and Low Failure Rate Case

Release Control Policy	Dispatching Rule	Mean Flow Time	Mean Waiting Time	Mean Tardiness
DRCP	FIFO	43.31(1.038)	0.76(1.027)	0.61(1.245)
	LWKR	40.01(1.005)*	0.74(1.000)*	0.52(1.061)
	SPT	39.81(1.000)	0.75(1.014)	0.57(1.163)
		40.38(1.014)	0.75(1.014)	0.57(1.156)
Poisson	FIFO	68.04(1.706)	1.90(2.568)	0.68(1.388)
	LWKR	59.83(1.503)	1.41(1.905)	0.54(1.102)
	SPT	61.44(1.543)	1.63(2.203)	0.60(1.224)
		63.10(1.585)	1.65(2.225)	0.61(1.238)
Deterministic	FIFO	53.67(1.348)	1.25(1.685)	0.61(1.245)
	LWKR	48.85(1.227)	1.14(1.541)	0.49(1.000)*
	SPT	50.31(1.264)	1.14(1.541)	0.54(1.102)
		50.94(1.280)	1.18(1.590)	0.55(1.116)
Workload Regulating	FIFO	41.86(1.051)	0.90(1.216)	0.63(1.212)
	LWKR	40.68(1.022)	0.89(1.203)	0.52(1.000)
	SPT	41.31(1.038)	0.90(1.216)	0.54(1.039)
		41.28(1.037)	0.90(1.212)	0.56(1.150)

(a) * indicates the best performance.
(b) entries inside parenthesis indicate percent deviation from the best performance.
(c) entries between thick lines represent performance averaged on three dispatching rules.

Table 3. Output of Tight Due Date and Low Failure Rate Case

Release Control Policy	Dispatching Rule	Mean Flow Time	Mean Waiting Time	Mean Tardiness
DRCP	FIFO	41.54(1.047)	0.70(1.000)*	2.95(1.412)
	LWKR	39.69(1.000)*	0.77(1.100)	2.09(1.000)*
	SPT	40.20(1.013)	0.79(1.129)	2.49(1.919)
		40.48(1.020)	0.75(1.076)	2.51(1.201)
Poisson	FIFO	68.25(1.720)	1.90(2.714)	5.56(2.660)
	LWKR	55.39(1.396)	1.35(1.929)	3.37(1.612)
	SPT	60.76(1.531)	1.63(2.329)	4.56(2.182)
		61.47(1.549)	1.63(2.324)	4.50(2.151)
Deterministic	FIFO	54.14(1.364)	1.23(1.757)	3.73(1.785)
	LWKR	48.55(1.223)	1.98(1.400)	2.56(1.225)
	SPT	50.91(1.283)	1.05(1.500)	3.15(1.507)
		51.20(1.290)	1.09(1.552)	3.15(1.506)
Workload Regulating	FIFO	41.88(1.055)	0.90(1.286)	3.60(1.723)
	LWKR	40.69(1.025)	0.89(1.271)	3.08(1.474)
	SPT	41.07(1.035)	0.89(1.271)	3.96(1.895)
		41.21(1.038)	0.89(1.276)	3.55(1.697)

(a) * indicates the best performance.
(b) entries inside parenthesis indicate percent deviation from the best performance.
(c) entries between thick lines represent performance averaged on three dispatching rules.

Table 4. Output of Loose Due Date and High Failure Rate Case

Release Control Policy	Dispatching Rule	Mean Flow Time	Mean Waiting Time	Mean Tardiness
DRCP	FIFO	45.78(1.050)	0.90(1.071)	1.11(1.144)
	LWKR	43.58(1.000)*	0.89(1.060)	0.97(1.000)*
	SPT	43.92(1.008)	0.84(1.000)*	1.05(1.083)
		44.43(1.019)	0.88(1.044)	1.04(1.075)
Poisson	FIFO	94.60(2.171)	3.50(4.167)	1.64(1.696)
	LWKR	71.83(1.648)	2.15(2.560)	1.24(1.278)
	SPT	82.74(1.899)	2.46(2.965)	1.48(1.526)
		83.06(1.906)	2.70(3.219)	1.45(1.498)
Deterministic	FIFO	92.66(2.126)	2.87(3.417)	1.28(1.320)
	LWKR	68.54(1.573)	1.89(2.250)	1.04(1.072)
	SPT	80.82(1.855)	2.35(2.798)	1.53(1.577)
		80.67(1.851)	2.37(2.822)	1.28(1.323)
Workload Regulating	FIFO	58.33(1.339)	1.14(1.357)	1.24(1.278)
	LWKR	52.87(1.213)	1.24(1.476)	1.07(1.103)
	SPT	53.55(1.229)	1.24(1.476)	1.46(1.506)
		54.92(1.260)	1.21(1.436)	1.26(1.295)

(a) * indicates the best performance.
(b) entries inside parenthesis indicate percent deviation from the best performance.
(c) entries between thick lines represent performance averaged on three dispatching rules.

Table 5. Output of Tight Due Date and High Failure Rate Case

Release Control Policy	Dispatching Rule	Mean Flow Time	Mean Waiting Time	Mean Tardiness
DRCP	FIFO	46.53(1.053)	0.97(1.128)	5.36(1.107)
	LWKR	44.18(1.000)*	0.86(1.000)*	4.84(1.000)*
	SPT	44.86(1.015)	0.91(1.058)	5.02(1.037)
		45.19(1.023)	0.91(1.062)	5.07(1.048)
Poisson	FIFO	85.25(1.930)	2.50(2.907)	9.83(2.031)
	LWKR	71.21(1.612)	2.16(2.512)	8.73(1.804)
	SPT	80.81(1.829)	2.55(2.965)	12.32(2.545)
		79.09(1.790)	2.40(2.795)	10.29(2.127)
Deterministic	FIFO	90.01(2.037)	2.64(3.070)	10.40(2.149)
	LWKR	70.12(1.587)	1.88(2.186)	7.23(1.494)
	SPT	79.23(1.793)	2.16(2.512)	8.82(1.822)
		79.79(1.806)	2.23(2.589)	8.82(1.822)
Workload Regulating	FIFO	57.10(1.292)	1.35(1.570)	7.24(1.496)
	LWKR	59.88(1.355)	1.49(1.733)	6.75(1.395)
	SPT	53.98(1.222)	1.19(1.384)	6.72(1.388)
		56.99(1.290)	1.34(1.562)	6.90(1.426)

(a) * indicates the best performance.
(b) entries inside parenthesis indicate percent deviation from the best performance.
(c) entries between thick lines represent performance averaged on three dispatching rules.

REFERENCES

1. P. B. Chevalier and L. M. Wein (1993), Scheduling networks of queues: heavy trafic analysis of a multistation closed network. *Operations Research* 41, 743-758.
2. R. C. Glassey and M. G. C. Resende (1988), Closed-loop release control for VLSI circuit manufcaturing. *IEEE Transactions on Semiconductor Manufacturing* 1, 36-46.
3. J. W. Lawton (1990), Wafer release in a GaAs environment. MSc. Thesis, Sloan School of Management, MIT, Cambridge, U.S.A.
4. R. C. Leachman, M. G. Solorzano and C. R. Glassey (1988), A queue management policy for the release of factory work orders. ESRC Report 88-19, University of California, Berkeley, U.S.A.
5. G. L. Ragatz and V. A. Mabert (1988), An evaluation of order release mechanism in a job-shop environment.. *Decision Sciences 19, 167-189.*
6. M. G. Solrzano (1989), Workload regulation of semiconductor fabrication facilities. ESRC Report 89-1, University of California, Berkeley, U.S.A.
7. L. M. Wein (1988), Scheduling semiconductor wafer fabrication. *IEEE Transactions on Semiconductor Manufacturing* 1, 115-130.
8. L. M. Wein (1992), Scheduling networks of queues: heavy traffic analysis of a multistation network with controllable inputs. *Operations Research* 40, S312-S334.