Correspondence

Modeling the Effect of Hot Lots in Semiconductor Manufacturing Systems

Y. Narahari and L. M. Khan

***Abstract*— The presence of hot lots or high-priority jobs in semiconductor manufacturing systems is known to significantly affect the cycle time and throughput of the regular lots since the hot lots get priority at all stages of processing. In this paper, we present an efficient analytical model based on re-entrant lines and use an efficient, approximate analysis methodology for this model in order to predict the performance of a semiconductor manufacturing line in the presence of hot lots. The proposed method explicitly models scheduling policies and can be used for rapid performance analysis. Using the analytical method and also simulation, we analyze two re-entrant lines, including a full-scale model of a wafer fab, under various buffer priority scheduling policies. The numerical results show the severe effects hot lots can have on the performance characteristics of regular lots.**

I. Introduction

In this paper, we model semiconductor manufacturing systems as *re-entrant lines* [1] and study the effect of hot lots or high-priority jobs through an approximate analysis of the re-entrant line model using mean value analysis (MVA) [2]. The MVA-based method facilitates explicit modeling of buffer priority based scheduling policies used in re-entrant lines [1] and is computationally much more efficient than simulation. We provide numerical results obtained using the analytical method and also simulation, to study the effect of hot lots on performance characteristics such as mean cycle time, variance of cycle time, and mean throughput rate, of regular lots and hot lots.

In semiconductor manufacturing systems, wafer fabrication constitutes the most important step. Wafer fabrication involves a large, complex sequence of processing steps. An important feature of wafer fabrication processing is *re-entrancy*, which refers to multiple visits by a wafer lot to the same processing center at various times. *Re-entrant lines* [1] constitute an appropriate queuing model for wafer fabrication lines. Fig. 1 shows a typical re-entrant line, with two processing centers—centers 1 and 2, and 4 buffers. Fig. 2 shows a re-entrant line of realistic size, with 12 processing centers and 60 buffers. This is a model of a real-life semiconductor fab, considered earlier by Lu, Ramaswamy, and Kumar [3].

In a general re-entrant line, each service center may have several machines or servers. In this paper, we assume, however, that there is only one machine in each service center. Also, there could be several types of parts (jobs or wafer lots), each one following a different route. The machines could be prone to failures in a random fashion. Further, substantial setup times may be required before operations on jobs can be initiated. Since the main emphasis of this paper is to bring out the effect of hot lots and also capture priority scheduling policies, we shall make some simplifying assumptions. In particular, we assume that there is only a single part type, that the machines do not fail, and that the setup times are negligible. The approach presented in this paper can be extended to handle these features, but in this paper, we emphasize secondary effects such as caused by hot lots and scheduling policies rather than primary effects such as due to setups and equipment failures.

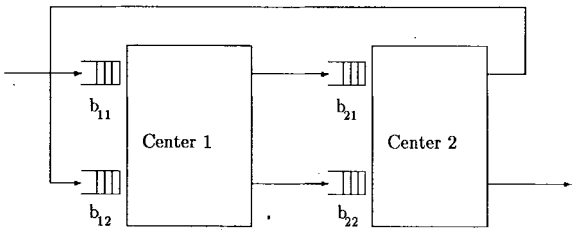

Fig. 1. A re-entrant line with two stations and four buffers.

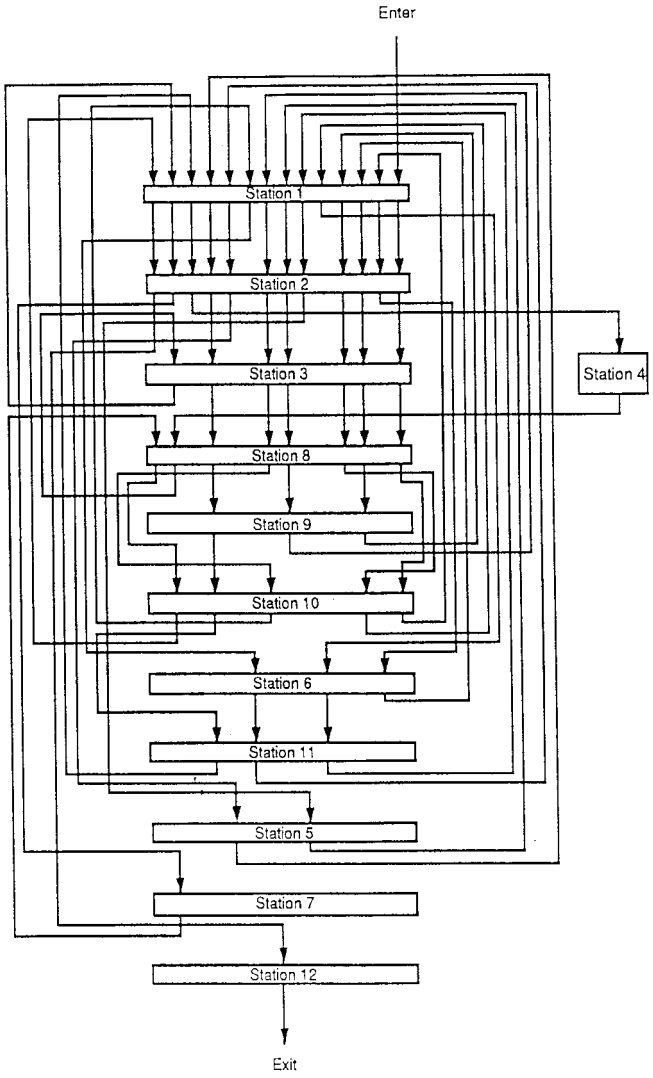

Fig. 2. A full-scale re-entrant line.

Manuscript received March 3, 1995; revised July 1, 1996. This work was supported in part by the Office of Naval Research and the Department of Science and Technology Grant N00014-93-1017.

The authors are with the Department of Computer Science and Automation, Indian Institute of Science, Bangalore 560012, India.

Publisher Item Identifier: S 0894-6507(97)01025-7.

The scheduling or dispatching problem in re-entrant lines becomes interesting because several parts at different stages of processing may be in contention with one another for service at the same machine. Several researchers have focused on the issue of scheduling in re-entrant lines [1], [3], [4]. Performance analysis of scheduling policies in re-entrant lines is the subject of study in [5]–[7]. In the first three of these articles, the emphasis is on finding bounds on the steady-state performance of stationary, nonidling, buffer priority-based scheduling policies by solving linear programs. Narahari and Khan [7] have proposed an efficient *mean value analysis*—based method for analyzing re-entrant lines. In all these studies however, hot lots or high priority jobs have not been considered.

In the semiconductor manufacturing terminology, hot lots refer to the class of jobs with the highest priority, often introduced into the system purely out of marketing and business considerations [8]. Hot lots often lead to irregular flow of parts and can drastically alter the cycle time and throughput of regular jobs. From a modeling point of view, hot lots have the same route and the same processing times at all processing centers as the regular lots but get priority over regular lots everywhere. Ehteshami, Petrakian, and Shabe [8] have carried out a simulation study to understand the impact of hot lots on the cycle time of regular lots in the system. Their study only considers the fixed work-in-process policy as the input release policy and FCFS as the scheduling policy, and does not account for any other scheduling policy. It is of much importance and practical interest to study the effects of hot lots under specific priority scheduling policies. This study attempts to address this problem.

In this paper we first outline, in Section II, an efficient method for approximate analysis of re-entrant lines in the presence of hot lots. The method is an extension of the one proposed by Narahari and Khan [7]. An important object of the analysis method is the ability to model buffer priority scheduling policies such as LBFS (Last Buffer First serve) and FBFS (First Buffer First Serve). We next present in Section III, detailed results obtained, through the proposed method and simulation, and study the effect of hot lots on the performance of regular lots. The performance indices considered are:

- Mean steady-state cycle time of hot lots and regular lots.
- Variance of steady-state cycle time of hot lots and regular lots.
- Mean steady-state throughput rates of hot lots and regular lots.

II. An Approximate Analysis Methodology

Without loss of generality, assume that the scheduling policy is LBFS (last buffer first serve). Let the re-entrant line have m processing centers. Center i has n_i logical or physical buffers, $b_{i1}, \cdots, b_{in_i}$, where for $j \in \{1, 2, \cdots, n_i\}$ the buffer b_{ij} contains regular lots visiting center i for the jth time. We also assume that center i, $i \in \{1, 2, \cdots, m\}$, has n_i logical buffers, $b'_{i1}, \cdots, b'_{in_i}$, to accommodate hot lots, so that the buffer b'_{ij} contains the hot lots visiting center i for the jth time.

Let the performance measures of the network be denoted as follows:

- $L_{ij}(k, h)$ Expected number of regular lots in stage (i, j) when the network has k regular lots and h hot lots.
- $L'_{ij}(k, h)$ Expected number of hot lots in stage (i, j) when network has k regular lots and h hot lots.
- $W_{ij}(k, h)$ Mean steady-state delay for regular lots in stage (i, j) (mean waiting time in buffer b_{ij}+ mean processing time).
- $W'_{ij}(k, h)$ Mean steady-state delay for hot lots in stage (i, j) (mean waiting time in buffer b'_{ij}+ mean processing time).
- $\lambda(k, h)$ Mean-steady state throughput rate of regular lots when the network has k regular lots and h hot lots.
- $\lambda'(k, h)$ Mean-steady state throughput rate of hot lots when the network has k regular lots and h hot lots.

If $W(k, h)$ and $W'(k, h)$ denote the mean total delay (mean cycle time) in the entire network, we immediately have

$$W(k, h) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} W_{ij}(k, h) \tag{1}$$

$$W'(k, h) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} W'_{ij}(k, h) \tag{2}$$

Using MVA, we compute $W(N, H)$, $W'(N, H)$, $\lambda(N, H)$, and $\lambda'(N, H)$ in a recursive way. The detailed recursive equations and their efficient solution using an iterative formulation are discussed in [9]. This approximate analytical method is quite efficient. Given that the regular lot population is N and that the hot lot population is H, we obtain performance measures for regular lots, namely $W(N, H)$, $\lambda(N, H)$, and $L_{ij}(N, H)$, in exactly N iterations. Each iteration involves $O(M)$ *operations* where M is the total number of buffers. Each *operation* refers to computing W_{ij}, L_{ij}, and λ terms. During the iterations we also get as intermediate values, $W(k, H)$, $\lambda(k, H)$, and $L_{ij}(k, H)$, for $k = 1, 2, \cdots, N$. This constitutes a very attractive feature of the method. Similarly, the performance measures for hot lots namely $W'(N, h)$, $\lambda'(N, h)$, and $L'_{ij}(N, h)$ for $h = 1, 2, \cdots, H$ are obtained in exactly H iterations.

III. Numerical Results

In order to validate the analytical method and verify the accuracy of its performance predictions, we have carried out detailed simulations of several re-entrant lines. It is found that simulations are much slower compared to the analytical method. Also there is a close agreement in the results given by the two methods. The simulations were carried out using SIMSCRIPT II.5 on a Intel 80 486-DX based machine. A single long run was used to compute the steady-state performance measures for each experiment. The initial transient period in each case was determined by making several pilot runs. Statistics were collected after removing the initial transients. Also, statistical tests were conducted to obtain a 0.95 level for all the experiments.

A. A Two Station, Four Buffer Re-Entrant Line

Consider the two station, four buffer system of Fig. 1. In this system, only four buffer priority policies are possible:

1) LBFS at station 1 and LBFS at station 2 (Policy 1).
2) LBFS at station 1 and FBFS at station 2 (Policy 2).
3) FBFS at station 1 and LBFS at station 2 (Policy 3).
4) FBFS at station 1 and FBFS at station 2 (Policy 4).

Let

$$\frac{1}{\mu_{11}} = \frac{1}{\mu_{12}} = \frac{1}{\mu_{21}} = \frac{1}{\mu_{22}} = 1$$

Assuming a constant regular lot population of 50 and varying the hot lot population from 0 to 10, we computed the mean cycle time (MCT) and the mean throughput rate (TR), in the steady state, of regular lots and hot lots, using the proposed analysis method and also simulation. Table I shows these values, assuming LBFS policy at both the stations.

We find that the presence of hot lots brings down significantly the throughput rate of regular lots and consequently the mean cycle time of the regular lots rises quite dramatically. From this table, we also see a close agreement between the values obtained using the analytical and simulation.

TABLE I
ANALYTICAL AND SIMULATION RESULTS FOR EXAMPLE 1 (FIG. 1)

Population		MCT-Reg		MCT-Hot		TR-Reg		TR-Hot	
Reg	Hot	SIM	MVA	SIM	MVA	SIM	MVA	SIM	MVA
50	0	101.95	102.97	0	0	0.49029	0.48558	0	0
50	1	137.37	141.93	7.89	7.88	0.36384	0.35228	0.12669	0.12683
50	2	189.22	197.83	8.83	8.92	0.26412	0.25275	0.22657	0.22417
50	3	263.13	270.63	9.97	10.27	0.18986	0.18475	0.30099	0.29201
50	4	355.64	355.36	11.41	11.89	0.14044	0.14070	0.35053	0.33651
50	5	461.45	477.42	13.05	13.66	0.10826	0.11175	0.38303	0.36596
50	6	576.21	545.47	14.82	15.52	0.08666	0.09166	0.40477	0.38657
50	7	699.74	649.29	16.66	17.42	0.07136	0.07701	0.42029	0.40174
50	8	825.56	785.74	18.54	19.35	0.06043	0.06590	0.43138	0.41335
50	9	960.55	873.77	20.46	21.30	0.05192	0.05722	0.43993	0.42253
50	10	1093.83	994.31	22.40	23.26	0.04558	0.05029	0.44646	0.42995

TABLE II
MEAN DELAYS AT INDIVIDUAL BUFFERS FOR VARIOUS SCHEDULING POLICIES

Population		Delay at b_{11}		Delay at b_{12}		Delay at b_{21}		Delay at b_{22}	
Reg	Hot	Policy 1	Policy 2	Policy 1	Policy 2	Policy 1	Policy 2	Policy 1	Policy 2
50	0	47.80	49.01	3.18	2.81	47.80	3.54	3.18	46.77
50	1	64.40	66.30	4.49	3.86	64.40	5.21	4.49	62.23
50	2	87.50	91.32	7.09	5.82	87.50	8.30	7.09	84.51
50	3	119.89	125.14	11.73	9.48	119.89	14.02	11.73	114.91
50	4	158.57	165.58	19.30	15.41	158.57	22.66	19.30	150.67
50	5	200.92	211.45	29.78	23.67	200.92	34.50	29.78	191.29
50	6	245.40	259.77	42.67	33.91	245.40	48.80	42.67	232.05
50	7	292.00	311.02	57.90	47.45	292.00	66.26	57.90	272.44
50	8	338.03	359.20	74.99	63.19	338.03	87.14	74.99	315.65
50	9	385.85	416.38	94.79	79.66	385.85	108.62	94.79	362.96
50	10	430.51	466.99	116.77	96.60	430.51	130.56	116.77	399.74

TABLE III
STANDARD DEVIATION OF CYCLE TIMES FOR REGULAR AND HOT LOTS

Population		Policy 1		Policy 2		Policy 3		Policy 4	
Reg	Hot	SD-Reg	SD-Hot	SD-Reg	SD-Hot	SD-Reg	SD-Hot	SD-Reg	SD-Hot
50	0	10.11	0	19.56	0	10.11	0	10.84	0
50	1	14.39	2.79	29.03	2.79	14.35	2.80	15.79	2.80
50	2	23.01	3.04	45.21	3.06	22.61	3.05	26.02	3.09
50	3	36.62	3.27	70.86	3.34	36.20	3.31	42.62	3.37
50	4	56.04	3.52	102.49	3.73	55.26	3.53	65.08	3.69
50	5	79.65	3.76	143.10	4.19	79.51	3.80	94.34	4.03
50	6	106.81	4.02	187.31	4.72	109.09	4.03	131.26	4.37
50	7	136.19	4.28	234.11	5.29	141.14	4.28	170.09	4.70
50	8	162.76	4.45	280.24	5.81	182.14	4.50	205.82	5.02
50	9	199.95	4.68	339.20	6.33	217.38	4.74	255.96	5.31
50	10	237.16	4.88	382.43	6.90	255.03	4.95	299.34	5.54

The MCT and TR values were obtained for the other three policies also and were found to be virtually the same for all four scheduling policies. This is because of the closed nature of the network. The mean steady state delay at individual buffers will however be different for different scheduling policies. Table II shows the delay of regular lots at buffers b_{11}, b_{12}, b_{21}, and b_{22}, as predicted by our analysis method, for two different scheduling policies: Policy 1 (LBFS at station 1 and LBFS at station 2); Policy 2 (LBFS at station 1 and FBFS at station 2).

Since the overall MCT and TR values for the policies were found to be virtually the same, we computed the standard deviation of cycle times as a function of the hot lot population. This enables us to capture the variability of performance offered by various buffer priority policies and assumes importance in the light of desirability of fluctuation smoothing [3]. Here different scheduling policies yield different trends. In all the cases however, the cycle time of the regular lots shows a very high standard deviation. Table III shows these results. The LBFS policy applied to both the stations (Policy 1) is found to yield the lowest values of the standard deviation. Also, in all the cases, it is found that the coefficients of variation increase with the hot lot populations.

B. A Full-Scale Re-Entrant Line

Here we present analytical and simulation results for a re-entrant line of realistic size (Fig. 2). Let $1/\mu_i$ be the mean processing time at center i ($i = 1, \cdots, 12$) on each visit to that center. We assume:

$$\frac{1}{\mu_1} = 0.125, \frac{1}{\mu_2} = 0.125, \frac{1}{\mu_3} = 0.250,$$
$$\frac{1}{\mu_4} = 1.800, \frac{1}{\mu_5} = 0.900, \frac{1}{\mu_6} = 0.600,$$
$$\frac{1}{\mu_7} = 1.800, \frac{1}{\mu_8} = 0.200, \frac{1}{\mu_9} = 0.600,$$
$$\frac{1}{\mu_{10}} = 0.333, \frac{1}{\mu_{11}} = 0.600, \frac{1}{\mu_{12}} = 1.250.$$

TABLE IV
PERFORMANCE MEASURES OF THE FULL-SCALE RE-ENTRANT LINE

Population		MCT		TR		SD of cycle time	
Reg	Hot	Reg	Hot	Reg	Hot	Reg	Hot
500	0	911.94	0	0.54566	0	60.73	0
500	1	957.04	38.80	0.51976	0.02577	67.85	5.20
500	5	1151.69	43.30	0.43154	0.11546	81.67	5.70
500	10	1462.85	48.82	0.33892	0.20479	120.11	6.23
500	15	1800.61	55.371	0.27509	0.27081	163.24	6.907
500	20	2184.37	62.46	0.22590	0.32013	211.96	7.39
500	25	2574.48	70.069	0.19133	0.35667	256.27	8.164
500	30	3062.38	77.63	0.16073	0.38628	341.87	8.55
500	35	3547.88	85.69	0.13776	0.40824	467.95	9.341
500	40	4082.63	93.79	0.11946	0.42623	499.81	10.46
500	45	4597.28	101.78	0.10588	0.44184	662.24	10.59
500	50	5189.61	110.13	0.09379	0.45370	800.41	10.87

The above parameters are the same as in [3]. Note that the hot lots also have the same mean processing times. Assuming the number of regular lots to be 500, and LBFS policy at all stations, we study the effect of hot lots by varying their number from 0–50. Table IV gives the mean cycle time of regular lots and hot lots (computed using our MVA-based method), and standard deviation of cycle times (computed using simulation). The same trends as in the previous two examples are discernible here.

IV. CONCLUSIONS

The analytical method, based on MVA approximation, presented in this paper has been found to be efficient and quite accurate in predicting the performance of semiconductor manufacturing lines in the presence of hot lots. The numerical results obtained on some illustrative re-entrant lines show the significant effect of hot lots on the mean cycle time, variance of cycle time, and throughput rate of regular lots.

Topics for future work include: modeling of due date based policies and fluctuation smoothing policies [1], [3] using the MVA approximation and modeling of multiple machines in each service or processing center, multiproduct types, machine failures and repairs, setup times, and presence of multiple equipment in the same processing center.

ACKNOWLEDGMENT

We have benifited immensely from several discussions with Prof. P. R. Kumar, University of Illinois, Urbana-Champaign.

REFERENCES

[1] P. R. Kumar, "Re-entrant lines," *Queuing Syst.: Theory Applicat.*, vol. 13, pp. 87–110, 1993.
[2] M. Reiser and S. S. Lavenberg, "Mean value analysis of closed multichain queuing networks," *J. ACM*, vol. 27, no. 2, pp. 313–322, Apr. 1980.
[3] S. H. Lu, D. Ramaswamy, and P. R. Kumar, "Efficient scheduling policies to reduce mean and variance of cycle-time in semiconductor manufacturing plants," *IEEE Trans. Semiconduct. Manufact.*, vol. 7, pp. 374–388, Aug. 1994.
[4] L. M. Wein, "Scheduling semiconductor wafer fabrication," *IEEE Trans. Semiconduct. Manufact.*, vol. 1, pp. 115–130, Aug. 1988.
[5] S. Kumar and P. R. Kumar, "Performance bounds for queuing networks and scheduling policies," *IEEE Trans. Automat. Contr.*, vol. 39, no. 8, pp. 1600–1611, Aug. 1994.
[6] H. Jin, J. Ou, and P. R. Kumar, "The throughput of closed queuing networks: Functional bounds, asymptotic loss, efficiency, and the Harrison-Wein conjectures," Tech. Rep., Coordinated Sci. Lab., Univ. Illinois-Urbana Champaign, 1995.
[7] Y. Narahari and L. M. Khan, "Performance analysis of scheduling policies in re-entrant manufacturing systems," *Comput. Oper. Res.*,vol. 23, no. 1, pp. 37–51, 1996.
[8] B. Ehteshami, R. G. Petrakian, and P. M. Shabe, "Trade-offs in cycle time management: Hot lots," *IEEE Trans. Semiconduct. Manufact.*, vol. 5, pp. 101–106, May 1992.
[9] L. M. Khan, "Performance analysis of scheduling policies in stochastic re-entrant lines," Ph.D. dissertation, Dept. Comp. Sci. Automat., Indian Institute of Science, Bangalore, Apr. 1995.