Modeling Green Fabs – A Queuing Theory Approach for Evaluating Energy Performance

Hyun Woo Jeon, Vittaldas V. Prabhu

Modeling Green Fabs – A Queuing Theory Approach for Evaluating Energy Performance

Hyun Woo Jeon[1] and Vittaldas V. Prabhu[2]

[1] the Marcus Department of Industrial and Manufacturing Engineering, Pennsylvania State University, University Park, PA 16802 USA
albert.jeon@psu.edu
[2] the Marcus Department of Industrial and Manufacturing Engineering, Pennsylvania State University, University Park, PA 16802 USA
(phone: 814-863-3212; fax: 814-865-4715)
prabhu@engr.psu.edu

Abstract. More than 30% of the total energy consumed in the U.S. is attributed to industrial sector which motivated improvements in energy efficiency of manufacturing processes and entire factories. Semiconductor fabrication (fab) represents an interesting challenge for energy efficiency because of their relatively high energy consumption to process a unit mass of material. The focus of this paper is to develop an energy-aware analytical model based on queuing theory that has re-entrant network structure commonly found in fabs to analyze the impact of reducing idle power consumption in individual equipment. The proposed analytical model based on BCMP network for re-entrant lines has the same mathematical form as serial lines and is tested for using detailed simulation of a generic CMOS fab with three processing steps. Results show that the energy consumption predicted by the analytical model differs from simulation typically within 10% and worst case of 14%, in the tested cases.

Keywords: Sustainability, Queueing Network, Re-entrant Networks, Semiconductor Manufacturing, Energy-aware Model

1 Introduction

Faced with a growing trend of increasing energy consumption worldwide, various organizations are trying to be more energy efficient [1]. Especially there is a particular need for the U.S., as one of the largest energy consumers among nations, to increase energy efficiency across all segments of the economy as the American industrial segment accounts for 31% of its total energy consumption [2]. Two benefits that can be expected from improved energy efficiency in the industry segment are direct cost savings and indirect environmental benefit, both stemming from reduced energy consumption. Consequently there has been increasing research on understanding energy consumption of various manufacturing processes and associated equipment [3]. This prior research has found that semiconductor manufacturing processes is among

the highest energy consumption per unit mass in the manufacturing industry. For example, the oxidation process in fabs typically consumes 1.E+13 – 14 J/kg compared to milling steel which consumes about 1.E+05 J/kg. Moreover, idling manufacturing machines typically consume as much as 40 to 60% of the power when they are busy for processing parts [4]. A logical progression of research effort is to build on the energy consumption models of unit processes to higher levels such as the factory level and the supply chain level. For analyzing busy/idle states of a system consisting of many machines, simulation and queuing modeling are attractive approaches. Good insights can be acquired when both approaches are used, including testing and validation of each other judiciously [5].

Our recent efforts in this direction include developing hybrid simulation models consisting of continuous and discrete process for predicting energy consumption in discrete manufacturing [6]. We have also proposed queuing models to predict energy savings in serial production lines when idling machines are switched to a lower power state in serial production lines consisting of machines with Poisson arrival and exponential service time [7].

This paper builds on our previous work [7] to extend and generalize the energy-aware queuing model to a re-entrant structure which is applied to model semiconductor fabs. Section 2 reviews the prior analytical model along with its parameters and limitations. In Section 3 a model for re-entrant lines is introduced and energy performance characteristics are derived. In Section 4 a generic semiconductor production system is used for comparing results predicted by the analytical model with those by detailed simulation. Section 5 summarizes this paper with possible future research directions.

2 Energy-aware Queuing Model

Our recently proposed approach for improving energy efficiency in manufacturing system is to use energy control policies that switch machines to a lower power consumption state when a machine is anticipated to be idle for longer than some threshold, τ [6, 7]. In order to predict the efficiency gained by using such energy control policies, the state transitions between the various energy consumption states and production states of the machines need to be modeled. Without any energy control policy (EC_0) the machine power consumption in its idle state will be W_1, the nominal power for idling. When energy control policy is used (EC_1), energy state will be switched to the low power idling of W_0 if the idle time duration is anticipated to be longer than τ; otherwise it will be W_1. In the busy state the power consumption will be W_p regardless of whether EC_0 or EC_1 used.

For an M/M/1 queue with an arrival rate of λ and a service rate μ, the probability that a new arrival arrives after the time threshold τ is given by

$$P(x > \tau) = \int_{\tau}^{\infty} \lambda e^{-\lambda x} dx = e^{-\lambda \tau} \quad (1)$$

Based on this, the total energy consumption with EC_0 and EC_1 over time T can be modeled as follows

$$E_{EC0} = \{W_p\rho + W_1(1-\rho)\}T \tag{2}$$

$$E_{EC1} = \{W_p\rho + W_1(1-\rho)(1-e^{-\lambda\tau}) + W_0(1-\rho)e^{-\lambda\tau}\} \tag{3}$$

In Equations 2 and 3 ρ is the machine utilization. Therefore the total energy saving can be expressed as

$$E_{total\ energy\ saved} = (W_1 - W_0)(1-\rho)e^{-\lambda\tau}T \tag{4}$$

It should be noted from Equation 4 that energy saving can be increased by decreasing W_0 and τ for a given ρ. The above model for a single queue can be extended to serial production lines [7].

3 Re-entrant System Modeling

3.1 Queueing Model for Re-entrant System

Semiconductor fabs are markedly different from serial production lines or job shops [8, 9]. A typical semiconductor fab has a re-entrant structure in which the products flow through a fixed route visiting a machine multiple times [10, 11]. Semiconductor fabs can be modeled as an open queuing network in which parts visit a machine multiple times. It should be pointed out that because the route in semiconductor fab is deterministic such systems cannot be modeled as Jackson networks [12]. Therefore in this paper we will model semiconductor fabs as a BCMP network, which is also the approach taken by several other researchers [9, 11, 13]. The main parameters and assumptions in modeling a semiconductor fab as a BCMP queuing network are as follows

- Arrival rate from outside is λ and V_m is the number of visits to machine *m* before a part exits the system.
- Arrival process is Poisson and mean service time follows exponential distribution with rates of λ_m and μ_m, respectively.
- Queueing discipline is FCFS (first come first serve) and parts are served regardless of buffer in which the part is located.
- Service rate μ_m is same for all parts in any buffer of the machine.
- Route of each part is fixed and deterministic.
- Idle time threshold for EC_1 is τ_m for machine *m*.

Based on the above model, the arrival rate at machine *m* can be expressed as

$$\lambda_m = V_m\lambda \tag{5}$$

The utilization of machine *m* can be also defined as

$$\rho_m = \lambda_m/\mu_m = V_m\lambda/\mu_m \tag{6}$$

The long run probability that there are n parts at a machine *m* is therefore given by

$$\pi_n(m) = (1-\rho_m)\rho_m^n \tag{7}$$

From (6), (7) the probability of machine *m* is busy, $P_B(m)$ can be expressed as

$$P_B(m) = \rho_m \tag{8}$$

The re-entrant structure of the model gives rise to multiple arrival streams at a machine that may not be independent of each other. Furthermore, summing these arrivals may not be an exact model for a composite Poisson process with an exponentially distributed inter-arrival time [14]. However treating the composite arrivals as a Poisson process can be a reasonable approximation [15], which we adopt for modeling the probability of a machine being in nominal power idling (P_N) and low-power idling (P_L) states when EC_1 is used. These probabilities for machine *m* are

$$P_N(m) \approx (1-\rho_m)\left(1-e^{-\lambda_m\tau_m}\right) \tag{9}$$

$$P_L(m) \approx (1-\rho_m)e^{-\lambda_m\tau_m} \tag{10}$$

From (8)-(10), the energy consumption for machine *m* of a re-entrant system for time period T can be modeled as

$$E_{EC0}(m) \approx \{W_p(m)\rho_m + W_1(m)(1-\rho_m)\}T \tag{11}$$

$$E_{EC1}(m) \approx \begin{Bmatrix} W_p(m)\rho_m + W_1(m)(1-\rho_m)\left(1-e^{-\lambda_m\tau_m}\right) \\ +W_0(m)(1-\rho_m)e^{-\lambda_m\tau_m} \end{Bmatrix} T \tag{12}$$

where $W_p(m)$, $W_1(m)$, and $W_0(m)$ are power consumption levels of machine *m* for busy, nominal power idling, and low power idling states respectively. It should be emphasized that this energy model for a re-entrant line has the same mathematical structure as the M/M/1 model reviewed in the previous section. Therefore the total spent energy in the entire re-entrant system can be modeled as

$$E_{Total(EC0)} \approx \sum_{m=1}^{n} E_{EC0}(m) \tag{13}$$

$$E_{Total(EC1)} \approx \sum_{m=1}^{n} E_{EC1}(m) \tag{14}$$

In the next section the energy consumption predicted by this analytical model is compared with a detailed simulation model to the efficacy of the model.

3.2 Simulation Experiments

Figure 1 illustrates a generic CMOS (Complementary Metal Oxide Semiconductor) fab with Deposition (D), Lithography (L), and Etching (E) processes and a re-entrant structure that follows an overall processing sequence indicated by the numerical values on the arrowed lines. The power consumption levels in these processes are hy-

pothesized based on published data [16] and shown in Table 1 along with other parameters used to simulate three different scenarios with varying arrival rates into the system.

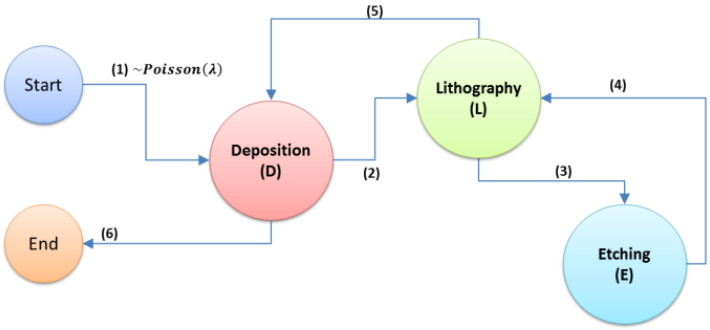

Fig. 1. Example re-entrant system for CMOS

The simulation model was implemented by using the SIMIO simulation software package. For each scenario, P_B, P_N, and P_L are estimated by fraction of the time a machine is in the corresponding state during a simulation run of 10,000 time units and averaged over 30 replications which is then used to compute the results in Tables 2 and 3.

Table 1. Simulation parameters

Parameters for each Simulation Scenario									
Scenario	1			2			3		
Process	D	L	E	D	L	E	D	L	E
λ	0.050	0.050	0.025	0.040	0.040	0.020	0.030	0.030	0.015
μ	0.100	0.067	0.029	0.100	0.067	0.029	0.100	0.067	0.029
ρ	0.500	0.750	0.875	0.400	0.600	0.700	0.300	0.450	0.525
τ	8.00	10.00	24.00	10.00	12.50	30.00	13.33	16.67	40.00
$\lambda\tau$	0.40	0.50	0.60	0.40	0.50	0.60	0.40	0.50	0.60
W_p	24857	9140	12238	24857	9140	12238	24857	9140	12238
W_1	8286	3047	4079	8286	3047	4079	8286	3047	4079
W_0	1657	609.3	815.9	1657	609.3	815.9	1657	609.3	815.9
Replication		30			30			30	

The Figure 2 illustrates energy consumed with EC_0 and EC_1 of each of the deposition process in scenario 1. As expected, energy savings in all three processes increases with time. Figure 3 shows the power consumption for the deposition processing step with EC_0 and EC_1 in scenario 1.

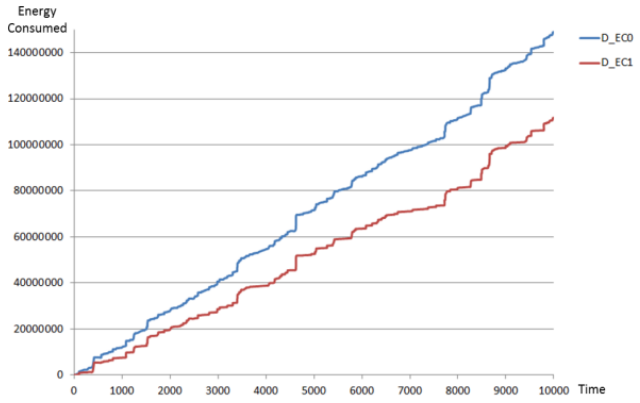

Fig. 2. Energy consumed in the deposition process in scenario 1

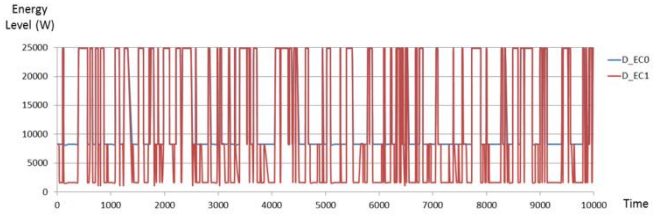

Fig. 3. Power consumption level for deposition process in scenario 1

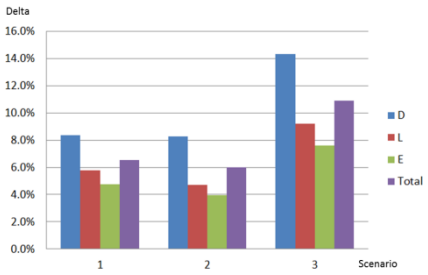

Fig. 4. Model difference with EC_1

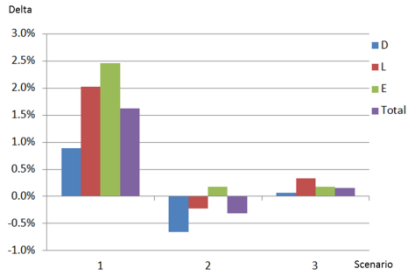

Fig. 5. Model difference with EC_0

Table 2. Comparison of Simulation and Analytical Models with EC_1

Scenario	Simulation				Analytical Approximation				Delta = (A-S)/S			
	D	L	E	Total	D	L	E	Total	D	L	E	Total
1	13245	6850	10494	30590	14350	7247	10994	32591	8.3%	5.8%	4.8%	6.5%
2	11315	5837	8903	26054	12248	6111	9253	27613	8.2%	4.7%	3.9%	6.0%
3	8876	4556	6982	20414	10147	4976	7512	22634	14.3%	9.2%	7.6%	10.9%

Table 3. Comparison of Simulation and Analytical Models with EC_0

Scenario	Simulation				Analytical Approximation				Delta = (A-S)/S			
	D	L	E	Total	D	L	E	Total	D	L	E	Total
1	16425	7465	10949	34839	16571	7617	11218	35406	0.9%	2.0%	2.5%	1.6%
2	15013	6717	9774	31504	14914	6703	9790	31407	-0.7%	-0.2%	0.2%	-0.3%
3	13248	5769	8348	27365	13257	5789	8363	27408	0.1%	0.3%	0.2%	0.2%

From Table 2 and 3, it can be observed that analytical and simulation estimate of total energy consumed typically differ within 10% with the worst case being 14.3%. Figures 4 and 5 graphically illustrate the difference between analytical and simulation models (delta) presented in Tables 2 and 3. The difference is much less with EC_0 compared to EC_1 and delta decreases as utilization ρ increases. We conjecture that the exponential distribution assumption for the composite inter-arrival times used in the analytical model underestimates the probability of being in low-power idling states, therefore overestimating the energy consumption. From energy efficiency perspective, EC_1 consumes less energy than EC_0 as expected. Energy savings $(EC_0 - EC_1)$ decrease as ρ increases because there is lesser opportunity for EC_1 as machines get busier. Moreover, for a given ρ, smaller τ will make EC_1 more energy efficient. In practice, the choice of τ will be constrained by the physics and economics of the process.

4 Conclusions

An energy-aware analytical model based on BCMP network is proposed in this paper for semiconductor fabs with re-entrant structure commonly. Focus of the analysis is to assess the energy efficiency gained by reducing idle power consumption by switching machines to a lower power state when idling. Simulation tests indicate that the energy consumption predicted by the analytical models are typically differ within 10% from detailed simulations with worst case being about 14%, which makes the approach promising for rapid evaluation of candidate energy control strategies. Future work could focus on analyzing the factors that can influence the accuracy of the analytical model including the composite inter-arrival distribution model. Another potential future direction is to generalize the queuing networks to include G/G/1 and GI/G/1 systems [17, 18].

References

1. The U.S. Energy Information Administration: International Energy Outlook 2011, DOE/EIA-0484(2011), http://www.eia.gov/forecasts/ieo/ (2011)
2. The U.S. Department of Energy: Annual Energy Review 2010, DOE/EIA-0384 (2010), http://www.eia.gov/totalenergy/data/annual/index.cfm (2010)
3. Gutowski, T. G., M. S. Branham, J. B. Dahmus, A. J. Jones and A. Thiriez, D. P. Sekulic: Thermodynamic analysis of resources used in manufacturing processes, Environ Sci and Technol, vol. 43, pp. 1584-1590 (2009)
4. Diaz, N. Redelsheimer, E. and Dornfeld, D.: Energy Consumption Characterization and Reduction Strategies for Milling Machine Tool Use. Glocalized Solutions for Sustainability in Manufacturing 2011, 263-267 (2011)
5. Connors, D. P., Feigin, G. E., and Tao, D. D.: A Queueing Network Model for Semiconductor Manufacturing, IEEE Transactions on Semiconductor Manufacturing, Vol. 9, No. 3, (1996)
6. Prabhu, V. and M. Taisch.: Simulation Modeling of Energy Dynamics in Discrete Manufacturing Systems, in Proceedings of 14th IFAC Symposium on Information Control Problems in Manufacturing (INCOM 2012), Bucharest, Romania, May 23-25 (2012)
7. Prabhu, V., Jeon, H. W., and M. Taisch: Modeling Green Factory Physics – An Analytical Approach, in Proceedings of IEEE CASE 2012 (2012)
8. Graves, S. C.: A review of production scheduling, Operations Research, 29, pp. 646-675 (1981)
9. Kumar, P. R.: Re-entrant lines, Queueing Systems 13, pp. 87-110 (1993)
10. Jaeger, R. C.: Introduction to Microelectronic Fabrication, 2nd edition, Volume V, ISBN 0-201-44494-7, Prentice Hall (2002)
11. Kumar, S. and Kumar, P. R.: Queueing Network Models in the Design and Analysis of Semiconductor Wafer Fabs, IEEE Transactions on Robotics and Automation, Vol. 17, No. 5, 548–561 (2001)
12. Jackson, J. R.: Networks of waiting lines, Operations Research, 5(4), (1957)
13. Baskett, F., Chandy, K. M., Muntz, R. R., and Palacios, F. G.: Open, Closed, and Mixed Networks of Queues with Different Classes of Customers, Journal of the Association for Computing Machinery, Vol. 22, No. 2, pp. 248-260 (1975)
14. Wolff, R. W.: Stochastic Modeling and the Theory of Queues, 1st edition, Prentice Hall, pp. 321-322 (1989)
15. Kuehn, P. J.: Approximate Analysis of General Queueing Networks by Decomposition, IEEE Transactions on Communications, Vol. Com-27, No.1 (1979)
16. Hu, S.-C. and Chuah, Y.K.: Power consumption of semiconductor fabs in Taiwan, Energy, Volume 28, Issue 8, pp. 895–907 (2003)
17. Whitt, W.: The Queueing Network Analyzer, The Bell System Technical Journal, Vol. 62, No. 9 (1983)
18. Daley, D. J.: Queueing Output Processes, Adv. Appl. Prob., 8, pp. 295-415 (1976)