Proceedings of the 2009 Winter Simulation Conference
M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin and R. G. Ingalls, eds.

A METHOD FOR CYCLE TIME ESTIMATION OF SEMICONDUCTOR MANUFACTURING TOOLSETS WITH CORRELATIONS

Raha Akhavan-Tabatabaei

Fitts Department of Industrial and
Systems Engineering
North Carolina State University
Raleigh, NC, USA

Shengwei Ding
J. George Shanthikumar

Department of Industrial Engineering
and Operations Research
University of California
Berkeley, CA, USA

ABSTRACT

This paper proposes a cycle time estimation method for typical toolsets in Semiconductor Fabrication Facilities (Fabs). Due to sophisticated process flows and requirements of the process, queueing models for toolsets can become very complicated and their performance has been unsatisfactory due to the low accuracy of their results. In this paper, we first study the performance of classical queuing models using a high volume manufacturing toolset as our case study and discuss the potential causes for failure of classical models in predicting its cycle time. Then we propose a new approach for estimating the cycle time of toolsets that have inherent correlation between their arrival and service processes. Finally we apply this method to our case study toolset and show that the accuracy of cycle time estimation is improved significantly compared to use of classical queueing models.

1 INTRODUCTION

Reducing cycle time (CT) and improving delivery performance has long been a key focus area for semiconductor manufacturers. Accurate cycle time estimation can play a large role in production planning and scheduling of fabs. The key to efficiently design a fab is to have proper capacity predictions based on accurate cycle time estimations. On the other hand a good estimation of cycle time and understanding the contributing components will help identify the most effective levers to reduce it.

For a fab with relatively stable production rates, we can apply Little's law to estimate the relationship between number of jobs in the system or work in process (WIP) and cycle time. Little's law states that the number of customers or jobs in a system is directly related to the arrival rate and to the amount of time that those jobs spend in the system. Thus cycle time estimation of different process steps can be translated to the WIP that is held in each of these steps. This is important information for fab managers to understand fab production capabilities and thus correctly set up production targets.

However, due to the complicated nature of semiconductor manufacturing systems (SMS) it is not easy to accurately estimate the cycle time for fab toolsets. The process flow of wafers through the fab usually involves hundreds of operations, many of which share the same toolset with a few other operations throughout the sequence of the process flow. This phenomenon is also known as the re-entrant process since lots of wafers visit or re-enter the same toolset multiple times during their life in a fab. Wafer scraps, reworks and lots that are put on hold waiting for test results further complicate the factory operations and make the environment more dynamic and unpredictable.

Usually fabs form a job-shop system where similar tools are grouped together as one toolset to perform similar operations. Most of the SMS toolsets are very complicated in design and their operations involve issues such as lot cascading or batching, setup and layer dedication. Scheduled and non-scheduled tool downtimes also add to the uncertainty of the service provided by these toolsets.

Shanthikumar et al. (2007) discuss the common approaches to cycle time estimation in SMS. They conclude that although simulation is the most common approach in cycle time estimation of SMS it has some inherent shortcomings. It requires enormous amounts of input data, including equipment details, WIP management policies, and product information. The simulation models normally require substantial resources to maintain and update. Based on the nature of the simulation modeling, multiple replications are needed to perform confident statistical analysis. Therefore, it can be a difficult and ex-

978-1-4244-5771-7/09/$26.00 ©2009 IEEE

tremely time consuming method in exploring what-if scenarios and evaluation of managerial decisions' effect on the cycle time.

Compared with simulation, analytical approaches based on Little's law and queueing theory can be much faster in achieving reasonable results. Queueing models make cycle time estimation based on the stochastic analysis of arrival process and service process. This theory has been successfully applied in communication networks, computer systems and manufacturing systems. SMS is one of the most complicated manufacturing systems and raises many interesting problems in modeling and analysis of the queuing systems.

The major factor to make queueing models inaccurate for SMS is the underlying assumptions of these models. In general queueing models assume the arrival process and the service process are independent. However this is not the case in SMS. Some process related factors as well as managerial policies that aim at smoothing the flow of the production cause dependencies between arrival and service process. In this paper we attempt to identify such correlations and incorporate them into our CT estimation procedure.

The remaining of the paper is structured as follows. Various classical queueing models for toolsets are introduced and studied in Section 2. Potential causes for inaccuracy of these models for SMS are discussed in Section 3. A novel approach to CT estimation of SMS toolsets with correlations is proposed in Section 4. In Section 5 we present open questions and potential solutions along with concluding remarks and the future research directions.

2 CLASSICAL QUEUEING MODELS FOR TOOLSETS

A toolset is defined as a group of tools that perform the same or similar operations. The arrival of lots to a toolset forms a stochastic process and inter-arrival times of lots usually follow certain distributions. Based on tool configurations and process requirements, the processing times also form a stochastic process. Jobs that have already arrived but cannot be processed immediately are placed in the waiting queue. Queueing theory applies the stochastic information to estimate the average lot waiting time in such a queue.

In this section, we introduce the classical queueing models for cycle time estimation that are relevant to SMS toolsets. We study the performance of these models through a case study on a typical toolset of a high volume manufacturing fab. This toolset consists of seven similar but not identical tools, also known as heterogeneous tools. The tools are similar in the sense that they all perform the same operations on the wafers. However, they are not identical in their service rates, more specifically in their time to breakdown and repair distributions. This difference is due to several factors including difference in generations or builders of the tools.

The following notations are commonly used in the classical queueing models. Arrival rate is denoted by λ and service rate by μ, while service time or processing time is denoted by s. Furthermore, the ratio of arrival rate to service rate is known as system utilization and is denoted by $\rho = \frac{\lambda}{\mu}$. Hopp and Spearman (2001) present the notion of effective processing time and denote it by $t_e = s / A,$ where A is the availability of the servers and is calculated as $A = m_f /(m_f + m_r)$. In this formulation m_f represents the mean time to failure of the server and m_r is the mean time to repair the server.

Furthermore, in some of these models the squared coefficients of variation (SCV) of service time and inter-arrival time appear. Squared coefficient of variation is defined as the ratio of variance to squared mean of a random variable and is defined as $C^2 = \frac{\sigma^2}{\mu^2}$. Following this formula we denote the squared coefficient variation of service time and effective service time as C_s^2, C_e^2, that of the inter-arrival time as C_a^2 and that of the repair times as C_r^2. Squared coefficient of variation of effective processing time is calculates as $C_e^2 = C_s^2 + (1 + C_r^2)(1 - A)Am_r / s$ (Hopp and Spearman 2001).

2.1 Single Server Systems

We begin with models that are developed for single server queues. To apply these models to a toolset in SMS we need to make the assumption that each tool in the toolset works independent of the other six and therefore we have seven separate queueing systems. We assume each tool has its own queue of lots waiting to be processed. To provide input to the model we make the following assumption. Lots that are eventually processed by a certain tool are assumed to have arrived to a queue designated to that particular tool. This is a deviation from the real world situation where a single queue is formed in front of the toolset and lots are randomly (or by choice) assigned to tools based on their availability.

Tables 1-4 show normalized data from the toolset which are required for CT analysis using the classical queueing models. These parameters are estimated through the analysis of historical performance of the toolset over a six month period in relatively steady state conditions.

Note that all the historical fab data in this paper are normalized to protect the intellectual property of the com-pany that is providing the data. However, the normalization is performed uniformly on all data; hence the outputs of these models are still comparable with each other as well as to the normalized actual fab cycle time.

Table 1: Normalized Actual % WIP and Queue Time Mean and Variance

Tool	% WIP Processed	Actual Queue Time	
		Mean	Variance
Tool 1	24.87%	11.09	373.35
Tool 2	25.39%	11.76	382.75
Tool 3	15.08%	13.55	566.87
Tool 4	7.24%	12.11	305.58
Tool 5	12.01%	19.44	1072.45
Tool 6	8.69%	18.38	813.34
Tool 7	6.72%	18.97	965.69

Table 2: Normalized Actual Mean and SCoV of Inter-arrival Times

Tool	Mean Inter-arrival Time	Squared Coefficient of Variation
Tool 1	0.92	3.02
Tool 2	0.90	2.58
Tool 3	1.51	9.18
Tool 4	3.15	4.43
Tool 5	1.90	3.19
Tool 6	2.62	1.72
Tool 7	3.39	2.17

Table 3: Normalized Actual Mean and SCoV of Service Times

Tool	Mean Processing Time	Squared Coefficient of Variation
Tool 1	0.92	3.02
Tool 2	0.90	2.58
Tool 3	1.51	9.18
Tool 4	3.15	4.43
Tool 5	1.90	3.19
Tool 6	2.62	1.72
Tool 7	3.39	2.17

Table 4: Actual Availability, Effective Utilization and Effective Processing Time

Tool	Availability *(A)*	Effective Utilization *(ρ)*	Effective PT	
			Mean	SCoV
Tool 1	0.80	0.85	0.78	3.08
Tool 2	0.78	0.86	0.77	3.48
Tool 3	0.58	0.88	1.34	14.02
Tool 4	0.44	0.93	2.92	8.24
Tool 5	0.69	0.80	1.53	5.16
Tool 6	0.63	0.92	2.40	2.78
Tool 7	0.41	0.94	3.19	10.77

Single server models with general arrival and service distributions are usually presented as G/G/1 queues where the first G indicates the general distribution of the inter-arrival time and the second show the distribution of service time. Value 1 show that there is a single server in the system. Hopp and Spearman (2001) present a queue time approximation for G/G/1 queues.

$$t_q = \frac{\rho({C_a}^2 + {C_e}^2)}{2(1-\rho)} t_e,$$

Using Tables 1-4 we apply this formula to the toolset of our case study and show the results in Table 5.

Table 5: Comparison of Actual CT and H&S G/G/1 Estimation

Tool	Actual CT (hrs)	H&S G/G/1 CT (hrs)	% Difference
Tool 1	11.09	14.63	32%
Tool 2	11.76	15.17	29%
Tool 3	13.55	120.26	787%
Tool 4	12.11	239.84	1880%
Tool 5	19.44	27.78	43%
Tool 6	18.38	62.92	242%
Tool 7	18.97	337.22	1678%

Buzacott and Shanthikumar (1993) propose three approximations for G/G/1 queues. Two of these formulas are appropriate for the queues that have $C_a^2 \le 2$ and the third approximation is proposed for cases where $C_a^2 \le 1$. Since in the case of SMS toolset C_a^2 is typically much larger than 1 we apply the more appropriate approximation,

$$t_q = \left\{\frac{\rho^2(1+{C_s}^2)}{1+\rho^2 C_s^2}\right\}\left\{\frac{(C_a^2 + \rho^2 C_s^2}{2\lambda(1-\rho)}\right\} + t_e$$

and show the results in Table 6.

Table 6: Comparison of Actual CT and B&S G/G/1 Estimation

Tool	Actual CT (hrs)	B&S G/G/1 CT (hrs)	% Difference
Tool 1	11.09	15.82	43%
Tool 2	11.76	16.12	37%
Tool 3	13.55	130.92	866%
Tool 4	12.11	248.96	1956%
Tool 5	19.44	30.67	58%
Tool 6	18.38	64.19	249%
Tool 7	18.97	340.72	1696%

2.2 Multi-Server Systems With General Arrival and Service Distributions

These models treat the entire toolset as a single stage queueing system with parallel servers and general distributions for arrival and service processes and are denoted as G/G/m queues. To apply the developed approximations for multi-server systems using the data in Tables 1-4 further calculations are necessary. Note that data in those tables are populated for each tool of this toolset individually. However, the G/G/m approximations require certain parameters such as availability, utilization and effective processing time be evaluated for the group of m tools.

To this end we use the weighted averages of these parameters. We define the weight as the percentage of the total WIP that is processed by each individual tool over the observation period. The WIP percentages are shown in Table 1. In the case of this toolset there are two groups of heterogeneous tools which are manufactured by two distinct companies; tools 1-4 belong to the first group 1 and tools 5-7 belong to the second group. Our study shows that dividing the toolset into two queues of G/G/4 and G/G/3 enhances the accuracy of cycle time estimations versus the case of a single G/G/7 queue. Hence in this section all the approximations are done for each group separately.

Table 7: Toolset data for multi-server systems

Tool Group	Inter-arrival Time		Effective Utilization	Effective Processing Time	
	$(1/\lambda)$	C_a^2	ρ	t_e	C_e^2
1-4	0.31	6.29	0.87	1.10	8.80
5-7	0.83	4.29	0.88	2.19	6.73

Hopp and Spearman (2001) propose the following formula to approximate the queue time of the G/G/m queues, the result of applying this formula to our case study toolset are presented in Table 8.

$$t_q = \left(\frac{C_a^{\,2} + C_e^{\,2}}{2}\right)\left(\frac{\rho^{(\sqrt{2(m+1)}-1)}}{m(1-\rho)}\right)t_e$$

Table 8: Comparison of Actual CT and H&S G/G/m Approximation

Tool Group	Actual CT (hrs)	G/G/m CT (hrs)	% Difference
1 (tools1-4)	11.94	13.50	13%
2 (tools 5-7)	18.99	28.36	49%

Buzacott and Shanthikumar (1993) propose the following approximation for the G/G/m queue which is based on the queue time of an M/M/m queue. An M/M/m queue is a multi server system whose arrival and service processes both follow the exponential distributions denoted by M.

$$t_q^{G/G/m} = \frac{C_a^2(1-(1-\rho)C_a^2)/\rho + C_e^2}{2}t_q^{M/M/m}$$

$$\text{where} \quad t_q^{M/M/m} = \left(\frac{1}{m\mu-\lambda}\right)\left(\frac{m^m\rho^m}{m!(1-\rho)}\right)p(0)$$

and *p(0)* denotes the steady state probability that we have zero lots in the queue. Applying this formulation to the data in Table 7 results in the cycle time estimates of Table 9 for each tool group.

Table 9: Comparison of Actual CT and B&S G/G/m Approximation

Tool Group	Actual CT (hrs)	G/G/m CT (hrs)	% Difference
1 (tools1-4)	11.94	9.46	-21%
2 (tools 5-7)	18.99	23.46	24%

2.3 Queueing Approximations for Semiconductor Toolsets

Based on the classical queuing approximations some customized models for cycle time estimation in SMS are developed in the literature. The attempt of such models has been to enhance the estimation accuracy of the classical models by making some modifications. In this section we study the performance of such models that are relevant and applicable to our case study toolset data.

2.3.1 Whitt (1993) GI/G/m Approximation

Whitt's approximation for the expected waiting time in queue of a GI/G/m system is based on the proportional relationship with the exact values for the M/M/m model. Whitt's formula estimates waiting time as follows.

$$t_q^{G/G/m}(\rho, C_a^2, C_e^2, m) \approx \phi(\rho, C_a^2, C_e^2, m)(\frac{C_a^2 + C_e^2}{2})t_q^{M/M/m}$$

where $\phi(\rho, C_a^2, C_e^2, m)$ is given by Whitt (1993) and he concludes that his approximation for the expected waiting time is "fairly accurate because it is relatively robust and extensively studied".

Based on Whitt's approximation for GI/G/m queue Hopp et al (2002) propose an optimized queuing network for capacity planning to support fab design. This model minimizes facility cost required to meet production volume and cycle time targets. They incorporated features specific to semiconductor manufacturing such as batching, re-entrant processes, multi-product classes and machine setups. This model is designed to analyze a network of queues where each node in the networks represents a toolset. To analyze the cycle time of each node or toolset they apply Whitt's approximation (1993). Applying Whitt's approximation to the case study toolset estimate the cycle time of each tool group as shown in Table 10.

Table 10: Comparison of Actual CT and Whitt G/G/m Approximation

Tool Group	Actual CT (hrs)	G/G/m CT (hrs)	% Difference
1 (tools1-4)	11.94	84.76	86%
2 (tools 5-7)	18.99	202.40	91%

2.3.2 Morrison and Martin (2007) Practical Extensions to Cycle Time Approximation for The G/G/m Queue

Morrison and Martin propose practical extensions to the G/G/m queue time approximations to estimate the cycle time with more accuracy. They try to address issues that rise in manufacturing and particularly in SMS such as tools with production parallelism, tools that are idle with work in progress, travel to queue and the tendency of lots to defect from a failed server and return to the queue. Through some examples they show that their approximations can significantly reduce the percentage of estimation error compared to popular intuitive closed form approximations for G/G/m queues.

They propose the Martin Approximation for the cycle time of a G/G/m queue as follows. Applying this formula to the data of our case study toolset approximates the cycle time as shown in Table 11.

$$E(CT) \approx \frac{1}{\mu} + \frac{1}{\mu}(\frac{C_a^2 + C_e^2}{2})\left(\frac{\rho^m}{(1-\rho^m)}\right)$$

Table 11: Comparison of Actual CT and Morrison G/G/m Approximation

Tool Group	Actual CT (hrs)	G/G/m CT (hrs)	% Difference
1 (tools1-4)	11.94	12.83	7%
2 (tools 5-7)	18.99	27.65	46%

3 POTENTIAL CAUSES FOR INACCURACY OF CLASSICAL QUEUEING MODELS

In Section 2, study of classical queueing approximations as well as models that are developed specifically for SMS shows that most of these formulas estimate the toolset cycle time with a relatively larger error and in some cases the magnitude of error is unreasonably large.

We believe this estimation error is due to the fact that the basic assumptions behind such models are far from the reality of semiconductor manufacturing. These assumptions include

1. *The sequence of service times is a sequence of independent and identically distributed random variables.*
2. *Arrival rate, service rate and WIP level are mutually independent.*
3. *Machine breakdown and maintenance occur independent of the WIP level.*

In real manufacturing systems and specifically in SMS these assumptions are violated frequently. One reason is that in real fab operations, in order to reduce queue times and WIP levels, line managers use real-time information collected from the floor to better control the flow of WIP through the entire line. For example, when one or a few tools in a toolset break down, line managers try to make decisions to decrease the arrival rate to the toolset in order to avoid large WIP levels and high queue times. Or when there is WIP build-up at one operation, the operation managers try to postpone the preventative maintenance of the corresponding tools to a later time so that they avoid more congestion. These interventions and adjustments in the queueing system eventually result in lower cycle times than what an independent model would predict.

Another case happens in batch tools or tools that take more time to produce the first unit of product because every time that the tool begins processing a new product type, some testing and qualification procedures are required on the first lot (lead lot) of the new type. If the results of the tests are satisfactory then they can run in large batches. In this case the more WIP of that product they have the more output they can produce which is counter-intuitive and not captured by classical queueing models.

Another case where classical queueing models fail to be accurate is when heterogeneous tools are grouped together in a toolset. These tools can theoretically perform the exact same operations however, some of them have faster processing rates, easier setups, longer times to failure or shorter repair times. The reasons for this variation can be different tool designs by different vendors, older versus newer generations of the same tool and the natural variation between two identical units. In this case if a multi-server model is assumed then the assumption of identical servers is violated. Although splitting tools of similar generation or vendor into subgroups, as we did in Tables 8-9 improves the accuracy of estimation, the error is still not negligible.

Comparing results of single server versus multi-server models in Section 2 shows that treating a toolset as a multi-server queueing system yields better cycle time estimation than separating tools in single server queues. However, it is also noteworthy that some of the multi-server approximations underestimate the cycle time, e.g., tool group 1 in Table 9. We believe that the reason for this underestimation is that in reality even within a specific tool group the machines are not identical in their characteristics and treating them as equally capable servers is a false assumption in the queueing models. For example, Table 4 shows that tools in group 1 are very different in their availability and processing times.

Another contributing factor is that is that a lot may not be able to be processed by all available tools in the toolset at any given time. Issues such as qualification or setup can explain some of this underestimation error too. Tool qualification means that a tool cannot work on one type of product until after a qualification process takes place. This qualification process can take a long time, hence we observe periods that one or more tools are available but not utilized during a busy period. Queueing approximations cannot capture these details.

Also for the consideration of better yield, one tool might be more preferred than the other. Since the models above assume no difference between the tools and treat them as identical, the estimated cycle time can be much lower than reality.

4 PROPOSED MEHODOLOGY FOR CYCLE TIME ESTIMATION OF TOOLSETS WITH CORRELATIONS

This Methodology aims at improving the accuracy of cycle time estimation through incorporating the existing correlations in arrival and service processes into the analysis. This method calls for observing the flow of lots through the toolset over a relatively long period of time (three to six months) and quantifying the existing correlation between the arrival process and service processes. This correlation is then used to generate more accurate forecasts of WIP levels which results in more accurate cycle time estimations through Little's Law. After finding the correlation, this method generates the profile of forecasted WIP based on the established correlation through an iterative algorithm. Then these WIP forecast values are used to estimate the cycle time by applying Little's Law.

4.1 Procedure

For this methodology to generate more accurate results collection of data over a relatively long period of time is necessary. The minimum duration recommended is data from three consecutive months of a system operating in the steady state. To reduce the computational complexity when larger periods of time are studied, the length of the period of study (*study-time*) is recommended to be divided into slots of shorter duration which are referred to as buckets. For example, for a three month period of study dividing the time into two-hour buckets might be appropriate. Appropriate length of the buckets (*bucket-length*) depends on how much data points falls into each bucket and can be found through trial and error.

Step 1 – Data Collection

For each bucket of time the following variables are estimated from historic data:

- Arrival (IN) - Calculate the total number of lots that have arrived to the system during the course of each bucket.
- Throughput (OUT) - calculate the total number of lots that have been processed during each bucket
- Queue Size (WIP) – at the end of each bucket record the WIP level or the total number of lots waiting in the queue at that instance.

Step 2 – Regression and Parameter Estimation

If the length of the buckets are chosen sufficiently short, to quantify the correlation between WIP and service process we need to correlate the WIP of the previous bucket (WIP_Lag) to the output of the current bucket. The reason is that WIP as is calculated through this procedure is a snapshot of the system at the instance when each bucket ends. Hence its effect on the number of outs can only be captured through the following period.

For the WIP_Lag, IN and OUT data we find the range (i.e., Max and Min of WIP_Lag/IN/OUT). It is safe to assume that any integer value within this range is a feasible value for WIP_Lag, IN or OUT value. If sufficient amount of data is collected, for each feasible value of WIP_Lag we should observe several values of OUT. If the data is not sufficient and some values of feasible WIP_Lag have no OUT data then we suggest to group together a few values for WIP and put them into a bin. For example, one can put every five consecutive values of feasible WIP_Lag into one bin. The appropriate bin size can be determined through trial and error on the specific set of data under study. In this case the WIP level for each bin can be represented by the middle point of each bin (e.g., 2.5 can represent the WIP level for a bin that contains values between 0 and 5).

The next step is to find the relationship between WIP_Lag and OUT. To this end we find the average value of OUT data for each WIP_Lag feasible value or WIP_Lag bin value. Then we use these average points for a regression analysis and define Avg_OUT as a quadratic function of WIP_Lag. We also find the standard deviation of OUT data for each bin and call it Std_OUT. Then we define the Std_OUT as a quadratic function of WIP through regression over the Std_OUT of each bin. The input data (IN) can also be expressed as a quadratic function of WIP_Lag only if the historical data suggests that there exist a significant correlation. At the end of this step we should have established the following relationships:

$$\begin{aligned}
&Avg\_OUT = a_1 + b_1 WIP\_Lag + c_1 WIP\_Lag^2\\
&Std\_OUT = a_2 + b_2 WIP\_Lag + c_2 WIP\_Lag^2\\
&Avg\_IN = a_3 + b_3 WIP\_Lag + c_3 WIP\_Lag^2\\
&Std\_IN = a_4 + b_4 WIP\_Lag + c_4 WIP\_Lag^2
\end{aligned} \tag{1}$$

Step 3 – Distribution Fitting

For each value of feasible WIP we can fit a distribution to OUT with the first moment Avg_OUT and the second moment Avg_OUT^2 + Std_OUT^2. Distribution fitting techniques and goodness of fit tests can be employed to find the best fitting distribution on OUT. Also if we find a significant correlation between WIP_Lag and IN a distribution needs to be fitted in a similar fashion.

Step 4 – Iterative WIP Generation Algorithm

1. $n = \dfrac{study - time}{bucke - length}$
2. Set $i = 0$.
3. Choose an arbitrary value of feasible WIP$_i$.
4. For WIP$_i$ calculate the first and second moments of OUT and IN using the regression formulas in Step 2.
5. Generate the corresponding values of OUT$_{i+1}$ and IN$_{i+1}$ by drawing a random sample of each from the fitted distributions in Step 3.
6. If OUT$_{i+1}$ or IN$_{i+1}$ are out of the feasible range repeat 5 until feasible values are reached.
7. Estimate WIP$_{i+1}$ = WIP$_i$ – OUT$_{i+1}$ + IN$_{i+1}$
8. If WIP$_{i+1}$ is out of feasible WIP range repeat 5 until an acceptable value is generated.
9. Set $i = i+1$.
10. If $i \geq n$ then go to 12 else go to 11.
11. Go to 2.
12. Calculate average WIP: $Avg\_WIP = \dfrac{\sum_{i=0}^{n} WIP_i}{n}$
13. Calculate Cycle Time: $CT = \dfrac{Avg\_WIP}{Avg1^*\_IN}$

(* Note that *Avg1_IN* is the average of IN of the collected data in step 1 and not the average IN as a function of WIP_Lag).

This iterative algorithm should be repeated for a sufficient number of replications to give the desired half width around average cycle time.

4.2 Case Study

This methodology is applied to data from the toolset described in section 2. This data is collected over six consecutive months when the fab was operating in relatively steady state. Studying the data shows that choosing buckets with length of two hours reduce the computational intensity while providing enough data in each bucket to establish the correlations. The WIP levels are also divided into bins of size 5. Table 12 shows the values of estimated parameters for the quadratic relationships of Step 2.

Table 12: Estimated Parameters

a	Estimated Value	b	Estimated Value	c	Estimated Value
a_1	5.2947	b_1	0.0798	c_1	-0.0003
a_2	31.3806	b_2	1.3881	c_2	-0.0061
a_3	8.7514	b_3	0	c_3	0
a_4	96.6828	b_4	0	c_4	0

Figure 1 shows the scatter plot of WIP_Lag versus OUT along with the quadratic regression line of the WIP bin averages as described in Step 2. In this figure the green dots show the real data, the red asteroids show the bin averages of OUT and the center line shows the linear fit. The blue curves around the center line represent [mean +/- 1 standard deviation] for the linear fit.

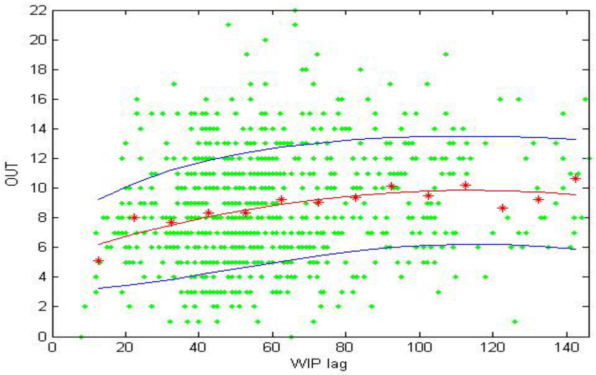

Figure 1: WIP_lag and OUT Relationship

Figure 2 shows the relationship between WIP_Lag and IN and points out that the correlations between IN and WIP_Lag in the case of this toolset is not significant and therefore can be ignored.

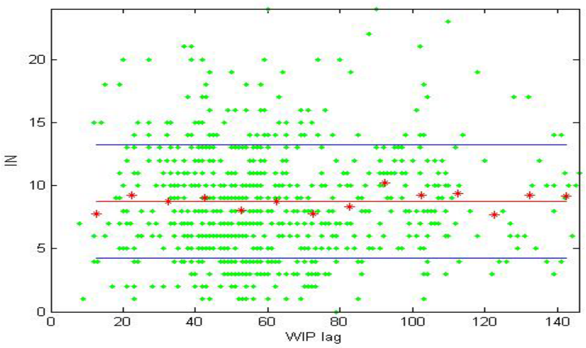

Figure 2: WIP_lag and IN Relationship

Distribution fitting of Step 3 shows that Normal distribution is the best fit for this data. Running 500 replications of the iterative algorithm in Step 4 produces the results in Table 13. The results show that the Flow Analysis Model yields more accurate results than classical queueing models.

Table 13: Results of the iterative algorithm

	Normal Fit		
	WIP	OUT	CT (hrs)
Actual	60.94	8.70	13.87
Sim-Mean	64.06	8.81	14.54
Half Width	3.12	0.15	0.71

Results in Table 13 show that the cycle time estimations through this method are statistically equal to the actual values for WIP and OUT since the confidence interval around the estimated WIP and OUT averages both include the actual values for these parameters. Now that WIP and OUT can be estimated accurately CT can be predicted through Little's Law. Figure 3 shows the accuracy predicting the WIP flow with model. This figure is generated in the same way as Figure 1, except that its data is generated through the iterative algorithm. The similarity between the two graphs shows the accuracy of our approximation approach graphically.

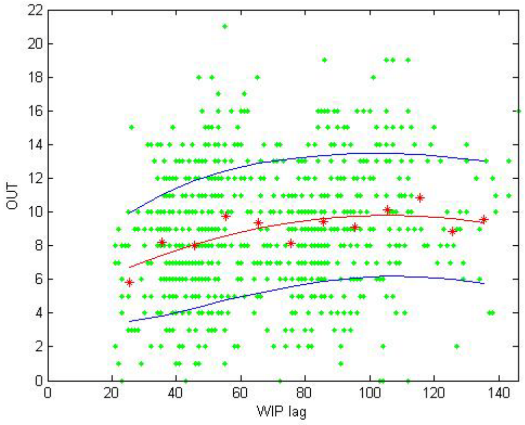

Figure 3: Normal Model

Now that the validity of the Flow Analysis Model is tested through this case study the estimated parameters in Table 12 can be safely used to predict the CT of this toolset given various WIP level situations. However, such predictions are accurate only if other conditions at the toolset such as the number of tools and wafer starts remain unchanged.

This model is more appropriate for strategic capacity planning purposes. The reason is that although the information about tactical availability of each tool is inherent in the OUT variation, this factor is not explicitly addressed. Hence this model is not appropriate for predictions about situations when one or more tools are down for a short period of time like one or a few shifts.

5 CONCLUDING REMARKS

This paper first studies the performance of classical queueing models in estimating the cycle time of a typical toolset in SMS and discusses the potential causes for low accuracy of such models. Then a novel approach to modeling the cycle time of toolsets with correlations in SMS is discussed and the results are compared to the classical models.

This model can be safely used for strategic capacity planning in SMS. However for tactical decision making an extension of this model with explicit tool availability shall be developed. This paper is expected to propose a new approach in queueing modeling for SMS. We hope that, through further investigative efforts, the queueing modeling can be applied to a variety of toolsets for faster and more accurate cycle time estimation and identifying the impacting factory on cycle time. This can eventually lead managers in SMS to make more informed decisions on cycle time reduction.

ACKNOWLEDGMENTS

This work is in part supported by Intel Fab and Sort Manufacturing Research (CNDA #3945776). We extend our appreciation to all of our colleagues in Intel FSM Research for their continuous support. We especially want to thank Erik Hertzler and Dr. Adar Kalir for their valuable insights that helped improve the quality of this work.

REFERENCES

Buzacott, J. A. and J. G. Shanthikumar. 1993. *Stochastic Models of Manufacturing Systems*. Englewood Cliffs, NJ: Prentice Hall.

W. J. Hopp, M. Spearman, S. Chayet, K. Donohue, and E. Gel. 2002. Using an optimized queueing network model to support wafer fab design. *IIE Transactions* 34 (2): 119–130.

Gel, E. S., J. W. Fowler, and K. Khowala. Queueing approximations for capacity planning under common setup rules. working paper.

Hopp, W. J. and M. L. Spearman. 2001. *Factory Physics: Foundations of Manufacturing Management*. New York: McGraw-Hill.

Morrison, J. R. and D. P. Martin. 2007. Practical extensions to cycle time approximations for the G/G/m-queue with applications. *IEEE Transactions on Automation Science and Engineering*. 4 (4): 523-532.

Shanthikumar, G., S. Ding and M. Zhang. 2007. Queueing Theory for Semiconductor Manufacturing Systems: A Survey and Open Problems. *IEEE Trans-actions on Automation Science and Engineering.* 4 (4): 321-335

Whitt, W. 1993. Approximations for the GI/G/m queue. *Production and Operations Management.* 2 (2): 114-161.

AUTHOR BIOGRAPHIES

RAHA AKHAVAN-TABATABAEI is a PhD candidate in the Fitts Department of Industrial Engineering at North Carolina State University. She received her master's degree co-majored in industrial engineering and operations research from the same university. She also works for Intel Corporation as a senior industrial engineers. Her email is <rakhava@ncsu.edu>.

SHENGWEI DING is currently with Leachman and Associates, LLC. When conducting this research, he was a post-doctoral fellow in the department of Industrial Engineering and Operations Research at University of California, Berkeley. He received his Ph.D. degree from University of California, Berkeley in 2004. His research interests include planning, scheduling and queueing analysis of semiconductor manufacturing. His email is <dingsw@cal.berkeley.edu>.

J. GEORGE SHANTHIKUMAR is a Professor in the department of Industrial Engineering and Operations Research at the University of California, Berkeley. He received the M.S. and Ph.D. degrees in industrial engineering from the University of Toronto, Toronto, ON, Canada. His email is <shanthikumar@ieor.berkeley.edu>.