Performance Evaluation of Serial Photolithography Clusters: Queueing Models, Throughput and Workload Sequencing

James R. Morrison
Engineering and Technology Dept.
Central Michigan University
Mount Pleasant, MI, USA
morri1j@cmich.edu

Beverly Bortnick
Integrated Supply Chain
IBM Corporation
Essex Junction, VT, USA
nickbort@us.ibm.com

Donald P. Martin
Productivity Characterization
IBM Corporation
Essex Junction, VT, USA
martindp@us.ibm.com

Abstract

For clustered configuration of a photolithography toolset, operating under a scheduling policy inducing serial processing, measures of system performance are deduced. Queueing models demonstrate that, due to the parallelism inherent in the system configuration, the normalized cycle time behavior is different than that of the standard single server queue. Cluster throughput is evaluated based on measures of the frequency and magnitude of events common in manufacturing operation. It is shown that the maximum throughput of a serial photolithography cluster tool is not influenced by the order in which two classes of lots with different wafer processing speeds are processed.

Keywords

Cluster tools, queueing models, photolithography performance evaluation, throughput, cycle time, workload sequencing.

I. INTRODUCTION

Cluster tools are common in the manufacture of semiconductor wafers and consist of a collection of processing modules, at which wafers may receive processing, grouped into a single chassis. Wafer processing at a cluster tool proceeds as follows: wafers queue for processing, enter the tool, receive processing from one or more of the modules within the tool in a prescribed order for prescribed lengths of time (dependent upon the wafer) and then exit the cluster tool. When multiple wafers are present in a cluster tool, there may be contention for the cluster's resources. One extreme of the class of possible cluster tool configurations is the serial processing cluster in which each module provides distinct processing and each wafer proceeds from one module to the next in sequential order. Our special toolset of motivation is the photolithography cluster, which is typically selected as the bottleneck toolset in semiconductor manufacturing, is prohibitively expensive to acquire and is often operated as a serial processing cluster.

As an increasing number of cluster tools are employed in semiconductor fabrication facilities, the need to design, control and evaluate the cycle time and throughput performance of such tools increases. There has been interest in the characterization of cluster tool throughput performance and the development of scheduling algorithms to control cluster tools for over a decade. The authors in [1] and [2] explore throughput models, [3] conduct analysis of dual and single blade robots, and [4], [5], [6] and [7] evaluate model features such as redundant chambers, systems of cluster tools (with and without reliable chambers), and modifications to module process parameters. Among others, simulation approaches to performance evaluation have been pursued in [8] and [9]. The modeling of cluster tools in the context of an entire semiconductor fabrication facility is considered in [9]. Perhaps the first use of Petri nets to study cluster tool performance appears in [10]. In [11] and [12], control and performance evaluation via a Petri net approach are conducted. Other approaches to the scheduling and control of cluster tools may be found in [13], [14], [15] and [16]. The measurement of cluster system behavior in a manufacturing environment is discussed in [17] and [18].

For elementary models of serial processing cluster tools, the following contributions, which have not appeared in the literature, are presented:

- Explicit formulae for the cycle time performance with a stochastic wafer arrival process (Theorem 1).
- Throughput analysis accounting for empty module and idle tool events (Theorems 2 and 3).
- Throughput analysis accounting for wafer processing speed diversity (Theorem 4 and Corollary 1).

In Section II, several parameters essential to the results of this paper are recalled, for a basic model of a photolithography cluster tool as in [1] and [2]. In Section III, queueing models are discussed which enable the evaluation of the expected cycle time performance. In Section IV, the throughput of a photolithography cluster tool in the presence of practical events not explicitly accounted for in our queueing models is determined. Concluding remarks are presented in Section V.

II. SYSTEM DESCRIPTION

A serial processing photolithography cluster consists of a sequence of M processing modules or stations (servers, in the queueing parlance), m_1, m_2, ..., m_M, which include the processing modules for resist application, pre-expose baking of the resist, the stepper staging and expose operations and post expose baking. To begin, it is assumed that there is only one class of wafer, so that all wafers require the same steps and processing durations. The processing time (including transport time and load and unload of the wafer) for a wafer in module m_j is deterministic with duration Δ_j. Let $\Delta = \max_i\{\Delta_j\}$ denote the production time of the slowest module(s). The effect of a limited number of loading ports is not considered.

When there are sufficient wafers in queue and production is uninterrupted, independent of the individual module processing durations, the rate at which wafers complete production is one wafer every Δ time units. This is because Δ is the bottleneck duration within the cluster. Justified by the fact that throughput is dictated by the duration Δ alone, the model utilized below simplifies the general model and assumes that $\Delta_j = \Delta$ for all modules m_i, $i = 1, \ldots, M$. Throughput evaluation is not influenced by this simplification, though minor cycle time deviations may result.

Wafers arrive to the serial processing photolithography cluster in lots (batches) consisting of W wafers (W is a positive integer) as a Poisson process of rate α. Lots are processed in a first come first served fashion (this assumption can be relaxed). Processing begins for a lot when its first wafer enters the first module in the photolithography cluster. Upon receiving processing for Δ time units from a module m_i a wafer next proceeds to module m_{i+1}, unless $i = M$ in which case the wafer exits the clustered tool. Wafers proceed sequentially through the cluster tool so that once the w^{th} wafer receives processing from the first module it proceeds into the second stage of processing at m_2. At that time the $w+1^{th}$ wafer enters module m_1 and begins production. Figure 1 depicts the M stages of processing and the flow of wafers.

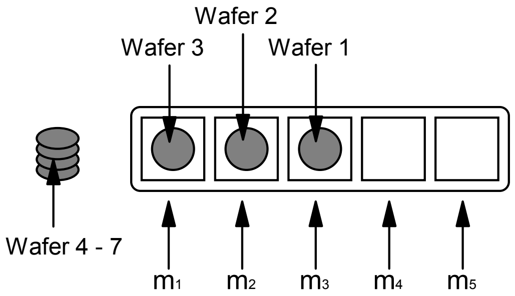

Figure 1. A five module (M = 5) serial processing cluster has three wafers of a seven wafer lot (W = 7) in production.

When the last wafer of a lot exits the first module, the first wafer of the subsequent lot enters module m_1 and begins processing, unless the queue is empty in which case module m_1 remains empty. Processing for a lot is complete when its last wafer exits the final module. Hereafter, the serial processing nature of the photolithography cluster modeled is assumed without mention.

The model detailed here can also incorporate the presence of multiple paths for wafers at a given stage of processing (multi-path processing) in which, for example, three modules may be devoted to one (bottleneck) process to increase throughput. If there are N modules devoted to a single stage of processing requiring δ time units, model the multiple path stage of processing as N modules with δ/N time units of processing each (which is then replaced with the bottleneck processing time Δ, as mentioned above).

The processing time for a *lot*, denoted as RPT (for raw process time) is the time required for the first wafer to traverse all modules plus the time for the remaining (W-1) wafers to exit the tool. Thus,

$$RPT = (M + W - 1)\Delta .$$

When a sufficient number of lots are queued for processing, and the tool is filled with wafers, a lot will exit the tool every WΔ units of time. Similar to [17], the minimum time between lot completions ($TBLC_{MIN}$) from the tool is thus WΔ. The maximum throughput rate of the tool λ_{MAX} in wafers per unit time is

$$\lambda_{MAX} = \frac{W}{W\Delta} = \frac{1}{\Delta} .$$

The measured throughput (obtained as the number of wafers completed divided by the time that at least one wafer of a lot is in the cluster) may not coincide with the maximum throughput for many reasons. For example, the inability or failure to load lots which results in the presence of a few empty modules will reduce the measured throughput as some portion of the cluster will be in production but it will not be fully utilized. Thus, while the tool is in production and appears busy, its capacity is not fully utilized. When a serial processing tool is fully utilized, [17] defines the maximum parallelism as

$$\|_{MAX} = RPT / TBLC_{MIN} = (M + W - 1) / W .$$

The maximum parallelism is only achieved when the tool is fully utilized and represents the mean number of lots that are receiving service from the tool.

III. QUEUEING MODELS

Using the model described in Section II, one may formalize the system as a network of tandem queues with deterministic processing times, a Poisson batch arrival process and blocking (the number of wafers in a module may not exceed one). Application of what [19] refers to as intermediate queueing theory (essentially the Pollaczek-Khinchin formula), enables us to obtain the expected cycle time behavior of a cluster tool. To our knowledge, this is the first application of such techniques to this problem. Theorem 1 provides the result without proof.

For reference, recall that the expected cycle time behavior for lots in an M/D/1 queue is given as

$$E[CT] = \frac{1 - \rho / 2}{1 - \rho} RPT ,$$

where $\rho = \alpha/\mu$ is the toolset loading, α is the arrival rate of lots, μ is the server process rate (lots per unit time) and RPT is the raw processing time of a lot.

Theorem 1: Mean cycle time behavior. *The steady state expected cycle time E[CT] for lots of W wafers arriving to a serial processing photolithography cluster via a Poisson arrival process of rate α with the requirement of deterministic processing time Δ for each wafer at each of the M processing modules is*

$$E[CT] = \frac{1-\rho[1-1/(2\cdot ||_{MAX})]}{1-\rho} RPT\ ,$$

where, $\rho = \lambda(W\Delta)$ is the loading on the toolset, $||_{MAX} = (W+M-1)/W$ and $RPT = \Delta(W+M-1)$. The steady state expected percent of time that the cluster is busy (stated as the proportion of time that at least one wafer is in production in any module) is

$$1-(1-\rho)e^{-\rho(M-1)/W}\ .$$

Figure 2 depicts the mean cycle time performance predicted by the theorem as a function of loading in comparison to the M/D/1 queue. Observe that the standard M/D/1 normalized cycle time performance is inferior to that of the serial processing cluster tool. This is a consequence of the fact that the queueing of lots to enter the cluster tool depends upon the throughput rate of wafers as opposed to the raw process time RPT (which is used for the normalization).

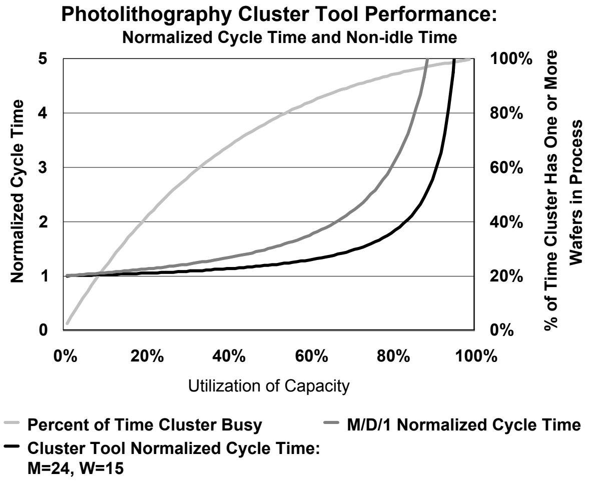

Figure 2. Cycle time performance for a cluster tool with 24 modules and 15 wafers per lot.

IV. THROUGHPUT AND WORKLOAD SEQUENCE

To incorporate events of practical importance such as tool interrupts and overheads unaccounted for in the model of Section II, the maximum throughput in the presence of disturbances is considered. The paper examines three classes of disturbances: empty modules, idle tool events and bottleneck module speed diversity.

Consider the maximum throughput in the presence of lots with different bottleneck module process speeds (i.e., a different Δ possible for each *lot*). First, assume there are N different classes of lots, with a lot consisting of W wafers independent of class. To each class i associate the (bottleneck) module processing time Δ_i (so that lots of different classes have different maximum throughput rates). Suppose that no module is ever wanting for a wafer (i.e., the tool is never short of a supply of lots) and that over a long horizon the fraction of lots of class i processed by the tool is given as f_i (with no $f_i = 0$). An upper bound on the maximum throughput, denoted as λ_{MAX}, is readily calculated as

$$\lambda_{MAX} = \frac{1}{\sum_{i=1}^{N} f_i \Delta_i} = \frac{1}{\underline{\Delta}},$$

where $\underline{\Delta} = \sum_{i=1}^{N} f_i \Delta_i$ denotes the mean module processing time.

IV.1. Empty Modules

Empty modules can occur for many reasons. Examples include the presence of a monitor wafer between product lots (so that one or more modules are empty of product) and processing interruptions which cause wafers before a certain module to temporarily cease movement. Figure 3 depicts an empty module event.

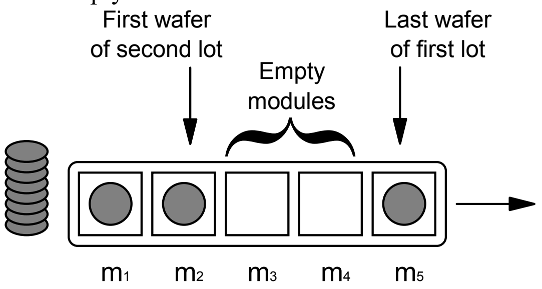

Figure 3. An empty module event with $t_{em}=2\Delta$.

Consider the system with a single class of lots with wafer processing time per module given by $\underline{\Delta}$. Let f_{em} and t_{em} denote the proportion of lots which follow an empty module event and the average duration of each empty module event, respectively. The duration of an empty module event is measured as the time between the completion of the last wafer of the previous lot and the exit of the first wafer of the subsequent lot from the cluster. Let the duration of each empty module event be an integer multiple of Δ (though the average need not be). Assume that at all times at least one module is processing a wafer (the case when a tool idles completely is covered later). Theorem 2 characterizes the throughput in the presence of empty module events.

Theorem 2: Throughput with empty modules. *Let λ be defined as*

$$\lambda := \lim_{t\to\infty}\frac{D(t)}{t},$$

where the variable D(t) is the number of wafers to have completed processing in the time period [0,t). Then

$$\lambda = \lambda_{MAX}\,\frac{1}{1+f_{em}t_{em}/(\underline{\Delta}W)} = \frac{1}{\underline{\Delta}+f_{em}t_{em}/W}.$$

In a manufacturing system, the data required to calculate $\underline{\Delta}$, f_{em} and t_{em} are (not necessarily readily) available from the logistics data bases, the internal tool data bases and process recipe databases. Note that the value of this approach is not in calculating the achieved throughput rate (which is more readily obtained from wafers out/production time) but rather in identification of the impact of the frequency and duration of empty module events to the overall throughput.

IV.2. Idle Tool Events

An idle tool event occurs every time the entire photolithography cluster tool is idle. When a lot begins processing following an idle event, the toolset must first be filled before a single wafer completes processing. Idle tool events are sometimes referred to as flush and fill events and can occur for reasons such as a lack of product or a tool failure. Let the measured throughput λ_{MEA} be defined as

$$\lambda_{MEA} := \lim_{t\to\infty}\frac{D(t)}{P(t)},$$

where D(t), as before, is the number of lots to complete processing in [0,t) and P(t) is the total time that there is at least one wafer in process on the cluster tool in time [0,t). Theorem 3 characterizes the throughput with idle tool events and Figure 4 depicts specific examples.

Theorem 3: Measured throughput and idle tool events. *Assume there is a single wafer processing time $\underline{\Delta}$ and all processed lots either begin processing on an idle tool or enter processing when the first module initially becomes available immediately following the last wafer of the preceding lot. With λ_{MEA} defined as above,*

$$\lambda_{MEA} = \frac{\lambda_{MAX}}{1+f_{idle}(||_{MAX}-1)},$$

where f_{idle} is the fraction of lots that begin processing on the tool when it is idle and $||_{MAX} = (M+W-1)/W$.

IV.3. Bottleneck Module Speed Diversity

The final event explored is the presence of wafers with different bottleneck processing duration requirements. As wafers move sequentially through the processing modules and only advance when the wafer ahead advances, slower moving wafers will decrease the rate at which a wafer with a need for less processing may move through the tool. Thus, a slow lot preceding a fast one may slow the rate at which the fast lot can move through the tool.

To formalize the discussion of bottleneck module speed diversity, assume that there can be no more than two lots on the tool at a given time (i.e., $W \geq M-1$) and that there is a single bottleneck module, denoted as the Bth module, which induces the reduced throughput for slower lots. That is, the time in a module for all wafers in the cluster is dictated by the wafer processing duration of the lot with a wafer in module B. Let Δ_i denote the wafer processing time required from module B for a lot of class i. Though it is assumed that all modules remain full, one can easily allow for a default wafer processing time when B is empty. Such behavior is quite common in practice as the robot control for many serial photolithography cluster tools enforces the restriction that all wafers advance at the rate dictated by the slowest module in process. Let there be N classes of wafers.

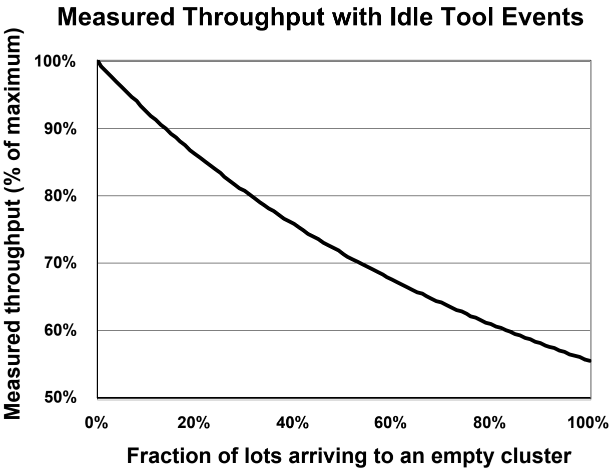

Figure 4. Impact of idle tool events when $||_{MAX}$=1.8.

Theorem 4: Throughput and workload sequencing of lots with diverse bottleneck module speeds. *Let f_{ij} denote the fraction of class i lots that follow class j lots and assume that there are no empty module events or idle tool events. With λ defined as in Theorem 2,*

$$\lambda = \frac{\lambda_{MAX}}{1+\dfrac{(M-B)}{\underline{\Delta}W}\sum_{i=1}^{N}\sum_{j>i}(\Delta_j-\Delta_i)(f_{ij}-f_{ji})}.$$

Corollary 1: Invariance of throughput to workload sequencing. *For any lot processing sequence satisfying $(f_{ij}\text{-}f_{ji}) = 0$, the throughput which can be achieved is*

$$\lambda = \lambda_{MAX} = \frac{1}{\Delta}.$$

Corollary 1 states that, for example, if one only has two classes of lots (with f_1 and f_2 not equal to 0), the impact of diversity is captured in the calculation for the average throughput λ_{MAX}. There is no additional degradation due to the order in which the lots are processed (note that a setup time when switching between lots of different classes is not included). This is because every time there is a fast lot slowed by a slow lot in front of it, there must also be a slow lot which increases speed when its final wafer passes the B^{th} module and the faster speed of the lot following begins to dictate the wafer pace. Note that the formula of Theorem 4 also gives insight into λ in the presence of idle tool events as f_{ij} may not equal to f_{ji} when idle tool events are present.

IV.4. Example of Method

Consider a photolithography cluster tool consisting of eleven stages of processing, one of which consists of a triple path module, with each stage requiring 60 seconds of processing excepting the triple path module which requires 150 seconds. To model the system let the number of modules be thirteen (M=13) - one for each location a wafer may reside - and determine the bottleneck throughput rate (λ_{MAX}), which is the maximum rate at which wafers can exit the cluster tool, as Δ = max(60,150/3) = max(60,50) = 60 seconds. Let there be fifteen wafers per lot (W=15). For this case, $||_{MAX}$ = (M+W-1)/W = (13+15-1)/15 = 1.8.

If one in five lots arrive to an empty tool (f_{idle}=0.2) due to maintenance generated flush and fill events, the maximum throughput is reduced by the multiplicative factor 1/(1+0.2*(1-1.8))=1/1.16=0.862.

If all lots arrive to a fully busy tool but there is a setup associated with the starting of each lot equal to 6*Δ (corresponding to the need to flush say six modules in the pre-expose portion of the track due to recipe contention issues), this can be modeled as an empty module event. Since all lots face these empty modules, f_{em}=1.0 and t_{em}=6*Δ=360 seconds. Thus, the maximum throughput is degraded from 1/Δ by a multiplicative factor of 1/(1+1.0*360/(60*15)) = 1/1.4 = 0.714.

Finally, consider a completely filled tool facing two classes of lots. Let the bottleneck module correspond to the stepper in say module seven, so that B=7. Let Δ_1=60 seconds, Δ_2=90 seconds, f_{12}=f_{21}=0.5. Then, by Corollary 4, the throughput is λ_{MAX} . Had there been idle instances, there might have appeared to be a speed loss, but the maximum throughput potential of the tool is not degraded unless setups are incurred. In the event of setup losses, one should apply Theorem 2.

V. CONCLUDING REMARKS

To better understand the cycle time and throughput performance of photolithography cluster tools the paper studied a model intended to provide us with a background for understanding essential performance evaluation concepts for photolithography cluster tools. Queueing models enabled the deduction of the normalized cycle time performance as a function of system loading. It was demonstrated that the maximum lot parallelism was a natural part of the expression for performance.

To incorporate features of practical importance in manufacturing, events that detract from the throughput performance of serial processing cluster tools were considered. Explicit throughput formulae for three classes of events encompassing many issues of practical importance were developed.

REFERENCES

[1] T. L. Perkinson, P. K. McLarty, R. S. Gyurcsik and R. S. Calvin III, Single-wafer cluster tool performance: "An analysis of throughput," *IEEE Transactions on Semiconductor Manufacturing*, Vol. 7, No. 3, pp. 369-373, August 1994.

[2] S. C. Wood, "Simple performance models for integrated processing tools," *IEEE Transactions on Semiconductor Manufacturing*, Vol. 9, No. 3, pp. 320-328, August 1996.

[3] S. Venkatesh, R. Davenport, P. Foxhaven and J. Nulman, "A steady-state throughput analysis of cluster tools: Dual-blade versus single-blade robots," *IEEE Transactions on Semiconductor Manufacturing*, Vol. 10, No. 4, pp. 418-424, November 1997.

[4] T. L Perkinson, R. S. Gyurcsik and P. K. McLarty, "Single-wafer cluster tool performance: An analysis of the effects of redundant chambers and revisitation sequences on throughput," *IEEE Transactions on Semiconductor Manufacturing*, Vol. 9, No. 3, pp. 384-400, August 1996.

[5] M. J. Lopez and S. C. Wood, "Systems of multiple cluster tools: Configuration and performance under perfect reliability," *IEEE Transactions on Semiconductor Manufacturing*, Vol. 11, No. 3, pp. 465-474, August 1998.

[6] J. W. Hermann, N. Chandrasekaran, B. F. Conaghan, M. Q. Nguyen, G. W. Rublof and R. Z. Shi, "Evaluating the impact of process changes on cluster tool performance," *IEEE Transactions on Semiconductor Manufacturing*, Vol. 13, No. 2, pp. 181-192, May 2000.

[7] M. J. Lopez and S. C. Wood, "Systems of multiple cluster tools: Configuration, reliability and performance," *IEEE Transactions on Semiconductor Manufacturing*, Vol. 16, No. 2, pp. 170-178, May 2003.

[8] H. T. LeBaron and R. A. Hendrickson, "Using emulation to validate a cluster tool simulation model," Proceedings of 2000 Winter Simulation Conference, Vol. 2, pp. 1417-1422, December 2000.

[9] S. T. Shikalgar, D. Fronckowiak and E. A MacNair, "Application of cluster tool modeling to a 300mm fab simulation," Proceedings of the 2003 Winter Simulation Conference, Vol. 2, pp. 1394-1397, December 2003.

[10] R. S. Srinivasan, “Modeling and performance analysis of cluster tools using Petri nets,” *IEEE Transactions on Semiconductor Manufacturing*, Vol. 11, No. 3, pp. 394-403, August 1998.

[11] Ja-Hee Kim, Tae-Eog Lee, Hwan-Yong Lee and Doo-Byeong Park, “Scheduling analysis of time-constrained dual-armed cluster tools,” *IEEE Transactions on Semiconductor Manufacturing*, Vol. 16, No. 3, pp. 521-534, August 2003.

[12] W. M. Zuberek, “Cluster tools with chamber revisiting – modeling and analysis using timed Petri nets,” *IEEE Transactions on Semiconductor Manufacturing*, Vol. 17, No, 3, pp. 333-344, August 2004.

[13] H. L. Oh, “Reducing complexity of wafer flow to improve quality and throughput in a single-wafer cluster tool,” Proceedings of the 24th IEEE/CPMT Electronics Manufacturing Technology Symposium, pp. 378-388, 1999.

[14] Yong-Jae Joo and Tae-Eog Lee, “Virtual control – A virtual cluster tool for testing and verifying a cluster tool controller and a scheduler,” *IEEE Robotics and Automation Magazine*, Vol. 11, No. 3, pp. 33-49, September 2004.

[15] S. Rostami and B. Hamidzadeh, “An optimal residency-aware scheduling technique for cluster tools with buffer module,” *IEEE Transactions on Semiconductor Manufacturing*, Vol. 17, No. 1, pp. 68-73, February 2004.

[16] S. Venkatesh and J. S. Smith, “An evaluation of deadlock-handling strategies in semiconductor cluster tools,” IEEE Transactions on Semiconductor Manufacturing, Vol. 18, No. 1, pp. 197-201, February 2005.

[17] K. Butler and J. Matthews, “How differentiating between utilization of effective availability and utilization of effective capacity leads to a better understanding of performance metrics,” 2001 IEEE/SEMI Advanced Semiconductor Manufacturing Conference, pp. 21-22, 2001.

[18] K. Connerney, D. Martin and R. Tomka, “Determining the capacity components of different classes of multichamber tools,” 2001 IEEE SEMI Advanced Semiconductor Manufacturing Conference, pp. 29-32, 2001.

[19] L. Kleinrock, Queueing Theory, Volume 1: Theory, John Wiley - Interscience, New York, N.Y., 1975.

BIOGRAPHY

James R. Morrison holds a Ph.D. degree in Electrical and Computer Engineering from the University of Illinois at Urbana-Champaign. He works as an Assistant Professor of Electrical Engineering at Central Michigan University.

Beverly Bortnick holds a BS in Industrial Engineering and an MS in Manufacturing Management from the Rochester Institute of Technology. She works as a Staff Engineer for the Integrated Supply Chain at the IBM Corporation.

Donald P. Martin received his Ph.D. degree in Chemical Engineering from the Massachusetts Institute of Technology (MIT). He works as a Senior Technical Staff Member in Industrial Engineering at the IBM Corporation.