Cycle Time Approximations for the G/G/m Queue Subject to Server Failures and Cycle Time Offsets with Applications

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

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

Abstract

Approximate queueing formulae are often employed for the practical evaluation of manufacturing system performance. Common approximations do not fully address practical issues such as idle tools with work in queue, travel time between stages of production, removal of lots from queue pending process issue resolution and the tendency of lots to defect from a failed server in favor of an equivalent available server. In this paper, approximate queueing formulae are proposed which extend popular existing formulae. To test the quality of the proposed approximations, data from production toolsets in IBM's 200mm semiconductor manufacturing fabricator is considered. It is demonstrated that the approximations perform well on the toolsets studied.

Keywords

Approximate performance evaluation, G/G/m queue, cycle time, server failures, cycle time offsets, hold time.

I. INTRODUCTION

Queueing, or waiting in line for an available server, is a common phenomenon in many industries including computing, manufacturing and customer service and has been studied for perhaps 100 years (Erlang published his first paper on queueing theory in 1909, [1]). Simple closed form approximation formulae for the mean cycle time behavior of the G/G/m queue (potentially subject to server failures) have been developed and employed over the course of decades to predict, understand and plan for manufacturing system behavior, see as a start [2, 3, 4] and the references contained therein. Rough cut approximations have been employed, as in [5], in the performance evaluation of networks of queues by the use of approximate coupling formulae intended to capture the interaction between the arrival and departure processes within the network. Approximations are often used in lieu of exact numerical procedures or performance bounds. The intuitive value of simple expressions for system performance coupled with their ease of use has resulted in their ascendance as the primary tool for understanding cycle time behavior in IBM's 200mm semiconductor wafer fabricator. Much work has been devoted to successfully bridging the gap between measures of actual system performance and the predictions obtained from the M/D/1 (or the more generic G/G/m) queue [6, 7, 8, 9, 10, 11].

While these simple approximations can serve well to predict the behavior of many manufacturing workstations, they may fall short in certain applications. When tool failure behavior does not imply that a lot (our term for a unit of work or product) must necessarily remain with a failed tool, the suggested approximations of [4] may not perform well. Further, cycle time offsets which are essentially independent of the queueing at the workstation, such as hold time (during which, for example, a lot is removed from production until a process issue is resolved) and travel from one workstation to another, can substantially alter the cycle time performance.

This paper develops closed form expressions to serve as simple approximations to the cycle time behavior of G/G/m queues subject to server failures and cycle time offsets which address many of the issues associated with existing formulae. By application of the independence of queue lengths in a Jackson network [12] (or a BCMP network), it is inferred that events such as travel and holds upstream of a workstation result in a cycle time offset. The offset is equivalent to the average time spent in travel or hold and independent of the queue length at the production workstation. For workstations subject to server failures, under the assumption that lots need not remain at the server with which they initially entered service, residual life arguments lead to approximation formulae for the mean cycle time behavior.

Further, once we have developed appropriate extensions to existing approximation formulae, we employ the approximations obtained to understand the cycle time performance of multi-server workstations in IBM's 200mm semiconductor wafer fabricator. The result is that we are able to quite accurately predict the cycle time performance of toolsets not previously amenable to such analysis using measured statistics such as the squared coefficient of variation of the interarrival times to the workstation C_A^2, the average hold and travel time of a lot, the time a tool is idle in the presence of available WIP (often termed idle with WIP, operator loading loss or deployment loss), tool availability and loading (utilization of capacity, or throughput achieved divided by throughput potential).

The paper is organized as follows. Section II develops an approximation formula for the mean cycle time in a G/G/m queue subject to server failures drawn from a consideration of the approximations of [2, 4]. Here, idle instances in which work is present (idle with WIP, see [6]) are incorporated. The presence of hold time and travel time prior to arrival at the server is considered in Section III. Section IV extends the approximations of Section II to the case where lots (units of work) are free to defect to another server in the event that their server fails. Residual life arguments are central to the development. Section V highlights the application of the approximations to two toolsets operating at IBM's 200mm semiconductor wafer fabrication facility. Section VI presents concluding remarks and suggests further directions for continued work on this subject.

II. CYCLE TIME APPROXIMATIONS FOR THE G/G/m QUEUE AND PRACTICAL EXTENSIONS

A G/G/m queue (described in, for example, [4, 12]) is depicted in Figure 1 and consists of a waiting room to which lots (unit of work) arrive and m identical servers from which lots receive service. The interarrival times between lots are given by a random variable with general distribution and mean 1/λ. The service times are dictated by a random variable with general distribution and mean 1/μ. All interarrival and service times are independent. The waiting room is of infinite size (there is no blocking) and customers are served in a first-come first-served manner (FCFS). Each server caters to only one customer at a time and devotes all of its resources to completing the transaction. If a server is idle it will immediately begin to serve a customer from the queue (if one is available and not in service with another).

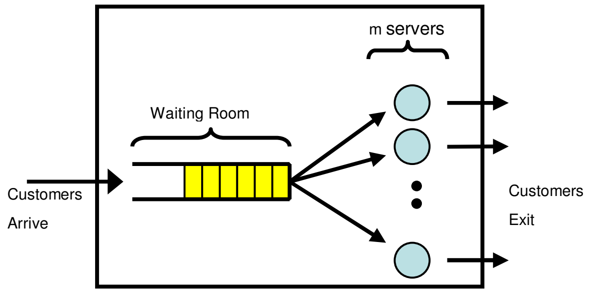

Figure 1. A depiction of a multiserver queue.

One important measure of system performance for a G/G/m queue is the expected cycle time, defined as the mean time that a lot spends in queue and receiving service. Drawing upon the work of [2], [4] proposed the following approximation for the expected cycle time in a G/G/m queue

$$E(CT) \approx \frac{1}{\mu} + \frac{1}{\mu}\left(\frac{C_S^2 + C_A^2}{2}\right)\left(\frac{\rho^{-1+\sqrt{2m+2}}}{m(1-\rho)}\right),$$

where $\rho = \lambda/(m\mu)$, C_S^2 is the squared coefficient of variation of the service time defined as $C_S^2 = \sigma_S^2/(1/\mu)^2$ (σ_S represents the standard deviation of the service time) and C_A^2 is the squared coefficient of variation of the arrival time defined as $C_A^2 = \sigma_A^2/(1/\lambda)^2$ (σ_A represents the standard deviation of the interarrival time). The approximation agrees exactly for the M/G/1 queue [4, 12].

Consider now that a server is subject to random failures. Once functional, the time until a failure occurs is exponentially distributed with mean m_F. Once a failure occurs, the time until repair is generally distributed with mean m_R, standard deviation σ_R and coeffiecient of variation C_R ($C_R = \sigma_R / m_R$). The mean availability of the server is denoted $A = m_F/(m_F + m_R)$. For a G/G/1 queue subject to server failures as just detailed, in which service is resumed (no work is lost) following a failure and in which a lot cannot defect to another server (there is only one), [4] suggested the following practical approximation:

$$E(CT) \approx \frac{1}{\mu^*} + \frac{1}{\mu^*}\left(\frac{C_{S,E}^2 + C_A^2}{2}\right)\frac{\rho^*}{1-\rho^*}.$$

Here, $\rho^* = \lambda/(m\mu A)$, $0 < \rho^* < 1$ ($m = 1$ in this case), and the following effective parameters incorporate the behavior imposed by the failures

$$\mu^* \doteq \mu A,$$

$$\sigma_{S,E}^2 \doteq \frac{\sigma_S^2}{A^2} + \frac{(m_R^2 + \sigma_R^2)(1-A)}{A m_R \mu},$$

$$C_{S,E}^2 \doteq C_S^2 + (1 + C_R^2)(1-A) A m_R \mu.$$

For the M/G/1 queue subject to server failures as described, this approximation is exact.

A natural generalization suggested by the above two formulae is the following approximation for a G/G/m queue subject to server failures as above:

$$E(CT) \approx \frac{1}{\mu^*} + \frac{1}{\mu^*}\left(\frac{C_{S,E}^2 + C_A^2}{2}\right)\frac{(\rho^*)^{\sqrt{2(m+1)}-1}}{m(1-\rho^*)}. \qquad (1)$$

All variables in this approximation have been defined previously. A simpler and more intuitive approximation, hereafter referred to as the Martin Approximation, suggests another form for the loading terms (containing ρ*):

$$E(CT) \approx \frac{1}{\mu^*} + \frac{1}{\mu^*}\left(\frac{C_{S,E}^2 + C_A^2}{2}\right)\frac{(\rho^*)^m}{1-(\rho^*)^m}.$$

Figure 2 compares the mean cycle time approximations to the exact mean cycle time for an M/M/2 queue (the M/M implies exponential interarrival and service times) subject to exponential server failure and repair. The failure parameters are $m_F = 16$ hours, $m_R = 4$ hours and $1/\mu = 1$ hour. The Factory Physics Style approximation refers to the approximation of equation (1). The approximations perform relatively well.

A matter of practical import is idle with WIP (work in process), discussed in [6]. To model this phenomenon, suppose that each lot experiences a random delay prior to loading (for the first time) with mean Ω and standard deviation σ_Ω, so that the server upon which the lot is being loaded is idle. Consider these loading events to be independent of all other random events. The additional time may be considered as *production time* and incorporated into the service distribution. By doing so, the capacity loss associated with the idle with WIP, described in [6], is accounted for (increased production time implies increased loading).

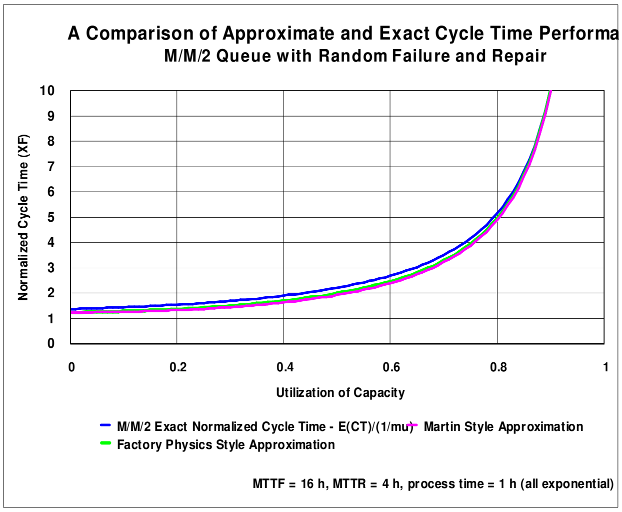

Figure 2. Approximate and exact cycle time behavior for an M/M/2 queue with imperfect server availability.

Combining idle with WIP and the approximation of (1) yields the following approximation for the expected cycle time in a G/G/m queue subject to server failures and idle with WIP:

$$E(CT) \approx \frac{1}{\mu_e} + \frac{1}{\mu_e}\left(\frac{C_{S,E}^2 + C_A^2}{2}\right)\frac{(\rho^*)^{\sqrt{2(m+1)}-1}}{m(1-\rho^*)}, \qquad (2)$$

where $\rho^* = \lambda(\Omega + 1/\mu)/(mA)$, $0 < \rho^* < 1$. The parameters to be used in the approximation are given as

$$\mu_e \doteq \left[\frac{1}{\mu A} + \frac{\Omega}{A}\right]^{-1},$$

$$C_{S,E}^2 \doteq \frac{\sigma_S^2 + \sigma_\Omega^2}{[(1/\mu)+\Omega]^2} + (1+C_R^2)(1-A)A\left[\frac{m_R}{(1/\mu)+\Omega}\right].$$

Note that $1/\mu_e$ is increased by the mean idle with WIP time Ω and the server mean availability A (since the it is assumed that a lot will not defect from a failed server).

III. CYCLE TIME OFFSETS

Events common in manufacturing systems include the travel of lots from one toolset to another, the holding of lots until a process issue is resolved and possibly delay in the unloading of a lot following its production. All of these events can generally be considered independent of the queue length or availability status of the tools at which they occur. As such, the elements of cycle time caused by these events is additive with the cycle time incurred queueing and in production.

If one considers the offset events as a separate queue upstream or downstream of the toolset itself, simplifying assumptions can be made to convert the queues into a Jackson network [12]. Then, the cycle time resulting from the offset events and the cycle time resulting from the toolset are decoupled (independent) and additive. More general assumptions converting the queues into a BCMP network would yield the same result.

Let T, H and P denote the mean travel, hold and post production unloading delays, respectively. The approximation of (2) becomes:

$$E(CT) \approx T + H + \frac{1}{\mu_e} + \frac{1}{\mu_e}\left(\frac{C_{S,E}^2 + C_A^2}{2}\right)\frac{(\rho^*)^{\sqrt{2(m+1)}-1}}{m(1-\rho^*)} + P,$$

where all parameters have been defined previously. The independence leading to the addition of the mean offset event times is a fundamental result of queueing network theory [12]. Further, this approximate form suggests that if hold time, travel time and post production delay time are reduced in a fabricator consisting of many toolsets, the overall cycle time for the entire fabricator will decrease.

IV. DEFECTION OF LOTS FROM FAILED SERVERS

The approximations of Section II may perform poorly in the low loading regime when lots are allowed to defect from a failed server and enter the head of the queue. Note that when $\rho^* = 0$ in equation (1), the resulting expected cycle time prediction is $1/\mu^* = 1/(\mu A)$. Thus, if the availability is, for example, 80%, the cycle time prediction yields $(1.25)/\mu$, independent of the number of servers!

What is really expected in the low loading regime with m servers and lots prone to defection from a failed server is that no delay is incurred due to a failed server unless *all* servers are down. The expected cycle time as ρ^* decreases to 0 can be roughly approximated by

$$\lim_{\rho \to 0^+} E(CT) \approx \left(\frac{1}{\mu} + \Omega\right) + (1-A)^m \frac{m_R}{m+1},$$

where the last term is the mean residual down time faced by a lot arriving to a system with all servers in failure (and assuming deterministic repair times).

The following approximation is thus suggested for G/G/m queues with exponential server failures, deterministic server repair, lot defection, idle no WIP and cycle time offsets:

$$E(CT) \approx (T+H+P) + (1-A)^m\left(\frac{m_R}{m+1}\right) + \left(\frac{1}{\mu} + \Omega\right) + \left(\frac{1}{\mu} + \Omega\right)\left(\frac{C_{S,E}^2 + C_A^2}{2}\right)\frac{(\rho^*)^{\sqrt{2(m+1)}-1}}{m(1-\rho^*)}, \quad (3)$$

where

$$C_{S,E}^2 \doteq \frac{\sigma_S^2+\sigma_\Omega^2}{[(1/\mu)+\Omega]^2} + 2(1-A)A\left[\frac{m_R}{(1/\mu)+\Omega}\right].$$

Note that the C_R^2 term has been set to unity in the $C_{S,E}^2$ definition, in agreement with the use of a deterministic repair time in the determination of the residual down time. To account for general repair time distributions, one could keep the C_R^2 term in the definition of $C_{S,E}^2$ and replace the mean residual down time $m_R/(m+1)$ by the calculated mean residual down time for a general repair distribution.

The mean cycle time approximation of equation (3) incorporates many features common in practical manufacturing systems. Further, as is demonstrated in the subsequent section, approximation equation (3) performs quite well in predicting measured cycle performance.

V. APPLICATION OF THE APPROXIMATIONS

The approximation of equation (3) has been tested on multiple toolsets within IBM's 200mm semiconductor wafer fabricator. Two points are worth mention. First, very few toolsets actually behave as an independent G/G/m queue with the features described. Lot process times are a function of their stage of production rather than being identically distributed according to a single distribution. Also, the toolsets are not operated in isolation, but rather in a queueing network. Second, at IBM's 200mm semiconductor wafer fabricator, few equivalent toolsets are grouped in a single geographic area. Even when geographically close, tools may be assigned to specific operators. In addition, most toolsets are purposely limited in deployment (so that a lot is not qualified to receive processing from all tools in a group) to ease the detection of yield problems. As a consequence, the practical number of servers that are available to a given lot is on the order of 3 for many toolsets. There are some notable exceptions, including the chemical-mechanical polishing tools, each of which is virtually indistinguishable.

To apply the approximation of equation (3), all parameters specified must be determined. Standard statistical analysis can be applied to the tool and lot logistics databases in a manufacturing facility to obtain the statistics. IBM's 200mm semiconductor wafer fabrication facility has developed many automated data analysis tools [6, 8] to enable the acquisition of needed data. In particular, idle with WIP time, travel time, hold time, post production unloading time, utilization (to some extent) and tool availability are generated automatically for each toolset. The other statistics were calculated separately. In addition, many parameters, such as idle with WIP and travel time, may be a function of fabricator loading. The approximation of equation (3) assumes that its parameters do not change with loading.

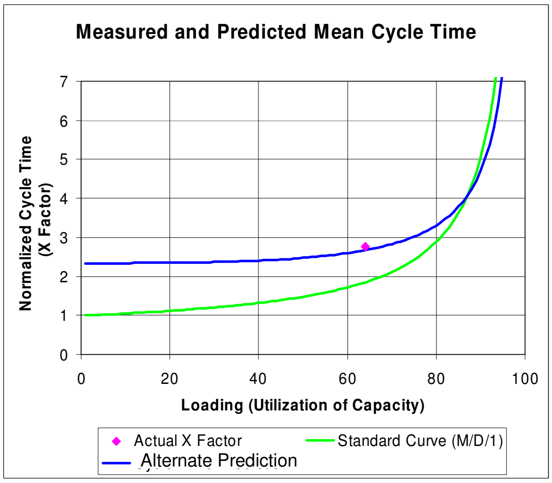

Figure 3. Measured and predicted cycle time for a toolset in IBM's 200mm wafer fabrication facility.

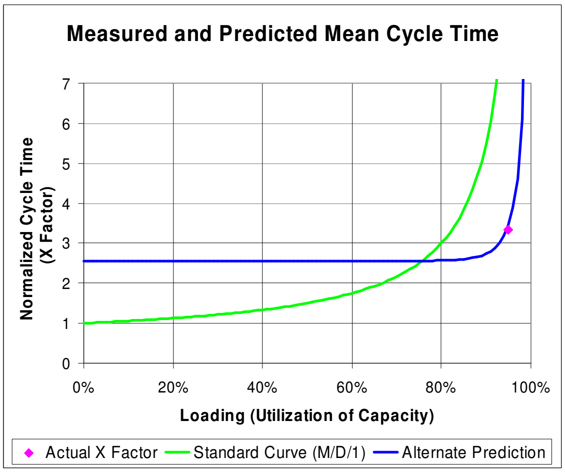

Figure 4. Measured and predicted cycle time for another toolset in IBM's 200mm wafer fabrication facility.

Figure 3 provides an example of the application of the approximation of equation (3) to a toolset with substantial cycle time offsets. The actual performance, measured from a production workstation, is very close to our predicted performance curve (referred to as the Alternate Prediction in the figure). Figure 4 provides a second example in which cycle time offsets and a large number of servers play key roles. Again, the measured cycle time performance is relatively close to the approximation of equation (3). In each figure, the cycle time performance for an M/D/1 queue is provided for comparison. The cycle time depicted has been normalized by dividing by (1/μ) and is referred to as X-Factor.

VI. CONCLUDING REMARKS

To approximate the mean cycle time behavior of practical models of manufacturing systems, extensions to the standard approximations for G/G/m queues were presented. A common manufacturing phenomenon known as idle with WIP was incorporated. Overhead which are independent of the queue length, such as travel, hold and post production unloading delays were explicitly incorporated. In addition, the incorporation of realistic lot defection behavior in the presence of server failures was considered.

The resulting approximations were employed to create an approximate mean cycle time performance curve for toolsets within IBM's 200mm semiconductor wafer fabricator. Using data drawn from actual toolset behavior, the approximations were generated and compared with the measured mean cycle time of the toolset. The results demonstrate that the approximations perform well.

Future directions could include the application of the approximation to networks of queues, development of an algorithm for the determination of the number of servers that are available to each lot and the modeling of idle with WIP as a function of workstation loading. Also, increased rigor could be applied to the approximations developed for lots prone to defection from a failed server.

REFERENCES

[1] A. K. Erlang, "The theory of probabilities and telephone conversations," *Nyt Tidsskrift for Matematik B*, Vol. 20, 1909.

[2] H. Sakasegawa, "An approximation formula $L_q = \alpha\beta^{\rho}/(1\text{-}\rho)$," *Annals for the Institute for Statistics Mathematics*, Vol. 29, pp. 67-75, 1977.

[3] W. Whitt, "Approximations for the GI/G/m Queue," *Production and Operations Management*, Vol. 2, No. 2, Spring 1993.

[4] Wallace J. Hopp and Mark L. Spearman, Factory Physics: Foundations of Manufacturing Management, Second Edition, publisher Irwin/McGraw-Hill, New York, NY, 2001.

[5] M. Segal and W. Whitt, "A queueing network analyzer for manufacturing," *Teletraffic Science for New Cost-Effective Systems, Networks and Services*, ITC-12, M. Bonatti, Editor, Elsevier-Science, Amsterdam, pp. 1146-1152, 1989.

[6] D. P. Martin, "Capacity and cycle time – throughput understanding system (CAC-TUS) an analysis tool to determine the components of capacity and cycle time in a semiconductor manufacturing line", *Proceedings of the IEEE/SEMI ASMC*, Boston, MA, September 8-10, 1999, pp. 127-131.

[7] 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," *Proceedings of the IEEE/SEMI ASMC*, Munich, Germany, 2001, pp. 21-22.

[8] K. Connerney, D. Martin and R. Tomka, "Determining the capacity components of different classes of multichamber tools," *Proceedings of the IEEE/SEMI ASMC*, 2001, Munich, Germany, pp. 29-32.

[9] B. S. Bortnick, P. P. Shirk, J. R. Morrison, and D. P. Martin, "Calculating the performance curve for batch tools," to appear in the *Proceedings of the IEEE/SEMI ASMC*, Boston, MA, May 2006.

[10] J. R. Morrison, B. S. Bortnick and D. P. Martin, "Performance evaluation of serial photolithography clusters: Queueing models, throughput and workload sequencing," to appear in the *Proceedings of the IEEE/SEMI* ASMC, Boston, MA, May 2006.

[11] J. H. Jacobs, L. F. P. Etman, J. E. Rooda and E. J. J. van Campen, "Quantifying operational time variability: The missing parameter for cycle time reduction," *Proceedings of the IEEE/SEMI ASMC*, Munich, Germany, 2001.

[12] 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.

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.