Proceedings of the 2014 Winter Simulation Conference
A. Tolk, S. Y. Diallo, I. O. Ryzhov, L. Yilmaz, S. Buckley, and J. A. Miller, eds.

MEASURING CYCLE TIME THROUGH THE USE OF THE QUEUING THEORY FORMULA (G/G/M)

DJ Kim
Lixin Wang
Robert Havey

Industrial Engineering Department
Micron Technology, Inc.
9600 Godwin Drive
Manassas, VA 20110, USA

ABSTRACT

Semiconductor manufacturers are required to reduce their product cycle times since many product embedded semiconductor devices often have a very short life cycle. One way to reduce cycle time is to purchase extra manufacturing tools. However, these tools cost several millions of dollars and facility space is limited. Another way to reduce cycle time is to improve performance of the critical tools. The second option is less costly and provides a significant cost savings for manufactures, which leads them to maximize efficiency. In order to determine which tools are critical and require analytical resources to optimize their performance, a system is needed to prioritize which are the critical tools. This paper will focus on Kendall's Classification of Queues, and it will focus on the G/G/m Queue (general distribution arrival process / general distribution service process / m servers) (Kendall 1953).

1 QUEUING FORMULAS

Queuing theory is the mathematical study of waiting lines or queues. A facility can be conceptualized as a set of products (wafers) traveling through a network of queues whose servers are tools. Optimizing the variation of arriving work in progress (WIP), WIP processing times, tool repair time, and the number of qualified tools will improve cycle time for the system and increase throughput of the critical tools.

Kendall's classification of a queuing station (A/B/m) (Kendall 1953), where:

A: Arrival process
B: Service process
m: number of machines

and distributions:

M: Exponential (Markovian) distribution
G: Completely general distribution
D: Constant (Deterministic) distribution

M/M/m
M/G/m
M/D/m
G/M/m

G/G/m
G/D/m
D/M/m
D/G/m
D/D/m

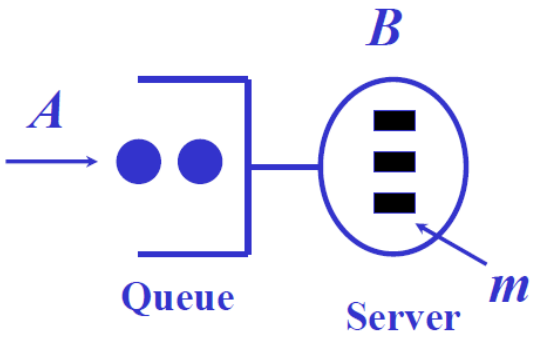

Figure 1: Characterization of a queuing station.

The station in Figure 1 can be described using the following parameters:

r_a: Rate of arrivals in job per unit time
t_a: Average time between arrivals ($t_a = 1/r_a$)
c_a: Coefficient of variation of inter-arrival times
m: Number of machines
r_e: Rate of the station in jobs per unit time
c_e: Coefficient of variation of effective process times
u: Utilization of station (r_a/r_e)

and the following measures:

CT_q: Expected waiting time spent in queue
CT: Expected time spent at the process center (queue time plus process time
WIP_q: Expected WIP (in jobs) in queue
WIP: Average WIP level (in jobs) at the station

with the following relationships:

$CT = CT_q + t_e$
$WIP = r_a * CT$
$WIP_q = r_a * CT_q$

If CT_q is known, WIP, WIP_q and CT can be calculated (Kendall 1953).

The Kingman's equation (Kingman 1961) modified for m servers (Hopp and Spearman 2001, Medhi 1991):

$$CT_q = V \times U \times t = \left(\frac{c_a^2 + c_e^2}{2}\right)\left(\frac{u^{\sqrt{2(m+1)}-1}}{m(1-u)}\right)t_e$$

1.1 The G/G/m Queue

The G/G/m queue is a completely general distribution of arrival and process times with *m* servers. No exact performance measures can be written, so approximation is used. Cases where approximation works poorly are where c_a and c_e are much larger than 1, and u is larger than 0.95 or smaller than 0.1. In addition, the assumptions of the G/G/m queue are First-Come, First-Served, infinite calling population and unlimited queue lengths are allowed.

1.1.1 Notations

A: Effective availability
b: Weighted average number of lots processed
c_0^2: Squared coefficient of variation of natual process time
c_a^2: Squared coefficient of variation of arrivals of lots or batches
c_e^2: Squared coefficient of variation of process time
c_r^2: Squared coefficient of variation of repair time
m: Number of qualified machines
m_r: Average length of mean time to repair (MTTR)
t_0: Average natural process time
t_e: Mean effective process time
u: Average utilization of machines

Step Cycle Time Formula (Hopp and Spearman 2001):

$$CT_q = \left(\frac{\frac{c_a^2}{b} + c_e^2}{2}\right)\left(\frac{u^{\sqrt{2(m+1)}-1}}{m(1-u)}\right)t_e$$

$$c_e^2 = c_0^2 + (1 + c_r^2)A(1-A)\frac{m_r}{t_0}$$

1.2 Limitations of Cycle Time Formula

The Cycle Time Formula is a static model because the model does not run over time. The accuracy of expected moves is very important for an accurate prediction due to the reentrance flow effect seen in semiconductor manufacturing. The formula also assumes that Part-Step (PS) jobs waiting at queue can be processed by any of the servers (M) (Hopp and Spearman 2001), Figure 2, which may not be correct if all servers are not qualified to process the WIP.

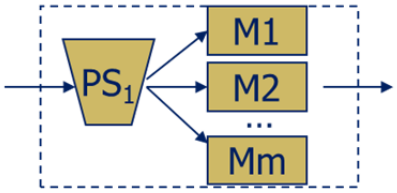

Figure 2: G/G/m job waiting queue.

2 VALIDATIONS OF INPUT PARAMETERS

Accuracy of cycle time estimation using G/G/m queue heavily relies on the input parameters. One way is to aggregate all variations (lot variation, equipment variation, and processe time variation) into one single parameter Variability V_B (Schelasin 2013b). A backward calculation method is used with formula

$$V_B = \frac{CT_q}{U \times t_e}$$

Where CT_q, U, and t_e are all historical data.

This paper adopts a different approach to determine the input parameters. All input parameters without aggregation used in G/G/m queue are generated based on historical data with time span of 1~3 months, namely A, c_a, c_o, c_r, m, t_e, and b. Certain filtering criteria is required to remove outlier data. When upstream tool has long term down, during that time period, time between lot arrivals will be peak points as shown in Figure 3. Thus c_a value will go up significantly. Another example is MTTR value. Due to system setup, certain tools log have large number of down event <1 minutes, which is not correct as tool is processing wafers during that time. These kind of filtering criteria has to be build on continuous monitoring of the parameters and help from area experts like tool engineers. Furthermore, the filtering criteria can be different across fabs. For this purpose, Best Known Methods (BKM) process has been widely used.

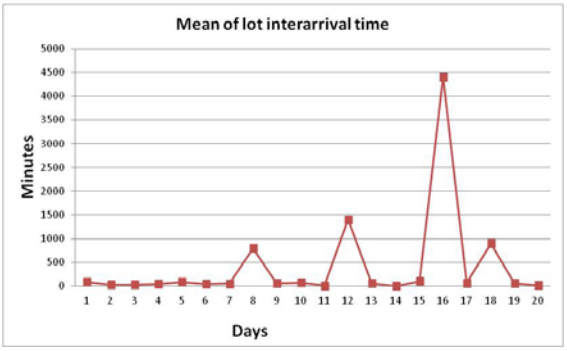

Figure 3: mean of lot interarrival time by day.

3 RESULTS – AREAS OF IMPROVEMENT

The first step is to identify which workstation needs to be optimized to decrease cycle time in the facility. Any method can be used to identify the bottleneck and near-bottleneck of the system. Bottleneck and near-bottleneck workstations should a have direct impact on facility loading or impact the Coefficient of Variation of the various operating curves for the facility.

The second step taken was to identify which variables within the Cycle Time formula will decrease cycle time for each workstation, and how much those variables need to be improved to achieve the desired reduction of cycle time for each workstation. The charts in Figures 4 to 6 show the values before optimization of the Photo Clusters, the Goal values based on expected gains for the Photo Clusters, and the In Progress values based on actual gains made by the Photo Cluster workstation. The circle on the chart shows where on the operating curve the Photo Clusters have been operating.

Figure 4 shows how Percent Downtime was improved on the Photo Clusters. Aggressive goals for reducing unscheduled and overall downtime was targeted for many workstations, including the Photo Clusters. Reduction of downtime has varied between the workstations, and the Photo Clusters have achieved ~20% reduction in downtime, with additional projects still being worked on for further cycle time improvement.

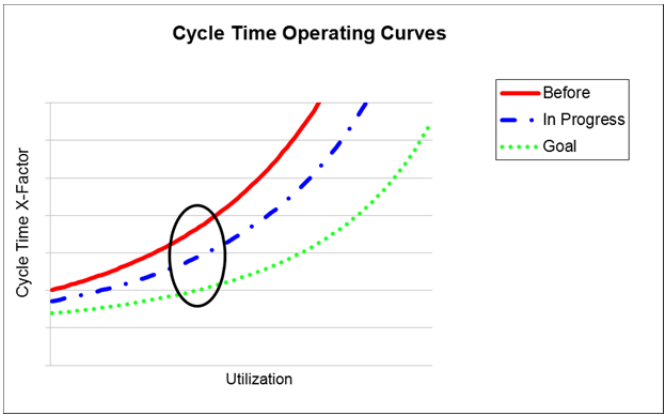

Figure 4: Photo cluster - % downtime improvement.

Figure 5 shows improvement to MTBF and Percent Downtime for the Photo Clusters. As Percent Downtime was decreased, the MTBF was increased for the Photo Clusters. The unscheduled downtime was reduced by ~55%, while scheduled downtime was increased by ~10%, resulting in a net reduction of ~20% for the Photo Cluster. By improving Preventative Maintenance procedures, a substantial gain was achieved in unscheduled down.

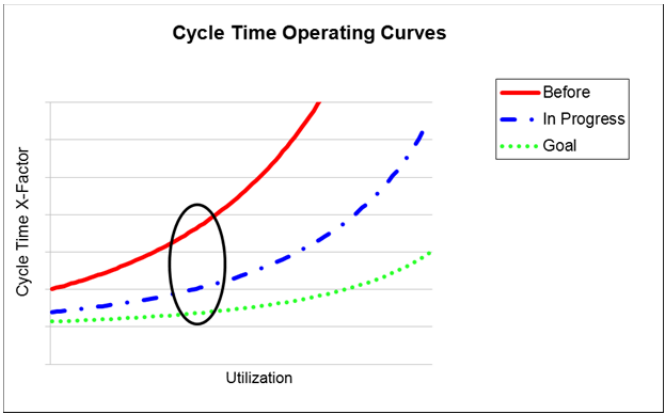

Figure 5: Photo cluster - % downtime and MTBF improvements.

Figure 6 shows improvement to Repair Time Variability, MTBF and Percent Downtime for the Photo Clusters. In addition to improving MTBF and Percent Downtime, improvement of the Repair Time Variability was also achieved. The majority of this gain was achieved through the reduction of unscheduled downtime on the Photo Clusters. In addition, improvements to Preventative Maintenance also reduced repair time variability, even though scheduled downtime did increase for the Photo Clusters. Overall, these improvements resulted in less troubleshooting of unscheduled issues and allowed the technicians to focus on preventative maintenance to improve workstation availability.

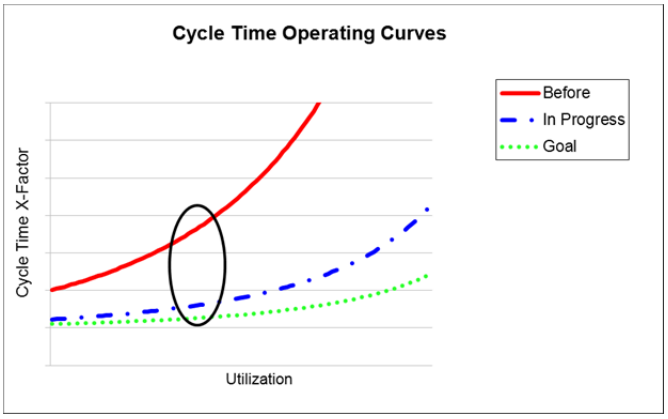

Figure 6: Photo cluster - % downtime, MTBF and repair time variability improvements.

The third step is where potential improvements can be further analyzed to achieve the Goal for the workstation that was set by the facility. As improvements were being made to the Photo Clusters, similar improvements to the upstream workstations affected interarrival times of WIP to the Photo Clusters, resulting in an ~18% decrease of interarrival time variation.

In addition to the improvments to interarrival times, there was also an improvement to the variation of process times on the Photo Clusters. The reduction in unscheduled downtime improved the consistency of the Photo Clusters, resulting in an ~8% decrease in process time variation. The improved consistency of the tools resulted in improved load balancing, which has resulted in more consistant throughput from each tool in the Photo Cluster workstation.

The fourth step is to combine all the analyzed data together and develop Squared Coefficient of Variance (SCV) curves. The SCV curves show the curve where the facility is currently at (Default), the target curve for the facility (Target) and the best case curve for the facility (Best). In addition to identifying these curves, there are four questions that need to be answered to help the facility achieve the goal of moving to a more desirable curve. The questions are: Where am I at? Where am I heading if I do not change anything? What is the best I can do in the future if I make improvements as planned? Do I need to make more improvements in order to hit my cycle time target? (Schelasin 2013a)

Where am I at? This question is asking what curve the facility is currently on and where is the facility on that curve. The question is important because it is the first read point for the facility. It gauges the facility's cycle time vs. the facility's current loading based on the (SCV) curves.

Where am I heading if I do not change anything? This question is asking where the facility is heading if the facility remains on the same curve. Increasing Fab Loading on the same curve will result in higher cycle time, and decreasing Fab Loading will decrease cycle time until the knee in the curve is surpassed and cycle time decreases will not keep up with Fab Loading decreases.

What is the best I can do in the future if I make improvements as planned? This question is asking which curve the facility can jump to if planned improvements are achieved. By jumping to another curve, the facility has the ability to keep cycle time constant and increase Fab Loading %, decrease cycle time and WIP while keeping current Fab Loading %, or any combination between. Attempting to achieve Best case shows where the facility can move towards if enough improvements are identified to hit this curve.

Do I need to make more improvements in order to hit my Cycle Time Target? This question is asking how close is the facility to hitting its Cycle Time Target. If additional work is required, opportunities will need to be identified, expected gains need to be verified by the cycle time formula, and projects need to be tracked and measured.

After the targets have been approved, the fifth step is to develop projects to achieve the desired cycle time based on Fab Loading %. Cycle time gains from the projects can be estimated through the Cycle Time Formula, and the projects can be tracked through any Program Management system.

Not all projects need to be identified at this stage. Future projects can be developed as spin-offs from current projects that do not fall within the scope of the project, or as independent projects that are identified at a later date. Achieving these cycle time improvements can take many months of work to achieve, so there is no rush to identify all potential projects to achieve target cycle time before starting the work to optimize cycle time.

As new projects are identified and scoped out, the process defined in this paper is repeated. The projects are analyzed to determine if they will generate a measurable impact to cycle time, the SCV curves are updated to measure how close to Target the facility is towards that goal, and the Waterfall chart is updated to measure the march to Target.

In addition, as projects are completed, the same process is applied as well. Actual data can be used to measure the cycle time impact on the workstation, and actual cycle time gains are measured against projected cycle time gains to ensure that the waterfall chart is accurately reflecting cycle time.

Finally, just because a project is completed does not mean that the issue is permanently fixed. In order to maintain the cycle time gains, routine work must be completed to keep the cycle time gains locked in. ignoring the workstation after the work is completed could lead to cycle time creep, which will cause the facility to miss its Target.

4 CONCLUSION AND FUTURE DEVELOPMENT

The processes discussed in this paper have been implemented at our facility and cycle time has been reduced by >10%. Since the project was started, the process areas are using the cycle time formula to help them identify the area bottleneck and near bottleneck, and work on methods to improve those workstations. Silo area project work is held to a minimum due to an alignment of process areas that naturally work together based on the process flow. This synergy is helping reduce variability in the line through better communication and a stronger understanding of the impact that line variation has on cycle time.

By understanding the components that go into cycle time, if becomes easier to understand how our previous actions have impacted cycle time. The process of understanding the complexities and interrelatedness of the various components that makeup cycle time is allowing for changes in behavior that have negatively impacted cycle time in the past. It also drove home the point that regular feedback was necessary to ensure cycle time does not get out of control in the future.

The Cycle Time operating curves show how throughput, utilization and cycle time are connected, but WIP levels still needs to be calculated from these values. Intel developed a method based on the operating curve and Little's Law (Hopp and Spearman 2001) to sow this relationship called O_L Graph (Li et al. 2005). The methods and formulas described in the Intel paper are being explored to determine if and how they can be used to help improve cycle time without decreasing facility output.

REFERENCES

Hopp, W. and Spearman, M. L. 2001. *Factory Physics*. Boston: Irwin.

Kendall, D. G. 1953. "Stochastic Processes Occurring in the Theory of Queues and their Analysis by the Methods of the Imbedded Markov Chain." *The Annals of Mathematical Statistics* 24 (3): 338-354.

Kingman, J. F. C. 1961. "The Single-Server Queue in Heavy Traffic." *Mathematical Proceedings of the Cambridge Philosophical Society* 57(4): 902-904.

Li, N., L. Zhang, M. Zhang, and L. Zheng. 2005. "Applied Factory Physics Study on Semiconductor Assembly and Test Manufacturing". In *Proceedings of the 2005 IEEE International Symposium on Semiconductor Manufacturing*, 307–311.

Medhi, J. 1991. *Stochastic Models in Queuing Theory*. Boston, MA: Academic Press.

Schelasin, R. E. A. 2013a. "Capacity Management Using Static Modeling, Queuing Theory, and Performance Curves". IIE Annual Conference & Expo.

Schelasin, R. E. A. 2013b. "Estimating Wafer Processing Cycle Time Using an Improved G/G/M Queue". In *Proceedings of the 2013 Winter Simulation Conference*, edited by R. Pasupathy, S.-H. Kim, A. Tolk, R. Hill, and M. E. Kuhl, 3789–3795. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

AUTHOR BIOGRAPHIES

DJ KIM is a senior member technical staff in the Industrial Engineering Department at Micron Technology Inc., Virginia. He received M.S. in Engineering Management from Queensland University of Technology, Australia, an MBA from Georgetown University, and has performed 16 years of simulation and modeling experiences in semiconductor industry. He is a member of INFORMS. His e-mail is djkim@micron.com.

LIXIN WANG is a senior industrial engineer in the Industrial Engineering department at Micron Technology Inc., Virginia. He received B.S. in Mechanical Engineering from Tsinghua University, Beijing, China and Ph.D. in Industrial and Systems Engineering from Virginia Tech. His interest is mathematical modeling and simulation of semiconductor manufacturing systems. His e-mail is stanleywang@micron.com**.**

ROBERT HAVEY is an industrial engineer in the Industrial Engineering Department at Micron Technology, Inc. and has worked 12 years in semiconductor and automated material handling systems planning. He received his B.S. in Industrial Engineering from Texas A&M University. His e-mail is rhavey@micron.com.