Proceedings of the 2024 Winter Simulation Conference
H. Lam, E. Azar, D. Batur, S. Gao, W. Xie, S. R. Hunter, and M. D. Rossetti, eds.

COMPARISON STUDY TO EVALUATE THE RELATIONSHIP BETWEEN EQUIPMENT UPTIME VARIABILITY METRICS AND CYCLE TIME

Chris Keith[1], Maryam Anvar[1], and Marino Arturo[1]

[1]Applied Materials, Santa Clara, CA, USA

ABSTRACT

This paper evaluates several measures of equipment uptime variability to determine which most closely track with the cycle time in a semiconductor wafer fab. The analysis is focused on a fleet of similar process tools performing the same function in parallel and the resulting cycle time of lots running through those tools. Discrete-event simulation is used to study the relationship between cycle time and several uptime variability metrics, including coefficient of variation (CV), A80, and A20-A80 for different combinations of fleet size (numbers of tools) and levels of uptime variability based on a range of tool mean time to fail (MTTF) and mean time to repair (MTTR) combinations. For the scenarios analyzed, the CV of shift uptime is found to correlate with cycle time as well or better than the other studied equipment uptime variability metrics.

1 INTRODUCTION

A semiconductor wafer fab processes thin disks, or wafers, of high-purity semiconducting material, typically Silicon, in highly specialized manufacturing plants with a large amount of floor space configured as a cleanroom environment with strict standards for particles and contaminants. The wafers move through many repeated sequential processing steps in order to produce integrated electronics or photonics circuits. In leading-edge wafer fabs, the total number of steps in a single flow, including both process and measurement, can easily exceed 1,000 with cycle times measured in months.

The manufacturing line is organized as a job shop with similar types of equipment grouped together in the same area of the factory. Large fabs, sometimes referred to as giga-fabs, can have 1,000 or more tools. Many different types of equipment are used depending on the exact type of processing or measurement required. Examples of processing steps include chemical vapor deposition (CVD), physical vapor deposition (PVD), plasma or wet etch, Ion Implant, thermal oxidation or deposition, chemical mechanical polish (CMP), surface clean, and photolithography.

The wafers are stored in special containers to isolate them from the environment as they move between the equipment groups, and often revisit the same tools multiple times. The containers holding the wafers are either carried by humans or transported by automated vehicles between equipment groups. A typical wafer will travel several miles before it completes all processing and measurement steps.

Wafer fabs are very capital intensive, with leading edge factories requiring investments of USD($) 20 – 40 billion or more over the first ten years of construction and operation depending on the fab size and process complexity. The primary driver of the spending is the cost of the processing and measurement tools, which drives certain operational behaviors, including round-the-clock operation of the factory in order to maximize the return on the capital investment.

.

2 BACKGROUND

2.1 Background

Variability is a key challenge to the successful operations of many manufacturing systems, including semiconductor wafer fabs. There are many sources of variability in a manufacturing line, including unpredictable arrivals of work-in-process (WIP) from upstream operations, variation in the service times for jobs, and unreliable equipment that is down 15% of the time or more. Managing each of these sources of variation is crucial to enabling the factory to consistently meet its output targets with reasonable cycle times.

Attempts to quantify the influence of variability on manufacturing indicators such as cycle time can be traced to queueing theory, where an approximation for steady-state wait time (CTq) for systems with a single server and with general interarrival and process times was first investigated by Kingman (1961) and is given in (1) from Hopp and Spearman (2001). The wait time separates into three terms: a dimensionless variability term V, a utilization term U, and a time term T. The variability term uses the coefficient of variation (CV) of both the interarrival (c_a) and process time (c_e) distributions, the utilization term uses the server average utilization (ρ), and the time term uses the average process time (t_e). For stations with no downtime or setup time or rework, t_e is assumed to be the natural process time (t_0) and c_e is the coefficient of variation of the natural process time. This equation is sometimes referred to as Kingman's equation or as the VUT equation.

$$CT_q(\mathrm{G/G/1}) = \left(\frac{c_a^2 + c_e^2}{2}\right)\left(\frac{\rho}{1-\rho}\right)t_e \tag{1}$$

The formula in (1) can be extended to systems with multiple parallel machines as in (2) based on Sakasegawa (1977) and appearing in Hopp and Spearman (2001), where an additional factor (m), is added for the number of stations. Note that (2) reduces to (1) in the special case where m is set equal to 1.

$$CT_q(\mathrm{G/G/m}) = \left(\frac{c_a^2 + c_e^2}{2}\right)\left(\frac{\rho^{\sqrt{2(m+1)}-1}}{\mathrm{m}(1-\rho)}\right)t_e \tag{2}$$

For stations with downtime, setup, rework, or any other sources of variability besides arrival times and natural process time of the job, the definitions of t_e and c_e can be expanded. The effective process time (t_e) is defined by adjusting the natural process time (t_0) to account for the fraction of time the machine is available (A) as in (3), where A is then defined as in (4) based on Mean Time to Fail (MTTF) and Mean Time to Repair (MTTR) exponential distributions. Similarly, the coefficient of variation of effective process time (c_e) is also calculated based on the effective process times of each unit processed.

$$t_e = \frac{t_0}{A} \tag{3}$$

The calculation of the coefficient of variation of effective process time requires knowledge of the effective process time of each individual unit, or lot, processed. Jacobs et al. (2003) provide a general algorithm for calculating effective process times for a multiple machine workstation using job start times and end times from factory manufacturing execution systems (MES) databases. This algorithm provides the logic for calculating the effective process time of each job and combines the effects of both job actual process times and tool-driven delays in job completions.

$$A = \frac{MTTF}{MTTF+MTTR} \tag{4}$$

While effective process time has been a useful metric for accurately forecasting cycle time, alternative measures of variability focusing only on equipment uptime have emerged. While these equipment uptime variability measures may not be able to be directly incorporated into the queueing formulas for predicting cycle time, they are useful for factory and operations managers as indicators for measuring relative performance and tracking improvement. Cunningham and Babikian (1998) initially proposed the application of the A80 metric to quantify variation in daily fleet uptime and the resulting impact to predictability of capacity. A80 is defined as the best availability reached in 80% of the time periods in a consecutive set of periods, with shifts or days being the most common time period. They track daily fleet uptime over a recent period and then take the 80th percentile of that set of data points. They use the resulting value to monitor and drive improvements in equipment capacity stability.

The concept of using percentiles of the uptime distribution is extended to include A20 in Gaboury (2001). A20 is the best availability achieved in at least 20% of the time periods. The difference between A20 and A80 can then be used as a direct measure of the variation, or spread, in equipment uptime.

In addition to the A80 KPI, another measurement of the spread is the coefficient of variation (CV), which is a standardized measure of dispersion of a frequency distribution. CV is a dimensionless number that is defined as the ratio of the standard deviation () to the mean () of the data population, where the data set is composed of a distribution of uptimes of equal time periods such as 12 or 24 hours. Note that the CV discussed here should not be confused with the coefficients of variation referenced in (1) and (2) and cannot be directly substituted into the queueing equations. To help graphically illustrate the differences in uptime variability, Figures 1 and 2 show two scenarios reflecting both high and low variability of uptime based on a daily average. Figure 1 shows daily uptime with a grand average of 88%, CV of 0.07, A80 of 83%, and A20 – A80 of 11.4%. Note in Figure 1 that the daily value of uptime is frequently at or near 100% and rarely below 80%. Figure 2 shows a distribution of values with the same average uptime of 88% but with a higher amount of variability as measured by CV with a value of 0.15, A80 of 77%, and A20 – A80 of 23%. In Figure 2 there are not only more data points at 100% compared to Figure 1, but also many more data points below 80%, reflecting the increase variability in the distribution. While both data sets have the same average uptime of 88%, Figure 1 shows a much tighter spread and would likely result in shorter cycle times.

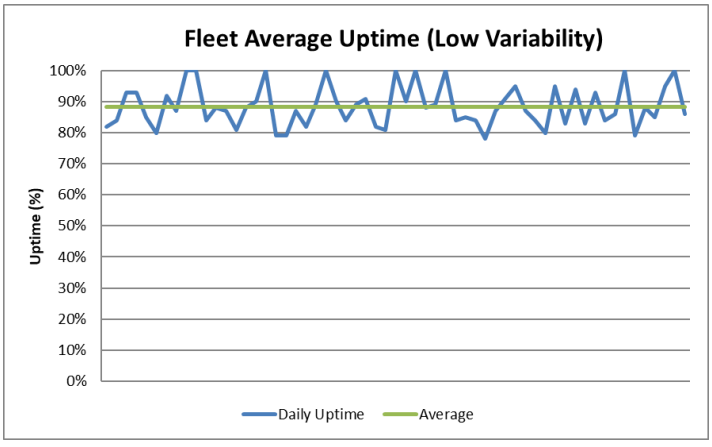

Figure 1: Low uptime variability (uptime CV = 0.07) example.

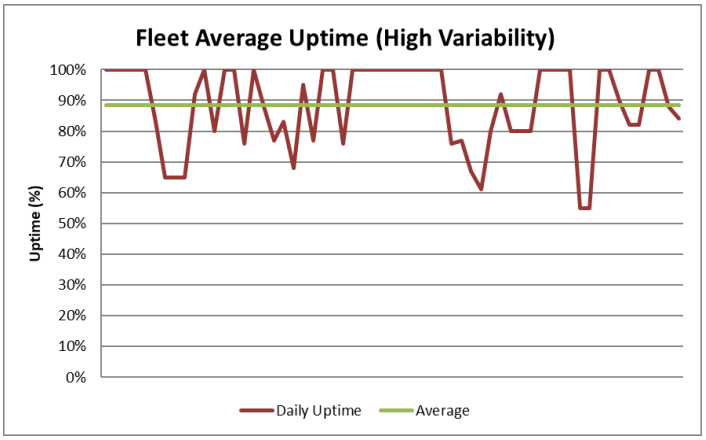

Figure 2: High uptime variability (uptime CV = 0.15) example.

3 STUDY METHODOLOGY

For the study, a discrete-event simulation model was built using AutoSchedAP® due to the availability of the software and familiarity of the analysts with the software. To study the impact of different fleet sizes (e.g. numbers of parallel tools) and levels of variation in the aggregated fleet uptime distributions, a single group of tools running similar steps in parallel was created. The simulation parameters were set as follows:

- Tools were set to process one lot of 25 wafers at a time, with the process time for each lot set to a normal distribution with an average process time of 1 hour and standard deviation of 0.1 hours, resulting in a natural process time CV of 0.10. Note that actual tool data typically shows higher variation, or CV, values but a lower value of CV was used to minimize the probability that the variation from tool process times would overwhelm and drown out the variation from equipment uptime, which was the focus of this study.
- The queue discipline was set to First-In-First-Out (FIFO), where lots were processed in the order of their arrival.
- System capacity for the number of lots accommodated either in service or queue is set to infinity.
- The number of tools was varied across three levels: 5 tools, 10 tools, and 20 tools.
- In order to maintain an expected value of 85% uptime across different levels of uptime variation, we established the following three combinations of tool-level MTTF and MTTR with exponential distributions. The low and medium variability levels were chosen to be consistent with levels of MTTR observed in typical tools in the semiconductor industry, with Figure 3 showing a representative histogram of downtime distributions with MTTR equal to 10 hours. The high variability MTTF and MTTR values were included as an extreme case that is occasionally observed with new tools running leading-edge process nodes.
    - Low variability: MTTF of 45.5 hours and MTTR of 8 hours
    - Medium variability: MTTF of 3.9 days and MTTR of 16 hours
    - High variability: MTTF of 15 days and MTTR of 2.55 days
- Material was started in front of the tools with an exponential inter-arrival time in order to achieve a target of 80% average utilization. The average lot inter-arrival time for each tool count was as follows: 5 tools – 0.249 hours, 10 tools – 0.12384 hours, and 20 tools – 0.06192 hours.

- For each combination of tool count (5, 10, and 20) and tool-level variability (low, medium, and high), a 365-day simulation warm-up period was used, after which the model was run for another 810 days and data from that time period was used for the KPI calculations.
- For calculating the independent variables, 12-hour time periods (shifts) were established and the average uptime of all tools during that shift was calculated.
- For the dependent variable, the average cycle time was calculated over the entire simulation period. This cycle time was then normalized to X-factor by dividing the cycle time by the process time.

Note the need to distinguish between the variability used as an input to the simulation based on tool-level MTTF and MTTR inputs versus the resulting variability of the fleet uptime as measured by the various KPI. The low and medium levels of

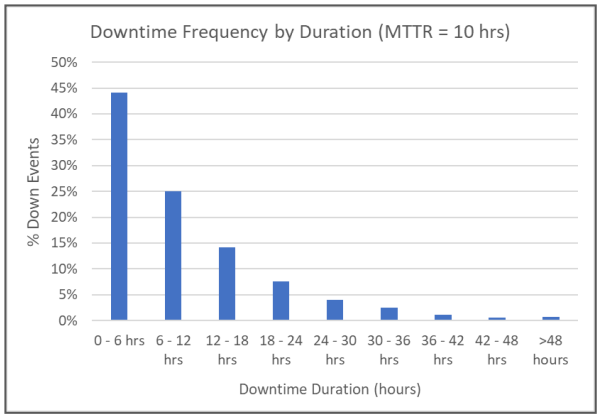

Figure 3: Downtime frequency for exponential distribution with MTTR of 10 hours.

Lastly, for each of the nine scenarios, three fleet-level uptime variability KPI were calculated based on the 1620 data points of the average uptime on each shift over the 810-day period. The KPI were CV, A80, and A20-A80.

4 STUDY RESULTS

For each of the nine scenarios, three uptime variability KPI were calculated based on the last 800 days of the simulation run. Figure 4 summarizes the key dependent variable (cycle time as measured by X-factor) and independent variables (uptime variability KPI) for all 9 simulation runs. As expected, for a given tool count increased levels of tool-level variability based on MTTF and MTTR lead to increasing cycle time as measured by X-factor. Additionally, for a given fleet size of tools, the cycle time increases as the machine-level variability level (low, medium, high) increases. The first question is, how does each uptime variability metric track the cycle time. For all cases, the CV of the shift uptime tracks well with the measured cycle time.

This relationship between cycle time and CV for each of the nine scenarios is shown graphically in Figure 5. Figure 5 illustrates the impact of variability pooling as the number of tools increases as illustrated by the fact that a high CV on a fleet of 10 tools can result in the same cycle time as a lower CV value with a fleet of 5 tools. Figure 6 shows the scatterplot of cycle time versus A80, and Figure 7 plots the cycle time versus A20-A80 values.

Scenario	Tool Count	Tool Variability Level	Cycle Time (X-Factor)	Uptime CV	A80	A20-A80
1	5	Low	6.62	0.14	74.9%	22.4%
2	5	Medium	10.33	0.16	74.9%	25.1%
3	5	High	27.2	0.18	77.6%	22.4%
4	10	Low	4.10	0.10	77.9%	14.7%
5	10	Medium	5.03	0.12	77.5%	16.7%
6	10	High	14.54	0.13	76.4%	21.0%
7	20	Low	2.54	0.07	79.7%	10.0%
8	20	Medium	2.71	0.08	79.8%	11.2%
9	20	High	6.12	0.09	79.3%	12.7%

Figure 4: Summary of simulation outputs for each scenario.

For Figures 6 and 7, the pattern of increasing cycle time with either higher A80 or lower A20-A80 is not supported in all scenarios. Specifically, the scenario with 5 tools and high tool-level uptime variability shows higher cycle time and higher A80 as in Figure 6 and higher cycle time and lower A20-A80 in Figure 7 (both scenarios are circled to help with identification). Investigation of the potential reasons for this led to a review of the actual shift uptime histogram for the scenario with 5 tools and high variability as shown in Figure 8. The histogram shows a clear multi-modal distribution with the peaks corresponding to the fractional values associated with three (60%), four (80%), or five (100%) of the tools being up for a given 12-hour shift. It seems likely that the A80 and A20 indicators may not be robust for distributions with numerous distinct peaks.

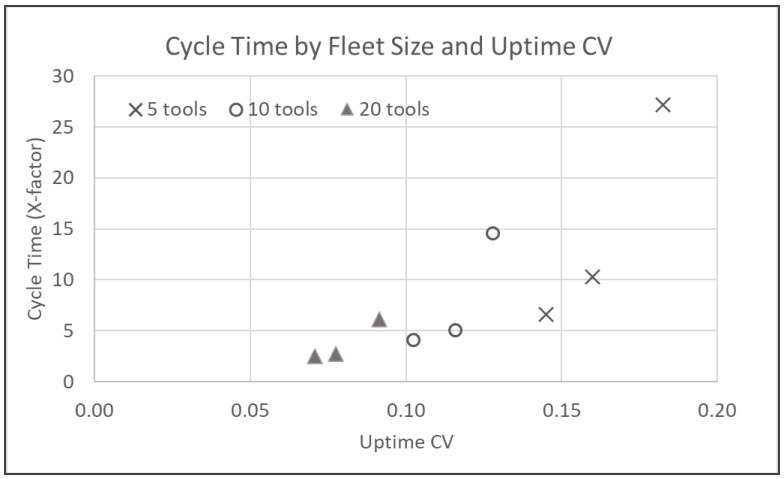

Figure 5. Cycle time x-factor versus CV for 9 scenarios.

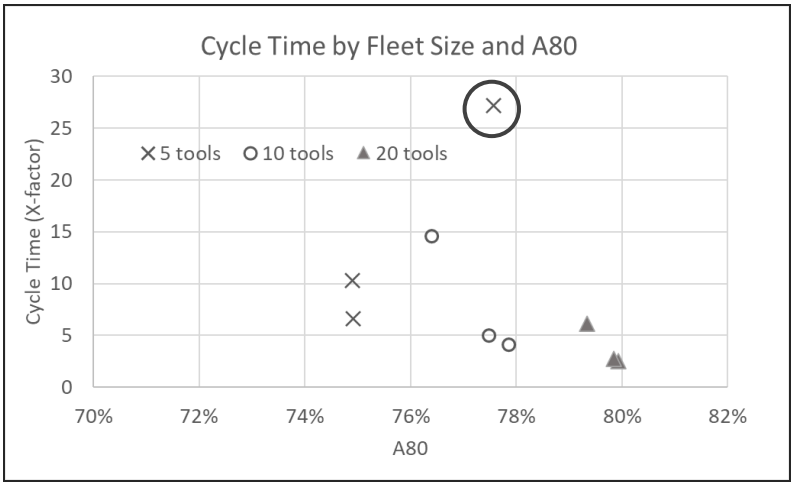

Figure 6: Cycle time x-factor versus A80 for 9 scenarios.

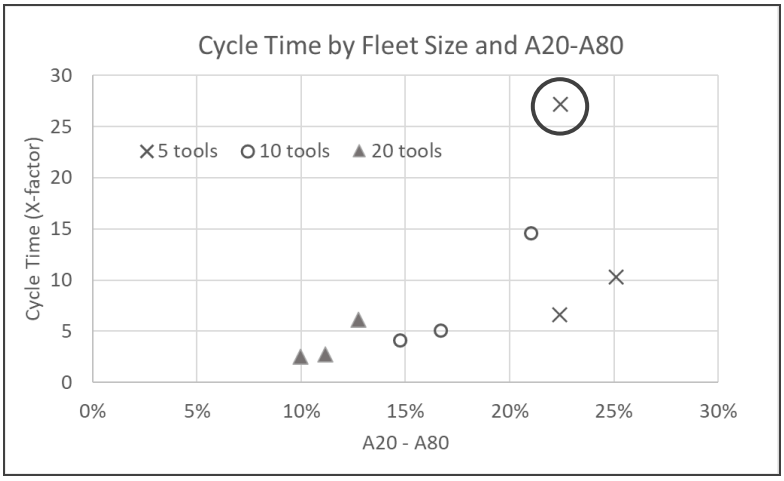

Figure 7: Cycle time x-factor versus A20-A80 for 9 scenarios.

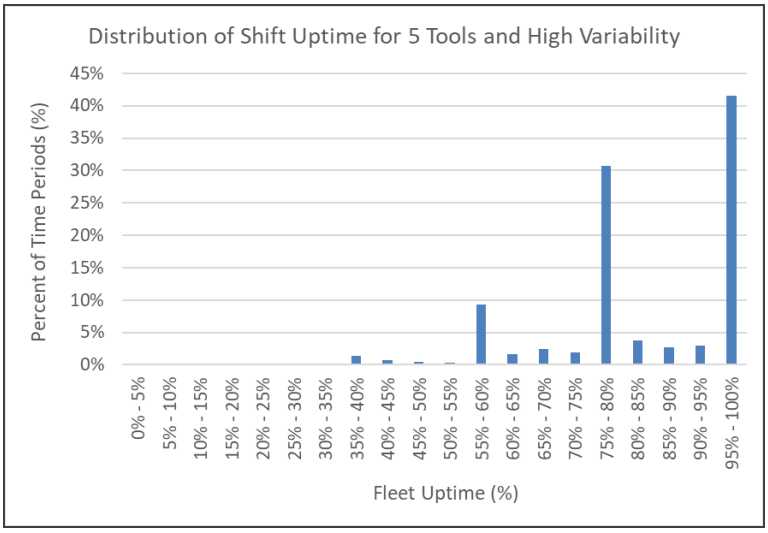

Figure 8: Histogram of shift uptimes for simulation with 5 tools and high variability.

5 CONCLUSIONS

The study demonstrates that under typical conditions in semiconductor wafer fabs for tool counts and tool-level variability that several of the uptime variability indicators identified track well with cycle time. For the nine scenarios analyzed, the CV of shift uptime is found to correlate with cycle time as well or better than other studied variability metrics based on visual examination of the scatterplots. It also shows that not only is the measured uptime variability predictive of cycle time but also that for a given level of measured uptime variability, cycle time can vary based on the number of tools running in parallel. We assume this is due to the pooling of the process time variation across a greater number of tools per equations (1) and (2), but that is left for future work.

Understanding the relationship between uptime variability and cycle time can be helpful in guiding decision making, leading to improved productivity and reduced cycle times. By identifying the uptime variability metrics that most closely relate to cycle time, fabs can not only implement targeted strategies and allocate resources judiciously, but also have reliable indicators to track the amount of improvement and the impact of these actions on both uptime variability and cycle time.

Proposed strategies for reducing the variability of uptime would fall into one of two categories: first, those that would reduce variability but either not change or even reduce the average uptime, and second those that would both reduce variability and increase uptime. Examples of the first group include proactive scheduling of preventive maintenance to minimize the overlap between unscheduled and scheduled downtime and also the use of predictive maintenance to enable proactively taking tools down for maintenance before premature failures. For the second group, any engineering or operations-driven activities that increase average uptime, such as reducing maintenance green-to-green (GtG) time, increasing parts availability, improving troubleshooting, or increased maintenance staffing and training would also likely lead to reduce variability of uptime.

6 FUTURE WORK

There are several opportunities to extend and improve on the study methodology. First would be to modify the tool-level variability from a simple MTTF and MTTR framework to include downtime distributions that are more reflective of actual tools, such as having separate distributions for equipment

failures versus preventive, or scheduled, maintenance. Another option for downtime distributions would be to use actual equipment downtimes and employ a bootstrap sampling methodology.

Second would be to simulate fleets with other tool counts, either less than five or more than 20. There are many fabs that have groups of equipment with either fewer than five tools running in parallel or 40 or more tools running similar operations.

Third would be attempt to further understand the observation that for a given level of fleet uptime variability (CV, A20-A80), why the cycle time of smaller fleets is longer than the cycle time of fleets with more tools. As mentioned previously, the current assumption is that this is at least due in part to the pooling of the process time variability across more tools, but future work would be required for testing.

Last would be to simulate under different levels of utilization (ρ) of the available uptime, giving the ability to construct operating curves of cycle time versus utilization for different combinations of tool count and tool-level variability. The resulting operating curves might provide additional insight into how the utilization impacts the extent to which the uptime variability metrics track with cycle time and how the level of uptime variability affects the utilization at which the cycle time curve starts to rise dramatically.

REFERENCES

Cunningham, C. and R. Babikian. 1998. "A80-a New Perspective on Predictable Factory Performance". In *Proceedings of the 1998 IEEE/SEMI Advanced Semiconductor Manufacturing Conference and Workshop,* September 23rd– 25th, Boston, MA, USA, 71-76.

Gaboury, P. 2001. "Equipment Process Time Variability: Cycle Time Impacts" *Future Fab International* 11: 45-47.

Hopp, W. J. and M. L. Spearman. 2001. *Factory Physics.* 2nd ed. New York: McGraw-Hill/Irwin.

Jacobs, J. H., L.F.P. Etman, E.J.J. van Campen, and J.E. Rooda, 2003. "Characterization of Operational Time Variability Using Effective Process Times". *IEEE Transactions on Semiconductor Manufacturing* 16(3): 511-520.

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

Sakasegawa, H. 1977. "An Approximation Formula $L_q \simeq \alpha\rho^{\beta}/(i-\rho)$". *Annals of the Institute of Mathematical Statistics*, Part A (29): 67-75.

AUTHOR BIOGRAPHIES

CHRIS KEITH is a Director and Distinguished Member of Technical Staff in the Applied Global Servies (AGS) group at Applied Materials in Santa Clara, CA. His reserarch interests include semiconductor manufacturing, dispatching and scheduling, semiconductor equipment productivity, and data-driven optimization. He has authored several research articles and papers on these topics. His email address is Chris_Keith@amat.com.

MARYAM ANVAR is an Industrial Engineer in the Applied Global Services (AGS) Group at Applied Materials in Santa Clara, CA. Her focus areas include discrete-event simulation, semiconductor equipment productivity, and tool-level modeling. She has co-authored several papers that utilize discreve-event simulation to model operations and optimize decisions related to lot dispatching and recipe dedication in semiconductor wafer fabs. Her email address is Maryam_Anvar@amat.com.

MARINO ARTURO is an Industrial Engineer in the Applied Global Services (AGS) Group at Applied Materials in Santa Clara, CA. His focus areas include data-driven optimization, semiconductor equipment productivity, and the application of machine learning in support of data analytics. His email address is Marino_Arturo@amat.com.