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

MEAN CYCLE TIME OPTIMIZATION IN SEMICONDUCTOR TOOL SETS VIA PM PLANNING WITH DIFFERENT CYCLES: A $G/G/m$ QUEUEING AND NONLINEAR PROGRAMMING APPROACH

James R. Morrison and Hungil Kim

Industrial and Systems Engineering, KAIST
291 Daehak-ro (373-1 Guseong-dong), Yuseong-gu
Daejeon, 305-701, South Korea

Adar A. Kalir

Fab/Sort Manufacturing Division
Intel Corporation
2 HaZoran St., Qiriat-Gat 82109, Israel

ABSTRACT

In semiconductor manufacturing, preventive maintenance (PM) activities are typically scheduled via a two tier hierarchical decomposition approach. The first decision tier determines a PM cycle plan while the second tier schedules these planned events into the manufacturing operations. Following recent work based on the use of $G/G/m$ queueing approximations for PM planning, we develop a method to allow for multiple PM cycles in a tool set. We formulate a nonlinear program with the PM cycle durations as continuous decision variables with the objective of minimizing the mean cycle time. We examine certain special cases and characterize the optimal solutions. Numerical studies are conducted with realistic multiple PM cycle data to assess the implications of the proposed approach. The results suggest that it may be possible to obtain significant improvements in the overall cycle time performance of tool sets in semiconductor manufacturing relative to existing PM planning procedures.

1 INTRODUCTION

Preventive maintenance (PM) activities conducted on semiconductor manufacturing tool sets increase tool availability, reduce unplanned down time and improve the reliability of tool set operations (Yao, Fernandez-Gaucherand, Fu, and Marcus 2004). With the increasing cost and complexity of semiconductor manufacturing equipment, PMs have become more complex, with more tasks and longer durations. In Tirkel (2013), it is demonstrated that tool down time and its variability play key roles in semiconductor fab cycle time. As such they should be planned and implemented with care. The PM planning and scheduling task is sufficiently complex that it is commonly partitioned into two phases; cf. Yao, Fernandez-Gaucherand, Fu, and Marcus (2004). This hierarchical decomposition considers the planning of the PM cycles in the first phase and the scheduling of those planned PMs into the operations in the second phase. Our focus in this paper is on the planning of PMs cycles for the first PM decision tier. Throughout, we assume that the PMs are time based rather than use based. However, there is a good way to approximate use based cycles by time based given in Ramirez-Hernandez and Fernandez-Gaucherand (2003).

1.1 The Problem and Fundamental Tradeoffs

The duration of the cycle for each PM task is typically predetermined by the equipment engineer or equipment vendor. They specify, for example, that task 1 should be performed every 30 days. There may be many such PM tasks, each with a different specified cycle duration. Often, many of them share the same cycle duration and can be grouped together. Our goal is to determine which of these tasks the PM plan should group together. There are two competing behaviors to balance when seeking an answer to this question.

First, the fewer activities that are grouped together, the greater the number of setups required. Each time a tool is taken down for a PM, there is an associated setup in which the tool is removed from production, returned to atmosphere, opened, closed, pumped to vacuum, qualified for production, etc. These activities require time, perhaps 6 hours. If many PM tasks are grouped together into a single PM event, there will be fewer setups per unit time; tool availability will be greater. If each PM task is planned to be conducted separately, there will be many setups per unit time; tool availability will be lower.

Second, the more activities that are grouped together, the longer the tool is unavailable for production each time it is taken down for service. If a PM consists of 40 tasks, each consisting of on average 3 hours of work, the tool will require about 120 hours of service when it is removed from production (there is also time for the setup activities). This is a long time and may lead to significant lot queues (WIP bubbles) if the tool set is not able to address the incoming WIP without the down tool. If a PM consists of only a few tasks at a time, the tool is down for a much shorter duration, and WIP bubbles are less likely on account of the PM.

1.2 Literature Review

Many efforts have been devoted to the study of preventive maintenance over the years; cf., the survey papers Cho and Parlar (1991) and Dekker (1996). Much of the work focuses on determining the frequency at which PMs should be conducted. For our purposes, we will assume that this decision is out of our scope; it has been dictated by equipment engineers or the tool vendor.

The scheduling of PMs given the PM task groupings for manufacturing operations or technician allocation have been considered by numerous authors; cf., van Dijkhuizen and van Harten (1998) and Marquez, Gupta, and Heguedas (2003). In the context of semiconductor manufacturing, papers focusing on the scheduling of PM activities given the PM task grouping and cycle duration plans have been considered; cf., Yao, Fernandez-Gaucherand, Fu, and Marcus (2004), Davenport (2010) and Ramirez-Hernandez and Fernandez (2010).

Fewer have focused on the first decision tier in PM planning. In Yao, Fernandez-Gaucherand, Fu, and Marcus (2004), a framework suggesting the use of a Markov decision process (MDP) in the first tier was presented. However, their focus is on the second decision tier. In Cassady and Kutanoglu (2005), early efforts to jointly optimize the planning and operations tier are conducted for a single machine.

A practical approach to addressing PM planning was proposed in Kalir (2013). There, the PM planning tier was addressed by the consideration of a $G/G/m$ approximation for the mean cycle time in failure prone queues. The goal was to determine how to split, or equivalently group, PM tasks with a single common cycle duration. They assumed that the PM tasks were initially all grouped into a single PM event occurring at the given cycle duration. The discrete decision variable determined the number of equal sized portions into which the PM should be split. They incorporated imperfect PMs and investigated properties of their optimization problem. Realistic examples were provided.

1.3 Contribution and Organization

Our work extends the approach of Kalir (2013) to allow for PM tasks with different cycle durations and continuous decision variables rather than integer splits. The benefit of continuous decision variables is that the number of PM tasks in a group need not be $1/2, 1/3, 1/4, \ldots$ of the total number of tasks. As such, a better solution may be obtained. Kalir (2013) did not consider PM tasks with different cycles (e.g., some tasks must be conducted once every month and some every three months). This is because the $G/G/m$ cycle time approximation formula relies on the fundamental assumption that there is a single renewal process for the time to failure durations and another single renewal process for the time to repair durations. That is, the $G/G/m$ approximation assumes the up and down durations form an alternating renewal process. When all PM tasks have the same cycle duration (e.g., once every four weeks), they can be readily modeled as an alternating renewal process. However, if the PM tasks have different cycle durations (e.g., some every four

weeks and some every seven weeks), it is not clear how to construct such an alternating renewal process. Here we develop a practical method to do so.

The organization and contributions of the work are as follows. In Section 2.1, we recall essential preliminary results and state model assumptions. We

- Formulate the PM plan as a function of real valued decision variables in Section 2.2;
- Develop a model that allows for PM tasks with different cycle durations in Section 2.3;
- Formulate our optimization problem, explore its convexity and state related consequences in Section 3; and
- Consider numerical studies with practically inspired data in Section 4.

Concluding remarks are presented in Section 5.

2 $G/G/m$ BASED MODELS FOR PM PLANNING

2.1 Failure Prone $G/G/m$ Queues

The underlying system employed is that of the failure prone $G/G/m$ queue. Refer to Figure 1. Customers arrive to the system as a renewal process with arrival rate λ. Use σ_A and $C_A := \lambda\sigma_A$ as the standard deviation and coefficient of variation of the interarrival times. There are m identical servers that provide service to the customers. Customers await service in a single infinite queue and are served from the queue in first come first served manner immediately when one of the servers is available. The service durations are drawn from a renewal process with mean μ, standard deviation σ_S and coefficient of variation $C_S := \mu\sigma_S$. The servers are subject to preempt resume failures. The uptime durations (time to fail) and failure durations (time to repair) form an alternating renewal process. Let m_F and m_R denote the mean time to fail and time to repair, respectively. The tool availability $A := m_F/(m_F + m_R)$. We assume that the independent identically distributed (IID) times to fail (the uptime durations) are exponentially distributed. The IID times to repair have general distribution with standard deviation σ_R and coefficient of variation $C_R := \sigma_R/m_R$. All of these expectations are assumed finite. All processes are independent of each other.

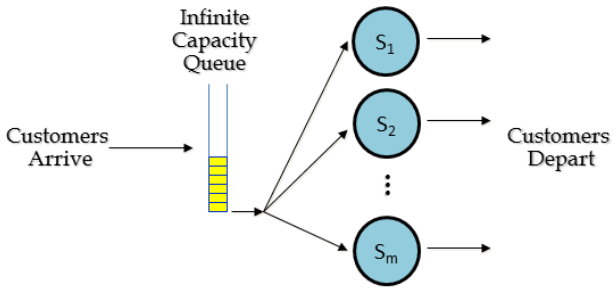

Figure 1: A $G/G/m$ queue.

When the system loading $\rho := \lambda/(m\mu A) < 1$, we are interested in the mean cycle time a customer will spend in the system $E[CT]$. This time includes both queueing and service time. In the special case of a failure prone $M/G/1$ queue (and recalling our assumption on exponential times between failures), $E[CT]$ can be obtained exactly. In Avi-Itzhak and Naor (1963), an $M/G/1$ queue subject to preempt-resume, time-based failures (with exponential times to fail and generally distributed times to repair) is referred to as Model A in which "station breakdowns occur homogeneously in time". We have from their equation (26) (after some algebra and the application of Little's Law):

$$E[CT] = \frac{1}{\mu A} + \frac{1}{\mu A}\left(\frac{\rho}{1-\rho}\right)\left(\frac{1}{2}\right)\left(1 + C_S^2 + \frac{(1+C_R^2)A(1-A)m_R\mu}{\rho}\right) \tag{1}$$

for $0 \leq \rho = \lambda/(\mu A) < 1$. There is no such exact expression for the failure prone $G/G/m$ queue. However, we shall employ the commonly used approximation:

$$E[CT] \approx \frac{1}{\mu A} + \frac{1}{\mu A}\Big(\frac{C_A^2 + C_{S,E}^2}{2}\Big)\frac{\rho^{(-1+\sqrt{2m+2})}}{m(1-\rho)}, \tag{2}$$

for $0 \leq \rho = \lambda/(m\mu A) < 1$, where $C_{S,E}^2 := C_S^2 + (1 + C_R^2)A(1-A)m_R\mu$ is typically interpreted as the squared coefficient of variation of the effective service time. Refer to Morrison and Martin (2007) for some discussion of the origin and interpretation of these formulae and their application to tool sets in semiconductor manufacturing. The reader is also encouraged to consult Wu, McGinnis, and Zwart (2011) and Wu (2014). They have recently developed improved approximations that enable differentiation between time-based and run-based failures that can be preemptive or not.

2.2 Problem Formulation with a Single PM Cycle

We first consider the case of a single PM as in Kalir (2013). We develop a different formulation of that problem that extends the search space from the integers to the real numbers. It also extends Kalir (2013) to allow for PM task grouping (aggregation) or ungrouping (splitting). Thus, the optimal value may be improved.

Throughout, for a random variable X we use m_X, σ_X and C_X as its mean, standard deviation and coefficient of variation, respectively.

To match the assumptions required in the $G/G/m$ queue, we consider $\{SU_j\}_{j=1}^{\infty}$, $\{PM_j\}_{j=1}^{\infty}$ and $\{U_j\}_{j=1}^{\infty}$ as renewal processes for the setup times, PM durations and uptime durations, respectively. They are independent of each other. The index j is for the j^{th} occurrence of that event. The uptime durations further form a Poisson process; the U_j are IID exponential. The duration of time between the start of PMs is thus also a renewal process $\{T_j\}_{j=1}^{\infty}$, where $T_j := SU_j + PM_j + U_j$. The duration of time the tool is down for the j^{th} repair is $R_j := SU_j + PM_j$. For convenience, we let SU be a random variable (RV) with the same distribution as SU_j (for any and all j since they are IID by virtue of our renewal process assumption). Similarly define the RVs PM, U and T. The second and third time lines depicted in Figure 2 provide examples of the notation and cycle.

Note that we assume the PM cycle (up times and down times) is an alternating renewal process. As such, each tool is available until it "fails" (to start the PM). It returns again to availability once the PM has been completed.

Suppose that some information is provided on the current PM plan in the fab (or some default data). It establishes the relationship between setup, PM and uptime durations within a cycle. For the default cycle, we use SU^0, PM^0 and T^0 as the RVs for the setup, PM and cycle durations, respectively. The first time line in Figure 2 depicts a cycle based on default data.

We will consider the mean cycle duration m_T as our decision variable in an effort to obtain a minimal mean cycle time for the tool set. The following assumptions are essential for use of the failure prone $G/G/m$ queue model.

- Assumption A1: The tool state is an alternating renewal process

$$\{SU_j, PM_j, U_j\}_{j=1}^{\infty}.$$

- Assumption A2: For any m_T value considered, the uptime random variable

$$U \sim Exp\Big(\frac{1}{m_T - (m_S + m_{PM})}\Big)$$

so long as the resulting $m_U > 0$. As such, $m_U = m_T - (m_S + m_{PM})$.

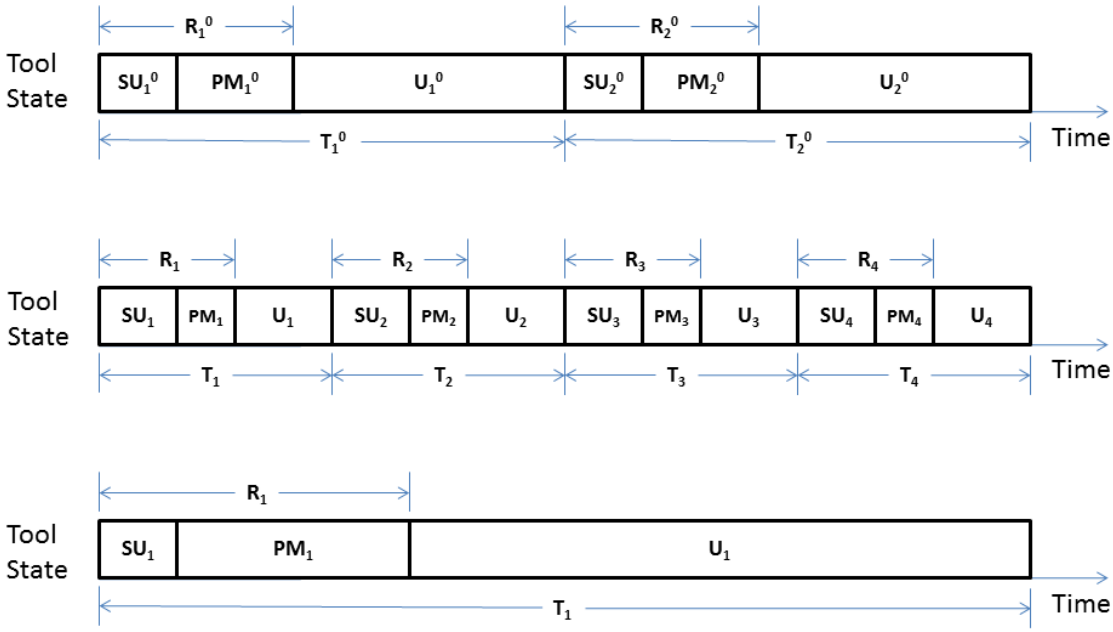

Figure 2: PM Cycles.

In failure prone $G/G/m$ queues, the up and down times form an alternating renewal process. Exponential times between failures are an underlying assumption for the mean cycle time results of equations (1) and (2).

The following assumptions are convenient for analysis (and key components are fundamental as detailed after the list).

- Assumption A3: $SU_j^0 = PM_j^0/k$, for all $j = 1, \cdots, \infty$, for some constant $k > 0$. More succinctly,

  $$SU^0 = PM^0/k$$

  Further, for any m_T value considered, $SU_j = PM_j^0/k, \forall j$. Succinctly,

  $$SU = PM^0/k.$$

- Assumption A4: For any m_T value considered, $PM_j = (m_T/m_{T^0})PM_j^0$, for all $j = 1, \cdots, \infty$. Succinctly,

  $$PM = (m_T/m_{T^0})PM^0.$$

In A3, what we want to require in any event is that the distribution of the setup times remains invariant with changes in m_T. In A4, we want to ensure that the mean PM duration scales exactly as m_T relative to m_{T^0}. A3 and A4 ensure this as well as simplifying some calculations as below.

These assumptions are convenient because they imply the following consequences. We label these consequences, for example, C3 as a consequence of the assumption A3.

- C3: A3 implies

  $$R^0 = (1 + 1/k)PM^0,$$

  so that $m_{R^0} = (1 + 1/k)m_{PM^0}$ and $C_{R^0} = C_{PM^0}$. Further, $m_{SU} = m_{SU^0} = (1/k)m_{PM^0}$.
- C4: Together with A3, $R_j = [(1/k) + (m_T/m_{T^0})]PM_j^0$, for all $j = 1, \cdots, \infty$. Succinctly,

  $$R = [(1/k) + (m_T/m_{T^0})]PM^0.$$

As such, $m_R = [(1/k) + (m_T/m_{T^0})]m_{PM^0}$ and $C_R = C_{PM^0}$.

As noted, A3 and A4 can be relaxed somewhat at the expense of a some of additional bookkeeping associated with calculating the coefficient of variation of the repair durations (down times). We elect to use them to simplify the resulting notation and results. If one elects to violate these assumptions, the setup variation must be specified and the manner in which the PM duration coefficient of variation changes as the mean PM value is controlled must be clearly elucidated.

The default cycle establishing the mean duration of the setup and the relationship between the means of the PMs and the entire cycle duration is depicted in the top time line of Figure 2. The middle and bottom time lines in Figure 2 depict the setup, PM and uptime cycles for two cases with PM activities grouped differently. In the middle time line, half the tasks are grouped into the PM so that $m_{PM} = m_{PM^0}/2$. The overall cycle duration follows this PM duration scaling so that $m_T = m_{T^0}/2$. The setup duration is unchanged $m_{SU} = m_{SU^0}$ and the mean uptime m_U is determined as a function of these. The figure explicitly denotes each RV in the renewal processes, e.g., $SU_1, SU_2, \ldots$, $PM_1, PM_2, \ldots$. In the bottom time line, the mean PM duration has been doubled (twice the tasks are grouped together relative to the default case). That is, $m_{SU} = m_{SU^0}$, $m_{PM} = 2m_{PM^0}$, $m_T = 2m_{T^0}$ and the mean uptime m_U depends on those.

2.3 Modeling Multiple PM Cycles

We will consider n different classes of PMs, each with its own cycle. To characterize the default settings for PMs of class $i = 1, \ldots, n$, let $m_{SU^{i,0}}$, $m_{PM^{i,0}}$ and $m_{T^{i,0}}$ denote the mean value of the setup times, PM durations and uptime durations, respectively. This represents the initial PM plan used in the fab (or just a default value to specify the relevant parameters). We assume they individually satisfy A3 (and thus enjoy the properties of C3) with their own constants $k_1, \ldots, k_n$ and RVs $PM^{1,0}, \ldots, PM^{n,0}$. Similarly, for a given m_{T^i}, let the setup, PM and cycles obey A4 and enjoy the C4 properties. In particular, $R^i_j = [(1/k^i) + (m_{T^i}/m_{T^{i,0}})]PM^{i,0}_j$, for all $j = 1, \cdots, \infty$, $i = 1, \ldots, n$. More simply,

$$R^i = [(1/k^i) + (m_{T^i}/m_{T^{i,0}})]PM^{i,0}, \forall i = 1, \ldots, n.$$

Thus, $m_{R^i} = [(1/k_i) + (m_{T^i}/m_{T^{i,0}})]m_{PM^{i,0}}$ and $C_{R^i} = C_{PM^{i,0}}$. Practically these PM cycles operate on the same time line, interfering with each other and the resulting up and down processes need not be renewal in general.

For the purpose of our $G/G/m$ model, we construct an alternating renewal process for the up and down times that retains many of the properties of these multiple PM cycles. Intuitively, the idea is to consider the interval of time $[0, P_n]$, where $P_n := m_{T^{1,0}} \cdot \ldots \cdot m_{T^{n,0}}$. During this interval, one expects there will be $h_i := P_n/m_{T^{i,0}}$ PMs of class i. The total number of PMs of all classes that one expects to occur in the interval is $H_n := (P_n/m_{T^{1,0}}) + \ldots + (P_n/m_{T^{n,0}})$. The proportion of these total PMs that are of class i is $\alpha_i := h_i/H_n$.

Let $f_{PM^{i,0}}(x)$, for $i = 1, \ldots, n$, denote the probability density functions of the default PM durations for each class. Immediately from C4, one obtains

$$f_{R^i}(x) = \beta_i f_{PM^{i,0}}(\beta_i x), \forall i = 1, \ldots, n,$$

where $\beta_i := [(1/k^i) + (m_{T^i}/m_{T^{i,0}})]$. We define an alternating renewal process for the tool down and up time durations as follows. As a function of $m_{T^1}, \ldots, m_{T^n}$, the down time renewal process $\{R_j\}_{j=1}^{\infty}$ is a sequence of IID RVs. Let R denote an RV with the same probability density function as the R_j. We define this density as

$$f_R(x) := \sum_{i=1}^{n} \alpha_i f_{R^i}(x). \tag{3}$$

The up time renewal process $\{U_j\}_{j=1}^{\infty}$ we define as a sequence of exponentially distributed IID RVs. Let U denote such an RV. As a consequence of these definitions, we can immediately calculate the statistics required

for use in the $G/G/m$ formulae of equations (1) and (2). Let $\alpha := (\alpha_1, \ldots, \alpha_n)^T$ and $\mathbf{m_R} := (m_{R^1}, \ldots, m_{R^n})^T$. Denote the vector of second moments of the R^i as $\mathbf{m_{2,R}} := (E[(R^1)^2], \ldots, E[(R^n)^2])^T$. Recall that the coefficient of variation of a random variable X is defined as $c_X := \sqrt{E[(X-m_X)^2]}/m_X = \sigma_X/m_X$, where $m_X := E[X]$ and σ_X is the standard deviation of X.

The mean m_R and coefficient of variation C_R of the time to repair (which we have called the down time R) and mean time to failure m_F (which we have called the up time U) can thus be obtained as follows. We explicitly highlight their dependence upon the PM class cycle durations, which is our vector of decision variables $\mathbf{m_T} = (m_{T^1}, \ldots, m_{T^n})^T$.

$$m_R(\mathbf{m_T}) = \alpha^T \mathbf{m_R} \tag{4}$$

$$C_R^2(\mathbf{m_T}) = \frac{\alpha^T \mathbf{m_{2,R}}}{(m_R(\mathbf{m_T}))^2} - 1 \tag{5}$$

$$m_F(\mathbf{m_T}) = \frac{P_n}{H_n} - m_R(\mathbf{m_T}) \tag{6}$$

$$A(\mathbf{m_T}) = m_F(\mathbf{m_T})/(m_F(\mathbf{m_T}) + m_R(\mathbf{m_T})) \tag{7}$$

$$\rho(\mathbf{m_T}) = \frac{\lambda}{m\mu A(\mathbf{m_T})}. \tag{8}$$

The availability A and system loading ρ are provided as well. These expressions hold when there is only a single class of PMs as well.

These up and down time renewal processes, together with the arrival process, service process, etc., serve to fully specify our failure prone $G/G/m$ queue model. The motivation for our choices of these processes is rooted in the underlying PM cycles we strive to study. As such, the model shares many properties in common with how one expects the multiple PM cycles to behave practically.

Note again that we have constructed the PM cycle (up times and down times) as an alternating renewal process. As such, each tool is available until it "fails" (to start the PM). It returns again to availability once the PM has been completed.

3 OPTIMIZATION MODELS AND PROPERTIES

3.1 Nonlinear Program for the G/G/m Multiple PM Cycles Case

For the case of multiple PM cycles, we will employ the mean cycle time approximation of equation (2). All processes for our failure prone $G/G/m$ queue have been previously defined. For convenience, let $\mathcal{O}_1(\mathbf{m_T})$ denote our objective function. Use

$$\mathcal{O}_1(\mathbf{m_T}) := \frac{1}{\mu A(\mathbf{m_T})} + \frac{1}{\mu A(\mathbf{m_T})}\left(\frac{C_A^2 + C_{S,E}^2(\mathbf{m_T})}{2}\right)\frac{\rho(\mathbf{m_T})^{(-1+\sqrt{2m+2})}}{m(1-\rho(\mathbf{m_T}))}, \tag{9}$$

where $C_{S,E}^2(\mathbf{m_T}) := C_S^2 + (1 + C_R^2(\mathbf{m_T}))A(\mathbf{m_T})(1 - A(\mathbf{m_T}))m_R(\mathbf{m_T})\mu$. Our nonlinear program for the multiple PM $G/G/m$ queue, which we refer to as N_1, seeking good PM cycle durations follows:

$$\text{Min } \mathcal{O}_1(\mathbf{m_T}) \tag{10}$$

$$\text{subject to} \tag{11}$$

$$0 \le \rho(\mathbf{m_T}) \le 1, \tag{12}$$

$$0 \le L_i^{min} \le m_{T^i} \le L_i^{max}, \forall i = 1, \ldots, n, \tag{13}$$

where L_i^{min} and L_i^{max} are upper and lower bounds on the possible values for m_{T^i} that may be imposed by the fab. The first constraint ensures the system loading $\rho \le 1$. The objective function is an approximation

for the mean cycle time, so that an optimal solution for N_1 need not be exactly optimal for the $G/G/m$ queue.

Though we have not been able to prove it, in all examples studied thus far, the objective function is convex.

Conjecture 1 $\mathscr{O}_1(\mathbf{m_T})$ is convex in the region of inequality (12).

For the case of a failure prone $M/G/1$ queue, we employ the objective function $\mathscr{O}_2(\mathbf{m_T})$ as

$$\mathscr{O}_2(\mathbf{m_T}) := \frac{1}{\mu A(\mathbf{m_T})} + \frac{1}{\mu A(\mathbf{m_T})}\Big(\frac{\rho}{1-\rho}\Big)\Big(\frac{1}{2}\Big)\Big(1+C_S^2+\frac{(1+C_R^2(\mathbf{m_T}))A(\mathbf{m_T})(1-A(\mathbf{m_T}))m_R(\mathbf{m_T})\mu}{\rho(\mathbf{m_T})}\Big). \tag{14}$$

This is exactly the expression for the mean cycle time. Let N_2 be the nonlinear program seeking to minimize $\mathscr{O}_2(\mathbf{m_T})$ with the same constraints as N_1. Any optimal solution to N_2 the will provide an optimal solution for the $M/G/1$ queue. We similarly observe convexity of this objective function in the examples we have studied.

Conjecture 2 $\mathscr{O}_2(\mathbf{m_T})$ is convex in the region of inequality (12).

3.2 Nonlinear Program for the G/G/m Single PM Cycle Case

The nonlinear programs N_1 and N_2 for the $G/G/m$ and $M/G/1$ with multiple PM cycles, respectively, simplify when there is a single PM cycle. In these cases it is possible to strongly characterize the convexity of the objective functions. The proofs are omitted for brevity.

Proposition 3 With a single PM cycle, $\mathscr{O}_1(\mathbf{m_T})$ is strictly convex in the region of inequality (12).

Proposition 4 With a single PM cycle, $\mathscr{O}_2(\mathbf{m_T})$ is strictly convex in the region of inequality (12).

As a consequence, one might be able to use derivatives to obtain the optimal solution. The objective functions are differentiable in the feasible region. If the inflection point lies in the feasible region, it is the global optimum. More practically, there are very efficient algorithms to search for optimal solutions. In the case of N_1, an optimal solution provides an approximately optimal mean cycle time for the $G/G/m$ queue, since $\mathscr{O}_1(\mathbf{m_T})$ is an approximation to the cycle time. In the case of N_2, an optimal solution is in fact optimal for the system, since $\mathscr{O}_2(\mathbf{m_T})$ is exact.

4 NUMERICAL STUDIES

In this section, we consider two cases with PM and tool set data based on realistic industry values. They are not however taken directly from real tool sets. We consider:

- $M/G/1$ queue with two PM cycles.
- $M/G/2$ queue with two PM cycles.

Throughout, we solve the nonlinear programs using the Matlab Optimization Toolbox on an Intel dual core CPU PC running at 3.4 GHz with 4GB RAM. Solutions are obtained for all cases within a second.

4.1 An $M/G/1$ Example with Two PM Cycles

We consider the system with data given in Table 1. There the "-" means the value is determined by other parameters in the table (based on the model assumptions detailed in Section 3). The uniform distribution is denoted as $U(a,b)$. The Erlang and exponential distributions use standard notation. The constants k_i relating $SU^{i,0}$ to $PM^{i,0}$ are $k_1 := m_{PM^{1,0}}/m_{SU^{1,0}} = 22$ and $k_2 := m_{PM^{2,0}}/m_{SU^{2,0}} = 7.5$. C_{RV} is the coefficient of variation of the relevant RV.

The results of the optimization are provided in Tables 2 and 3. The cycle of PM 1 has been reduced from 240 hours to 55 hours. The cycle of PM 2 has been reduced from 720 hours to 425 hours. While the tool availability decreased by about four percentage points, and the loading increased correspondingly, the mean cycle time significantly decreased from 73.57 hours to 42.22 hours. This is because the effect

Table 1: Input parameters for an $M/G/1$ queue with two PM cycles.

RV	Distribution	Mean (hours)	C_{RV}
$SU^{1,0}$	-	3	-
$SU^{2,0}$	-	4	-
$PM^{1,0}$	Erlang$(1/33,2)$	66	$1/\sqrt{2}$
$PM^{2,0}$	Erlang$(1/15,2)$	30	$1/\sqrt{2}$
$T^{1,0}$	-	240	-
$T^{2,0}$	-	720	-
U^0	Exp$(1/119.75)$	119.75	1
Service	U(3.425,4.075)	3.75	0.05
Interarrival	Exp(0.15)	$6\frac{2}{3}$	1

of the WIP bubbles caused by the initial settings for the PM durations was quite large. Reducing those PM durations improves performance even though the availability decreased somewhat. The new PM plan promises to improve tool set performance. Since this is an $M/G/1$ model, the result is optimal for that system.

Table 2: Cycle results in an $M/G/1$ queue with two PM cycles.

Parameter	Original PM 1	Optimal PM 1	Original PM 2	Optimal PM 2
Mean cycle duration (hours)	$m_{T^{1,0}} = 240$	$m_{T^1} = 55$	$m_{T^{2,0}} = 720$	$m_{T^2} = 425$
Mean PM duration (hours)	$m_{PM^{1,0}} = 66$	$m_{PM^1} = 30$	$m_{PM^{2,0}} = 30$	$m_{PM^2} = 18$
Mean setup duration (hours)	$m_{SU^{1,0}} = 3$	$m_{SU^1} = 3$	$m_{SU^{2,0}} = 4$	$m_{SU^2} = 4$

Table 3: Overall results in an $M/G/1$ queue with two PM cycles.

Parameter	Original Value	Optimal Value
Mean down time (hours)	$m_{R^0} = 60.25$	$m_R = 18.62$
Mean up time (hours)	$m_{U^0} = 119.75$	$m_{PM} = 30.35$
Availability A	0.6653	0.6197
Loading ρ	0.7328	0.7866
Mean Cycle Time (hours) $E[CT]$	73.57	42.22

4.2 An $M/G/2$ Example with Two PM Cycles

We now consider the system of Subsection 4.1 with two servers $m = 2$ and arrival rate $\lambda = 0.25$. Otherwise, all parameters are as given in Table 1.

The results of the optimization are provided in Tables 4 and 5. The cycle of PM 1 has been reduced from 240 hours to 66 hours. The cycle of PM 2 has been reduced from 720 hours to 507 hours. While the tool availability decreased by about three percentage points, and the loading increased correspondingly, the estimated mean cycle time significantly decreased from 28.18 hours to 19.67 hours. Here again the initial PM durations were causing WIP bubbles and a reduction in PM durations, coupled with a corresponding decrease in availability, will actually improve performance. The new PM plan promises to improve tool set performance. Since this is an $M/G/2$ model, the result need not be optimal since the objective function $\mathscr{O}_1(\mathbf{m_T})$ is an approximation of the mean cycle time. However, the result should be good since the approximation is typically of good accuracy.

Table 4: Cycle results in an $M/G/2$ queue with two PM cycles.

Parameter	Original PM 1	Optimal PM 1	Original PM 2	Optimal PM 2
Mean cycle duration (hours)	$m_{T^{1,0}} = 240$	$m_{T^1} = 66$	$m_{T^{2,0}} = 720$	$m_{T^2} = 507$
Mean PM duration (hours)	$m_{PM^{1,0}} = 66$	$m_{PM^1} = 18$	$m_{PM^{2,0}} = 30$	$m_{PM^2} = 21$
Mean setup duration (hours)	$m_{SU^{1,0}} = 3$	$m_{SU^1} = 3$	$m_{SU^{2,0}} = 4$	$m_{SU^2} = 4$

Table 5: Overall results in an $M/G/2$ queue with two PM cycles.

Parameter	Original Value	Optimal Value
Mean down time (hours)	$m_{R^0} = 60.25$	$m_R = 21.65$
Mean up time (hours)	$m_{U^0} = 119.75$	$m_{PM} = 36.88$
Availability A	0.6653	0.6301
Loading ρ	0.7328	0.7737
Mean Cycle Time (hours) $E[CT]$	28.18	19.67

5 CONCLUDING REMARKS

We have proposed a practical method for PM planning that allows for multiple PM cycles. The approach seeks to minimize the cycle time of a $G/G/m$ queue by the selection of PM cycle durations. It extends previous work in this direction by allowing for the PM cycle durations to be continuous decision variables. It allows for multiple PM cycles by creating an alternating renewal process that can be used in the $G/G/m$ queue model and yet retains many of the key properties of the PM cycle data from which it is inspired.

The problem of finding the optimal cycle time and optimal PM cycle duration solutions is formulated as a nonlinear program consisting of a nonlinear objective function with linear constraints. The convexity of the objective function is explored. Two numerical examples are considered. They demonstrate that it may be possible to obtain significant improvements in mean cycle time compared to typical industry PM plans.

There are numerous directions for future work. While the equations used for the $M/G/1$ cases are exact, the $G/G/m$ models are approximate. Recently, Wu, McGinnis, and Zwart (2011) and Wu (2014) have shown that the common cycle time approximation formula we use may introduce systematic errors. These may be corrected by careful consideration of the types of failure modes under consideration and use of improved approximations appropriate to the context. The integration of plans obtained by the approaches considered here into fab operations optimization algorithms should be considered. The application of these ideas to fab level PM planning would be of interest. Working with an equipment supplier and fab to create plans that promise to improve fab performance is highly desirable.

ACKNOWLEDGMENTS

The authors are grateful for helpful discussions with Sanjay Rajguru of SEMATECH, USA.

REFERENCES

Avi-Itzhak, B., and P. Naor. 1963. “Some queueing problems with the service station subject to breakdown”. *Operations Research* 11 (3): 303–320.

Cassady, C. R., and E. Kutanoglu. 2005. “Integrating Preventive Maintenance Planning and Production Scheduling for a Single Machine”. *IEEE Transactions on Reliability* 54 (2): 304–309.

Cho, D. I., and M. Parlar. 1991. “A Survey of Maintenance Models for Multi-Unit Systems”. *European Journal of Operational Research* 51:1–23.

Davenport, A. 2010. "Integrated Maintenance Scheduling for Semiconductor Manufacturing". In *Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems*. Heidelberg, Berlin, Germany: Springer.

Dekker, R. 1996. "Applications of Maintenance Optimization Models: A Review and Analysis". *Reliability Engineering and System Safety* 51 (3): 229–240.

Kalir, A. 2013. "Segregating Preventive Maintenance Work for Cycle Time Optimization". *IEEE Transactions on Semiconductor Manufacturing* 26 (1): 125–131.

Marquez, A. C., J. N. D. Gupta, and A. S. Heguedas. 2003. "Maintenance Policies for a Production System with Constrained Production Rate and Buffer Capacity". *International Journal of Production Research* 41:1909–1926.

Morrison, J. R., and D. P. Martin. 2007. "Practical Extensions to Cycle Time Approximations for the $G/G/m$-Queue with Applications". *IEEE Transactions on Automation Science and Engineering* 4 (4): 523–532.

Ramirez-Hernandez, J. A., and E. Fernandez. 2010. "Optimization of Preventive Maintenance Scheduling in Semiconductor Manufacturing Models Using a Simulation-Based Approximate Dynamic Programming Approach". In *Proceedings of the 49th IEEE Conference on Decision and Control*, 3944–3949. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Ramirez-Hernandez, J. A., and E. Fernandez-Gaucherand. 2003. "An Algorithm to Convert Wafer to Calendar-Based Preventive Maintenance Schedules for Semiconductor Manufacturing Systems". In *Proceedings of the 42nd IEEE Conference on Decision and Control*, 5296–5931. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Tirkel, I. 2013. "The Effectiveness of Variability Reduction in Decreasing Wafer Fabrication Cycle Time". In *Proceedings of the Winter Simulation Conference*, edited by R. Pasupathy, S.-H. Kim, A. Tolk, R. Hill, and M. E. Kuhl, 3796–3805. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

van Dijkhuizen, G., and A. van Harten. 1998. "Two Stage Generalized Age Maintenance of a Queue Like Production System". *European Journal of Operational Research* 108 (2): 363–378.

Wu, K. 2014. "Classification of queueing models for a workstation with interruptions: a review". *International Journal of Production Research* 52 (3): 902–917.

Wu, K., L. McGinnis, and B. Zwart. 2011. "Queueing models for a single machine subject to multiple types of interruptions". *IIE Transactions* 43 (10): 753–759.

Yao, X., E. Fernandez-Gaucherand, M. C. Fu, and S. I. Marcus. 2004. "Optimal Preventive Maintenance Scheduling in Semiconductor Manufacturing". *IEEE Transactions on Semiconductor Manufacturing* 17 (3): 345–356.

AUTHOR BIOGRAPHIES

JAMES R. MORRISON is an Associate Professor in the Department of Industrial and Systems Engineering, KAIST, South Korea. He holds a B.S. in Mathematics and a B.S. in Electrical Engineering from the University of Maryland at College Park, USA. He received the M.S. and Ph.D. in Electrical and Computer Engineering from the University of Illinois at Urbana-Champaign, USA. He is a co-Chair of the IEEE RAS Technical Committee on Semiconductor Manufacturing Automation. His email address is james.morrison@kaist.edu.

HUNGIL KIM participated in this work during his M.S. studies at KAIST. He is currently a Researcher with the Defense Agency for Technology and Quality, South Korea. He holds a B.S. in Industrial Engineering from Pusan National University, South Korea and an M.S. in Industrial and Systems Engineering from KAIST, South Korea. His email address is khk12khk3@kaist.ac.kr.

ADAR A. KALIR is a Principal Engineer with the Fab/Sort Manufacturing Division of Intel Corporation, Qiriat-Gat, Israel. He holds a B.S. and M.S. in Industrial Engineering from Tel-Aviv University, Israel, and a Ph.D. in Industrial Engineering and Operations Research from Virginia Tech, USA. He also serves as an Adjunct Professor at Ben-Gurion University, Israel. His e-mail address is adar.kalir@intel.com.