A Real-time Maintenance Policy for Multi-stage Manufacturing Systems Considering Imperfect Maintenance Effects

Jing Huang¹, Student Member, IEEE, Qing Chang¹*, Member, IEEE, Jing Zou¹, Student Member, IEEE, and Jorge Arinez², Member, IEEE

¹Department of Mechanical Engineering, Stony Brook University, Stony Brook, NY 11794 USA
²General Motors Research and Development Center, Warren, MI 48090 USA

Corresponding author: Qing Chang (e-mail: qing.chang@stonybrook.edu).

This work is supported by the National Science Foundation Grants CMMI 1351160.

ABSTRACT Machines are subject to deterioration due to usage and aging. Once one machine fails, maintenance action has to be imposed in order to resume its operation. In practice, maintenance actions are not always able to restore the machine as new. Multi-level maintenance is a more realistic scenario. It is important to select a proper maintenance level upon failure since it directly relates to the duration of a maintenance action and the deterioration status after maintenance and therefore determines the overall system performance. In this paper, a real-time maintenance policy is proposed to select an optimal maintenance level aiming at reducing maintenance-related costs. The maintenance cost includes resource cost and production loss due to machine stoppage. The maintenance cost rate, which is the cost per unit time after the maintenance, is used to select the optimal maintenance level. The maintenance effect is modeled with a virtual-age approach. With the available data collected by distributed sensors, a data-driven modeling of serial production lines is established for analyzing production dynamics. The opportunity window of machine stoppage in the stochastic scenario is derived, which is further used to estimate the production loss caused by the machine stoppage due to maintenance. A numerical experiment is conducted to validate the proposed maintenance policy.

INDEX TERMS Data-driven modeling, Maintenance cost rate, Real-time maintenance policy, Multi-stage manufacturing system, Imperfect maintenance

I. INTRODUCTION

Manufacturers are seeking to operate the production system in a more efficient and cost-effective way when facing the growing competition and globalization. Since capital expenditure incurred by maintenance accounts for a large portion of the overall production cost, optimal maintenance decision making is one of the fundamental aspects in achieving this end.

Random failure occurs when a machine deteriorates to a certain level due to usage and aging. Upon the random failure, a maintenance action has to be carried out in order to restore the machine to an acceptable operating condition. In industrial practice, maintenance is not necessarily to replace the failed machine with a new one. Based on the structure of the machine, multi-level maintenance options could be available. A perfect maintenance, or replacement, is recovering the machine ‘as good as new’, while a minimal maintenance is recovering the machine ‘as bad as old’, which only resumes its operation without changing the deterioration status. Imperfect maintenances are recovering the machine to somewhere between old and new.

The options of different maintenance levels largely expand the researches on maintenance. Based on the multi-level maintenance effects, a lot of maintenance policies have been proposed in the past decades [1]–[3], most of which were based on the single-unit system. Decision making on maintenance is essentially a trade-off between production performance and maintenance cost. For single-unit systems, the unavailability of the system during maintenance can be directly counted towards the production loss. Therefore, the maintenance scheduling, which is modeled as a stochastic process, can be conveniently optimized such that some long-
term criterion, such as maintenance cost rate or system unavailability, is minimized.

In contrast with the policies for the single-unit system, according to Wang [4], the traditional maintenance policies for multi-unit systems fall into two categories, i.e. group maintenance policies [5], [6] and opportunistic policies [7], [8]. These policies generally originate from the observation that in a multi-unit system the failure of any one component will immediately halt the whole system, during which other components can be maintained simultaneously without incurring extra production losses. However, it is not necessarily the case in a general multi-stage manufacturing system, where machines work asynchronously and intermediate buffers reduce the spread of stoppages. It is noted that some studies [9], [10] on maintenance in multi-stage manufacturing systems are also based on the assumption that there are no buffers between machines. These studies, along with the policies for single-unit and multi-unit systems, might not work well for a general multi-stage manufacturing system.

Multi-stage manufacturing systems are characterized by their complex structures of strongly interconnected machines and stochastic dynamics. [11] It is difficult to obtain the optimal maintenance policy in a general multi-stage manufacturing system. First, the relationship between machine stoppage due to maintenance and system production loss is unclear. Our previous studies [12] reveal that system production loss due to machine stoppage heavily depends on the system states. For example, given the same maintenance duration on the same failed machine at a different time, the production loss could be very different since the system states are dynamic. But the system state at a point of time is very difficult to evaluate with transition probability, because it is well known that the closed-form solution of system states only exists for two-machine-one-buffer system and systems with infinite buffers or no buffers [13]. As a result, it is nearly impossible to obtain a globally optimal maintenance policy in the long run for a multi-stage manufacturing system.

To address the complexity, some simulation-based studies have been conducted to find the optimal maintenance schedule [14]. For instance, Arab et al. [15] incorporated remaining reliability of machines and work-in-process inventories into the simulation model to search for the optimal maintenance schedule. However, changes in the process and equipment, which are the norm in today’s manufacturing industry, will lead to corresponding changes and reconstructions in the simulation models, and subsequently a higher resource consumption. Nahas et al. [16] attempted to allocate the buffers subject to a limited total buffer space aiming at minimizing the maintenance costs. This is applicable in system design stage rather than the maintenance control on an existing system. References [17], [18] both study the integration of maintenance control and quality control in multi-stage manufacturing systems, but they focus on the relationship between machine status and product quality. Some works [12], [19], [20] try to find the opportunities for maintenance in a multi-stage manufacturing system, but the machines’ reliability and aging are ignored. Therefore, a systematic method is desired to manage maintenance through careful analyses of both the machine’s reliability and production dynamics [21].

Since information about production systems has become increasingly transparent, detailed, and real-time, some researchers attempted to obtain optimal control strategies on the system by utilizing the real-time data collected by distributed sensors. Zou et al. [22] was able to establish a data-driven model for production system analysis, in which the diagnosis and prognosis of production losses were established. The model has been applied to energy control [23], gantry assignment [24] and etc. in multi-stage manufacturing systems. Consequently, it is feasible to utilize a maintenance policy based on real-time system information and implemented in a real-time manner instead of a globally optimal maintenance policy. In this paper, the production losses caused by maintenance actions are properly estimated based on the data-driven model of the manufacturing system. The real-time maintenance cost rate is established. A real-time maintenance policy of selecting the proper level of maintenance upon machine failure is proposed.

The remainder of this paper is organized as follows. In Section II, we illustrate system assumptions and introduce notations used in this paper. In Section III, the maintenance effect is modeled with the virtual-age approach and the maintenance cost is analyzed, based on which the concept of real-time maintenance cost rate is proposed. In Section IV, the mathematical model and analysis of manufacturing systems are presented, based on which Section V estimates the production losses incurred by maintenance actions. In Section VI, the real-time maintenance policy used to select the optimal maintenance level is presented. The setup and results of numerical experiments are given in Section VII. Finally, the paper is concluded in Section VIII.

II. SYSTEM DESCRIPTION
A serial production line with $M$ machines (represented by rectangles) and $M - 1$ buffers (represented by circles) is as shown in Fig. 1. The continuous flow model is adopted in this paper. In contrast with the discrete-flow model, the continuous-flow models have buffer levels that vary from zero to its maximum capacity continuously.

![General structure of a serial production line](image)

**FIGURE 1. General structure of a serial production line**

The following notations are used in this paper:
- $S_i$ denotes the $i^{th}$ machine, where $1 \leq i \leq M$;
- $B_i$ denotes the $i^{th}$ buffer, where $2 \leq i \leq M$;
- $b_i(t)$ denotes the buffer level of $B_i$ at time $t$;
We make the following assumptions in this paper:

1) Each machine has a rated speed $1/T_i$, $i = 1, ..., M$ where $T_i$ is the cycle time of machine $S_i$. Each machine in the production line is deteriorating starting from time $t_{ij}$ and lasting for $d_{ij}$ units of time, and the maintenance type is $m_{ij}$.

2) Each buffer has a finite capacity. With the abuse of notations, the capacity of each buffer is denoted as $B_i$, $i = 2, 3, ..., M$.

3) The first machine $S_1$ is never starved and the last machine $S_M$ is never blocked.

4) Each machine in the production line is deteriorating with time, and the time to failure of machine $S_i$ follows a known distribution with probability density function $p_i(t)$.

5) The duration of each maintenance level is assumed to be deterministic.

6) For machine $S_i$, a replacement $(m = 1c)$ recovers the machine “as good as new”. A minimal maintenance $(m = N_i c)$ only resumes the machine without changing its aging status. An imperfect maintenance $(m = nc, n = 2, ..., N_i - 1)$ recovers the machine to somewhere between old and new. The recovery effect is not stochastic and can be described by a deterministic age reduction factor.

7) The maintenance staff and resources are adequate. The response and travel time of the maintenance personnel have been incorporated in the duration of each maintenance level.

III. MAINTENANCE MODELING AND COST ANALYSIS

A. MAINTENANCE MODELING

Given the failure-time distribution $p_i(t^*)$ of machine $S_i$, the failure rate is

$$\lambda_i(t^*) = \frac{p_i(t^*)}{1 - \int_0^{t^*} p_i(\tau)d\tau} \quad (1)$$

It can be shown that the failure rate $\lambda_i(t^*)$ is increasing when the failure-time distribution of the machine has the positive aging property. In this scenario, the imperfect maintenance or perfect maintenance options are preferred. Otherwise, if the failure rate $\lambda_i(t^*)$ of the machine is decreasing with respect to its age, a minimal maintenance is always preferred. The imperfect maintenance has been modeled through several approaches, including failure-rate reduction [25] and virtual-age reduction [26]. For the ease of derivation, we adopt the virtual-age approach to model the maintenance effects.

![FIGURE 2. Maintenance actions taken at a machine](image)

The age of machine $S_i$ is continuously increasing with time until it breaks down and a maintenance action has to be imposed. If the maintenance effect is not perfect, after maintenance the machine will behave as if it already has an initial age, which is referred to as virtual age. The virtual age of machine $S_i$ right after its $j^{th}$ maintenance action is denoted as $v_{ij}^m$. The superscript $m$ is used to denote a value derived by assuming the maintenance level to be $m$. When the maintenance level has been determined, the superscript would be forsaken hereafter. The virtual age $v_{ij}^m$ is

$$v_{ij}^m = A_i^m(v_{i(j-1)} + z_{ij(j-1)}) \quad (2)$$

where $v_{i(j-1)}$ is the virtual age of machine $S_i$ right after its $(j - 1)^{th}$ maintenance action, $z_{ij(j-1)}$ is its survival time after $(j - 1)^{th}$ maintenance action, and $A_i^m$ is the age reduction factor of maintenance level $m$. Clearly, $A_i^m = 0$ corresponds...
to a replacement, since the virtual age of machine is reduced to zero. \( A'^m_i = 0 \) corresponds to a minimal maintenance while \( 0 < A'^m_i < 1 \) relates to imperfect maintenance.

The time to failure \( t^* \) follows the distribution conditioning on the machine’s virtual age \( v'^m_{ij} \), i.e.

\[
p_i(t^*|v'^m_{ij}) = \frac{p_i(t^* + v'^m_{ij})}{1 - \int_0^{v'^m_{ij}} p_i(t) dt}, t^* > 0
\tag{3}
\]

At any time \( t \) after the \( j^{th} \) maintenance, the age of machine \( S_i \) is denoted as \( g_i(t) \).

\[
g_i(t) = v'^m_{ij} + (t - t_{ij} - d_{ij}), t > t_{ij} + d_{ij}
\tag{4}
\]

Let \( p_i(t, t^*) \) denote the probability that machine \( S_i \) fails at time \( t^* \) given it is on at current time \( t \). Then \( p_i(t, t^*) \) can be evaluated as

\[
p_i(t, t^*) = \frac{p_i(t^* + g_i(t))}{1 - \int_0^{g_i(t)} p_i(t) dt}, t^* > 0
\tag{5}
\]

The probability \( p_i(t^*|v'^m_{ij}) \) will be used to derive the expected maintenance cost rate in the following section, and the probability \( p_i(t, t^*) \) is important in evaluating the production loss risk in Section V.

### B. MAINTENANCE COST ANALYSIS

The maintenance cost consists of resource cost and production loss due to machine stoppage. The resource cost includes part replacement and other consumable expenses, which vary with maintenance level. A maintenance action \( \tilde{e}_{ij} = (i, m, t_{ij}, d_{ij}) \), depending on the maintenance level \( m \), the cost of \( j^{th} \) maintenance action on machine \( S_i \) is evaluated as

\[
C_{ij}^m = c^m + c_p (PL_{\tilde{e}_{ij}} + PLR_{\tilde{e}_{ij}})
\tag{6}
\]

where \( c^m \) is the resource cost, \( c_p \) is the profit per part, \( PL_{\tilde{e}_{ij}} \) and \( PLR_{\tilde{e}_{ij}} \) are the permanent production loss and production loss risk caused by \( \tilde{e}_{ij} \) respectively. The latter two terms heavily rely on system states and will be derived in following sections through careful analysis on the production line dynamics.

An optimal maintenance decision cannot be made solely based on the cost, since a replacement cost will probably be much higher than a minimal maintenance cost, but it ensures that the machine can operate for a longer period of time. Therefore the real-time maintenance cost rate \( R_{ij}^m \) is introduced as maintenance cost per unit time before next the failure arrives.

\[
R_{ij}^m = \frac{C_{ij}^m}{z_{ij}^m + d_{ij}^m}
\tag{7}
\]

where \( d_{ij}^m \) is the duration of maintenance, which is a deterministic value related to the maintenance type \( m \), and \( z_{ij}^m \) is the lifetime of machine \( S_i \) after the maintenance, which is a random variable following the distribution \( p_i(t^*|v'^m_{ij}) \) in Equation (3). The expected cost rate after its \( j^{th} \) maintenance is

\[
E[R_{ij}^m] = \int_0^\infty \frac{C_{ij}^m}{t^* + d_{ij}^m} p_i(t^*|v'^m_{ij}) dt^*
\tag{8}
\]

\( E[R_{ij}^m] \) is the expected cost per unit time before the next failure. When a random failure occurs at a machine, a proper level maintenance should be chosen to minimize the expected cost rate. This cost rate formulation will be used to develop the real-time maintenance policy in Section VII.

### IV. MANUFACTURING SYSTEM MODELING

#### A. DATA-DRIVEN MATHEMATICAL MODEL

The manufacturing system is a stochastic dynamic system that can be modeled by a state space equation:

\[
\dot{X}(t) = F(X(t), U(t), W(t))
\tag{9}
\]

The physical meanings of \( X(t) \), \( U(t) \), \( W(t) \) and \( F(s) \) are as following:

1. \( X(t) = [X_1(t), X_2(t), ..., X_M(t)]' \), where \( X_i(t) \) is the accumulated production count of machine \( S_i \) up to time \( t \)
2. \( U(t) = [u_1(t), u_2(t), ..., u_M(t)]' \), where \( u_i(t) \) is the control input at machine \( S_i \) at time \( t \). In this paper, it is a binary variable indicating whether or not machine \( S_i \) is under maintenance or not.
3. \( W(t) = [W_1(t), W_2(t), ..., W_M(t)]' \), where \( W_i(t) \) is the random interruption at machine \( S_i \) at time \( t \). It also is a binary variable describing whether the machine is suffering a random failure at time \( t \).
4. \( F(s) = [f_1(s), f_2(s), ..., f_M(s)]' \), where \( f_i(s) \) is the dynamic function of machine \( S_i \)

Machine \( S_i \) is operational only when there is neither random failure nor undergoing maintenance. We use \( \theta(t) = \{\theta_1(t), \theta_2(t), ..., \theta_M(t)\} \) to indicate whether or not machines are operational at time \( t \), where

\[
\theta_i(t) = u_i(t)(1 - W_i(t)) \tag{10}
\]

Based on the conservation of flow, the accumulated production counts of any two machines \( S_i \) and \( S_j \), \( \forall i, j \in \{1, 2, ..., M\}, i \neq j \), satisfy

\[
X_i(t) - X_j(t) = \left \{ \begin{array}{ll}
\sum_{k=i+1}^{j-1} b_k(t) - \sum_{k=i+1}^{j} b_k(0), & i < j \\
\sum_{k=j+1}^{i} b_k(t) - \sum_{k=j+1}^{i} b_k(0), & i > j
\end{array} \right . \tag{11}
\]

Such production difference cannot exceed a certain boundary, i.e.

\[
X_i(t) - X_j(t) \leq \beta_{ij} \tag{12}
\]

where \( \beta_{ij} \) is the condition that the buffer levels between two machines are full \((i < j)\) or empty \((i > j)\).

\[
\beta_{ij} = \left \{ \begin{array}{ll}
\sum_{k=i+1}^{j} b_k(t) - \sum_{k=i+1}^{j} b_k(0), & i < j \\
\sum_{k=j+1}^{i} b_k(t) - \sum_{k=j+1}^{i} b_k(0), & i > j
\end{array} \right . \tag{13}
\]

If machine \( S_i \) is on, when \( X_i(t) - X_j(t) = \beta_{ij} \), machine \( S_i \) will be constrained by machine \( S_j \), i.e. the speed of machine \( S_i \) will be reduced to that of machine \( S_j \); when \( X_i - X_j < \beta_{ij} \), machine \( S_i \) will not be constrained by \( S_j \).
\[
\dot{X}_i(t) = \min \left\{ \frac{\zeta \left( (X_i(t) - X_1(t)) - \beta_{1i} \right) \theta_1(t)}{T_1}, \ldots, \frac{\zeta \left( (X_i(t) - X_2(t)) - \beta_{2i} \right) \theta_2(t)}{T_2}, \ldots, \frac{\zeta \left( (X_i(t) - X_{M(t)}(t)) - \beta_{Mi} \right) \theta_M(t)}{T_M} \right\}
\]

where \( \zeta(x) = 1 \) for \( x > 0 \) and \( \zeta(x) = 0 \) otherwise, is a segment function used to determine whether machine \( S_j \) constrains machine \( S_i \) or not. Therefore, an operational machine will operate either at its own speed (when it is not constrained) or at the speed of other machines (when it is constrained). We can extend Equation (16) by comparing machine \( S_i \) with all other machines in the line.

\[
\dot{X}_i(t) = \min \left\{ \frac{\zeta \left( (X_i(t) - X_1(t)) - \beta_{1i} \right) \theta_1(t)}{T_1}, \ldots, \frac{\zeta \left( (X_i(t) - X_2(t)) - \beta_{2i} \right) \theta_2(t)}{T_2}, \ldots, \frac{\zeta \left( (X_i(t) - X_{M}(t)) - \beta_{Mi} \right) \theta_M(t)}{T_M} \right\}
\]

(17)

In this way, the dynamic function \( F(*) \) of the production system is defined. The buffer level \( b(t) = [b_1(t), b_2(t), \ldots, b_M(t)]^{T} \) of buffer \( B_i \) at any time \( t \) is

\[
b_i(t) = b_i(0) + X_i(t) - X_{i-1}(t)
\]

(18)

The machine deterioration will impact the system in the form of random failures. The random failure immediately stops the machine and the machine resumes operation only when the maintenance action is completed. The maintenance level chosen upon failure directly determines the duration of downtime and leads to different system states and production performance. It also determines the machines’ deterioration process in the future. Therefore, it is critical to determine which level of maintenance should be taken upon random failures.

**B. OPPORTUNITY WINDOW**

In order to select the optimal maintenance level, it is necessary to identify the impact of machine stoppages on the whole system. In a multi-stage production line, a stoppage at one machine doesn’t necessarily lead to permanent production loss [12]. Our previous studies proposed that there might exist a threshold for machine stoppage time, below which the system can eventually recover without losing any production throughput. To keep the paper self-contained, we briefly summarize the related basic concepts without detailed proof (Equations (19) to (21)). We need to use the conclusion from our previous works on maintenance policy analysis.

**Definition 1.**[27], [28] Opportunity window is the longest possible disruption time on machine \( S_i \) at time \( t \) that does not result in permanent production loss at the end-of-line machine, i.e.

\[
OW_i(t) = \sup \left\{ d \geq 0: \text{s.t. } \exists T^*(d), \int_{0}^{T} X_M(\tau) d\tau \right\}
\]

(19)

where \( \int_{0}^{T} X_M(\tau) d\tau \) and \( \int_{0}^{T} X_M(\tau; \tilde{e}) d\tau \) are the production counts of the end-of-line machine \( S_M \) at time \( T \) with and without disruption event \( \tilde{e} = (i, t, d) \), respectively. \( T^*(d) \) signifies the potential dependency of \( T^* \) on \( d \).

If the duration of a downtime event is shorter than the machine’s opportunity window, the temporary stoppage is not leading to any production loss. If the downtime event lasts longer than its corresponding opportunity window, the whole production line will permanently lose some amount of production throughput. Once the opportunity window at a certain time has been identified, the permanent production loss of various maintenance options can be effectively estimated. To derive the opportunity window at time \( t \), we further introduce Proposition 1.

**Proposition 1.**[28] Given a realization of a production process subject to a sequence of disruption event \( E = [\tilde{e}_1, \tilde{e}_2, \ldots, \tilde{e}_n] \) and suppose \( \max \{t_i + d_i\} < T \) if the last slowest machine \( S_{M^*} \) stops for a duration of \( D \) during time \([0, T)\), then for the end-of-line machine \( S_M \), \( \exists T^* \geq T \), s.t.

\[
P_M(T^*) - P_M(T^*; E) = \frac{D}{T_{M^*}}, \forall T' > T^*
\]

(20)

where \( P_M(T^*) \) and \( P_M(T^*; E) \) are the output of the end-of-line machine \( S_M \) without and with disruption events \( E = [\tilde{e}_1, \tilde{e}_2, \ldots, \tilde{e}_n] \) during \([0, T')\), and \( T_{M^*} \) is the cycle time of \( S_{M^*} \).

Proof. See Section 4 of reference [28].

According to Proposition 1, any stoppage of the slowest machine \( S_{M^*} \) will lead to a corresponding permanent production loss. The permanent production loss caused by downtimes \( E = [\tilde{e}_1, \tilde{e}_2, \ldots, \tilde{e}_n] \) is denoted as \( PL_E \).

\[
PL_E = D \cdot \frac{1}{T_{M^*}}
\]

(21)

Note that the stoppage of slowest machine refers to not only the downtimes at machine \( S_{M^*} \) itself but also the starvation and blockage of machine \( S_{M^*} \) caused by other stopped machines. For the ease of further derivation, the buffer levels (if \( i < M^* \)) or buffer vacancies (if \( i > M^* \)) between machine \( S_i \) and machine \( S_{M^*} \) are denoted as \( \sigma_i(t) \).

\[
\sigma_i(t) = \begin{cases} 
\sum_{k=1}^{M^*} b_k(t), & i < M^* \\
0, & i = M^* \\
\sum_{k=M^*+1}^{M^*} (B_k - b_k(t)), & i > M^* 
\end{cases}
\]

(22)

The opportunity window of machine \( S_i \) is the time duration before all the buffers between machine \( S_i \) and \( S_{M^*} \) are exactly empty (if \( i < M^* \)) or full (if \( i > M^* \)) assuming machine \( S_i \) stops at time \( t \). In other words, during time span \([t, t + OW_i(t)]\), the production count of machine \( S_{M^*} \) is exactly \( \sigma_i(t) \). Therefore, \( OW_i(t) \) satisfies

\[\vdots\]
In the clean case, where we assume all stations operate without downtimes, the speed of the slowest machine is constant, i.e., $\dot{X}_M(t) = 1/T_M$. Insert the constant speed into Equation (23), yields

$$\overline{OW}_i(t) = T_M \cdot \alpha_i(t)$$

(24)

where $\overline{OW}_i(t)$ denotes the opportunity window of machine $S_i$ at time $t$ in clean-case scenario. However, in stochastic scenario, $OW_i(t)$ is not a deterministic term since $\dot{X}_M(t)$ depends on future control inputs and unexpected random failures. But it will be helpful to relate $OW_i(t)$ with its special realization under clean-case assumption.

**Proposition 2.** Given a stochastic production line subject to a sequence of downtime events $E$, let $\overline{OW}_i(t)$ and $\overline{OW}_i(t)$ be the opportunity window of machine $S_i$ at time $t$ with and without $E$ respectively, then

$$\overline{OW}_i(t) = \overline{OW}_i(t) + PL_E[t, t + \overline{OW}_i(t)] \cdot T_M.$$  

(25)

where $PL_E[t, t + \overline{OW}_i(t)]$ is the permanent production loss caused by $E$ during a time range from $t$ to $t + \overline{OW}_i(t)$.

Proven. See Appendix.

Note that $PL_E[t, t + \overline{OW}_i(t)]$ is a non-negative value, therefore the opportunity window of machine $S_i$ in clean-case scenario is the lower bound of the opportunity window in stochastic scenario at the same time instant, i.e. $\overline{OW}_i(t) \leq OW_i(t)$. According to Proposition 2, the size of the opportunity window depends not only on the instantaneous buffer status at time $t$ but also on current and future downtime events $E$.

**V. PRODUCTION LOSS EVALUATION OF MAINTENANCE ACTION**

The maintenance action causes machine stoppage and leads to different system states. The impacts of a maintenance action $\tilde{e}_{ij}$ on the system are twofold, i.e., permanent production loss $PL_{\tilde{e}_{ij}}$ and production loss risk $PLR_{\tilde{e}_{ij}}$. The permanent production loss is the direct outcome of the machine stoppage, and the production loss risk results from the disturbance to overall system states caused by $\tilde{e}_{ij}$.

**A. PERMANENT PRODUCTION LOSS EVALUATION**

Opportunity window is the largest possible stoppage duration of a machine before such stoppage induces permanent production loss. As discussed in the previous section, the opportunity window depends not only on the current buffer status but also on the downtime events $E$. In the stochastic scenario, the full downtime event list $E$ is unknown since there will be unexpected random failures in the future. However, in the context of maintenance, a subset $\tilde{E}$ of downtime events $E$ is known, namely, those maintenance actions already initiated before or right at current time $t$ and lasting beyond time $t$.

Since $\tilde{E} \subset E$, it can be concluded that $PL_{\tilde{E}} \leq PL_E$. Then the opportunity window estimated with $\overline{OW}_i(t)$ is the lower bound of that computed with $E$. Therefore the subset $\tilde{E}$ can be used instead of $E$ in Proposition 2 to safely estimate $OW_i(t)$.

$$\overline{OW}_i(t) = \overline{OW}_i(t) + PL_{\tilde{E}}[t, t + \overline{OW}_i(t)] \cdot T_M.$$  

(26)

Consider a maintenance action $\tilde{e}_{ij} = (i, m_{ij}, t_{ij}, d_{ij})$, if $d_{ij} \leq \overline{OW}_i(t)$, the maintenance won’t incur permanent production loss; however, if $d_{ij} > \overline{OW}_i(t)$, then the slowest machine will be stopped and the production time loss is the excessive time that $\tilde{e}_{ij}$ lasts beyond $\overline{OW}_i(t)$. To conclude, the permanent production loss of a maintenance action $\tilde{e}_{ij}$, denoted as $PL_{\tilde{e}_{ij}}$, is

$$PL_{\tilde{e}_{ij}} = \max \left\{ \frac{d_{ij} - \overline{OW}_i(t_{ij})}{T_M}, 0 \right\}.$$  

(27)

**B. PRODUCTION LOSS RISK EVALUATION**

Given a maintenance $\tilde{e}_{ij}$, if it lasts longer than its opportunity window, it will cause permanent production loss $PL_{\tilde{e}_{ij}}$ and the loss will last until the maintenance action is completed. Hence $PL_{\tilde{e}_{ij}}$ only measures the impact of $\tilde{e}_{ij}$ on the system within its presence. But the downtime event $\tilde{e}_{ij}$ will impact the system in a more profound manner since it leads to totally different system states. The impact of a maintenance action beyond the "current" permanent production loss will be evaluated in this section.

The stoppage at machine $S_i$ due to maintenance action $\tilde{e}_{ij}$ spread to nearby machines sequentially when it causes blockage or starvation. The opportunity windows of nearby machines gradually alter. Considering a subsequent downtime event, the permanent production loss incurred by $\tilde{e}_{ij}$ may also be changed due to the altered opportunity windows. In principle, the total loss cannot be directly attributed to the...
initial action $\tilde{e}_{ij}$, but $\tilde{e}_{ij}$ does indirectly impact the loss of the subsequent downtime event.

To ensure the resilience of the system to future random failures, we take the difference of production losses of the very first subsequent downtime event with and without $\tilde{e}_{ij}$ as production loss risk, denoted as $PLR_{\tilde{e}_{ij}}$.

Suppose that after time $t_{ij}$, the very first random failure occurs on machine $S_k$ ($k = 1, 2, ..., M$) at time $t^*$ and a replacement is taken. Since $\tilde{e}_{ik}$ is the first random failure after time $t_{ij}$, there is no other downtime event between $t_{ij}$ and $t_{ij} + t^*$. Only downtime events $\tilde{e}_{ij}$ and $\tilde{e}_{ik}$ are appended to the known downtime list $\tilde{E}$. The full downtime list $E$ is

$$E = [\tilde{E}, \bar{e}_{ik}, \bar{e}_{kj}]$$

With the current system state and fully observed downtimes $E$ in the future, using Equation (17)-(18) the buffer levels $b(t_{ij} + t^*)$ at time $t_{ij} + t^*$ can be derived, which further can be used to compute opportunity window $OW_k(t_{ij} + t^*)$. Then the permanent production loss $PL_{\tilde{e}_{k*}}(t^*)$ caused by $\tilde{e}_{k*}$ can be evaluated as

$$PL_{\tilde{e}_{k*}}(t^*) = \max\left\{ \frac{d_{kc}^l - OW_k(t_{ij} + t^*)}{T_{M^*}}, 0 \right\}$$

The probability associated with $PL_{\tilde{e}_{k*}}(t^*)$ is $p(k,t_{ij},t^*)$, which is the probability that the first random failure arrives at machine $S_k$ at time $t_{ij} + t^*$. The machines are independent with each other regarding reliability. Therefore $p(k,t_{ij},t^*)$ is the joint probability of $M$ machines, i.e.

$$p(k,t_{ij},t^*) = p_k(t_{ij},t^*) \prod_{i=1, i \neq k}^M \left[ 1 - \int_0^{t^*} p_i(t_{ij},\tau) d\tau \right]$$

The expected value of production loss of first downtime event with $\tilde{e}_{ij}$ can be evaluated as

$$E[PL_{\tilde{e}_{ij}}] = \sum_{k \neq 1}^M \int_0^\infty p(k,t_{ij},t^*)PL_{\tilde{e}_{k*}}(t^*)dt^*$$

Note that $PL_{\tilde{e}_{k*}}(t^*)$ is directly caused by the potential random failure $\tilde{e}_{k*}$, not the initial downtime event $\tilde{e}_{ij}$. The impact of $\tilde{e}_{ij}$ lies in that it might alter the value of $PL_{\tilde{e}_{k*}}(t^*)$. The production loss of $\tilde{e}_{k*}$ without $\tilde{e}_{ij}$ should also be evaluated in order to identify the real impact of $\tilde{e}_{ij}$. Considering a scenario without the downtime event $\tilde{e}_{ij}$, the full downtime list $\tilde{E}$ is

$$\tilde{E} = [\tilde{E}, \bar{e}_{kj}]$$

Following the similar aforementioned procedure, at time $t_{ij} + t^*$, the buffer levels $\bar{b}(t_{ij} + t^*)$, opportunity window $\bar{OW}_k(t_{ij} + t^*)$, and production loss $\bar{PL}_{\tilde{e}_{k*}}(t^*)$ can be computed in sequence. Therefore the expected production loss of first downtime event $\bar{e}_{k*}$ without $\tilde{e}_{ij}$ is

$$E[\bar{PL}_{\bar{e}_{ij}}] = \sum_{k \neq 1}^M \int_0^\infty \bar{p}(k,t_{ij},t^*)\bar{PL}_{\bar{e}_{k*}}(t^*)dt^*$$

where $\bar{p}(k,t_{ij},t^*)$ is the probability that machine $S_k$ fails at time $t_{ij} + t^*$. Note that $\bar{p}(k,t_{ij},t^*) \neq p(k,t_{ij},t^*)$, since $\bar{p}(k,t_{ij},t^*)$ is derived by assuming that machine $S_i$ doesn’t receive $\tilde{e}_{ij}$. Finally, $PLR_{\tilde{e}_{ij}}$ can be estimated by the difference of expected production losses with and without $\tilde{e}_{ij}$.

$$PLR_{\tilde{e}_{ij}} = E[PL_{\tilde{e}_{ij}}] - E[\bar{PL}_{\bar{e}_{ij}}]$$

The term $PLR_{\tilde{e}_{ij}}$ is the impact of $\tilde{e}_{ij}$ on the production loss of next random failure. By incorporating it into the maintenance cost rate function, a maintenance decision at the current time would always consider its impact on the whole system in future time.

VI. REAL-TIME MAINTENANCE POLICY

From the evaluation of production losses, we can conclude that the maintenance cost for machines in a manufacturing system heavily depends on the real-time system state. As discussed in Section I, it is extremely difficult to find an optimal maintenance policy for a global time horizon. Therefore, a feasible approach is to develop a control policy implemented on a real-time basis to obtain the near-optimal maintenance decisions.

![FIGURE 4. The framework of real-time maintenance control policy](image)

In this problem, the decision time is whenever a random failure is detected and the decision variable is the proper maintenance level that should be imposed on the failed machine. The structure of the control framework is as shown in Fig. 4. One can refer to Section IV for the physical meaning of each component. The feedback maintenance control inputs $U(t)$ are determined by the real-time states of the production system, i.e. $U(t) = G(X(t), W(t))$.

At any time $t$, distributed sensors monitor machine operation status in the system. If a random failure at machine $S_i$ is detected, i.e. $W_i(t) = 1$, then a maintenance action should be taken on machine $S_i$. Depending on the structure of the machine, maintenance of multiple levels $m_i = 1c, 2c, ..., N_i c$ might be available for the failed machine. For each eligible maintenance level $m_i$, the duration of maintenance $d_i^m$ and age reduction factor $A_i^m$ are known. By assuming a maintenance action $\tilde{e}_{im} = (i, m, t, d_i^m)$ at machine $S_i$, the expected maintenance cost rate $E[R_i^m]$ can be derived according to Equation (8). Then the optimal maintenance level $m_i$ should minimize the expected cost rate, i.e.

$$m_{is} = \arg\min_m E[R_i^m], m = 1c, 2c, ..., N_i c$$

VII. CASE STUDY

To demonstrate the effectiveness of the proposed real-time maintenance policy, extensive numerical studies are performed. In baseline policies, upon random failures, the
maintenance level is static. The other policy is the real-time maintenance policy proposed in this paper. The overall profit of the system, denoted as $\mathcal{P}(T)$, is taken as the main performance measure.

$$\mathcal{P}(T) = \text{Total Revenue} - \text{Total Cost}$$

$$= c_p \cdot X_M(T) - \mathcal{M}(T)$$

(36)

where $c_p \cdot X_M(T)$ is the production revenue and $\mathcal{M}(T)$ is the total maintenance resource cost during time span $[0, T]$. Let $K^m(T), m = 1c, 2c, ...$ denotes the total number of level-$m$ maintenance received by the machine $S_i$ up to time $T$. The total maintenance resource cost is

$$\mathcal{M}(T) = \sum_{i=1}^{M} \sum_{m=1c, ... N_c} K^m_i(T)c^m_i$$

(37)

We construct 50 different serial production lines by randomly selecting machine and buffer parameters from the following sets:

$$M \in \{3, 20\}$$

$$T_i \in [1, 5]\text{ min}, i = 1, 2, ..., M$$

$$B_i \in [2, 40], i = 2, 3, ..., M$$

$$b_i(0) \in [0, B_i], i = 2, 3, ..., M$$

In this case study, failure-time of the machines are assumed to follow Weibull distribution, which is a typical increasing-rate distribution widely used in machine reliability analysis. The probability density function of failure-time of machine $S_i$ is given as

$$p_i(t^*) = \exp \left(-\left(\frac{t^*}{\alpha_i}\right)\right)$$

(38)

where $\alpha_i$ is the scale parameter, and $\beta_i$ is the shape parameter. The conditional probability of fail-time given its virtual age $v_{ij}$ is

$$p_i(t^*|v_{ij}^m) = \frac{\beta_i}{\alpha_i} \left(\frac{v_{ij}^m}{\alpha_i}\right)^{\beta_i} \exp \left(-\left(1 + \left(\frac{v_{ij}^m}{\alpha_i}\right)^{\beta_i}\right)\right)$$

(39)

For every machine $S_i$ in randomly generated production lines, the shape parameter is $\beta_i = 2$, and the scale parameter $\alpha_i$ and initial virtual age $v_{i0}$ are randomly generated according to following set:

$$\alpha_i \in [500, 2000]\text{ min}$$

$$v_{i0} \in [0, 1000]\text{ min}$$

For the ease of implementation and further analysis, the maintenance parameters of each machine are set to be identical. Four maintenance levels for each machine are given, namely replacement ($1c$), minimal maintenance ($4c$), and two levels of imperfect maintenance ($2c, 3c$). The parameters regarding each maintenance level are shown in Table I.

<table>
<thead>
<tr>
<th>Maintenance Level $m$</th>
<th>Age Reduction Factor $A^m_i$</th>
<th>Maintenance Duration $d^m_i$ (min)</th>
<th>Resource Cost $c^m_i$ (US Dollar)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1c</td>
<td>0</td>
<td>30</td>
<td>2000</td>
</tr>
<tr>
<td>2c</td>
<td>0.3</td>
<td>20</td>
<td>1200</td>
</tr>
<tr>
<td>3c</td>
<td>0.6</td>
<td>15</td>
<td>900</td>
</tr>
<tr>
<td>4c</td>
<td>1.0</td>
<td>5</td>
<td>300</td>
</tr>
</tbody>
</table>

Fig. 5 illustrates the simulation results comparison. In all 50 production lines, the overall profits using real-time policy are greater than those using static maintenance policies. On average, the overall profit using real-time policy is $31.36\%$, $11.48\%$ and $15.37\%$ greater than those using policy $1c$, policy $2c$, and policy $3c$ respectively.

To further analyze our policy, a special production line is constructed. All six machines in the line have identical reliability parameters, but different cycle time due to process planning. The slowest machine in the line is $S_4$. We calculate the total numbers of replacement and imperfect maintenance (including minimal maintenance) under the real-time maintenance policy during one-time simulation.

As shown in Table II, although all the machines are identical regarding reliability parameters, it is noted that the slowest machine $S_4$ received more replacements and less imperfect maintenances than any other machines. The replacement number decreases and the imperfect maintenance number increases for the machine further away from the slowest machine. The machine far away from the slowest machine is more likely to take imperfect maintenance than
replacement. Besides the stochastic factors, the reason for this phenomenon is mainly that the opportunity window tends to be larger for the machines far away from the slowest machine. Imperfect maintenances, which typically last shorter than a replacement, is less likely to induce permanent production loss on these machines and therefore is preferable to replacement.

<table>
<thead>
<tr>
<th>Machine Number</th>
<th>Total Replacement Number</th>
<th>Total Imperfect Maintenance Number</th>
<th>Total Maintenance Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>S_1</td>
<td>27</td>
<td>66</td>
<td>93</td>
</tr>
<tr>
<td>S_2</td>
<td>27</td>
<td>66</td>
<td>93</td>
</tr>
<tr>
<td>S_3</td>
<td>28</td>
<td>63</td>
<td>91</td>
</tr>
<tr>
<td>S_4 (slowest)</td>
<td>32</td>
<td>45</td>
<td>77</td>
</tr>
<tr>
<td>S_5</td>
<td>29</td>
<td>54</td>
<td>83</td>
</tr>
<tr>
<td>S_6</td>
<td>27</td>
<td>79</td>
<td>106</td>
</tr>
</tbody>
</table>

One may also note that the total maintenance number is increasing with the distance from the slowest machine. The decrease in replacement times would inevitably impair the reliability of the machine, and thus more random failures can be expected. However, since these machines have larger opportunity windows, they are more resilient to random failures. It is reasonable for these machines to take more imperfect maintenances in order to reduce resource costs. This case illustrates that the maintenance decisions could be very different for identical machines when they are placed at different locations in a multi-stage manufacturing system.

To conclude, the real-time maintenance policy proposed in this paper is effective to choose proper maintenance levels in accordance with the production system dynamics.

VIII. CONCLUSION

In this paper, the real-time maintenance cost rate is established to facilitate the real-time decision making on multi-level maintenance in a multi-stage manufacturing system. Based on the data-driven mathematical model of the manufacturing system, the permanent production loss and production loss risk incurred by the maintenance action is derived. The proposed control framework is able to make cost-effective maintenance decisions considering multiple maintenance levels. In the future, we will consider the stochastic duration and effects of maintenance. The maintenance task prioritization with limited maintenance staff and resources also needs to be discussed in future works.

APPENDIX

**Proof of Proposition 2.**

Compare Equation (23) and (24), we have

\[ \int_{0}^{\tau} \dot{X}_{M}^{(t)}(\tau) d\tau = \bar{W}_{1}(t) \cdot \frac{1}{T_{M}^{*}} \]  
(A.1)

Since machine \( S_{M}^{*} \) is the slowest machine, i.e. \( 1/T_{M}^{*} < 1/\tau \), \( \forall \tau \neq M^{*} \), according to Equation (17), it can only operate at a speed of either \( 1/T_{M}^{*} \) or zero, i.e.

\[ \dot{X}_{M}^{(t)}(\tau) \in \left\{ 0, \frac{1}{T_{M}^{*}} \right\}, \forall \tau > 0 \]  
(A.2)

We can divide the time span into segments according to the actual speed of machine \( S_{M}^{*} \). During time span \([ t, t + OW_{1}(t)]\), let \( ow_{1}^{0}(t) \) denotes the accumulated time length when \( X_{M}^{(t)}(\tau) = 0 \), and \( ow_{1}^{1}(t) \) denotes the accumulated time length when \( X_{M}^{(t)}(\tau) = 1/T_{M}^{*} \), i.e. \( \tau \in [t, t + OW_{1}(t)] \). Note that

\[ OW_{1}(t) = ow_{1}^{0}(t) + ow_{1}^{1}(t) \]  
(A.3)

The left-handed term of Equation (A.1) can be rewritten as

\[ \int_{t}^{t+OW_{1}(t)} \dot{X}_{M}^{(t)}(\tau) d\tau = ow_{1}^{0}(t) \cdot 0 + ow_{1}^{1}(t) \cdot \frac{1}{T_{M}^{*}} \]  
(A.4)

Obviously, by comparing Equation (A.4) and (A.1), it can be concluded that

\[ ow_{1}^{1}(t) = \bar{W}_{1}(t) \]  
(A.5)

During \([ t, t + OW_{1}(t)]\), the slowest machine \( S_{M}^{*} \) stops (when \( X_{M}^{(t)}(\tau) = 0 \)) for totally \( ow_{1}^{0}(t) \) units of time. According to Proposition 1, the permanent production loss during time span \([ t, t + OW_{1}(t)]\) is

\[ PL_{E}(t,t + OW_{1}(t)) = ow_{1}^{0}(t) \cdot 1/T_{M}^{*}, \text{ i.e.} \]

\[ ow_{1}^{0}(t) = PL_{E}(t,t + OW_{1}(t)) \cdot T_{M}^{*} \]  
(A.6)

Therefore, we have

\[ OW_{1}(t) = \bar{W}_{1}(t) + PL_{E}(t,t + OW_{1}(t)) \cdot T_{M}^{*} \]  
(A.7)

End of proof.

ACKNOWLEDGMENT

This work was supported by the U.S National Science Foundation (NSF) Grant No. CMMI 1351160.

REFERENCES


