



Queueing Theory Model Based on Bivariate Birth/Birth-Death Stochastic Processes to Analyze Manufacturing Systems under Limited Maintenance Resources

Hasan Rasay ¹ , Jalal Taji ² , Amir-Mohammad Golmohammadi ³ , Mohammad Biglari Kami ⁴ 

1. Corresponding author, Department of Industrial Engineering, Kermanshah University of Technology, Kermanshah, Iran. E-mail: H.Rasay@kut.ac.ir

2. Department of Industrial Engineering, Faculty of Engineering, University of Bojnord, Bojnord, Iran. E-mail: j.taji@ub.ac.ir

3. Department of Industrial Engineering, Faculty of Engineering, Arak University, Arak, Iran. E-mail: a-golmohammadi@arak.ac.ir

4. Department of Industrial Engineering, Kermanshah University of Technology, Kermanshah, Iran. E-mail: M.Biglari.k@kut.ac.ir

Article Info

Article type:

Research Article

Article history:

Received March 29, 2026

Received in revised form June 12, 2026

Accepted August 17, 2026

Available online August 25, 2026

Keywords:

multi-unit production systems

maintenance

queueing theory

birth-death process

operational state

transition rate

ABSTRACT

Objective: This paper analyzes maintenance scheduling in systems comprising multiple identical parallel production units using queueing theory. Most queueing-based studies of such systems assume two states for each unit: operational (healthy) and failed. We model each unit with several operational states in addition to a failed state, with the system state represented by a bivariate state vector.

Methods: Given the state of the units, different maintenance actions are conducted by a crew capable of addressing a limited number of units concurrently. Based on birth-death process and queueing theory with a limited population and number of servers, the transition rate diagram (TRD) and equilibrium equations are derived. The system TRD combines two TRDs: (i) degradation without maintenance effects and (ii) maintenance actions. Using Little's laws and the TRD, essential performance metrics and steady-state probabilities are computed.

Results: The developed model provides steady-state probabilities for different system states and enables the calculation of key performance measures related to production-unit degradation, failure, and maintenance. Numerical examples demonstrate the proposed framework and quantify system performance under the considered maintenance conditions.

Conclusion: The proposed approach provides a framework for analyzing maintenance scheduling in systems with identical parallel production units, considering multiple operational states, degradation, and limited maintenance capacity.

Cite this article: Rasay, H., Taji, J., Golmohammadi, A. M., and Biglari Kami, M. (2026). Queueing Theory Model Based on Bivariate Birth/Birth-Death Stochastic Processes to Analyze Manufacturing Systems under Limited Maintenance Resources. *International Journal of Supply and Operations Management*, XX(X), pages. <https://doi.org/10.22034/ijsum.2026.110975.3529>



© Author(s) retain the copyright.

Publisher: Kharazmi University

DOI: <https://doi.org/10.22034/ijsum.2026.110975.3529>

1. Introduction

Multi-unit manufacturing systems operate through the concurrent functioning of several identical units within a production sector, forming a complex and highly interconnected structure that drives overall output. These systems rely on parallel processing, with each element executing similar tasks independently. Effective coordination and synchronization among the units are essential to maximize efficiency and productivity. Managing such systems presents challenges related to maintenance scheduling, reliability assurance, and resource distribution, which demand sophisticated analytical methods to capture system dynamics and support better decision-making for sustainable performance. In this context, Taji et al. (2022, 2023) advanced the field by introducing integrated maintenance frameworks designed to improve the performance of repairable, multi-unit systems. Wakiru et al. (2021) highlights the significance of harmonized maintenance across units, complementing the findings of Salari and Makis (2017), who examined maintenance strategies for multi-unit systems under varying production and demand conditions. Additionally, Wakiru et al. (2021) extended their work by investigating combined approaches that integrate remanufacturing, maintenance, and spare management to prolong the operational lifespan of multi-component systems. The survey conducted by Cho and Parlar (1991) contributes historical perspectives, encompassing maintenance models specifically tailored for multi-unit systems. Furthermore, the study by Salari and Makis (2020) addresses integrated strategies for maintenance scheduling and synchronized inventory replenishment, shedding light on effective strategies for multi-unit production systems. Sarkar and Suresh Kumar (2012) further contribute by examining maintenance policies, both historical and contemporary, for both single and multi-unit systems, providing a comprehensive understanding of the evolution in this critical field of study.

Expanding on the evolution of industrial frameworks, Lee et al. (2011) investigate self-healing mechanisms and 'engineering immune systems' as pathways toward autonomous manufacturing. Their work introduces proactive strategies for managing resource limitations through intelligent machine design. Complementing this, Colledani et al. (2014) explore the integration of system design and operational management, emphasizing how production quality is inextricably linked to the broader maintenance and management context, and underscoring the importance of an integrated approach to resource optimization for sustained operational excellence in multi-unit manufacturing environments with limited maintenance resources.

To address the complexities arising from limited maintenance resources in multi-unit manufacturing environments, queueing theory emerges as a powerful analytical tool for modeling stochastic interactions between system failures, repairs, and resource availability. Production systems comprising multiple components operating under constrained maintenance capacity present a challenging operational scenario where the simultaneous operation of multiple identical units is constrained by finite maintenance resources. In this intricate environment, the availability of maintenance crews, repair capacities, and spare parts becomes a critical factor influencing the overall system performance. The efficient allocation of these limited resources becomes paramount to ensure uninterrupted production, minimize downtime, and enhance the reliability of the manufacturing process. Addressing the complexities of such systems requires advanced analytical frameworks, such as queueing theory, to model and optimize maintenance strategies in the face of resource constraints. This intersection of multi-unit manufacturing and limited maintenance resources introduces a nuanced perspective, emphasizing the need for adaptive and effective maintenance planning to sustain optimal productivity and operational resilience. The study by Iyoo et al., (2006) on determining optimal maintenance capacity through the application of selective repair strategies provides insights into the strategic allocation of resources, crucial in scenarios where limitations exist. Additionally, the work by de Castro and Cavalca (2006) on optimizing the allocation of maintenance assets to enhance manufacturing system performance contributes to the understanding of resource management strategies.

The queueing theory approach in multi-unit manufacturing systems draws inspiration and theoretical underpinnings from seminal works in the field, particularly addressing the nuanced challenges associated with limited maintenance resources. Beyond production and manufacturing contexts, queueing theory has also been applied to analyze systems that are not essentially production/manufacturing systems (Bekker et al., 2024; Comert et al., 2024; Jayamani et al.,

2024). For instance, Ambika et al. (2024) employed an M/M/1 queue model to analyze a production system under the effects of maintenance actions. Similarly, Fallahnezhad and Pourgharibshahi (2017) addressed a three-state machine replacement system and optimal maintenance planning. In a broader scope, the survey by Shanthikumar et al. (2007) on queueing theory for semiconductor manufacturing systems provides foundational insights, opening avenues for understanding system dynamics and presenting open problems for future research.

Extending these ideas further, Ghalekhondabi and Suer (2018) utilized queueing-theoretic models to analyze the performance of production lines within a make-to-stock/make-to-order manufacturing framework. In a complementary direction, the hybrid dependability queueing model proposed by Chiacchio et al. (2022) adds a sophisticated dimension to the practical application of queueing theory, particularly in the context of repairable production lines. Moreover, Lee et al. (2021) and Federgruen and So (1990) explored stochastic queueing frameworks to refine preventive maintenance scheduling strategies and to optimize workforce sizing, respectively. Collectively, these studies highlight the adaptability and versatility of queueing theory approaches across diverse manufacturing and operational settings.

More recently, Luo et al., (2025) proposed a queueing theory model based on M/M/1 queueing for CBM planning of service systems. Using queueing theory systems, maintenance planning of excavators in construction sites are investigated in Antala and Sanga (2025). Game theory along with queueing theory is employed in Jackson et al. (2025) for maintenance planning and contract design. In another study, maintenance planning for inventory systems is investigated using queueing theory models (Jeganathan et al., 2025).

A significant body of research in queueing theory has adopted the Poisson arrival process as a foundational assumption for modeling system dynamics and optimizing maintenance decisions. For instance, Federgruen and So (1990) investigated the behavior of single-server queues with Poisson-distributed arrivals, breakdowns, and repair operations. They formulated an optimal corrective maintenance policy, characterized by a single threshold parameter, aimed at minimizing long-run average operating costs, with an extension that allowed for repair postponement. Extending this line of inquiry, Koyanagi and Kawai (2003) explored an optimal age maintenance strategy for a queueing system with Poisson arrivals and a general service distribution. They applied renewal theory to determine the optimal preventive maintenance age designed to minimize the expected rate of customer loss over an infinite time horizon. Taking a further step toward real-time intervention, Ejaz et al. (2019) investigated the incorporation of condition monitoring within a queueing framework, with a particular emphasis on server-side deterioration. Their study utilized a Poisson arrival process to model a single-server system undergoing Markovian degradation, providing a foundation for real-time maintenance intervention with general distributions assumed for service and repair times. The authors proposed an optimization approach, the model evaluates total operational expenditure by accounting for repair expenses, catastrophic failure penalties, quality-related losses, and inventory holding costs.

This rich body of literature collectively contributes to the comprehensive understanding and application of queueing theory in addressing the challenges posed by limited maintenance resources in multi-unit manufacturing systems. The birth-death process plays a crucial role in determining optimal maintenance capacity through the application of selective repair strategies by capturing the dynamic transitions between states in the queueing model. Additionally, the M/M/m/C queueing models in maintenance resources optimization provide a structured framework to analyze system behavior, considering factors such as server capacities and arrival rates.

Building on the queueing-theoretic foundations discussed above, this paper analyzes and evaluates maintenance strategies within multi-unit, identical parallel manufacturing systems using the powerful tool of queueing theory. The analysis centers on a queueing framework designed for the parallel operation of identical units subject to stochastic failures. By utilizing the birth-death process and the M/M/m/C queueing model, this study evaluates a maintenance system characterized by a finite repair capacity and a closed population of components. The paper introduces the fundamental equations governing the probability distribution of units in different states, laying the groundwork for subsequent analysis. To address the practical applicability of the proposed model, numerical examples are presented. The examples showcase the model's versatility in handling different system configurations, unit scales, and maintenance crew capacities. The derived performance metrics, specifically the average number of failed units, mean

maintenance lead times, and aggregate system failure rates provide a quantitative basis for stakeholders to refine maintenance scheduling and resource allocation.

As the literature review reveals, most studies about the application of queueing theory to analyze multi-unit production systems assume that components have a binary state space (operational or failed); this study adopts a three-state framework for each unit: healthy, unhealthy, and failed. The degradation process is modeled as a sequential transition, from healthy to unhealthy, and subsequently to failure, governed by independent exponential distributions. This multi-state approach allows for the implementation of a bifurcated maintenance strategy: Preventive Maintenance (PM) for units in the degraded (unhealthy) state and Corrective Maintenance (CM) for those in the failed state. There are limitations for the PM maintenance crew and for the CM maintenance crew. More specifically, a limited number of units in the failed state can concurrently be maintained using CM action. Also, a limited number of units in the unhealthy state can be concurrently maintained using PM action.

The remainder of this study is structured as follows: Section 2 introduces a motivating example and describes in detail the novelty aspects of the paper. Section 3 presents the problem statement; roadmap of the models and nomenclatures used in developing the models. In Section 4, Birth-death processes, M/M/1 and M/M/m/C Queueing models are described. We refer to the model and TRD discussed in this section as a two-state model because it is assumed that a unit of the system has two states: a healthy state and a failure state. In Section 5, a three-state stochastic system comprising two distinct operational modes and a singular failure state is discussed. Section 6, numerical examples, a real case, analyses and validation of the models are presented. Finally, Section 7 concludes the paper.

2. Motivating Example and novelty aspects of the proposed model

In this section, first a motivating example is provided regarding wind turbine farms and then novelty aspects of the model are explained.

2.1 Motivating example

To demonstrate the practical utility of the developed framework, a wind farm installation is utilized as a case study. Given the central role of wind power in the global renewable energy transition, optimizing maintenance protocols for these fleets is essential to ensure high operational availability and mitigate the significant costs associated with turbine downtime.

In the context of our model, each wind turbine in the farm represents an identical unit in a multi-unit parallel manufacturing system. The turbines operate independently, generating electricity, and component failures within the turbine fleet are assumed to occur stochastically, governed by a Poisson arrival process. The failures could be caused by various factors such as mechanical issues, electrical malfunctions, or environmental wear and tear. To address these failures, a maintenance crew is available to perform repairs, with a limited capacity to handle a specific number of turbines simultaneously.

The birth-death process and M/M/m/C queueing model align with the dynamics of the wind turbine farm. The transition from operational to failure states corresponds to a turbine experiencing a malfunction, while maintenance actions represent the repair processes conducted by the crew.

The proposed model enables a detailed examination of system behavior, capturing metrics such as the overall failure rate within a wind farm, maintenance waiting times and the average number of turbines across various states. We attempt to illustrate the model's flexibility in addressing practical scenarios by applying this framework to the motivating example. Transition rate diagrams derived from the model depict turbine movements between operational, maintenance conditions and failed states, offering valuable insights into the dynamic behavior of the system. In addition, numerical case studies highlight the model's adaptability to diverse configurations, the capacity of the maintenance workforce, and the scale of production systems. This underscores its significance for decision-makers seeking to optimize maintenance strategies in parallel manufacturing systems, particularly within the renewable energy domain.

2.2 Novelty aspects of the proposed model

As the literature review reveals, most studies about the application of queueing theory to analyze multi-unit production systems assume that components have a binary state space (operational or failed); this study adopts a three-state framework for each unit: healthy, unhealthy, and failed. The degradation process is modeled as a sequential transition, from healthy to unhealthy, and subsequently to failure, governed by independent exponential distributions. This multi-state approach allows for the implementation of a bifurcated maintenance strategy: Preventive Maintenance (PM) for units in the degraded (unhealthy) state and Corrective Maintenance (CM) for those in the failed state. There are limitations for the PM maintenance crew and for the CM maintenance crew. More specifically, a limited number of units in the failed state can concurrently be maintained using CM action. Also, a limited number of units in the unhealthy state can be concurrently maintained using PM action. Thus, the primary contributions of this study are outlined as follows:

I. Novel three-state degradation model for system components

Unlike conventional approaches that assume a binary (operational/failed) state space for system components, we propose a three-state framework comprising healthy, unhealthy/degraded, and failed states. The degradation process is modeled through sequential transitions from healthy to unhealthy to failed states, with the possibility of direct healthy-to-failed transitions, providing a flexible and realistic representation of failure mechanisms observed in practice. This enables more realistic modeling of system behavior and more effective maintenance planning

II. Integrated preventive and corrective maintenance under resource constraints

Two types of maintenance actions can be carried out: **PM** for the units in the degraded state, and **CM** for the units in the failed state. Both maintenance actions are subject to limited resource availability, with m_1 and m_2 representing the maximum number of units that can undergo PM and CM concurrently. This dual-maintenance approach with explicit resource constraints addresses a critical gap in the literature.

III. Bivariate birth/birth-death stochastic modeling framework

We introduce a bivariate stochastic process (n_2, n_1) to represent the system state, where n_2 and n_1 denote the number of units in the failed and degraded states, respectively. This bivariate formulation, combined with the M/M/m//C queueing framework, provides a mathematically rigorous and computationally efficient approach to:

- Capture the coupled dynamics of degradation and maintenance
- Derive the system's transition rate diagram and steady-state probabilities
- Compute key performance metrics including average failed units, maintenance lead times, and crew utilization

To the best of our knowledge, this is one of the few studies employing bivariate birth/birth-death processes along with queueing theory for the evaluation of maintenance strategies in multi-unit parallel manufacturing systems.

IV. Practical applicability to real manufacturing systems

The proposed model is applicable to diverse system configurations, unit scales, and maintenance crew capacities. Also, we provide quantitative performance metrics (e.g., average failed units, maintenance waiting times, system failure rates) that enable decision-makers to evaluate maintenance scheduling, resource allocation, and staffing decisions. Applications of the models are shown through motivating example and numerical analyses that reflect realistic industrial scenarios.

Also, with respect to work of Azizi and Salari (2023), the main novelties and distinguishing features are as follows:

- While in Azizi and Salari (2023) paper, constraints regarding maintenance resources are not addressed, in the current study, two types of maintenance actions including PM and CM can be conducted and resources limitations for these maintenance actions are also taken into consideration.
- While Azizi and Salari (2023) develop their model only by focusing on Birth/Birth-Death stochastic modeling framework, in the current study the main framework is queueing theory concepts and equations along with Birth/Birth-Death stochastic modeling framework.

Our framework is designed to address the inherent complexities of multi-unit parallel manufacturing systems, providing a comprehensive approach to maintenance planning. This study contributes to the literature on maintenance scheduling and queueing theory by presenting a structured methodology for tackling the challenges associated with such systems. The presented framework equips decision-makers with practical tools to support informed maintenance strategies, thereby enhancing operational efficiency and promoting long-term sustainability in manufacturing environments.

3. Problem statement, roadmap of the models and nomenclatures

A multi-unit system consisting of C identical units are considered while the units are operating in parallel. First, in Section 4, it is assumed that the system's units have two states: an operational state and a failure state. The units independently fail according to a Poisson process with rate λ and each failed unit enters a maintenance queue. For this structure, only one type of maintenance, i.e., corrective maintenance (CM) is conducted and it is assumed that concurrently only m units can be repaired. The maintenance crew consists of m members. The required time to conduct a CM action follows an exponential distribution with rate μ . The basic queueing structure for such type system is denoted as M/M/m//C.

In Section 5, to extend the proposed binary-state model of Section 4, it is assumed that the units of the system have three states including two operational states and a failure state. These states are designated as healthy, unhealthy/degraded and failure states. A unit can directly degrade from state 0 to the failed state or degrades as follows: $0 \rightarrow 1 \rightarrow 2$. In this section a dual maintenance strategy is proposed. More specifically, for the units operating in the state 1, a PM action can be conducted while for the units in state 2, CM should be conducted. Both types of maintenance actions are perfect and bring the unit back to state 0. In order to represent the system's state a two-dimensional vector is employed as follows (n_2, n_1) so that n_2 and n_1 represent the number of units in the failure and degraded states respectively. The TRD for such a structure consists of two main parts: transition due to the effects of degradation and transition due to the effects of CM/PM actions. Combining these two effects results in a TRD for the system.

The system state evolves as a continuous-time Markov chain (CTMC) with finite state space $\mathcal{S} = \{(n_2, n_1) \mid n_2, n_1 \geq 0, n_2 + n_1 \leq C\}$. All transition times are exponentially distributed, and the process is time-homogeneous, implying that transition rates depend only on the current state and not on absolute time. The exponential distribution is the cornerstone of CTMC and birth-death processes. Its memoryless property enables us to derive closed-form equations for the system's state probabilities, leading to the steady-state solutions presented in the paper. In many real-world systems, particularly when considering the behavior of complex equipment, the exponential assumption is a widely accepted and practical. Its simplicity is key for providing a clear and intuitive understanding of the system's overall dynamics, enabling analytical performance evaluation and sensitivity analysis without the need for complex numerical methods.

In the following, Table 1 is provided to better represent the comparison of the models of Section 4 and 5. Also, Table 2 represents the notations used throughout the paper.

Table 1. Comparison of two-state model with three-state model

	Section 4 (two-state model)	Section 5 (three-state model)
Number of unit states	A unit has two states: operational or failure	A unit has three states: two operational states and a failure state
Maintenance action	Only one type of maintenance actions, i.e., CM, is conducted on the failed units	Two types of maintenance actions are conducted: PM for the units in state 1 and CM for the units in state 2.
Limitation of the maintenance resources	m members for conducting CM action	m_1 and m_2 members for conducting PM and CM, respectively
System state	A single variable state, i.e., number of units in the failed or healthy state	Bivariate variable state (n_2, n_1)
Failure mechanism	A unit transits from operational state to the failure state according to exponential distribution	A unit may degrade from state 0 to 1 and then to state 2, or may directly transit from state 0 to 2.

Table 2. Notations/Nomenclature

Symbol	Description
System Parameters	
C	Total number of identical units in the system
n_0	Number of healthy units
n_1	Number of degraded/unhealthy units
n_2	Number of failed units
m_1	Number of PM crew members
m_2	Number of CM crew members
C_n	Coefficient used in the computation of steady-state probabilities (Equation 3)
n	Number of failed units (used in two-state model as a state variable)
Transition Rates	
$\lambda_{0,1}$	Rate of transition from healthy (State 0) to degraded (State 1)
$\lambda_{0,2}$	Rate of transition from healthy (State 0) to failed (State 2)
$\lambda_{1,2}$	Rate of transition from degraded (State 1) to failed (State 2)

Symbol	Description
System Parameters	
λ	Failure rate of a unit (used in two-state model)
λ_n	Aggregate failure rate when n units are failed (two-state model)
Transition Rates (Maintenance/Repair)	
μ_1	Service rate for PM action
μ_2	Service rate for CM action
μ	Maintenance/service rate of a failed unit (used in two-state model)
μ_n	Aggregate repair rate when n units are failed (two-state model)
Steady-State Probabilities	
π_n	Steady-state probability of having n failed units (two-state model)
π_{n_2, n_1}	Steady-state probability of having n_2 failed units and n_1 degraded units (three-state model)
Performance Metrics	
L	Long-run average number of failed units (including those in service)
L_q	Long-run average number of failed units waiting in the queue for maintenance
W	Average time a failed unit spends in the maintenance system (total downtime)
W_q	Average time a failed unit waits in the queue before maintenance begins
$\bar{\lambda}$	Long-run average system failure rate

4. Birth-death processes, M/M/1 and M/M/m/C Queueing models

In this section, as a preliminary for Section 4, the birth-death process and M/M/1 queue are firstly introduced. In the following, a queueing system with m servers and a limited population size, C , is discussed. We refer to the model and

TRD discussed in this section as a two-state model because it is assumed that a unit of the system has two states: a healthy state and a failure state.

4.1 Birth-death processes and M/M/1 queueing model

Within the framework of a birth-death process, the system experiences arrivals following a Poisson distribution with a constant rate λ . The facility operates with a single server where service durations are characterized by an exponential distribution at a rate of μ . In queueing theory nomenclature, this configuration is standardly identified as an M/M/1 model.

Upon entry, a customer is processed immediately if the server is available; if the server is occupied, the customer joins a queue to await the next available service window. A fundamental assumption of this model is the statistical independence between the inter-arrival times and the service durations.

Accordingly, the TRD of the process is shown in Figure 1. By setting the rate of departure equal to the rate of arrival for a given state, the following equilibrium equations are derived:

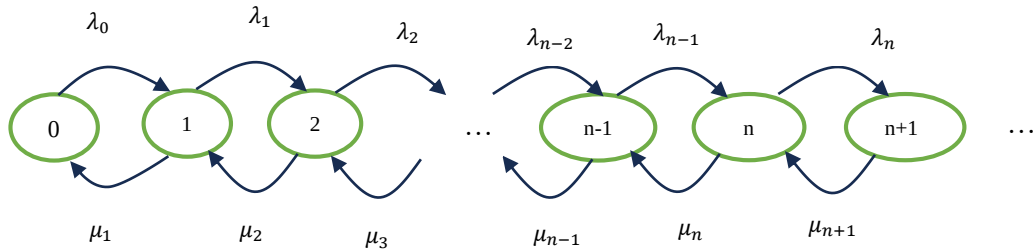


Figure 1. Transition rate diagram (TRD) of a birth/death process

$$\lambda_0 \pi_0 = \mu_1 \pi_1 \quad (1)$$

$$(\lambda_n + \mu_n) \pi_n = \lambda_{n-1} \pi_{n-1} + \mu_{n+1} \pi_{n+1}; \quad n \geq 1 \quad (2)$$

Following a series of algebraic transformations, the resultant equations are established as follows:

$$\pi_n = C_n \pi_0; \quad n \geq 1 \quad (3)$$

While:

$$C_n = \frac{\lambda_0 \lambda_1 \dots \lambda_{n-1}}{\mu_1 \mu_2 \dots \mu_n}; \quad n \geq 1 \quad (4)$$

As $\pi_0 + \sum_{n=1}^{\infty} C_n \pi_0 = 1$, the following equation holds for π_0 :

$$\pi_0 = \frac{1}{1 + \sum_{n=1}^{\infty} C_n} \quad (5)$$

4.2 M/M/m/C Queueing model

Assume a manufacturing setup featuring C uniform machines arranged in parallel. These units function independently, with no underlying physical or stochastic correlations. The transition of each unit from a functional to a failed state is governed by a Poisson process with an intensity of λ . Furthermore, maintenance durations for malfunctioning units are characterized by independent and identically distributed (i.i.d.) exponential variables with a service rate of μ . The maintenance facility is restricted to a maximum capacity of m concurrent repairs, where $m < C$. It can be inferred that the maintenance crew has m members. The TRD of this system is presented in Figure2.

Let n represent the number of units currently in a failed state, implying that $C - n$ components remain functional. At this juncture, the aggregate failure rate, which transitions the system toward an increased count of failed units, is defined as follows:

$$\lambda_n = (C - n)\lambda \quad (6)$$

The system's transition rate from a failed state back to an operational status, facilitated by the maintenance crew, is expressed by the following:

$$\mu_n = \begin{cases} n\mu, & n \leq m \\ m\mu & m < n \leq C \\ 0 & C < n \end{cases} \quad (7)$$

Through subsequent algebraic simplification, the following expressions are derived:

$$C_n = \begin{cases} \frac{C!}{(C-n)!n!} \left(\frac{\lambda}{\mu}\right)^n; & n \leq m \\ \frac{C!}{(C-n)!m!m^{n-m}} \left(\frac{\lambda}{\mu}\right)^n; & m \leq n \leq C \\ 0; & C < n \end{cases} \quad (8)$$

The expected number of failed components, including those currently being serviced and those in the queue awaiting maintenance is represented by L and is determined using the following expression:

$$L = \sum_{n=0}^C n\pi_n \quad (9)$$

While π_0 is computed using Equation (5) and π_i , for $i \geq 1$ is computed using Equation (3). Furthermore, the expected number of malfunctioning units specifically awaiting the commencement of maintenance is determined by the following expression:

$$L_q = \sum_{n=m+1}^C (n-m) \cdot \pi_n \quad (10)$$

Average rate of the system failure is determined using the following calculation:

$$\bar{\lambda} = \sum_{n=0}^C \lambda_n \pi_n = \sum_{n=0}^C (C-n)\lambda \pi_n = \lambda(C-L) \quad (11)$$

Consequently, by applying Little's Law, we can determine both the average time of a failed unit within the maintenance system (W) and the average queue duration prior to the commencement of maintenance (W_q). These metrics, representing the total downtime and the pure waiting time respectively, are calculated as follows:

$$W = \frac{L}{\bar{\lambda}}; \quad W_q = \frac{L_q}{\bar{\lambda}} \quad (12)$$

Finally, it should be noted that if the time-to-failure for each component is independent with a mean of $\frac{1}{\lambda}$, the derivations in this section remain valid regardless of the specific failure distribution. The requirement for exponentiality applies to the maintenance durations.

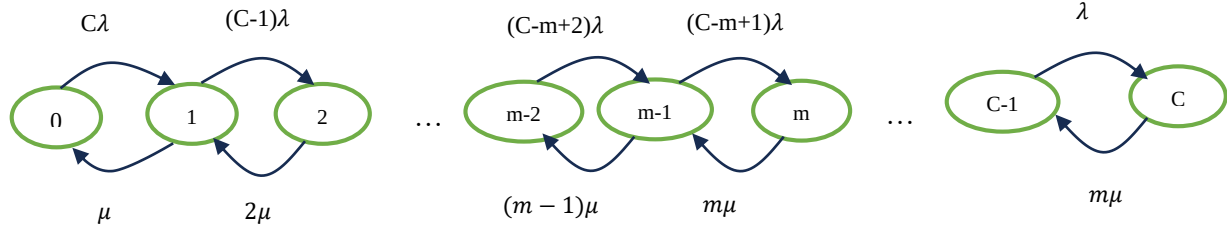


Figure 2. TRD of an M/M/m//C Queueing model

5. Three-state stochastic system comprising two distinct operational modes and a singular failure state

This section is organized into three subsections. The first provides a detailed discussion of the multi-unit production system across its various operational states, followed by the presentation of the TRD in the second subsection. To clarify the TRD, an illustrative example is also presented for a system with 3 units. In the third subsection, the equilibrium equations and details of TRD for the general case are discussed.

5.1 Multi-unit production system while the units have different operational states

This section considers a three-state operational model for each system component: operational (State 0), degraded (State 1), and failed (State 2). Transitions are governed by independent exponential distributions, where units can shift from State 0 to either 1 or 2 with intensities $\lambda_{0,1}$ and $\lambda_{0,2}$, respectively, or from State 1 to 2 at rate $\lambda_{1,2}$. The maintenance strategy comprises two distinct policies: Preventive Maintenance (PM) for degraded units and Corrective Maintenance (CM) for those in the failed state. The durations for PM and CM are exponentially distributed with rates μ_1 and μ_2 . Furthermore, the maintenance facility allocates m_1 personnel for preventive tasks and m_2 for corrective actions. It means that only m_1 unhealthy units can be simultaneously repaired using PM action, and for the failed units, only m_2 failed units can be concurrently maintained using CM action.

The system's state is represented by a two-dimensional state vector (n_2, n_1) , where n_1 and n_2 denote the number of units in the degraded and failed states, respectively. Consequently, the number of healthy units is expressed as $n_0 = C - n_1 - n_2$. Based on this state-space representation, a multi-dimensional birth-death process can be utilized for system analysis. The cardinality of the state space corresponds to the total number of non-negative integer solutions of the following constraint: $n_0 + n_1 + n_2 = C$, which is $\frac{(C+2)(C+1)}{2}$. Accordingly, the total departure rate due to the degradation is as follows:

$$n_0\lambda_{0,1} + n_0\lambda_{0,2} + n_1\lambda_{1,2} \quad (13)$$

And the total departure rate due to the maintenance actions:

$$\min(n_1, m_1)\mu_1 + \min(n_2, m_2)\mu_2 \quad (14)$$

The TRD is shown in Figures 3 and 4 so that Figure 3 illustrates exclusively the transitions resulting from the stochastic degradation of the system units, while Figure 4 shows the transition diagram due to the maintenance actions. The Combination of Figures 3 and 4 illustrates the transition diagram of the system. To better explain the state transition diagram, an example is presented in the following.

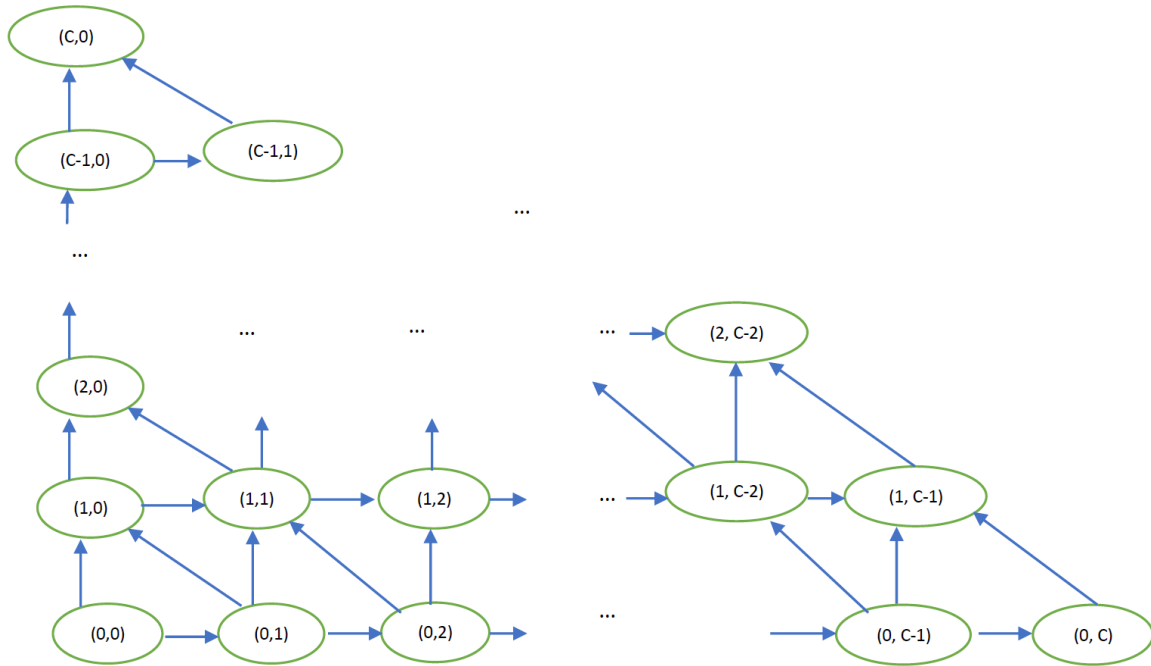


Figure 3. Degradation-only TRD (TRD under the deterioration and without the effect of maintenance actions)

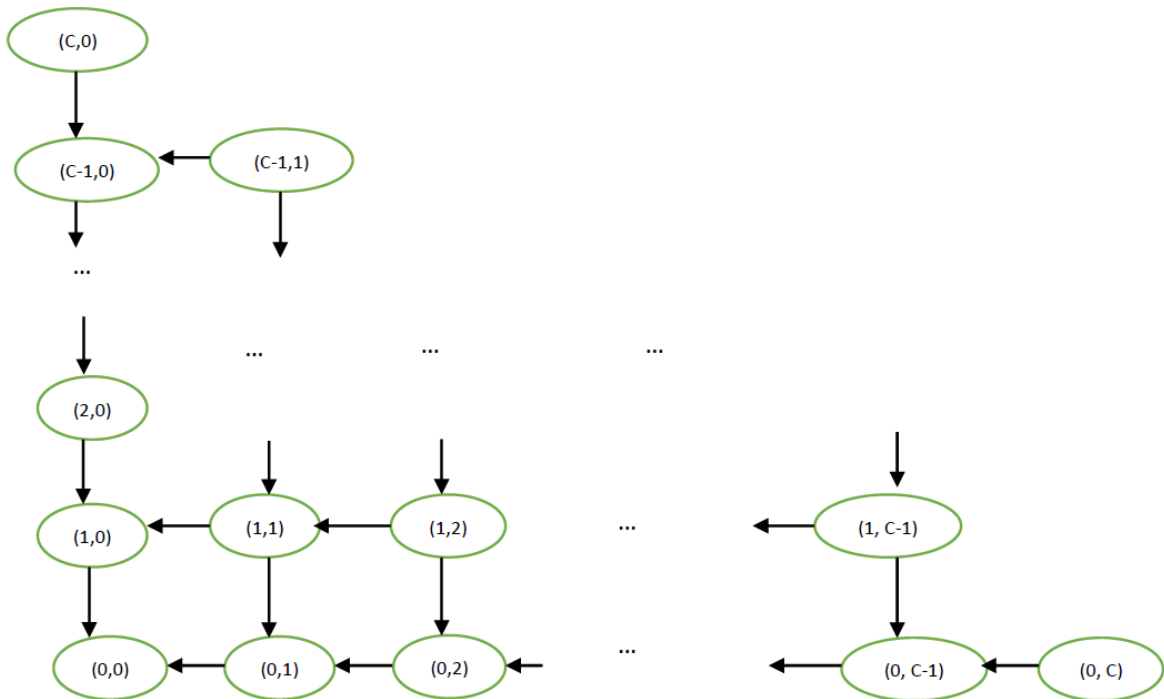


Figure 4. Maintenance-only TRD, i.e., TRD under the effects of maintenance actions and without the effects of deterioration

Example. Consider a system that has 3 units. The maintenance crew consists of 2 members so that 1 member of PM and 1 for CM actions. It means that simultaneously only one failed unit can be repaired using CM action, or similarly, only one unhealthy unit can be simultaneously repaired using PM action. Similar to Figure 3, Figure 5 illustrates the transition diagram due to the system degradation and without the effects of PM/CM actions. While Figure 6 corresponds to Figure 4 only shows the effect of PM/CM actions. Combining Figures 5 and 6 shows the system transition diagram as depicted in Figure 7.

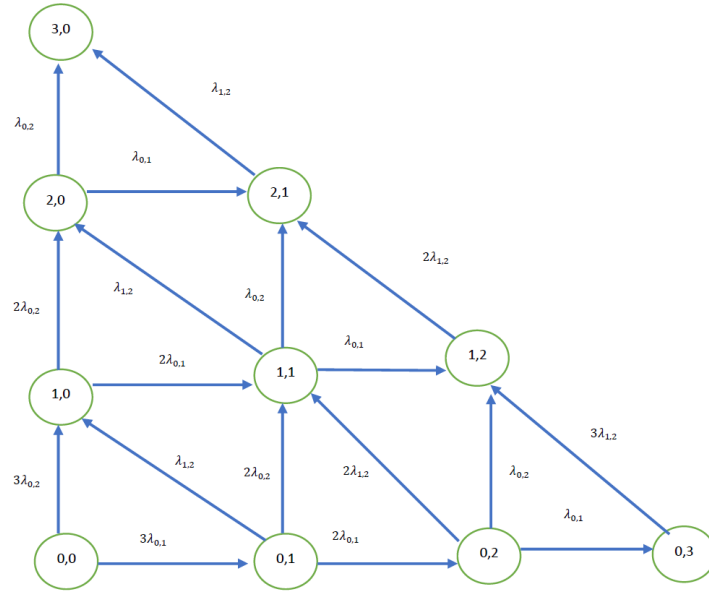


Figure 5. TRD under the deterioration and without the effect of maintenance actions for $C = 3, m_1 = 1, m_2 = 1$

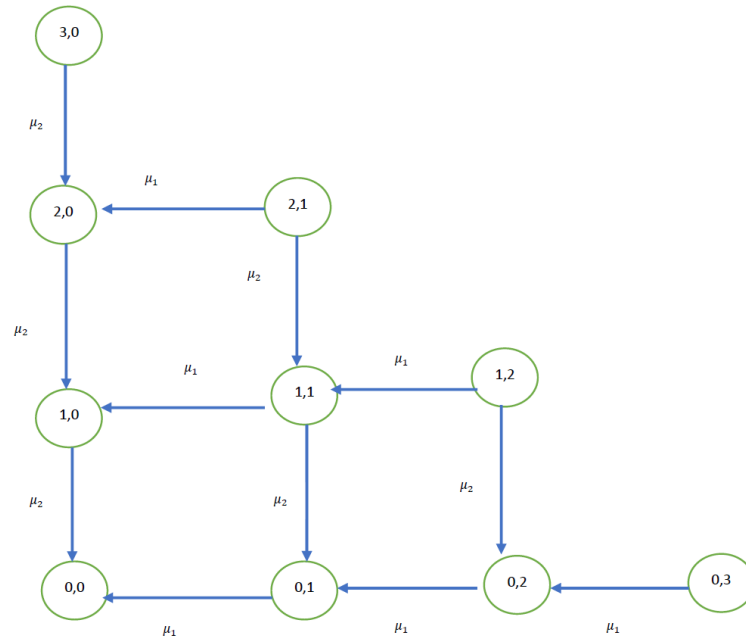


Figure 6. TRD under the effect of maintenance actions, $C = 3, m_1 = 1, m_2 = 1$

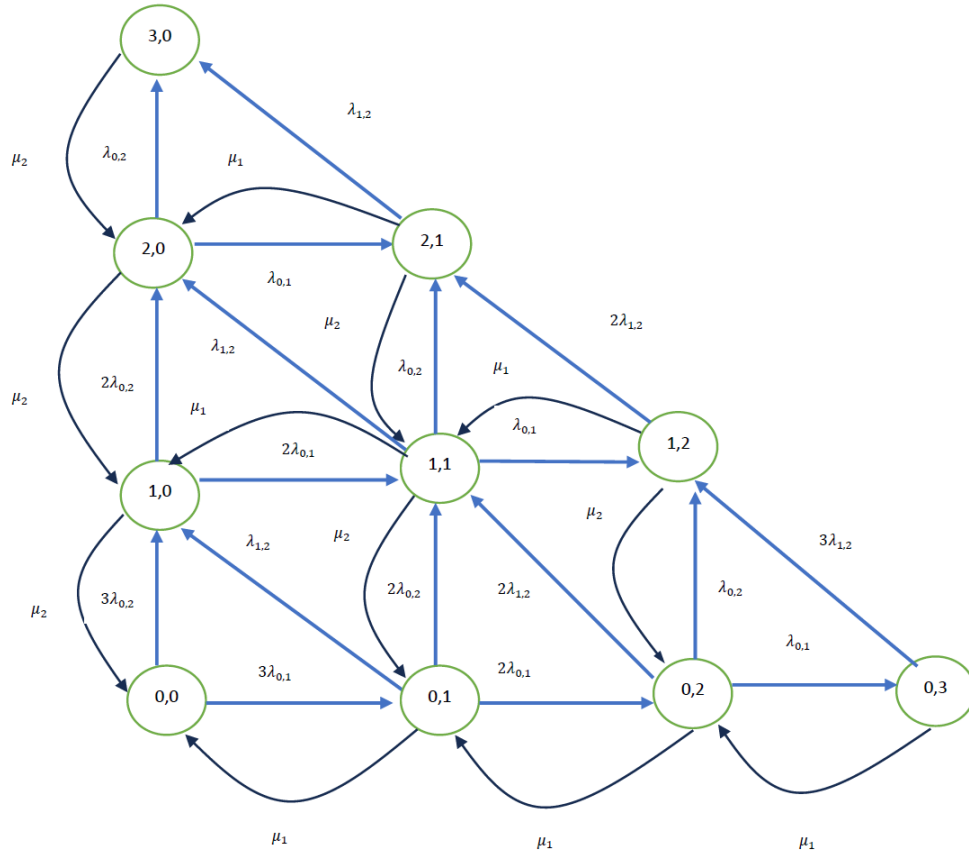


Figure 7. TRD for a system with $C = 3, m_1 = 1, m_2 = 1$ (it is a combination of Figures 5 and 6)

In the following subsection, detailed descriptions about construction of TRD are provided.

4.2 Construction of TRD

We will define the degradation-only TRD as a directed graph $G_d = (V, E_d)$, where: $V = \{(n_2, n_1) \mid n_1, n_2 \geq 0, n_1 + n_2 \leq C\}$ is the set of all feasible system states and $E_d \subset V \times V$ is the set of directed edges representing transitions due to the unit degradation, with the following transition rates:

- From state (n_2, n_1) to $(n_2, n_1 + 1)$ at rate $n_0 \lambda_{0,1}$ (healthy $\rightarrow$ unhealthy).
- From state (n_2, n_1) to $(n_2 + 1, n_1)$ at rate $n_0 \lambda_{0,2}$ (healthy $\rightarrow$ failed).
- From state (n_2, n_1) to $(n_2 + 1, n_1 - 1)$ at rate $n_1 \lambda_{1,2}$ (degraded $\rightarrow$ failed).

This formalizes the transitions illustrated in Figures 3 and 5.

For the definition of the maintenance TRD, we will define the maintenance-only TRD as a directed graph $G_m = (V, E_m)$, where V is the same state space as above and $E_m \subset V \times V$ represents transitions due to PM and CM actions, with rates:

- From (n_2, n_1) to $(n_2, n_1 - 1)$ at rate $\min(n_1, m_1) \mu_1$ (PM action, only if $n_1 > 0$).
- From (n_2, n_1) to $(n_2 - 1, n_1)$ at rate $\min(n_2, m_2) \mu_2$ (CM action, only if $n_2 > 0$).

These transitions are illustrated in Figures 4 and 6.

Overall TRD is obtained by the union of degradation-only TRD and Maintenance-only TRD which means: $G = G_d \cup G_m$, where the edge set $E = E_d \cup E_m$. Transition rates are additive when multiple edges connect the same pair of states in the same direction. Since the degradation and maintenance processes are independent and occur concurrently, the total rate of leaving a state is the sum of the rates of all outgoing edges from G_d and G_m . Similarly, the total rate of entering a state is the sum of incoming rates from both subgraphs.

This superposition principle is standard in Markov chain modeling for systems with independent concurrent processes (e.g., failures and repairs). It ensures that the combined TRD correctly represents the stochastic evolution of the system. According to the aforementioned discussions, it is noteworthy that Figures 3 and 5 show degradation-only TRDs for the general case and a specific example, respectively. Figures 4 and 6 show maintenance-only TRDs for the general case and the same example and Figures 7 illustrate the superposition of these two processes, i.e., the full system dynamics.

4.2. Equilibrium Formulation Using the Generator/infinitesimal Matrix

Let π be the row vector of steady-state probabilities, where π_{n_2, n_1} denotes the probability of being in state (n_2, n_1) . The generator matrix Q is a $|S| \times |S|$ matrix whose off-diagonal elements $q_{i,j}$ represent the transition rate from state i to state j , and whose diagonal elements are defined as $q_{i,i} = -\sum_{j \neq i} q_{i,j}$, ensuring that each row sums to zero. The steady-state probabilities are then obtained by solving the system of linear equations:

$$\pi Q = 0 \quad (15)$$

subject to the normalization condition:

$$\sum_{(n_2, n_1) \in S} \pi_{n_2, n_1} = 1 \quad (16)$$

The off-diagonal elements of Q are derived from the superposition of degradation and maintenance transitions as discussed in the previous subsection. The diagonal elements are then given by:

$$q_{(n_2, n_1), (n_2, n_1)} = -[(C - n_1 - n_2)(\lambda_{0,1} + \lambda_{0,2}) + n_1 \lambda_{1,2} + \min(n_1, m_1) \mu_1 + \min(n_2, m_2) \mu_2] \quad (17)$$

In the appendix of the paper, the detailed state-by-state equations are presented in Groups I–VII, which is equivalent to this compact formulation. It is obvious that this compact formulation offers significant advantages in terms of clarity, mathematical rigor, and ease of implementation in computational software. As an example, the equilibrium equation for a generic interior state (n_2, n_1) with $n_1, n_2 \geq 1$ and $n_1 + n_2 < C$ (Group I in the Appendix) can be derived from $\pi Q = 0$ as:

$$\begin{aligned} \pi_{n_2, n_1} & (n_0 \lambda_{0,1} + n_0 \lambda_{0,2} + n_1 \lambda_{1,2} + \min(n_1, m_1) \cdot \mu_1 + \min(n_2, m_2) \mu_2) \\ & = \min(n_2 + 1, m_2) \cdot \mu_2 \cdot \pi_{n_2+1, n_1} + \min(n_1 + 1, m_1) \cdot \mu_1 \cdot \pi_{n_2, n_1+1} \\ & + (n_1 + 1) \lambda_{1,2} \cdot \pi_{n_2-1, n_1+1} + n_0 \lambda_{0,2} \cdot \pi_{n_2-1, n_1} + n_0 \lambda_{0,1} \pi_{n_2, n_1-1} \end{aligned} \quad (18)$$

The system $\pi Q = 0$ with $\sum \pi_i = 1$ can be solved using standard numerical methods including direct methods, iterative methods or matrix-analytic methods. For moderate state spaces (e.g., $C \leq 200$), Gaussian elimination or LU decomposition with sparse matrix representations can be used. This approach yields an exact solution and is efficient for systems with up to approximately 20,000 states. For larger systems (e.g., $C > 200$), iterative methods such as the Gauss-Seidel method, Successive Over-Relaxation (SOR), or the Power Method can be employed. These methods exploit the sparsity of Q and require only $O(|S|)$ memory. Given the quasi-birth-death (QBD) structure of the state space, specialized algorithms such as the Matrix-Geometric method can be developed to further enhance computational efficiency for very large systems. For larger systems (e.g., $C > 500$), the solution of the linear system may become

computationally demanding. In such cases, we recommend simulation-based approaches. For extremely large systems, discrete-event simulation or Markov Chain Monte Carlo (MCMC) can be used to estimate key performance metrics without requiring the full steady-state distribution.

The state space of the proposed model is defined by the bivariate vector (n_2, n_1) , where $n_1 + n_2 \leq C$. The total number of feasible states is $|S| = \frac{(C+1)(C+2)}{2}$, which grows quadratically with the number of units C . This is a significant improvement over a full three-dimensional representation, which would have $O(C^3)$ states. Hence, the choice of solution method depends on the system size, accuracy requirements, and available computational resources. The analytical framework provided in this study remains valid regardless of the solution method employed.

5. Numerical Examples and Analyses

In this section, some numerical examples and analyses are provided about the applications of M/M/m/C for analyzing an identical, multi-unit production system under the limited maintenance resources. In the first two simple examples, corresponding to the discussion of Section 4, it is assumed that there is a binary state space for each unit, consisting of operational and failed states. In contrast, the third example incorporates a three-state model for each unit, comprising two distinct operational levels and one failure state, consistent with the framework established in Section 5.

Example 1 (small-scale system). To illustrate the model, consider a manufacturing system comprised of five identical components serviced by two maintenance crews. Following the two-state framework established in Section 4, we assume each unit's time-to-failure is exponentially distributed with a mean of 30 hours. Furthermore, the repair duration for a non-operational unit is also governed by an exponential distribution, with an average service time of 3 hours. In other words, $\lambda = \frac{1}{30}$ and $\mu = \frac{1}{3}$ machine per hour. According to Equations 2 and 3, the values of C_n and π_n are computed, and the results are provided in Table 3.

Table 3. Steady state probabilities and C_n for Example 1

	0	1	2	3	4	5
C_n	-	0.5	0.1	0.015	0.0015	0.00075
π_n	0.62	0.31	0.062	0.0093	0.00093	0.000465

With $L = 0.47$, it follows that an average of 4.53 units remain operational at any given time. The average time spent in the failure state is $W = 3.11$ hours, with a corresponding $\pi_0 + 0.5\pi_1 = 0.78$ idleness probability for each member of the maintenance crew.

Example 2 (Large-scale system). We extend the analysis to a large-scale configuration consisting of 100 identical units and a maintenance team of 5 members. Consistent with the two-state model in Section 4, historical data indicates a mean time to failure (MTTF) of 168 hours and a mean time to repair (MTTR) of 10 hours, both following exponential distributions. Applying Equations 8 and 9, the steady-state results yield $L = 5.68$ and $L_q = 0.069$, indicating that an average of 5.68 units are non-functional at any given time. Furthermore, the average downtime (W) is 10.122 hours, while the expected delay before maintenance begins (W_q) is 0.122 hours.

Example 3 (Three-state system). In this example, a system with 3 units is assumed while there exist 1 member for PM and 1 for CM. The Transition diagram of the example is shown in Figure 7. Unlike the previous two examples, in this example, each individual system component is characterized by three states: healthy, unhealthy, and failure. By setting the arrival rate equal to the departure rate, the following equations are obtained:

Node (0,0)	$\pi_{0,0}(3\lambda_{0,2} + 3\lambda_{0,1}) = \mu_2\pi_{1,0} + \mu_1\pi_{0,1}$
Node (0,1)	$\pi_{0,1}(\lambda_{1,2} + 2\lambda_{0,2} + 2\lambda_{0,1} + \mu_1) = \mu_1\pi_{0,2} + \mu_2\pi_{1,1} + 3\lambda_{0,1}\pi_{0,0}$
Node (0,2)	$\pi_{0,2}(2\lambda_{1,2} + \lambda_{0,2} + \lambda_{0,1} + \mu_1) = \mu_1\pi_{0,3} + \mu_2\pi_{1,2} + 2\lambda_{0,1}\pi_{0,1}$
Node (0,3)	$\pi_{0,3}(3\lambda_{1,2} + \mu_1) = \lambda_{0,1}\pi_{0,2}$
Node (1,0)	$\pi_{1,0}(2\lambda_{0,2} + 2\lambda_{0,1} + \mu_2) = \mu_2\pi_{2,0} + \mu_1\pi_{1,1} + \lambda_{1,2}\pi_{0,1} + 3\lambda_{0,2}\pi_{0,0}$
Node (1,1)	$\pi_{1,1}(\lambda_{1,2} + \lambda_{0,2} + \lambda_{0,1} + \mu_2 + \mu_1) = \mu_1\pi_{1,2} + \mu_2\pi_{2,1} + 2\lambda_{0,1}\pi_{1,0} + 2\lambda_{0,2}\pi_{0,1} + 2\lambda_{1,2}\pi_{0,2}$
Node (1,2)	$\pi_{1,2}(2\lambda_{1,2} + \mu_2 + \mu_1) = 3\lambda_{1,2}\pi_{0,3} + \lambda_{0,2}\pi_{0,2} + \lambda_{0,1}\pi_{1,1}$
Node (2,0)	$\pi_{2,0}(\lambda_{0,2} + \lambda_{0,1} + \mu_2) = \mu_2\pi_{3,0} + \mu_1\pi_{2,1} + \lambda_{1,2}\pi_{1,1} + 2\lambda_{0,2}\pi_{1,0}$
Node (2,1)	$\pi_{2,1}(\lambda_{1,2} + \mu_1 + \mu_2) = \lambda_{0,1}\pi_{2,0} + 2\lambda_{1,2}\pi_{1,2} + \lambda_{0,2}\pi_{1,1}$
Node (3,0)	$\mu_2\pi_{3,0} = \lambda_{0,2}\pi_{2,0} + \lambda_{1,2}\pi_{2,1}$
$\pi_{0,0} + \pi_{0,1} + \pi_{1,0} + \pi_{1,1} + \pi_{2,0} + \pi_{0,2} + \pi_{2,1} + \pi_{1,2} + \pi_{3,0} + \pi_{0,3} = 1$	

This system of linear equations consists of 11 equations and 10 variables. Hence, one equation is redundant and can be deleted. Solving this system of linear equations yields the steady state probabilities. To solve the example, data of Table 4 are used:

Table 4. The transition rate and PM/CM rate of Example 3

$\lambda_{0,1}$	$\lambda_{0,2}$	$\lambda_{1,2}$	μ_1	μ_2
0.005	0.001	0.01	0.04	0.01

The findings are summarized in Table 5. As illustrated in Table 5, for example, in the long term the system occupies state (0,0) with a probability of 41%. Figure 8 shows the steady-state probabilities. The idleness probability of the PM crew is $\pi_{0,0} + \pi_{1,0} + \pi_{2,0} + \pi_{3,0} = 0.79$, while the idleness probability for the CM crew is $\pi_{0,0} + \pi_{0,1} + \pi_{0,2} + \pi_{0,3} = 0.548$.

Table 5. Steady state probabilities of the third Example

$\pi_{0,0}$	$\pi_{1,0}$	$\pi_{0,1}$	$\pi_{1,1}$	$\pi_{2,0}$	$\pi_{2,1}$	$\pi_{0,2}$	$\pi_{1,2}$	$\pi_{0,3}$	$\pi_{3,0}$
0.407	0.254	0.12	0.053	0.107	0.011	0.02	0.005	0.001	0.022

In the following, some sensitivity analyses are provided. To this end, let's change the data of the degradation rates as in the following, while keep the values of μ_1 and μ_2 unchanged: $\lambda_{0,1} = 0.05, \lambda_{0,2} = 0.01, \lambda_{1,2} = 0.01$. It means that while the mean time to perform CM and PM actions does not change, the speed of degradation increases significantly. The results are provided in Figure 9. As the figure shows, the system spends around 40% of time in state (3,0). The percentage of time the system spends in state (0,0) drastically decreases with respect to the base example.

The findings show that as the system is experiencing a higher rate of degradation (e.g., due to more demanding operating conditions or aging components), the current maintenance strategy is insufficient to keep the system in a healthy state. This would necessitate a strategic decision: either increase the speed of maintenance (e.g., by hiring faster technicians or stocking more spare parts) or implement a more aggressive PM schedule to prevent the accelerated transition from healthy to unhealthy states. The model provides a quantitative basis for evaluating the cost-effectiveness of these decisions.

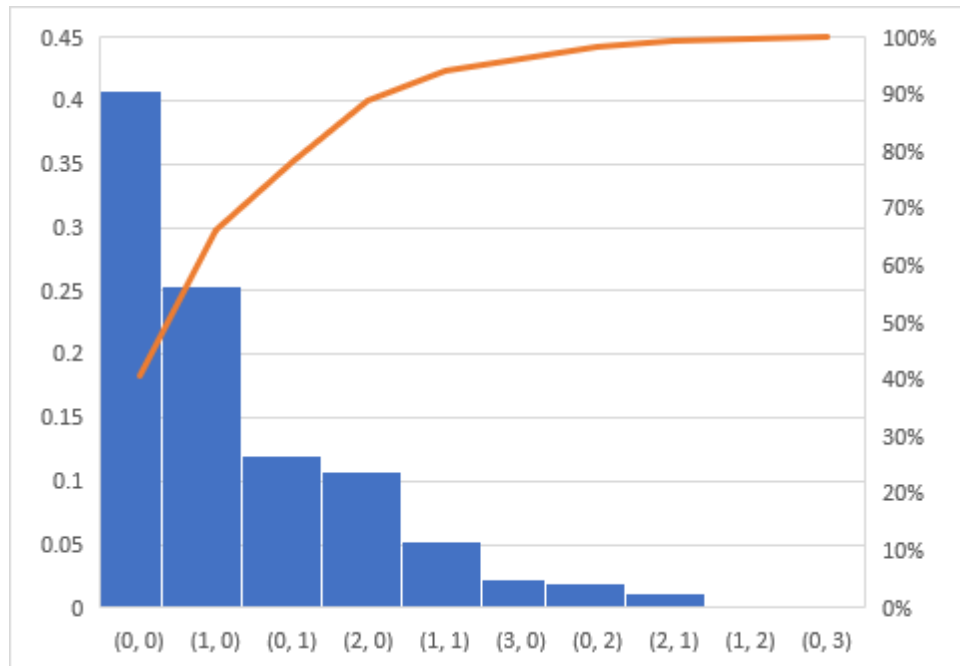


Figure 8. Steady-state probabilities for Example 3

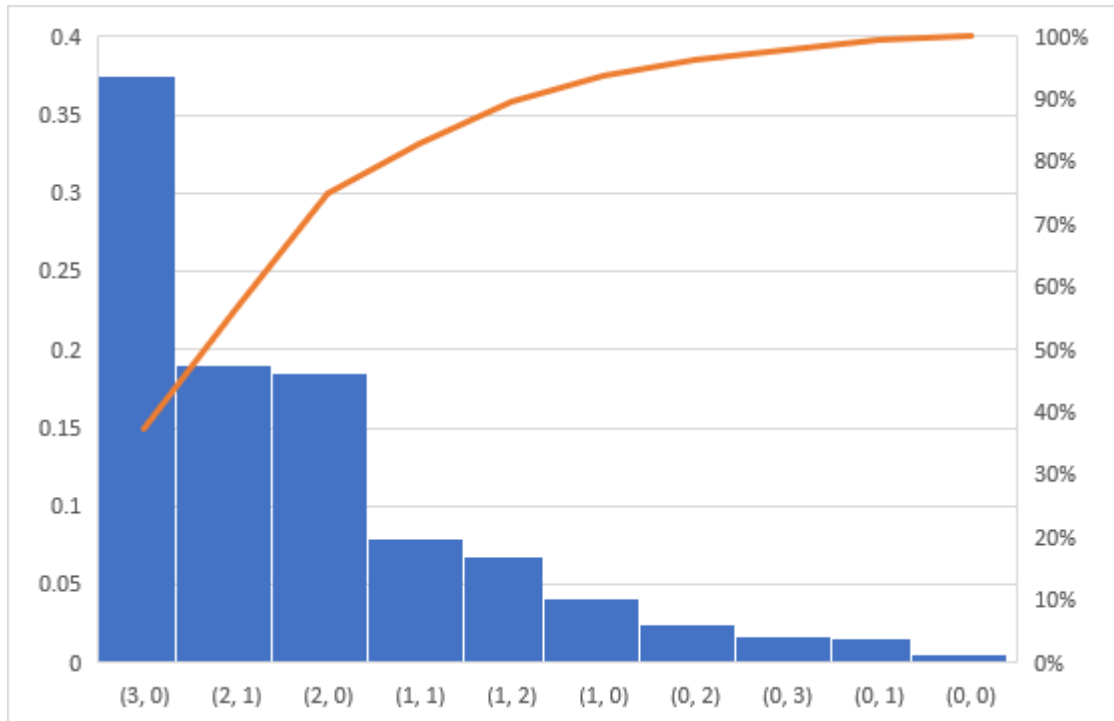


Figure 9. Steady state probabilities for the case that $\lambda_{0,1} = 0.05, \lambda_{0,2} = 0.01, \lambda_{1,2} = 0.01$

In the following, let's assume the speed of maintenance actions is halved which means $\mu_1 = 0.02$ and $\mu_2 = 0.005$. The steady state probabilities are shown in Figure 10. As the figure shows, the steady state probability of (0,0) is around 0.13 which significantly decreases with respect to the base example due to the speed reduction of the maintenance actions. For this case, the idleness probabilities of PM and CM crew are 79% and 20%. It means that decreasing the speed of the maintenance actions significantly decreases the idleness probability of the CM crew, while the corresponding value for PM crew does not significantly change with respect to the base example.

Slower maintenance directly leads to lower system availability. The model allows decision-makers to quantify the impact of a slower repair process and set a performance baseline. It can be used to justify investments in better equipment, training, or increased staffing to improve the repair rate, as the model shows a direct link between maintenance rate, e.g., μ_1 and μ_2 , and system performance metrics like L (average number of failed units) or W (average downtime).

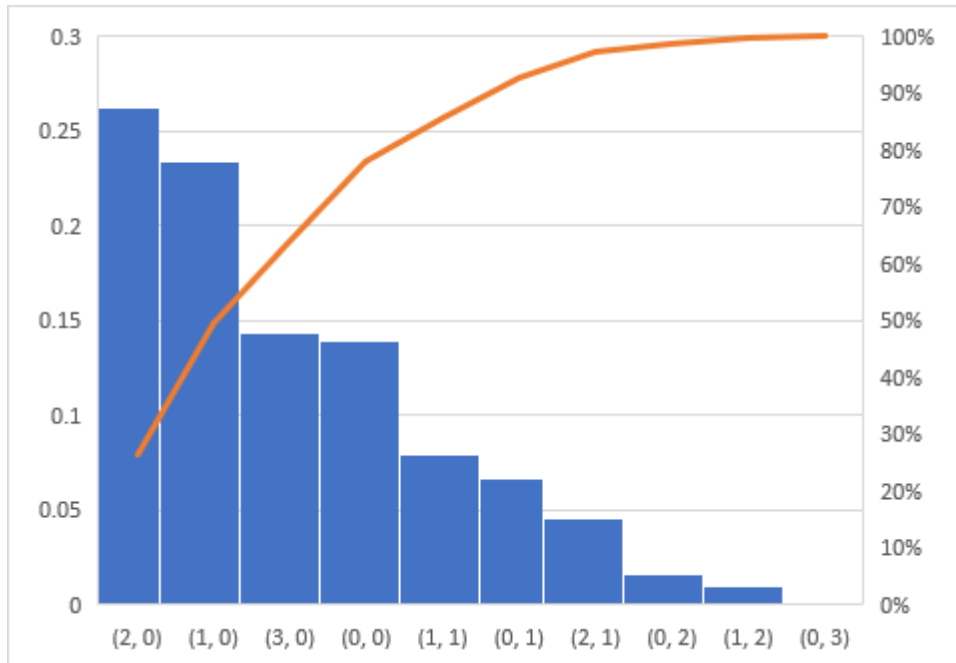


Figure 10. Steady state probabilities for the case that $\mu_1 = 0.02$ and $\mu_2 = 0.005$

Example 4. A numerical example inspired by wind turbine systems. We consider a wind farm comprising ten wind turbines equipped with condition monitoring capabilities. Each turbine consists of four main components: the main shaft, gearbox, generator, and transformer. Among these, the gearbox is particularly critical, as gearbox-related failures represent the most severe and frequent issues in wind turbine operations. Accordingly, the present analysis focuses on condition monitoring data specifically collected from the gearbox, whose primary function is to convert the rotor's low-speed rotation into the higher-speed rotation required by the electrical generator. The deterioration process of each gearbox is modeled as a CTMC comprising two operating states and one absorbing failure state, with sojourn times in the operating states following exponential distributions, consistent with the approach of Makis and Salari (2017). For the numerical illustration, the model parameters are assigned the values summarized in Table 6.

Table 6. Parameter values used in Example 4

$\lambda_{0,1}$	$\lambda_{0,2}$	$\lambda_{1,2}$	m_1	m_2	C	μ_1	μ_2
0.2619×10^{-3}	0.1726×10^{-3}	0.3036×10^{-3}	1	1	10	0.01	0.005

The results of the python code are reported in Table 7 and for all other states $\pi_{i,j} < 0.01$.

Table 7. Steady-state probabilities of Example 4

$\pi_{0,0}$	$\pi_{1,0}$	$\pi_{0,1}$	$\pi_{2,0}$	$\pi_{1,1}$	$\pi_{0,2}$	$\pi_{3,0}$	$\pi_{2,1}$
0.501931	0.182196	0.126998	0.059643	0.042036	0.028043	0.017354	0.012298

In the following, some descriptions about the model's validation are presented. To solve the system of linear equations of steady state probabilities, we use Python code. After obtaining π , we perform four verification checks:

1. Normalization: $\sum_i \pi_i = 1.0000$
2. Global balance: $\sum_i \pi_i q_{ij} = 0$ for all states j
3. Physical consistency: All probabilities are non-negative and less than 1 (computed to machine precision)
4. Flow balance verification: For each state i , we verify that the rate of probability inflow equals the rate of outflow:

$$\text{Inflow}_i = \sum_{j \neq i} \pi_j q_{ji}, \text{Outflow}_i = \pi_i \sum_{j \neq i} q_{ij} \quad (19)$$

These analyses confirm the numerical solution's accuracy. Furthermore, a Monte Carlo simulation is conducted. For example 3, the number of replications is considered 20, total events per replication are 10^7 , and 10^5 events are considered as warm up period. The mean of simulation replications, 95% confidence intervals along with analytical results are reported in Table 8. The data of this table clearly shows closeness of the two approaches results.

Table 8. Comparative study of the Monte Carlo simulation and the analytical results for Example 3 (values are rounded to four decimal places)

	$\pi_{0,0}$	$\pi_{1,0}$	$\pi_{0,1}$	$\pi_{1,1}$	$\pi_{2,0}$	$\pi_{2,1}$	$\pi_{0,2}$	$\pi_{3,0}$	$\pi_{1,2}$	$\pi_{0,3}$
Analytical Results	0.407	0.254	0.12	0.053	0.107	0.011	0.02	0.022	0.005	0.001
Mean of Simulation Results	0.4069	0.2546	0.1197	0.0527	0.1070	0.0114	0.0198	0.0220	0.0047	0.0014
Lower Confidence Interval	0.4060	0.2539	0.1194	0.0525	0.1064	0.0112	0.0197	0.0217	0.0046	0.0014
Upper Confidence Interval	0.4079	0.2553	0.1199	0.0528	0.1076	0.0115	0.0199	0.0222	0.0047	0.0015

To further validate the model, a Monte Carlo simulation is also conducted for the larger system of Example 4. It consists of 10 units and consequently has 66 feasible states. For this case, 30 replications are considered while each replication consists of 5×10^6 total events and the first 5×10^5 events are considered as the warm up period. The

results are reported in Table 9. For the other states that are not shown in Table 9, i.e., states that $\pi_{i,j} < 0.01$, the results of simulation and analytical method also show close conformance.

Table 9. Comparison of analytical and Monte Carlo simulation for Example 4

State (n_2, n_1)	Analytical results	Simulation Mean	95% confidence interval for simulation results
(0, 0)	0.501931	0.5022	[0.5008, 0.5036]
(1, 0)	0.182196	0.1818	[0.1805, 0.1831]
(0, 1)	0.126998	0.1272	[0.1260, 0.1282]
(2, 0)	0.059643	0.0594	[0.0585, 0.0603]
(1, 1)	0.042036	0.0419	[0.0412, 0.0426]
(0, 2)	0.028043	0.0278	[0.0271, 0.0285]
(3, 0)	0.017354	0.0172	[0.0165, 0.0179]
(2, 1)	0.012298	0.0121	[0.0115, 0.0127]

In the following, according to the steady state probabilities of Example 4, some performance measures are provided in Table 10. The equation employed to compute of each measure is also provided in this table.

Table 10. Some performance measures for the system of Example 4

Performance measure	Equation	Result
Average failed units/turbines	$E[n_2] = \sum n_2 \pi_{n_2, n_1}$	0.4742
Average degraded units/turbines	$E[n_1] = \sum n_1 \pi_{n_2, n_1}$	0.2907
System operational availability	$1 - \frac{E[n_2]}{C}$	95.26%
Probability of at least one failed unit	$1 - P(n_2 = 0)$	33.64%
Probability ≥ 2 failed	$1 - P(n_2 = 0) - P(n_2 = 1)$	10.21%
Average failed units waiting for CM	$\sum \max(n_2 - 1, 0) \cdot \pi_{n_2, n_1}$	0.1377
CM utilization	$P(n_2 \geq 1)$	33.64%
PM utilization	$P(n_1 \geq 1)$	23.30%

According to the performance measures reported in Table 10, the system operational availability is 95.26%. It means that in the long term, on average, around 95% of the turbines are operational, i.e., they are either in State 0 or 1. The table shows that in the long-term average number of failed units is 0.4742. However, the probability of having at least one failed unit or two failed units at an arbitrary time point are 0.34 and 0.10, respectively. As we have 1 CM crew member, whenever more than one unit fails a queue for CM actions is formed. The CM crew utilization is 33.64% which means that around 66.36% of the time, the CM crew is idle and around 76.7% the PM crew is idle. To reduce PM/CM crew idle time and improve productivity, the following suggestions can be provided. Using more proactive PM actions, and utilizing condition-based maintenance can improve PM crew productivity so that early detection of unhealthy states occurs and the transition from a degraded state to the failed state is prevented as much as possible.

For CM actions, employing flexible or on-call CM resources can be advised. Having more inventory for the spare parts of the wind turbine gearboxes can also be recommended to improve the system availability and reduce the system downtime.

6. Conclusion

In the current study, a production system consisting of multiple identical units was studied. The system units are characterized by both structural and stochastic independence so that the units are independently operating. Each system unit is modeled with a three-state state space, specifically distinguishing between healthy and unhealthy operational modes alongside a terminal failure state. The unit undergoes a progressive degradation process, transitioning from a healthy state to an unhealthy state before ultimately reaching a state of total failure. Transitions among the unit states are governed by independent exponential distributions. For the units in the unhealthy state, PM action is conducted to return them to the healthy state, and for the units in the failure state, CM action is conducted to return them to the healthy state. There are limitations for CM and PM actions so that a limited number of unhealthy units can be concurrently maintained using the PM actions, and similarly, a limited number of failed units can be simultaneously maintained using CM action. A two-element vector is defined to show the state of the system. Using the concepts of the birth-death process and the queueing theory with a limited population size and a limited number of servers, which is abbreviated as M/M/m//C, the transition rate diagram and the equilibrium equations are derived. We provide several numerical examples and conduct sensitivity analyses to examine how changes in key parameters influence the system's behavior.

The research can be extended from two main directions as follows:

- Extending the study so that the units of the system have more than two operational states
- Providing mathematical or simulation model to optimize the maintenance resources

Author Contributions

“Conceptualization, Hasan Rasay and Jalal Taji; methodology, Hasan Rasay; software, Hasan Rasay; validation, Hasan Rasay and Jalal Taji ; formal analysis, Hasan Rasay and Jalal Taji.; investigation, Jalal Taji.; resources, Jalal Taji.; data curation, Hasan Rasay; writing Hasan Rasay and Jalal Taji; writing—review and editing, All authors; visualization, Amir-Mohammad Golmohammadi and Mohammad Biglari Kami; supervision, Hasan Rasay; project administration, Hasan Rasay.

All authors have read and agreed to the published version of the manuscript.”

Data Availability Statement

Not applicable here

Acknowledgements

The authors would like to thank anonymous reviewers for their valuable suggestions in manuscript revision.

Ethical considerations

The authors avoided data fabrication, falsification, and plagiarism, and any form of misconduct.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest

The authors declare no conflict of interest.

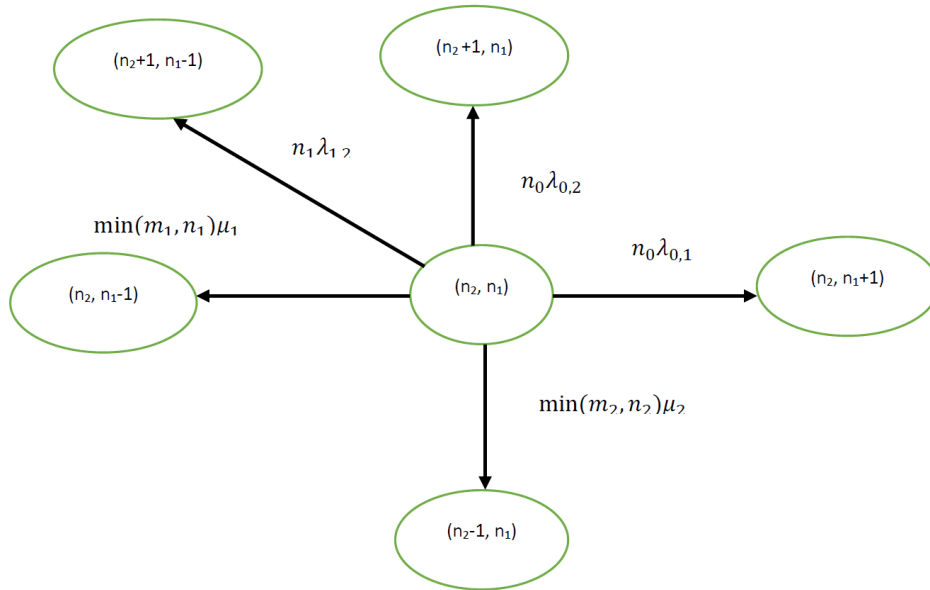
Appendix. Equilibrium equations and details of TRD

In the following, for the general case, i.e., a system with C units, m_1 and m_2 maintenance crews for the PM and CM respectively, the details of the transition diagram and equilibrium equations are provided. Depending on the system state, the transition diagram may be different to some extent as discussed in the following. In each case, the equilibrium equation for each specific state is derived by balancing the total transition rate out of the state with the total transition rate into it. Instead of providing a general equilibrium equation or a general TRD, we prefer to classify the system states into 7 groups and discuss each one separately below.

Group I. states (n_2, n_1) ; $\forall n_2, n_1 \geq 1, n_1 + n_2 < C$

Figure A.1 shows the TRD for this type of states. To avoid crowding of the figure, it is shown in two parts, so that Figure A.1(a) only shows the arrows leaving the state while Figure A.1(b) only shows the arrows entering the state. It is worth noting that for the example presented in subsection 5.1., state (1,1) corresponds to this class of states. Equating the rate at which the process leaves a state to the rate at which the process enters that state leads to the following equation:

$$\begin{aligned} \pi_{n_2, n_1} (n_0 \lambda_{0,1} + n_0 \lambda_{0,2} + n_1 \lambda_{1,2} + \min(n_1, m_1) \cdot \mu_1 + \min(n_2, m_2) \mu_2) \\ = \min(n_2 + 1, m_2) \cdot \mu_2 \cdot \pi_{n_2+1, n_1} + \min(n_1 + 1, m_1) \cdot \mu_1 \cdot \pi_{n_2, n_1+1} \\ + (n_1 + 1) \lambda_{1,2} \cdot \pi_{n_2-1, n_1+1} + n_0 \lambda_{0,2} \cdot \pi_{n_2-1, n_1} + n_0 \lambda_{0,1} \pi_{n_2, n_1-1} \end{aligned} \quad (\text{Eq. A1})$$



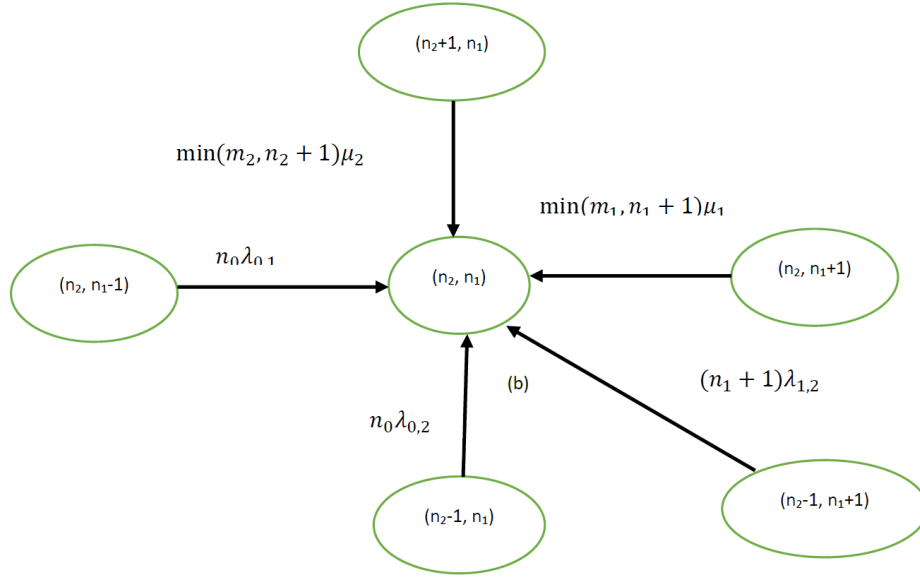


Figure A.1. TRD for Group I. (a) arrows leaving the state (b) arrows entering the state

Group II. States $(0, n_1)$; $0 < n_1 < C$

Figure A.2 shows TRD for such type of states and the equilibrium equation is as follows. For the example of Subsection 5.1, $(0,1)$ and $(0,2)$ belong to this type of states.

$$\begin{aligned} \pi_{0,n_1} (n_0\lambda_{0,2} + n_0\lambda_{0,1} + n_1\lambda_{1,2} + \min(n_1, m_1) \cdot \mu_1) \\ = n_0\lambda_{01}\pi_{0,n_1-1} + \mu_2\pi_{1,n_1} + \min(n_1+1, m_1) \cdot \mu_1\pi_{0,n_1+1} \end{aligned} \quad (\text{Eq. A2})$$

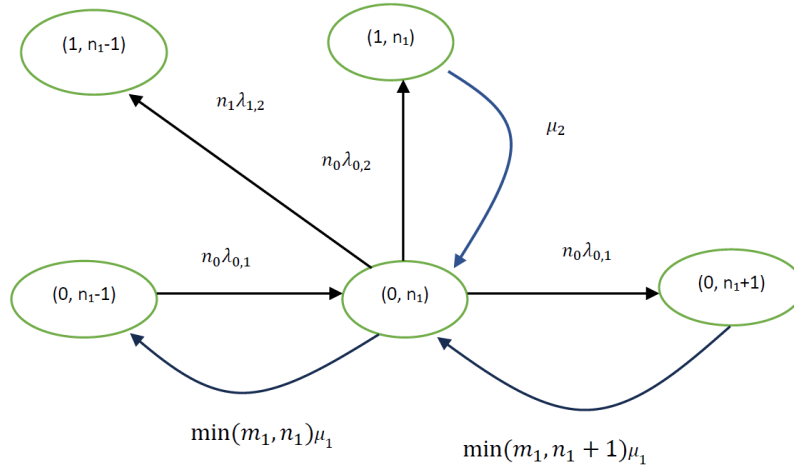


Figure A.2. TRD for Group II

Group III. States $(n_2, 0)$; $0 < n_2 < C$

Figure A.3 shows TRD for such type of states and the equilibrium equation is as follows. Also, for the example of Subsection 5.1, (1,0) and (2,0) belong to this group.

$$\begin{aligned} \pi_{n_2,0} & (n_0\lambda_{0,2} + n_0\lambda_{0,1} + \min(n_2, m_2)\mu_2) \\ & = \min(n_2 + 1, m_2) \cdot \mu_2 \cdot \pi_{n_2+1,0} + n_0\lambda_{0,2} \cdot \pi_{n_2-1,0} + \lambda_{1,2}\pi_{n_2-1,1} \\ & \quad + \mu_1 \cdot \pi_{n_2,1} \end{aligned} \quad (\text{Eq. A3})$$

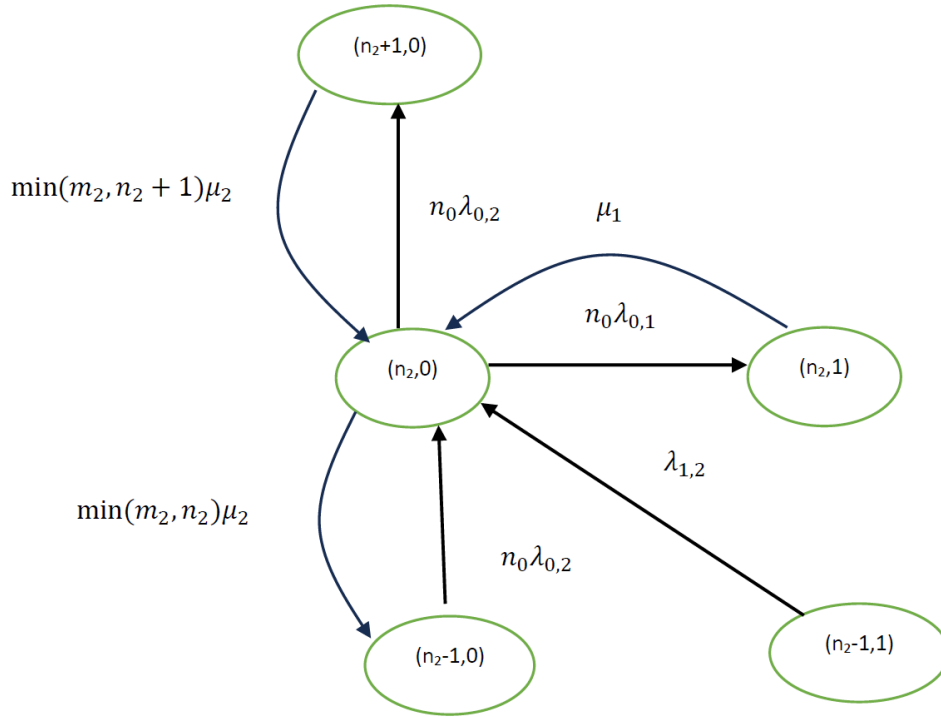


Figure A.3. TRD for Group III

Group IV. States (n_2, n_1) ; $n_2 + n_1 = C, n_1, n_2 \neq 0$

Figure A.4 shows TRD for such type of states and the equilibrium equation is as follows. In the example discussed, (1,2) and (2,1) belong to this group.

$$\begin{aligned} \pi_{n_2,n_1} & (n_1\lambda_{1,2} + \mu_1 \cdot \min(n_1, m_1) + \mu_2 \cdot \min(n_2, m_2)) \\ & = \lambda_{0,1} \cdot \pi_{n_2,n_1-1} + \lambda_{0,2} \cdot \pi_{n_2-1,n_1} + (n_1 + 1)\lambda_{1,2} \cdot \pi_{n_2-1,n_1+1} \end{aligned} \quad (\text{Eq. A4})$$

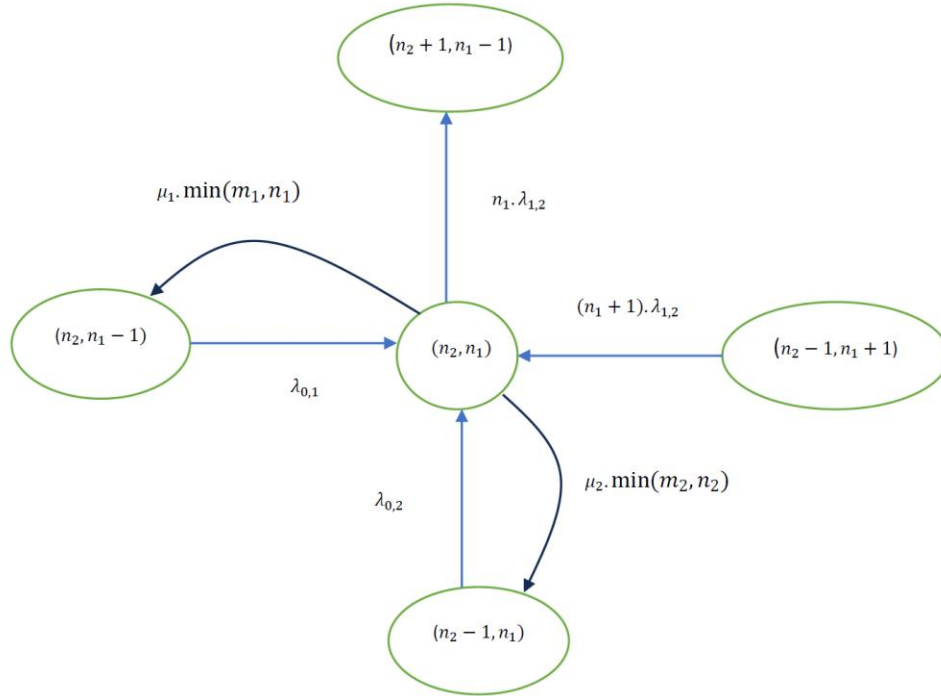
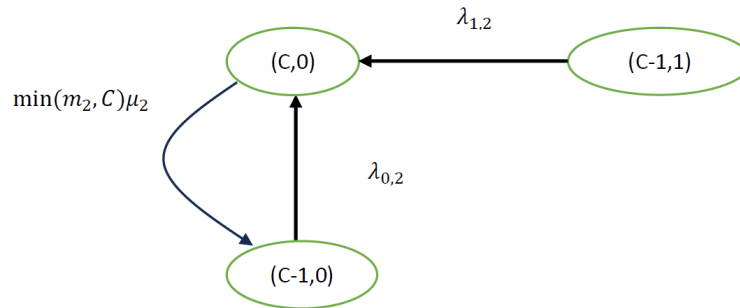


Figure A.4. TRD for Group IV

Group V. State $(C, 0)$

In this state, all units are in the failed state (state 2). In the example, $(3,0)$ has this situation. TRD is shown in Figure A.5 and the equilibrium equation is as follows:

$$\lambda_{1,2}\pi_{C-1,1} + \lambda_{0,2}\pi_{C-1,0} = \min(C, m_2) \cdot \mu_2 \cdot \pi_{C,0} \quad (\text{Eq. A5})$$

Figure A.5. TRD for $(C, 0)$

Group VI. State (0, C)

In this case, all units are in the unhealthy state (state 1). State (0,3) has this condition in the example presented in Subsection 5.1. TRD is shown in Figure A.6 and the equilibrium equation is as follows:

$$\pi_{0,C}(C\lambda_{1,2} + \min(C, m_1) \cdot \mu_1) = \lambda_{0,1} \cdot \pi_{0,C-1} \quad (\text{Eq. A6})$$

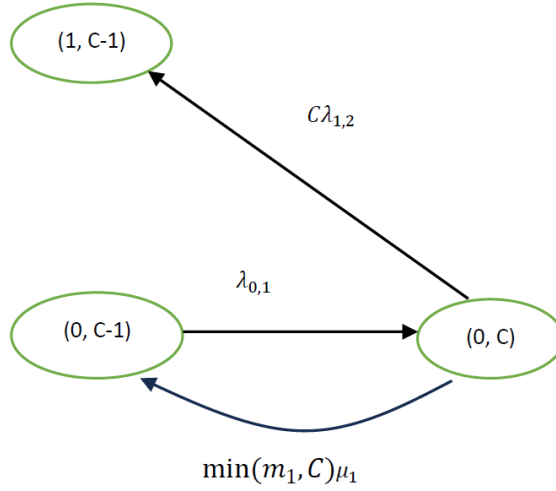


Figure A.6. TRD for (0,C)

Group VII. State (0, 0)

In this case, all units are in state 0. TRD and Equilibrium equation for this state are as follows:

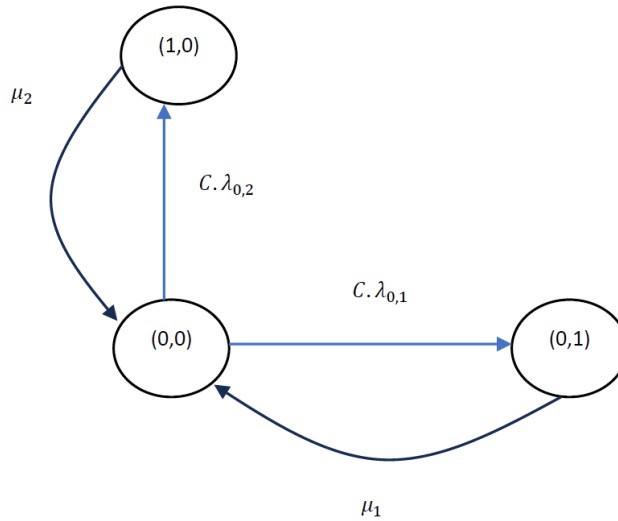


Figure A.7. TRD for (0,0)

$$\pi_{0,0}(C\lambda_{0,2} + C\lambda_{0,1}) = \mu_2\pi_{1,0} + \mu_1\pi_{0,1} \quad (\text{Eq. A7})$$

References

- Ambika, K., Vijayashree, K. V., & Janani, B. (2024). Modelling and analysis of production management system using vacation queueing theoretic approach. *Applied Mathematics and Computation*, 479, 128856. <https://doi:10.1016/j.amc.2024.128856>.
- Antala, K. S., & Sanga, S. S. (2025). Cost optimization, reliability, and MTTF analysis for failed excavators in hydraulic repair center using queueing theory. *Swarm and Evolutionary Computation*, 92, 101824. <https://doi:10.1016/j.swevo.2024.101824>.
- Azizi, F. and Salari, N. (2023). A novel condition-based maintenance framework for parallel manufacturing systems based on bivariate birth/birth–death processes. *Reliability Engineering & System Safety*, 229, 108798. <https://doi:10.1016/j.res.2022.108798>.
- Bekker, R., Bharti, B., Lan, L. and Mandjes, M. (2024). A queueing-based approach for integrated routing and appointment scheduling. *European Journal of Operational Research*, 318(2), 534–548. <https://doi:10.1016/j.ejor.2024.05.038>.
- de Castro, H.F. and Cavalca, K.L. (2006). Maintenance resources optimization applied to a manufacturing system. *Reliability Engineering & System Safety*, 91(4), 413–420. <https://doi:10.1016/j.res.2005.02.004>.
- Chiacchio, F., Oliveri, L., Khodayee, S.M. and D’Urso, D. (2022). Performance Analysis of a Repairable Production Line Using a Hybrid Dependability Queueing Model Based on Monte Carlo Simulation. *Applied Sciences*, 13(1), 271. <https://doi:10.3390/app13010271>.
- Cho, D.I. and Parlar, M. (1991). A survey of maintenance models for multi-unit systems. *European Journal of Operational Research*, 51(1), 1–23. [https://doi:10.1016/0377-2217\(91\)90141-H](https://doi:10.1016/0377-2217(91)90141-H).
- Colledani, M., Tolio, T., Fischer, A., Iung, B., Lanza, G., Schmitt, R. and Váncza, J. (2014). Design and management of manufacturing systems for production quality. *CIRP Annals*, 63(2), 773–796. <https://doi:10.1016/j.cirp.2014.05.002>.
- Comert, G., Amdeberhan, T., Begashaw, N., Medhin, N.G. and Chowdhury, M. (2024). Simple analytical models for estimating the queue lengths from probe vehicles at traffic signals: A combinatorial approach for nonparametric models. *Expert Systems with Applications*, 252, 124076. <https://doi:10.1016/j.eswa.2024.124076>.
- Ejaz, I., Alvarado, M., Gautam, N., Gebraeel, N. and Lawley, M. (2019). Condition-Based Maintenance for Queues With Degrading Servers. *IEEE Transactions on Automation Science and Engineering*, 16(4), 1750–1762. <https://doi:10.1109/TASE.2019.2893870>.
- Fallahnezhad, M.S. and Pourgharibshahi, M. (2017). Determining an Optimal Maintenance Policy for Three State Machine Replacement Problem Using Dynamic Programming. *International Journal of Supply and Operations Management Vcvc*, 7(2), 180–192. <https://doi:10.5267/j.msl.2017.3.001>.
- Federgruen, A. and So, K.C. (1990). Optimal Maintenance Policies for Single-Server Queueing Systems Subject to Breakdowns. *Operations Research*, 38(2), 330–343. <https://doi:10.1287/opre.38.2.330>.
- Ghalekhondabi, I. and Suer, G. (2018). Production line performance analysis within a MTS/MTO manufacturing framework: a queueing theory approach. *Production*, 28, e20180024. <https://doi:10.1590/0103-6513.20180024>.

- Iyoob, I.M., Cassady, C.R. and Pohl, E.A. (2006). Establishing Maintenance Resource Levels Using Selective Maintenance. *The Engineering Economist*, 51(2), 99–114. <https://doi:10.1080/00137910600695627>.
- Jackson, C., Pascual, R. and Kristjanpoller, F. (2025). Performance-based contract price negotiation with maintenance policies and limited resource allocation based on queueing and game theory. *International Journal of Production Economics*, 285, 109596. <https://doi:10.1016/j.ijpe.2025.109596>.
- Jayamani, V.N., Pavai Madheswari, S., Suganthi, P. and Josephine, S.A. (2024). An N - policy M/M/1 queueing model for energy saving mechanism in Networks. *Sustainable Computing: Informatics and Systems*, 44, 101016. <https://doi:10.1016/j.suscom.2024.101016>.
- Jeganathan, K., Koffer, V.A., Madheswari, S.P. and Venkatesan, T. (2025). An Investigation on Retrial Queueing-Inventory System with Server Maintenance Based on Customer Satisfaction. *International Journal of Applied and Computational Mathematics*, 11(5), 193. <https://doi:10.1007/s40819-025-02023-4>.
- Koyanagi, J. and Kawai, H. (2003). An optimal age maintenance for an M/G/1 queueing system. *Mathematical and Computer Modelling*, 38(11–13), 1333–1338. [https://doi:10.1016/S0895-7177\(03\)90135-X](https://doi:10.1016/S0895-7177(03)90135-X).
- Lee, J., Ghaffari, M. and Elmeligy, S. (2011). Self-maintenance and engineering immune systems: Towards smarter machines and manufacturing systems. *Annual Reviews in Control*, 35(1), 111–122. <https://doi:10.1016/j.arcontrol.2011.03.007>.
- Lee, M., Morrison, J.R. and Kalir, A.A. (2021). Practical Queueing Models for Preventive Maintenance Plan Optimization: Multiple Maintenance Types and Numerical Studies. *IEEE Transactions on Semiconductor Manufacturing*, 34(1), 104–114. <https://doi:10.1109/TSM.2020.3041789>.
- Luo, Z., Lian, Z., Li, J. and Yang, Z. (2025). Maintenance strategies for a queueing system with a standby equipment: An integrated approach using condition-based monitoring. *Reliability Engineering & System Safety*, 263, 111219. <https://doi:10.1016/j.ress.2025.111219>.
- Salari, N. and Makis, V. (2017). Comparison of two maintenance policies for a multi-unit system considering production and demand rates. *International Journal of Production Economics*, 193, 381–391. <https://doi:10.1016/j.ijpe.2017.08.003>.
- Salari, N. and Makis, V. (2020). Joint maintenance and just-in-time spare parts provisioning policy for a multi-unit production system. *Annals of Operations Research*, 287(1), 351–377. <https://doi:10.1007/s10479-019-03371-3>.
- Sarkar, A. and Suresh Kumar, E. (2012). Maintenance Policies of Single and Multi-Unit Systems in the Past and Present. *Asian Review of Mechanical Engineering*, 1(1), 13–24. <https://doi:10.51983/arme-2012.1.1.2510>.
- Shanthikumar, J.G., Ding, S. and Zhang, M.T. (2007). Queueing Theory for Semiconductor Manufacturing Systems: A Survey and Open Problems. *IEEE Transactions on Automation Science and Engineering*, 4(4), 513–522. <https://doi:10.1109/TASE.2007.906348>.
- Taji, J., Farughi, H., & Rasay, H. (2022). Economic-statistical design of fully adaptive multivariate control charts under effects of multiple assignable causes. *Computers & Industrial Engineering*, 173(108676). <https://doi:10.1016/j.cie.2022.108676>.
- Taji, J., Farughi, H., & Rasay, H. (2023). An integrated Markov chain model for the economic-statistical design of adaptive multivariate control charts and maintenance planning. *European Journal of Industrial Engineering*, 17(1), 1-28. <https://doi:10.1504/EJIE.2023.127739>.

- Wakiru, J., Pintelon, L., Muchiri, P.N. and Chemweno, P.K. (2021). Integrated remanufacturing, maintenance and spares policies towards life extension of a multi-component system. *Reliability Engineering & System Safety*, 215, 107872. [https://doi: 10.1016/j.ress.2021.107872](https://doi.org/10.1016/j.ress.2021.107872).