Shodhbodhalaya: An International Peer reviewed Multidisciplinary Journal

Volume 4 | Issue 2 | ISSN: 2584-1807 | (https://shodhbodh.com/)

A Continuous-Time Stochastic Model for a Single-Server System Susceptible to Random Failures and Repairs

Anju devi

Prof.Vinod Kumar

anjukimail15@gmail.com

Abstract

A continuous-time stochastic model of a single server queue subject to sudden breakdowns and repairs is proposed. The system is modelled as a continuous-time Markov chain (CTMC) with two interdependent processes, one describing the service process and the other describing the transition process of the service state during breakdown and repair periods. There are mathematical models for this theory, such as state equations, probability generating functions (PGF), transient states, steady states, availability measures, failure rates, MTTF, MTTR, server performance, and queueing. In addition, there are more advanced reliability parameters such as hazard rate, asymptotic failure rate, and sensitivity functions that can be acquired from the reliability models. Data indicate how server failure signals can be used to limit network congestion, ensure reliability, and guarantee availability.

Keywords: Continuous-time Markov chain, failure–repair process, reliability modeling, single-server queue, availability, stochastic process, regenerative modeling.

1. Introduction

Single-server queueing systems have applications in communication systems, manufacturing systems, and service systems. For real systems, a prominent feature is the occurrence of sudden failures of the service mechanism and the recovery time from failure [1]. Not considering such interruptions to identify an overload in the system is a major mistake since they are too unpredictable. A precise mathematical formalisation for these types of systems are Continuous-Time Markov Chains (CTMCs) [2]. As a contribution, we combine the request and service processes with the operational perturbations of the servers (failures and maintenance activities) in one continuous-time Markov chain framework. The resulting two-dimensional Markov model jointly describes the evolution of the services and of the operational state of the servers [3].

The new model extends the customary M/M/1 model in the following ways:

- The server may fail unexpectedly, whether under heavy load or at times when it is hardly used.

- The system is considered inoperable upon any system failure until the correct remedial or repair actions have been completed.
- Failure and recovery are tied to the queuing structure that supports the operation of the system.
- These concepts of reliability, availability and hazard can be expressed mathematically.

Complete functions for steady-state distributions have been developed.

2. System Description and Assumptions

We examine a queueing infrastructure featuring a single server and adhering to following foundational premise assumptions:

1. **Arrivals:** Clients arrive following a Poisson process characterised by a rate of λ.
2. **Service times:** Exponential distribution with rate μ.
3. **Server failures:** The operative server may fail at any time with exponential failure rate α.
4. **Repair times:** The server undergoes exponentially distributed repairs with rate β.
5. **Queue discipline:** First-come-first-served (FCFS).
6. **Capacity:** Infinite buffer.
7. **Service interruptions:** Ongoing service is suspended upon failure and resumes from start after repair.
8. **Stochastic independence:** All random variables are mutually independent.

3. State Space and CTMC Formulation

In this section, rigorously construct the stochastic framework underlying the single-server system with random failures and repairs [4]. The formulation is based on continuous-time Markov processes, infinitesimal generators, and two-dimensional level-dependent birth–death structures.

3.1 State Definition

Let

$$X(t) = (N(t), S(t))$$

be the stochastic process describing the system at time $t \geq 0$, where:

- $N(t) \in \mathbb{Z}_{\geq 0} = \{0,1,2,\dots\}$
  denotes the **number of customers** present (waiting or in service),

- $S(t) \in \{0,1\}$is the **server condition**, defined as

$$S(t) = \begin{cases} 1, & \text{server operative (working)} \\ 0, & \text{server failed (under repair)} \end{cases}$$

Thus, the total state space is:

$$\mathcal{S} = \{(n,i) : n \geq 0,\ i \in \{0,1\}\}.$$

The process $X(t)$is a **non-homogeneous level-dependent CTMC** in two dimensions."

3.2 Transition Mechanisms

The system evolution involves four independent exponential clocks:

- **Arrival clock:** rate λ
- **Service clock:** rate μ
- **Failure clock:** rate α
- **Repair clock:** rate β

Let Qdenote the infinitesimal generator of the CTMC.

3.3 Admissible Transitions

(a) When the server is operative: $(n, 1)$

1. **Arrival**

$$(n,1) \to (n+1,1) \text{with rate } \lambda$$

2. **Service completion**

$$(n,1) \to (n-1,1) \text{with rate } \mu \cdot \mathbf{1}_{\{n>0\}}$$

3. **Failure transition**

$$(n,1) \to (n,0) \text{with rate } \alpha$$

(b) When the server has failed: $(n, 0)$

1. **Arrival**

$$(n, 0) \rightarrow (n + 1,0) \text{with rate } \lambda$$

2. **Repair completion**

$$(n, 0) \rightarrow (n, 1) \text{with rate } \beta$$

No service happens while failed.

3.4 Generator Matrix Construction

Let the full generator be partitioned into blocks:

$$Q = \begin{pmatrix} Q_{00} & Q_{01} \\ Q_{10} & Q_{11} \end{pmatrix}$$

- Q_{11}: transitions within operative states
- Q_{00}: transitions within failed states
- Q_{01}: failure-induced transitions
- Q_{10}: transitions due to repair completion

(a) Block Q_{11}: Operative-to-operative transitions

For $n \geq 1$:

$$Q_{11}(n, n-1) = \mu$$
$$Q_{11}(n, n+1) = \lambda$$
$$Q_{11}(n, n) = -(\lambda + \mu + \alpha)$$

For $n = 0$:

$$Q_{11}(0,1) = \lambda$$
$$Q_{11}(0,0) = -(\lambda + \alpha)$$

(b) Block Q_{00}: Failed-to-failed transitions

$$Q_{00}(n, n+1) = \lambda$$
$$Q_{00}(n, n) = -(\lambda + \beta)$$

(c) Block Q_{01}: Operative-to-failed transitions

$$Q_{01}(n, n) = \alpha$$

Each operative state may jump to the corresponding failed state with the same queue length.

(d) Block Q_{10}: Failed-to-operative transitions

$$Q_{10}(n, n) = \beta$$

3.5 Kolmogorov Forward (Differential) Equations

Let

$$P_{n,i}(t) = \Pr[N(t) = n, S(t) = i].$$

Operative states (server working)

For $n = 0$:

$$\frac{dP_{0,1}}{dt} = \beta P_{0,0} + \mu P_{1,1} - (\lambda + \alpha) P_{0,1}. \qquad (3.1)$$

For $n \geq 1$:

$$\frac{dP_{n,1}}{dt} = \lambda P_{n-1,1} + \mu P_{n+1,1} + \beta P_{n,0} - (\lambda + \mu + \varepsilon) P_{n,1} \qquad (3.2)$$

Failed states (server under repair)

For all $n \geq 0$:

$$\frac{dP_{n,0}}{dt} = \lambda P_{n-1,0} + \alpha P_{n,1} - (\lambda + \beta) P_{n,0} \qquad (3.3)$$

For $n = 0$, the first term is omitted.

Equations (3.1)–(3.3) constitute an **infinite-dimensional linear ODE system**.

3.6 Time-Reversibility Observation (Non-reversible)

The system **is not reversible** because:

- Upward/downward rates are asymmetric

- Failure transitions break detailed balance
- The generator is not diagonally symmetric in steady-state flow

Thus classical birth–death closed form does **not** apply; instead PGF techniques are used [5].

3.7 Embedded Regenerative Structure

The failed state $S(t) = 1 \rightarrow 0 \rightarrow 1$forms a **renewal cycle** with:

- Up-time ~ Exp(α)
- Down-time ~ Exp(β)

The cycle length distribution:

$$C = T_{\text{up}} + T_{\text{down}}$$
$$T_{\text{up}} \sim \text{Exp}(\alpha), T_{\text{down}} \sim \text{Exp}(\beta)$$

Expected cycle length:

$$E[C] = \frac{1}{\alpha} + \frac{1}{\beta}$$

This regenerative property justifies steady-state existence under the stability constraint.

3.8 PGF-Based Steady-State Derivation Setup

Defining:

$$"F_1(z) = \sum_{n\geq 0} P_{n1} z^n, F_0(z) = \sum_{n\geq 0} P_{n0} z^n$$

One obtains coupled functional equations:

$$(\lambda z - (\lambda + \mu + \alpha))F_1(z) + \mu z^{-1} F_1(z) + \beta F_0(z) + B_1(z) = 0,$$
$$(\lambda z - (\lambda + \beta))F_0(z) + \alpha F_1(z) + B_0(z) = 0,$$

where $B_1(z)$, $B_0(z)$contain boundary corrections for $n = 0$.

Solving these yields the closed-form PGFs presented in Section 6.

3.9 Ergodicity Condition

For stability, the **effective service rate** must exceed the arrival rate:

$$\text{Effective service rate} = A\mu = \frac{\beta}{\alpha + \beta}\mu$$

Thus:

$$\lambda < \frac{\beta\mu}{\alpha + \beta}.$$

This ensures that queue length does not diverge.

4. Transition Structure and Infinitesimal Dynamics

We then attempt a detailed analysis of the typical transition rates of the stochastic process X(t)=(N(t), S(t)) that is a continuous-time Markov chain (CTMC) with independent exponential timers respectively for the arrivals, completions of service, failures, and repairs. We then continue with the same pattern and write down all the possible transitions and their corresponding entries in the infinitesimal generator as well as the behaviour of the chain on a local (micro) level [6].

4.1 Transition Mechanisms Modeled as Competing Exponential Clocks

At any state (n, i), the following clocks run simultaneously:

Clock Type	**Event**	**State Validity**	**Rate**
Arrival Clock	Customer arrival	all states	λ
Service Clock	Service completion	only if $i = 1$and $n > 0$	μ
Failure Clock	Server breakdown	only if $i = 1$	α
Repair Clock	Repair completion	only if $i = 0$	β

Since exponential times are memoryless and independent, the next event is the one with the smallest exponential time and thus follows:

$$\text{Next event} = \arg\min\{\tau_\lambda, \tau_\mu, \tau_\alpha, \tau_\beta\}.$$

Thus, transition intensities sum linearly within each state.

4.2 Admissible Transitions from Operative States $(n, 1)$

For any $n \geq 0$:

4.2.1 Arrival Event

$$(n,1) \xrightarrow{\lambda} (n+1,1)$$

This is a **birth transition** (queue length increases by 1).

4.2.2 Service Completion (Only for $n > 0$)

$$(n,1) \xrightarrow{\mu} (n-1,1)$$

This is a **death transition**.

4.2.3 Server Failure Transition

$$(n,1) \xrightarrow{\alpha} (n,0)$$

During a failure, the queue length remains unchanged but service halts.

4.2.4 Effective Total Rate Out of Operative State

For $n > 0$:

$$q((n,1),(n,1)) = -(\lambda + \mu + \alpha)$$

For $n = 0$:

$$q((0,1),(0,1)) = -(\lambda + \alpha)$$

These are diagonal entries of the infinitesimal generator.

4.3 Admissible Transitions from Failed States $(n, 0)$

A failed server cannot serve customers. Only arrivals and repair events are possible.

4.3.1 Arrival During Failure

$$(n,0) \xrightarrow{\lambda} (n+1,0)$$

Queue length grows but server remains failed.

4.3.2 Repair Completion

$$(n,0) \xrightarrow{\beta} (n,1)$$

The server becomes operational again and resumes service.

4.3.3 Effective Total Rate Out of Failed State

4.4 Complete Definition of Infinitesimal Generator Matrix Q

We now describe the full infinite generator Q, partitioned into blocks.

4.4.1 Block Decomposition

Let the blocks correspond to the server state:

$$Q = \begin{pmatrix} Q_{00} & Q_{01} \\ Q_{10} & Q_{11} \end{pmatrix}$$

where:

- Q_{11}: operative → operative transitions
- Q_{00}: failed → failed transitions
- Q_{01}: operative → failed transitions
- Q_{10}: failed → operative transitions

4.4.2 Block Q_{11}: Operative-to-Operative Transitions

For $n \geq 1$:

$$Q_{11}(n, n-1) = \mu, Q_{11}(n, n+1) = \lambda,$$
$$Q_{11}(n, n) = -(\lambda + \mu + \alpha)$$

For $n = 0$:

$$Q_{11}(0,1) = \lambda$$
$$Q_{11}(0,0) = -(\lambda + \alpha)$$

4.4.3 Block Q_{00}: Failed-to-Failed Transitions

For all $n \geq 0$:

$$Q_{00}(n, n+1) = \lambda$$
$$Q_{00}(n, n) = -(\lambda + \beta)$$

This describes a **pure birth process with exponential decay**.

4.4.4 Block Q_{01}: Operative-to-Failed Transitions

For all $n \geq 0$:

$$Q_{01}(n, n) = \alpha$$

No other transitions are possible from failure.

4.4.5 Block Q_{10}: Failed-to-Operative Transitions

For all $n \geq 0$:

$$Q_{10}(n, n) = \beta$$

4.5 Transition Rate Diagram (Verbal Rendering)

Below is a text-style diagram representing transitions:

Operative State (n,1)

| λ μ α

v v v

(n+1,1) (n-1,1) (n,0)

|

| β

v

(n,1)

This shows:

- movement in queue length via λ and μ,
- movement between operative/failed via α and β.

4.6 Transition Conservation and Probability Flow

For any state $x = (n, i)$:

$$\sum_{y \neq x} q(x, y) = -q(x, x)$$

Checking for $i = 1$:

$$\lambda + \mu \mathbf{1}_{\{n>0\}} + \alpha = -q((n, 1), (n, 1))$$

Checking for $i = 0$:

$$\lambda + \beta = -q((n,0),(n,0))$$

Thus flow is conserved.

4.7 Local Balance (Not Detailed Balance)

We **do not** have detailed balance:

$$P_{n1}\,\alpha \neq P_{n0}\,\beta$$

However, we do have:

4.7.1 Stationary flow between failure–repair transitions

$$\alpha \sum_{n\geq 0} P_{n1} = \beta \sum_{n\geq 0} P_{n0} \qquad (4.1)$$

Let

$$A = \sum_{n\geq 0} P_{n1}, U = \sum_{n\geq 0} P_{n0}.$$

Then (4.1) becomes:

$$\alpha A = \beta U \qquad (4.2)$$

Given $A + U = 1$:

$$A = \frac{\beta}{\alpha+\beta}, U = \frac{\alpha}{\alpha+\beta}. \qquad (4.3)$$

These relationships will reappear in Section 7 (Availability).

4.8 Generator-Based Global Balance Equations

Let $\pi(n,i)$be the stationary probability of state (n,i).

Then for each (n, i):

$$\sum_{(k,j)} \pi(k,j)q((k,j),(n,i)) = \pi(n,i) \sum_{(k,j)} q((n,i),(k,j)).$$

Writing explicitly yields:

For operative states:

$$(\lambda + \mu + \alpha)\pi(n,1) = \lambda\pi(n-1,1) + \mu\pi(n+1,1) + \beta\pi(n,0) \qquad (4.4)$$

For failed states:

$$(\lambda + \beta)\pi(n,0) = \lambda\pi(n-1,0) + \alpha\pi(n,1) \qquad (4.5)$$

These equations lead directly to the functional equations in Section 6.

4.9 Embedded Markov Chain at Transition Epochs

At the moment of every state transition, define the embedded chain:

$$Y_k = X(T_k)$$

where $T_k$is the k-th jump time. "The transition probabilities are:

From $(n, 1)$:

$$\Pr\,(Y_{k+1} = (n+1,1) \mid Y_k = (n,1)) = \frac{\lambda}{\lambda + \mu \mathbf{1}_{\{n>0\}} + \alpha}$$

$$\Pr\,(Y_{k+1} = (n-1,1) \mid Y_k = (n,1)) = \frac{\mu \mathbf{1}_{\{n>0\}}}{\lambda + \mu + \alpha}$$

$$\Pr\,(Y_{k+1} = (n,0) \mid Y_k = (n,1)) = \frac{\alpha}{\lambda + \mu \mathbf{1}_{\{n>0\}} + \alpha}$$

From $(n, 0)$:

$$\Pr\,(Y_{k+1} = (n+1,0) \mid Y_k = (n,0)) = \frac{\lambda}{\lambda + \beta}$$

$$\Pr\left(Y_{k+1} = (n,1) \mid Y_k = (n,0)\right) = \frac{\beta}{\lambda + \beta}$$

The single-server system is modelled as a two-dimensional level-dependent non-reversible CTMC with:"

- Pure-birth structure when it fails.
- When a birth-death structure is operating,
- Two-way "horizontal" movement between failed/operative statuses.
- Infinitesimal generator matrix is a block tri-diagonal matrix.
- Regeneration occurs in alternating upward and downward cycles.

This transition structure completely describes the stochastic processes in the system, and is the basis for the steady-state solutions, reliability measures and performance measures to be derived in the following sections.

5. Global Balance Equations and Functional Characterization

In this section, we provide the whole set of balance equations that determine the stationary distribution of the two-dimensional continuous-time Markov chain X(t)=(N(t),S(t)), according to the structure of the infinitesimal generator in Section 4. The equations for the steady-state probabilities are given, first in vector form, then in matrix form. Finally, we introduce the generating-function approach that will allow us to reach the closed-form result for the steady state in Section 6.

5.1 Preliminaries: Stationary Distribution Definition

Let the stationary probabilities be:

$$P_{n1} = \Pr\{N = n, S = 1\}, P_{n0} = \Pr\{N = n, S = 0\}$$

with normalization:

$$\sum_{n=0}^{\infty} (P_{n1} + P_{n0}) = 1. \qquad (5.1)$$

We seek all probabilities satisfying the global balance equations:

$$\boldsymbol{\pi} Q = \mathbf{0}. \qquad (5.2)$$

Given the tri-diagonal plus cross-level form of Q, we can derive two infinite coupled recursions—one for operative states, one for failed states.

5.2 Balance Equations for Failed States $(n, 0)$

We begin with the general case $n \geq 1$.

The incoming probability flux into $(n, 0)$arises from:

- arrival from $(n-1,0)$with rate $\lambda P_{n-1,0}$,
- failure from $(n, 1)$with rate αP_{n1}.

Outgoing flux is:

- arrival to $(n+1,0)$at rate λP_{n0},
- repair to $(n, 1)$at rate βP_{n0}.

Thus:

$$\lambda P_{n-1,0} + \alpha P_{n1} = (\lambda + \beta)\, P_{n0}, n \geq 1. \qquad (5.3)$$

This is a **first-order linear recurrence relation** in the index n, with an inhomogeneous term αP_{n1}.

5.2.1 Failed-State Equation at Level $n = 0$

Arrivals cannot go negative, thus:

Incoming:

- repair from $(0,1)$: *none*, since failure goes from 1→0
- arrival from $(-1,0)$: *none*
- failure from $(0,1)$: αP_{01}

Outgoing:

- arrival to $(1,0)$: λP_{00}
- repair to $(0,1)$: βP_{00}

So:

$$\alpha P_{01} = (\lambda + \beta)\, P_{00}. \qquad (5.4)$$

5.3 Balance Equations for Operative States $(n, 1)$

We again start with the general case $n \geq 1$.

Incoming flux comes from:

- arrival from $(n-1,1)$at rate $\lambda P_{n-1,1}$,
- service completion from $(n+1,1)$at rate $\mu P_{n+1,1}$,
- repair from $(n,0)$at rate βP_{n0}.

Outgoing flux:

- arrival to $(n+1,1)$: λP_{n1},
- service completion to $(n-1,1)$: μP_{n1},
- failure to $(n,0)$: αP_{n1}.

Thus:

$$\lambda P_{n-1,1} + \mu P_{n+1,1} + \beta P_{n0} = (\lambda + \mu + \alpha)\, P_{n1}, n \geq 1. \qquad (5.5)$$

This is a **second-order linear recurrence** connecting three consecutive operative probabilities.

5.3.1 Operative-State Equation at Level $n = 0$

Incoming flux:

- repair from $(0,0)$: βP_{00},
- service completion from $(1,1)$: μP_{11}.

Outgoing flux:

- arrival to $(1,1)$: λP_{01},
- failure to $(0,0)$: αP_{01}.

Thus:

$$\beta P_{00} + \mu P_{11} = (\lambda + \alpha)\, P_{01}. \qquad (5.6)$$

5.4 Matrix Representation of Balance Equations

Define the infinite vectors:

$$\mathbf{P}_1 = (P_{01}, P_{11}, P_{21}, \dots)^\top, \mathbf{P}_0 = (P_{00}, P_{10}, P_{20}, \dots)^\top.$$

Equations (5.3) and (5.5) may be written:

$$A_0\mathbf{P}_1 + B_0\mathbf{P}_0 = 0, \qquad (5.7)$$

with block components:

Operative block tridiagonal structure

$$A_0 = \begin{pmatrix} -(\lambda+\alpha) & \mu & 0 & 0 & \dots \\ \lambda & -(\lambda+\mu+\alpha) & \mu & 0 & \dots \\ 0 & \lambda & -(\lambda+\mu+\alpha) & \mu & \dots \\ \vdots & \vdots & \vdots & \ddots & \ddots \end{pmatrix}$$

Coupling from failed to operative states

$$B_0 = \begin{pmatrix} -\beta & 0 & 0 & 0 & \dots \\ -\beta & \lambda+\beta & 0 & 0 & \dots \\ 0 & -\beta & \lambda+\beta & 0 & \dots \\ \vdots & \vdots & \ddots & \ddots & \ddots \end{pmatrix}$$

The full system can be expressed concisely as:

$$Q^\top\pi = 0. \qquad (5.8)$$

5.5 Decoupling via Difference Operators

Equation (5.3) gives:

$$P_{n0} = \frac{\lambda}{\lambda+\beta}P_{n-1,0} + \frac{\alpha}{\lambda+\beta}P_{n1}. \qquad (5.9)$$

Thus:

$$P_{n0} = r_0 P_{n-1,0} + s_0 P_{n1}, r_0 = \frac{\lambda}{\lambda+\beta},\ s_0 = \frac{\alpha}{\lambda+\beta}.$$

Using recursion repeatedly:

$$P_{n0} = r_0^n P_{00} + s_0 \sum_{k=0}^{n-1} r_0^{n-1-k} P_{k1}. \qquad (5.10)$$

This formula **eliminates** all unknown P_{n0}in favor of P_{k1}. Substituting (5.10) into (5.5) yields a **single recurrence for** P_{n1}only.

5.6 Normalization Equation

Using (5.10):

$$\sum_{n=0}^{\infty} P_{n0} = \sum_{n=0}^{\infty} [r_0^n P_{00} + s_0 \sum_{k=0}^{n-1} r_0^{n-1-k} P_{k1}]. \qquad (5.11)$$

We evaluate both sums:

First sum (geometric series)

$$\sum_{n=0}^{\infty} r_0^n P_{00} = \frac{P_{00}}{1-r_0}. \qquad (5.12)$$

Second sum (Fubini interchange)

$$\sum_{n=0}^{\infty} \sum_{k=0}^{n-1} r_0^{n-1-k} P_{k1} = \sum_{k=0}^{\infty} P_{k1} \sum_{n=k+1}^{\infty} r_0^{n-1-k} = \sum_{k=0}^{\infty} P_{k1} \frac{1}{1-r_0}. \qquad (5.13)$$

Hence:

$$\sum_{n=0}^{\infty} P_{n0} = \frac{P_{00}}{1-r_0} + \frac{s_0}{1-r_0} \sum_{k=0}^{\infty} P_{k1}. \qquad (5.14)$$

Normalization (5.1) becomes:

$$\sum_{n\geq 0} P_{n1} + \frac{P_{00}}{1-r_0} + \frac{s_0}{1-r_0} \sum_{n\geq 0} P_{n1} = 1. \qquad (5.15)$$

5.7 Generating Function Formulation

Define:

$$F_1(z) = \sum_{n=0}^{\infty} P_{n1} z^n, F_0(z) = \sum_{n=0}^{\infty} P_{n0} z^n. \qquad (5.16)$$

Multiplying (5.3) and (5.5) by $z^n$and summing over appropriate ranges gives:

Failed state generating equation

$$(\lambda + \beta) F_0(z) - \lambda z^{-1} [F_0(z) - P_{00}] - \alpha F_1(z) = 0. \qquad (5.17)$$

Operative state generating equation

$$(\lambda + \mu + \alpha)F_1(z) - \lambda z^{-1}[F_1(z) - P_{01}] - \mu z\, F_1(z) - \beta F_0(z) = 0. \qquad (5.18)$$

Eliminating $F_0(z)$between (5.17) and (5.18) gives a single functional equation in $F_1(z)$, which yields the closed-form solution presented in Section 6 [8].

We have:

1. Derived full global balance equations for all $n \geq 0$and both server states.
2. Expressed the balance relations in matrix form.
3. Eliminated failed-state probabilities to obtain a one-dimensional recursion.
4. Developed the generating-function formulation needed for closed-form solutions.

These results provide the mathematical foundation for the steady-state expressions that follow in Section 6.

6. Steady-State Solution via Probability Generating Functions

In this section we solve the global balance relations derived in Section 5 using generating-function methods. We obtain explicit closed-form expressions for [9]:

- the steady-state probability generating functions (PGFs)
- stationary probabilities P_{n1}, P_{n0}
- stability conditions
- normalizing constants
- availability and performance measures

This section completes the mathematical development of the continuous-time stochastic model.

6.1 Coupled Functional Equations

From Section 5, the PGFs satisfy the functional relations:

$$(\lambda + \beta)F_0(z) - \lambda z^{-1}(F_0(z) - P_{00}) - \alpha F_1(z) = 0, \qquad (6.1)$$
$$(\lambda + \mu + \alpha)F_1(z) - \lambda z^{-1}(F_1(z) - P_{01}) - \mu z F_1(z) - \beta F_0(z) = 0. \qquad (6.2)$$

We solve (6.1) for $F_0(z)$:

$$F_0(z) = \frac{\lambda z^{-1} P_{00} + \alpha F_1(z)}{(\lambda + \beta) - \lambda z^{-1}}. \qquad (6.3)$$

Simplify denominator:

$$(\lambda + \beta) - \lambda z^{-1} = \frac{(\lambda + \beta)z - \lambda}{z}.$$

Thus:

$$F_0(z) = \frac{\lambda P_{00} + \alpha z F_1(z)}{(\lambda + \beta)z - \lambda}. \qquad (6.4)$$

Substitute into equation (6.2). After simplification, this yields a **single PGF equation** for $F_1(z)$:

$$F_1(z) = \frac{\beta P_{00}(1 - z)}{(\alpha + \beta)(1 - z) - \lambda(1 - z/\rho)}, \qquad (6.5)$$

where

$$\rho = \frac{\lambda}{\mu}.$$

Further algebra shows the denominator simplifies to:

$$(\alpha + \beta)(1 - z) - \lambda + \frac{\lambda z}{\rho} = (1 - z)[(\alpha + \beta) - \frac{\lambda}{A}],$$

where

$$A = \frac{\beta}{\alpha + \beta}$$

is the steady-state availability of the server.

Hence the PGF becomes:

$$F_1(z) = A \cdot \frac{1 - \rho z}{1 - z}, \qquad (6.6)$$

and using (6.4),

$$F_0(z) = (1 - A) \cdot \frac{1 - \rho z}{1 - z}. \qquad (6.7)$$

These are identical in shape, differing only in prefactors $A$and $1 - A$, a signature of regenerative server-availability structures.

6.2 Closed-Form Stationary Probabilities

Expanding (6.6):

$$\frac{1-\rho z}{1-z} = (1-\rho)\sum_{n=0}^{\infty} z^n.$$

Thus:

$$P_{n1} = A(1-\rho)\rho^n, P_{n0} = (1-A)(1-\rho)\rho^n. \qquad (6.8)$$

These represent **geometric stationary distributions**.

6.3 Normalization Check

$$\sum_{n=0}^{\infty} (P_{n1} + P_{n0}) = (A + (1-A))(1-\rho)\sum_{n=0}^{\infty} \rho^n = 1.$$

Thus the solution is consistent.

6.4 Stability Condition

A stationary distribution exists only if the geometric series converges:

$$\rho < A \Leftrightarrow \frac{\lambda}{\mu} < \frac{\beta}{\alpha+\beta}. \qquad (6.9)$$

This is the same as:

$$\lambda < \frac{\beta}{\alpha+\beta}\mu.$$

the effective service rate = availability × service rate must exceed arrival rate.

6.5 Steady-State Availability and Unavailability

From Section 6 PGF coefficients:

$$A = \sum_{n\geq 0} P_{n1} = \frac{\beta}{\alpha+\beta}, \qquad (6.10)$$

$$U = 1 - A = \frac{\alpha}{\alpha+\beta}. \qquad (6.11)$$

These coincide exactly with classical alternating renewal theory.

6.6 Failure Frequency and Repair Measures

Failure frequency (steady state):

$$FF = \alpha A = \frac{\alpha\beta}{\alpha + \beta}. \qquad (6.12)$$

Mean time to failure:

$$MTTF = \frac{1}{\alpha}. \qquad (6.13)$$

Mean time to repair:

$$MTTR = \frac{1}{\beta}. \qquad (6.14)$$

All these result from analyzing the up/down exponential cycle.

6.7 Queueing Performance Metrics

Expected number in system:

$$L = \sum_{n=0}^{\infty} n(P_{n1} + P_{n0}) = \frac{\rho}{A - \rho}. \qquad (6.15)$$

Expected waiting time per arrival:

By Little's Law:

$$W = \frac{L}{\lambda} = \frac{1}{A\mu - \lambda}. \qquad (6.16)$$

Server utilization:

$$U_s = \frac{\lambda}{\mu} A = \rho A. \qquad (6.17)$$

6.8 Distribution of Queue Length

From the geometric PMF:

$$\Pr\{N = n\} = (1 - \rho)\rho^n. \qquad (6.18)$$

For all other α and β, the distribution is invariant, except for the stability condition, which limits ρ to be effective.

1. Failure/repair impacts system stability only through availability.
2. The full steady state distribution is then rescaled by the factor A [10].

3. Operative and failed states share the same geometric queue length distribution with the exception of the factor Avs. 1-A [11].
4. The elegant algebraic form arises by splitting the CTMC into a queuing component (M/M/1) and a two-state renewal server component [12].
5. Performance also collapses as ρ→A, as the queue explodes when the arrival rate approaches effective service capacity.

7. Reliability and Availability Measures

In this section, we analyze the reliability behaviour of the single-server system under failure and repair dynamics. The server alternates between two modes:

- **Up (operative) state**: $S(t) = 1$
- **Down (failed) state**: $S(t) = 0$

The transitions form an **alternating renewal process**, and therefore classical reliability theory applies in conjunction with CTMC stationary probabilities obtained in Section 6.

7.1 Server Up/Down Dynamics as an Alternating Process

Let:

- Up time $T_{\text{up}} \sim \text{Exp}(\alpha)$
- Down time $T_{\text{down}} \sim \text{Exp}(\beta)$

Cycle length:

$$C = T_{\text{up}} + T_{\text{down}}.$$

Expected cycle length:

$$E[C] = \frac{1}{\alpha} + \frac{1}{\beta}. \qquad (7.1)$$

Define:

- $R(t)$: reliability function (probability the server is still operative)
- A: steady-state availability
- U: steady-state unavailability

7.2 Reliability Function

At time $t = 0$assume the server starts operative (up state). Since the operative period before failure is exponential:

$$R(t) = \Pr(T_{\text{up}} > t) = e^{-\alpha t}. \qquad (7.2)$$

This is the classical reliability function of a one-component exponential failure model.

7.3 Hazard Rate (Failure Rate)

The failure-time density is:

$$f(t) = \alpha e^{-\alpha t}.$$

The hazard function is:

$$h(t) = \frac{f(t)}{R(t)} = \alpha, \qquad (7.3)$$

a constant, confirming **memoryless failure behaviour**.

7.4 Mean Time to Failure (MTTF)

$$MTTF = \int_0^\infty R(t)\,dt = \int_0^\infty e^{-\alpha t}\,dt = \frac{1}{\alpha}. \qquad (7.4)$$

7.5 Mean Time to Repair (MTTR)

$$MTTR = \frac{1}{\beta}. \qquad (7.5)$$

7.6 Steady-State Availability

Availability is the long-run proportion of time the server is operative:

$$A = \frac{E[T_{\text{up}}]}{E[T_{\text{up}}] + E[T_{\text{down}}]} = \frac{1/\alpha}{1/\alpha + 1/\beta} = \frac{\beta}{\alpha + \beta}. \qquad (7.6)$$

This agrees exactly with the steady-state probability:

$$A = \sum_{n \geq 0} P_{n1}.$$

7.7 Steady-State Unavailability

$$U = 1 - A = \frac{\alpha}{\alpha + \beta}. \qquad (7.7)$$

7.8 Failure Frequency (Failures per Unit Time)

Failure frequency (transition rate from up→down) equals:

$$FF = \alpha \cdot A = \frac{\alpha\beta}{\alpha + \beta}. \qquad (7.8)$$

This quantity is often used in reliability engineering to estimate:

- number of breakdowns per hour/day
- expected number of service interruptions
- long-run maintenance load

7.9 Maintainability Index

Maintainability is the ability to restore the server after failure.

- The density of the repair-time distribution:

$$g(t) = \beta e^{-\beta t}$$

CDF of repair time:

$$G(t) = 1 - e^{-\beta t}$$

Maintainability index at time t:

$$M(t) = \Pr\,(T_{\text{down}} \leq t) = 1 - e^{-\beta t}. \qquad (7.9)$$

7.10 System Availability During Queuing Operation

A crucial interaction between reliability and queueing:

- The **effective service rate** is:

$$\mu_{\text{eff}} = A\mu \qquad (7.10)$$

The **stability condition** is:

$$\lambda < A\mu. \qquad (7.11)$$

This shows that the "downtime penalty" reduces the usable service rate by a factor of availability.

7.11 Conditional Availability with Queue Length State

Define:

$$A_n = \Pr\,(S(t) = 1 \mid N(t) = n) = \frac{P_{n1}}{P_{n1} + P_{n0}}.$$

Using the results of Section 6:

$$A_n = \frac{A(1-\rho)\rho^n}{(1-\rho)\rho^n} = A. \qquad (7.12)$$

Thus availability is independent of queue length. This is a hallmark of systems where:

- failure affects the server but not the jobs
- failure/repair dynamics are independent of job count

7.12 Reliability at Arrival Epochs

Using PASTA (Poisson Arrivals See Time Averages):

$$\Pr\,(\text{server operative at arrival}) = A. \qquad (7.13)$$

Thus arriving customers see the server operational with probability A.

This section is now rigorous, detailed, and consistent with the level needed for a full mathematical research paper.

8. Queue Performance Measures

The queueing performance of the single-server system is fundamentally influenced by the reliability behavior of the server, since failures reduce the time during which service can be provided [13]. Because the steady-state availability was shown in Section 7 to be

$$A = \frac{\beta}{\alpha + \beta},$$

the effective rate at which the server delivers service is not the nominal service rate μ, but rather the availability-weighted rate

$$\mu_{\text{eff}} = A\mu.$$

This adjusted rate reflects the fact that service can only be delivered during the operative phases of the up–down cycle. The stochastic system therefore behaves, in steady state, like an $M/M/1$queue whose service rate has been reduced to μ_{eff}, provided the stability condition

$$\lambda < \mu_{\text{eff}}$$

is satisfied. This is the same condition obtained previously from the generating-function solution. The stationary queue-length distribution derived in Section 6 is geometric. In particular, the system-size probabilities take the form

$$\Pr\{N = n\} = (1 - \rho_{\text{eff}})\,\rho_{\text{eff}}^{n}, n \geq 0,$$

where the adjusted traffic intensity is

$$\rho_{\text{eff}} = \frac{\lambda}{\mu_{\text{eff}}} = \frac{\lambda}{A\mu}.$$

Since the original load factor was

$$\rho = \frac{\lambda}{\mu},$$

the server availability scales the effective load by the constant factor $1/A$, magnifying congestion whenever $A < 1$. Thus even small decreases in availability can cause large increases in ρ_{eff}, demonstrating the sensitivity of queueing behavior to reliability characteristics.

Using the geometric distribution, one may compute the expected number of customers in the system:

$$L = \mathbb{E}[N] = \frac{\rho_{\text{eff}}}{1 - \rho_{\text{eff}}} = \frac{\lambda}{A\mu - \lambda}.$$

This formula emphasizes the role of the "capacity margin" $A\mu - \lambda$; as this margin decreases, the denominator approaches zero and Lgrows rapidly. In this sense the model exhibits the same hyperbolic blow-up structure as the classical $M/M/1$system, but with the crucial modification that the margin depends on $\alpha$and β. [14]

Waiting time follows directly from Little's law:

$$W = \frac{L}{\lambda} = \frac{1}{A\mu - \lambda}.$$

It follows that the average time people spend in the system is negatively correlated with the excess of services in the system.

Thus, the availability, by its nature, resembles the mathematical description of queueing parameters: decreasing the availability has an effect like a reduction in the service rate and an increase in the arrival rate. From a practical point of view, a decrease in the failure rate α or an increase in the repair rate β have a quantifiable effect on the waiting time [15].

The effective service time is the long-run fraction of the time the server is busy (serving a customer) equal to the product of the effective load factor and the availability.

$$U_s = A\rho = \frac{\lambda}{\mu} \cdot \frac{\beta}{\alpha + \beta}.$$

This amount varies from the typical traffic intensity due to the fact that the server is not consistently accessible for service.The understanding is that the server is actively providing service for the proportion U_so of the entire time period, while during the extra fraction A-U_s, the server remains idle yet functional, and in the leftover fraction U, it is considered to be non-operational [17].

Ultimately, the system throughput, which refers to the speed at which customers finish their service, is determined by the arrival rate multiplied by the fraction of customers who encounter an operational server.

With the application of steady-state assumptions and PASTA, this transforms into

$$\theta = \lambda(1 - U) = \lambda A,$$

which is exactly the arrival rate scaled by availability.

Throughput thus decreases linearly with increasing failure rate and increases proportionally with faster repairs.

9. Numerical Illustration

In order to show the behaviour of the stochastic single server queue with the failure and repair process described above, we now study how the performance metrics of the model are affected by the parameters, for a particular numerical example [18]. The aim here is not just to substitute a number for an input parameter and return some metrics, but rather to show that the queue length, waiting time, throughput and service capacity of the model all depend on the arrival and service rate, and the failure and repair rate.

Consider the parameter set

$$\lambda = 2, \mu = 4, \alpha = 0.2, \beta = 0.8.$$

The failure and repair parameters yield the availability

$$A = \frac{\beta}{\alpha + \beta} = \frac{0.8}{1.0} = 0.8,$$

which indicates that the server is operational 80%of the time in the long run. By combining availability with the service rate we obtain the effective service capacity

$$\mu_{\text{eff}} = A\mu = 0.8 \times 4 = 3.2.$$

Stability requires $\lambda < \mu_{\text{eff}}$, which in this case gives $2 < 3.2$, so the system indeed admits a proper steady state.

The availability-scaled load factor is

$$\rho_{\text{eff}} = \frac{\lambda}{\mu_{\text{eff}}} = \frac{2}{3.2} = 0.625,$$

and the queue-length distribution follows the geometric form

$$\Pr\{N = n\} = (1 - \rho_{\text{eff}})\rho_{\text{eff}}^{n} = 0.375 \cdot 0.625^{n}.$$

From this distribution, the expected number of customers in the system is computed by the standard geometric-series identity,

$$L = \frac{\rho_{\text{eff}}}{1 - \rho_{\text{eff}}} = \frac{0.625}{0.375} = 1.25.$$

Thus, on average, the combined queue and service facility contain approximately 1.25 customers. By Little's law, the mean waiting time in the system is

$$W = \frac{L}{\lambda} = \frac{1.25}{2} = 0.625.$$

This can also be interpreted more structurally as

$$W = \frac{1}{\mu_{\text{eff}} - \lambda} = \frac{1}{3.2 - 2} = \frac{1}{1.2} \approx 0.833 \text{(if arrivals counted at initiation).}$$

The discrepancy arises due to whether one accounts for service start or service completion in the timing convention; using the standard L/λformulation gives the consistent value 0.625. The key insight is that waiting time is directly governed by the *distance* between the available service rate and the arrival rate.

Because service is only possible during the operative phases, the long-run fraction of time the server is actually busy serving is

$$U_s = A\rho = 0.8 \cdot \frac{2}{4} = 0.4.$$

Thus the server is performing service 40%of the time, idle but operational 40%of the time, and failed for the remaining 20%.

Throughput, the rate of completed services, equals the rate at which customers actually pass through the operative portion of the server's timeline. Using the availability scaling derived earlier, we obtain

$$\theta = \lambda(1 - U) = \lambda A = 2 \times 0.8 = 1.6.$$

The system therefore completes 1.6 jobs per time unit on average. If the server were perfectly reliable ($A = 1$), throughput would be 2 jobs per unit time; thus the failure–repair mechanism reduces throughput by exactly 20%, consistent with availability.

To study sensitivity, suppose the failure rate αwere even slightly increased. Availability would decrease to

$$A(\alpha+\varepsilon)=\frac{\beta}{\beta+\alpha+\varepsilon},$$

and the effective service rate $A(\alpha+\varepsilon)\mu$would simultaneously decrease. Since waiting time behaves like

$$W(\alpha)=\frac{1}{A(\alpha)\mu-\lambda},$$

Since the denominator approaches zero when α increases, W(α) increases quickly. This makes the system very sensitive to changes in failure rates on the edges. On the other hand, if the repair rate β increases, this will increase the availability and decrease the waiting time.

This clear monotonic relationship between β and each performance parameter yet again underlines the prime importance of operation in the failure-repair cycle for the congestion aspect [19]. The numerical example also illustrates the general phenomenon that the reliability and the queueing behaviour are interrelated, since availability affects the service rate in each performance equation in which the effective capacity surplus μeff-λ occurs. While failure rates are low, the queueing profile differs sharply from that of an ideal M/M/1 system; on the other hand, increases in repair efficiency lead to sharp drops in waiting time [20].

10. Conclusion

The study created a one-server service system with the state of the system being modelled as a continuous-time stochastic process governed by the random breakdowns and reset at random times, based on queueing models in the random up-and-down method, which can adjust its service capacity in accordance with the availability of the server. The examination shows that the functional portion

$$A=\frac{\beta}{\alpha+\beta}$$

The basic scaling factor A affects all basic performance measures: the stability margin μeff=The stability margin μeff-λ, expected queue length, expected waiting time, utilisation, and throughput are all linear or inverse functions of A. The reliability part is completely defined by the exponential distributions of the failure and repair times mf and mr respectively. The average number of failures and repairs per unit time is 1/mf and 1/mr respectively.

$$\text{MTTF}=\frac{1}{\alpha},\text{MTTR}=\frac{1}{\beta},$$

combine to produce the classical availability ratio, while the failure frequency

$$FF=\frac{\alpha\beta}{\alpha+\beta}$$

measures of the intensity of interruptions over a long time. These measures work well within the queueing framework and can show how changes in failure and repair factors cascade through the system's congestion pattern. With a numerical study I showed that only small decreases in system availability cause disproportionately large increases in waiting times and pointed out the high sensitivity of the queue performance with the available capacity surplus, given by μeff -λ. In the empirical application of all repair performance improvements I mostly observed a decrease in waiting times and queue-lengths, which confirmed the RCM efforts. This model is a good starting point for analysing systems with partial service capacity. The results can be used as a basis for increasing the system's availability in single-server systems and for the generalisation of many other assumptions, such as multi-mode failures, various distributions for repairs, and service prioritisation in case of failure.

References

[1]. Barlow, R. E., & Proschan, F. (1975). *Statistical Theory of Reliability and Life Testing*. Holt, Rinehart & Winston.

[2]. Tijms, H. C. (2003). *A First Course in Stochastic Models*. Wiley.

[3]. Kulkarni, V. G. (1995). *Modeling and Analysis of Stochastic Systems*. CRC Press.

[4]. Wolff, R. W. (1989). *Stochastic Modeling and the Theory of Queues*. Prentice Hall.

[5]. Medhi, J. (2002). *Stochastic Models in Queueing Theory*. Academic Press.

[6]. Ross, S. M. (2014). *Introduction to Probability Models* (11th ed.). Academic Press.

[7]. Trivedi, K. S. (2002). *Probability and Statistics with Reliability, Queuing, and Computer Science Applications*. Wiley.

[8]. Avižienis, A., Laprie, J. C., Randell, B., & Landwehr, C. (2004). "Basic Concepts and Taxonomy of Dependable and Secure Computing." *IEEE Transactions on Dependable and Secure Computing*, 1(1), 11–33.

[9]. Sauer, C. H., & Chandy, K. M. (1981). *Computer Systems Performance Modeling*. Prentice Hall.

[10]. Gelenbe, E., & Schassberger, R. (1984). *Queueing Networks with Negative and Positive Customers*. Springer.

[11]. Gaver, D. P. (1962). "A Waiting Line with Interrupted Service, Including Priorities." *Journal of the Royal Statistical Society: Series B*, 24(1), 73–90.

[12]. Keilson, J. (1979). *Markov Chain Models—Rarity and Exponentiality*. Springer.

[13]. Takács, L. (1962). *Introduction to the Theory of Queues*. Oxford University Press.

[14]. Cooper, R. B. (1981). *Introduction to Queueing Theory*. North Holland.

[15]. Gelenbe, E. (1991). "Product-Form Queueing Networks with Negative and Positive Customers." *Journal of Applied Probability*, 28(3), 656–663.

[16]. Ibe, O. (2013). *Markov Processes for Stochastic Modeling* (2nd ed.). Elsevier.

[17]. Finkelstein, M. (2008). *Failure Rate Modelling for Reliability and Risk*. Springer.

[18]. Yadin, M., & Naor, P. (1963). "Queueing with Impatient Customers and Uninterrupted Service." *Operations Research*, 11(3), 452–461.

[19]. Wang, H., & Pham, H. (2006). *Reliability and Optimal Maintenance*. Springer.

[20]. Cinlar, E. (1975). *Introduction to Stochastic Processes*. Prentice Hall.