The M/M/1 Queue: A First-Principles Derivation

Abstract

This document presents a rigorous derivation of the M/M/1 queueing system from first principles. We begin by defining the continuous-time Markov chain model, derive the balance equations using probabilistic arguments, solve the resulting recurrence relation, and conclude with key performance metrics including expected queue length and waiting times via Little's Law.

Contents

1 Introduction: The Model

Consider a single-server queueing system where customers arrive, wait in line if the server is busy, receive service, and then depart. We wish to characterize the long-run behavior of such a system.

Definition 1.1 (State Variable)**.** Let $X(t)$ denote the number of customers in the system at time t, including any customer currently being served. The state space is

$$X(t) \in \{0, 1, 2, 3, \ldots\}.$$

The evolution of $X(t)$ is governed by two types of events:

1. **Arrivals:** Customers arrive randomly according to a Poisson process with rate $\lambda > 0$. Upon arrival, the state transitions as $n \to n+1$. A key property of the Poisson process is that the interarrival times are independent and exponentially distributed, and the arrival rate λ is independent of the current state n.

2. **Service Completions:** Service times are exponentially distributed with rate $\mu > 0$. When service completes, the state transitions as $n \to n-1$ (provided $n \geq 1$).

Remark 1.1. The notation "M/M/1" encodes this structure: the first M stands for *Markovian* (memoryless) arrivals, the second M for Markovian service times, and the 1 indicates a single server.

Under these assumptions, $X(t)$ is a **continuous-time Markov chain** on states $\{0, 1, 2, \ldots\}$. Our goal is to find the *steady-state distribution*—the long-run fraction of time the system spends in each state.

2 Steady-State Probabilities: Setup

Suppose the system runs for a long time. We define the steady-state probabilities:

Definition 2.1 (Steady-State Distribution)**.**

$$\pi_n = \lim_{t \to \infty} P(X(t) = n)$$

represents the long-run fraction of time the system has exactly n customers.

In particular:

- π_0 = fraction of time the system is empty
- π_1 = fraction of time exactly one customer is present
- π_n = fraction of time exactly n customers are present

The fundamental principle we use to find these probabilities is the **balance principle**: in steady state, the rate at which probability flows into each state must equal the rate at which it flows out.

3 Deriving the Balance Equations

3.1 Transition Probabilities in a Small Interval

Consider the system in state $n \geq 1$ and examine what happens over a small time interval Δt. The possible events are:

Event	**State Transition**	**Competing Events**
Arrival	$n \to n+1$	
Service Completion	$n \to n-1$	

These events are *independent* and *mutually exclusive* in an infinitesimal interval (only one can occur). Using the properties of exponential distributions:

$$P(\text{arrival in } \Delta t) = \lambda\Delta t + o(\Delta t) \tag{1}$$

$$P(\text{service completion in } \Delta t) = \mu\Delta t + o(\Delta t) \tag{2}$$

where $o(\Delta t)$ denotes terms that vanish faster than Δt as $\Delta t \to 0$.

3.2 Probability of Leaving a State

The probability of leaving state n (via either event) is:

$$\begin{aligned} P(\text{leave state } n \text{ in } \Delta t) &= P(\text{arrival or service}) \\ &= (\lambda + \mu)\Delta t + o(\Delta t) \end{aligned} \tag{3}$$

Meanwhile, since π_n is the fraction of time spent in state n:

$$P(\text{system is in state } n \text{ and leaves in } \Delta t) = \pi_n \cdot [(\lambda + \mu)\Delta t + o(\Delta t)]$$

3.3 Deriving the Rate of Leaving

We also know that if the system is in state n and doesn't leave, the probability of staying is:

$$P(\text{in state } n \text{ and doesn't leave in } \Delta t) = \pi_n\left[(1 - \lambda\Delta t)(1 - \mu\Delta t) + o(\Delta t)\right] = \pi_n\left[1 - (\lambda + \mu)\Delta t\right] + o(\Delta t)$$

Taking the limit as $\Delta t \to 0$:

$$\text{Rate of leaving state } n = \lim_{\Delta t \to 0} \frac{P(\text{system is in state } n \text{ and leaves in } \Delta t)}{\Delta t}$$

More formally, define the rate as:

$$\text{rate} = \lim_{\Delta t \to 0} \frac{P(\text{event occurs in } \Delta t)}{\Delta t}$$

Then:

$$\text{Rate of leaving state } n = \lim_{\Delta t \to 0} \frac{\pi_n(\lambda + \mu)\Delta t + o(\Delta t)}{\Delta t} = (\lambda + \mu)\pi_n$$

3.4 Total Exit Rate

Thus, the **total exit rate** from state n (for $n \geq 1$) is:

$$\text{Total exit rate from state } n = (\text{arrival rate} + \text{service rate}) \times \pi_n = (\lambda + \mu)\pi_n$$

4 Flow Balance: The Global Balance Equations

4.1 Ways to Enter a State

Consider state $n \geq 1$. The system can enter state n in two ways:

1. **From state $n-1$ via an arrival:**
   - Probability of being in state $n-1$: π_{n-1}
   - Arrival rate: λ
   - Contribution to entry rate: $\lambda\pi_{n-1}$
2. **From state $n+1$ via a service completion:**
   - Probability of being in state $n+1$: π_{n+1}
   - Service rate: μ
   - Contribution to entry rate: $\mu\pi_{n+1}$

Therefore:

$$\text{Total rate of entering state } n = \lambda\pi_{n-1} + \mu\pi_{n+1}$$

4.2 The Balance Equation for $n \geq 1$

In steady state, **flow in = flow out**:

$$\boxed{\lambda\pi_{n-1} + \mu\pi_{n+1} = (\lambda + \mu)\pi_n, \quad n \geq 1}$$

4.3 Special Case: State 0

State 0 is special because:

- The system *cannot go from state 0 to state* -1 (there's no such thing as -1 customers)
- When the system is empty, there is no service completion—only arrivals can occur

Therefore:

$$\begin{aligned} \text{Flow out of state } 0 &= \lambda\pi_0 \quad \text{(arrivals only)} \\ \text{Flow into state } 0 &= \mu\pi_1 \quad \text{(service completion from state 1)} \end{aligned}$$

The balance equation for state 0:

$$\boxed{\lambda\pi_0 = \mu\pi_1}$$

5 Solving the Recurrence Relation

5.1 Traffic Intensity

Before solving, we introduce a crucial parameter:

Definition 5.1 (Traffic Intensity)**.** The **traffic intensity** is defined as

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

This represents the fraction of time the server is busy on average. For stability (so the queue doesn't grow without bound), we require $\rho < 1$.

Remark 5.1. The condition $\rho < 1$ means arrivals occur slower than service completions on average ($\lambda < \mu$). If $\rho \geq 1$, the queue length grows to infinity—the system is unstable.

5.2 From Balance Equation to Recurrence

From the state 0 balance equation:

$$\pi_1 = \frac{\lambda}{\mu}\pi_0 = \rho\pi_0$$

Now we hypothesize a geometric form and verify. The balance equations for $n \geq 1$:

$$\lambda\pi_{n-1} + \mu\pi_{n+1} = (\lambda + \mu)\pi_n$$

This is a **linear recurrence relation** with constant coefficients. Rearranging:

$$\mu\pi_{n+1} - (\lambda + \mu)\pi_n + \lambda\pi_{n-1} = 0$$

5.3 Characteristic Equation

To solve, we try a solution of the form $\pi_n = a^n$ for some constant a. Substituting:

$$\mu a^{n+1} - (\lambda + \mu)a^n + \lambda a^{n-1} = 0$$

Dividing by a^{n-1}:

$$\mu a^2 - (\lambda + \mu)a + \lambda = 0$$

Using the quadratic formula:

$$a = \frac{(\lambda + \mu) \pm \sqrt{(\lambda + \mu)^2 - 4\mu\lambda}}{2\mu}$$

Simplifying the discriminant:

$$\begin{aligned}(\lambda + \mu)^2 - 4\mu\lambda &= \lambda^2 + 2\lambda\mu + \mu^2 - 4\mu\lambda \\ &= \lambda^2 - 2\lambda\mu + \mu^2 \\ &= (\mu - \lambda)^2\end{aligned}$$

Therefore:

$$a = \frac{(\lambda + \mu) \pm (\mu - \lambda)}{2\mu}$$

This yields two roots:

$$a_1 = \frac{\lambda + \mu + \mu - \lambda}{2\mu} = \frac{2\mu}{2\mu} = 1 \tag{4}$$

$$a_2 = \frac{\lambda + \mu - (\mu - \lambda)}{2\mu} = \frac{2\lambda}{2\mu} = \frac{\lambda}{\mu} = \rho \tag{5}$$

5.4 General Solution and Boundary Conditions

The general solution is:

$$\pi_n = C_1 \cdot 1^n + C_2 \cdot \rho^n = C_1 + C_2\rho^n$$

Now we apply constraints:

1. **Normalization:** $\displaystyle\sum_{n=0}^{\infty} \pi_n = 1$

   For $a_1 = 1$: The sum $\sum_{n=0}^{\infty} 1^n$ *diverges*. Hence we must have $C_1 = 0$.

2. **Convergence:** For $a_2 = \rho$: The geometric series $\sum_{n=0}^{\infty} \rho^n$ converges if and only if $\rho < 1$, giving $\frac{1}{1-\rho}$.

Therefore:

$$\pi_n = C \cdot \rho^n$$

for some constant C.

5.5 Finding the Normalization Constant

Using normalization:

$$\sum_{n=0}^{\infty} \pi_n = 1 \implies \sum_{n=0}^{\infty} C\rho^n = C \cdot \frac{1}{1-\rho} = 1$$

Thus $C = 1 - \rho$, giving us:

$$\pi_0 = 1 - \rho$$

5.6 The Steady-State Distribution

Theorem 5.1 (M/M/1 Steady-State Distribution)**.** *For the M/M/1 queue with arrival rate λ and service rate μ, where $\rho = \lambda/\mu < 1$, the steady-state probability of having n customers in the system is:*

$$\boxed{\pi_n = (1-\rho)\rho^n, \quad n = 0, 1, 2, \ldots}$$

This is a **geometric distribution** with parameter ρ.

Remark 5.2. As $\rho \to 1$ (the system approaches saturation), $\pi_0 \to 0$—the system is rarely empty. Conversely, when ρ is small, the system is empty most of the time.

6 Verification: Sanity Check

Let us verify that our solution satisfies the balance equations.

6.1 Checking the Balance Equation for $n \geq 1$

We need to verify:

$$\lambda\pi_{n-1} + \mu\pi_{n+1} = (\lambda + \mu)\pi_n$$

Substituting $\pi_n = \pi_0\rho^n$ where $\rho = \lambda/\mu$:

Left-hand side:

$$\begin{aligned} \lambda\pi_{n-1} + \mu\pi_{n+1} &= \lambda(\pi_0\rho^{n-1}) + \mu(\pi_0\rho^{n+1}) \\ &= \pi_0\left(\lambda\rho^{n-1} + \mu\rho^{n+1}\right) \end{aligned}$$

Factor out ρ^{n-1}:

$$= \pi_0\rho^{n-1}\left(\lambda + \mu\rho^2\right)$$

Substitute $\rho = \lambda/\mu$, so $\mu\rho = \lambda$:

$$\begin{aligned} &= \pi_0\rho^{n-1}\left(\lambda + \mu \cdot \frac{\lambda^2}{\mu^2}\right) \\ &= \pi_0\rho^{n-1}\left(\lambda + \frac{\lambda^2}{\mu}\right) \\ &= \pi_0\rho^{n-1} \cdot \lambda\left(1 + \frac{\lambda}{\mu}\right) \\ &= \pi_0\rho^{n-1} \cdot \lambda(1+\rho) \end{aligned}$$

Since $\lambda = \mu\rho$:

$$\begin{aligned} &= \pi_0\rho^{n-1} \cdot \mu\rho(1+\rho) \\ &= \pi_0\rho^n \cdot \mu(1+\rho) \\ &= \pi_0\rho^n(\mu + \mu\rho) \\ &= \pi_0\rho^n(\mu + \lambda) \end{aligned}$$

Right-hand side:

$$(\lambda + \mu)\pi_n = (\lambda + \mu)\pi_0\rho^n$$

Since LHS = RHS, our solution is verified. ✓

7 Performance Metrics

7.1 Expected Number of Customers in the System

Theorem 7.1. *The expected number of customers in the system is:*

$$E[L] = \frac{\rho}{1-\rho}$$

Proof. Using the definition of expectation for a discrete random variable:

$$\begin{aligned} E[L] &= \sum_{n=0}^{\infty} n \cdot \pi_n = \sum_{n=0}^{\infty} n(1-\rho)\rho^n \\ &= (1-\rho)\sum_{n=0}^{\infty} n\rho^n \end{aligned}$$

We use the standard result for the sum $\sum_{n=0}^{\infty} nx^n = \frac{x}{(1-x)^2}$ for $|x| < 1$:

$$E[L] = (1-\rho) \cdot \frac{\rho}{(1-\rho)^2} = \frac{\rho}{1-\rho}$$

□

Remark 7.1 (Side Derivation). To derive $\sum_{n=0}^{\infty} nx^n = \frac{x}{(1-x)^2}$: Start with the geometric series $\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}$. Differentiate both sides with respect to x:

$$\sum_{n=1}^{\infty} nx^{n-1} = \frac{1}{(1-x)^2}$$

Multiply both sides by x:

$$\sum_{n=1}^{\infty} nx^n = \frac{x}{(1-x)^2}$$

7.2 Little's Law

Theorem 7.2 (Little's Law)**.** *For any stable queueing system in steady state:*

$$E[L] = \lambda \cdot E[W]$$

where:

- $E[L]$ *= expected number of customers in the system*
- λ *= arrival rate*
- $E[W]$ *= expected time a customer spends in the system*

This remarkable result is *universal*—it holds regardless of the arrival distribution, service distribution, or number of servers!

7.3 Expected Time in System

Using Little's Law:

$$E[W] = \frac{E[L]}{\lambda} = \frac{1}{\lambda} \cdot \frac{\rho}{1-\rho} = \frac{1}{\lambda} \cdot \frac{\lambda/\mu}{1-\lambda/\mu}$$

Simplifying:

$$E[W] = \frac{1}{\lambda} \cdot \frac{\lambda/\mu}{(\mu-\lambda)/\mu} = \frac{1}{\lambda} \cdot \frac{\lambda}{\mu-\lambda} = \frac{1}{\mu-\lambda}$$

Theorem 7.3. *The expected time a customer spends in the system is:*

$$\boxed{E[W] = \frac{1}{\mu-\lambda}}$$

7.4 Expected Waiting Time in Queue

The total time in the system consists of:

$$W = W_q + S$$

where:

- W = total time in system
- W_q = waiting time in queue (before service begins)
- S = service time

Since service time is exponentially distributed with rate μ: $E[S] = 1/\mu$.
Taking expectations:

$$\begin{aligned} E[W] &= E[W_q] + E[S] \\ \frac{1}{\mu-\lambda} &= E[W_q] + \frac{1}{\mu} \end{aligned}$$

Solving for $E[W_q]$:

$$E[W_q] = \frac{1}{\mu-\lambda} - \frac{1}{\mu} = \frac{\mu-(\mu-\lambda)}{\mu(\mu-\lambda)} = \frac{\lambda}{\mu(\mu-\lambda)}$$

Theorem 7.4. *The expected waiting time in the queue (before service) is:*

$$\boxed{E[W_q] = \frac{\lambda}{\mu(\mu-\lambda)} = \frac{\rho}{\mu-\lambda}}$$

7.5 Behavior at Extremes

Examining the behavior of $E[W_q]$:

- **As $\lambda \to 0$:** $E[W_q] \to 0$ (no arrivals means no waiting)
- **As $\lambda \to \mu$:** $E[W_q] \to \infty$ (as the system approaches saturation, waiting times explode)
- **Large μ:** Fast service leads to shorter queues and wait times

8 Summary of Key Results

For an M/M/1 queue with arrival rate λ and service rate μ (where $\rho = \lambda/\mu < 1$):

Quantity	**Formula**
Steady-state probability	$\pi_n = (1-\rho)\rho^n$
Probability system is empty	$\pi_0 = 1 - \rho$
Expected customers in system	$E[L] = \dfrac{\rho}{1-\rho}$
Expected time in system	$E[W] = \dfrac{1}{\mu - \lambda}$
Expected waiting time in queue	$E[W_q] = \dfrac{\lambda}{\mu(\mu - \lambda)}$

These formulas provide the foundation for analyzing single-server queueing systems and extend naturally to more complex models like M/M/k (multiple servers) and M/G/1 (general service distributions).