Understanding Queueing Theory for ER Modeling: A Beginner's Guide to M/M/1 and M/M/c Systems

June 5, 2025

What Is Queueing Theory?

Queueing theory is the mathematical study of **waiting lines**. It helps us understand and analyze how items (like patients) arrive, wait, receive service, and then leave a system like an Emergency Room (ER).

In this guide, we focus on two models:

- **M/M/1**: 1 doctor (server)
- **M/M/c**: c doctors (servers)

Basic Terms and Inputs

- λ (lambda): **Arrival rate** — average number of patients arriving per hour.
- μ (mu): **Service rate** — average number of patients 1 doctor can serve per hour.
- c: Number of doctors (servers).
- ρ: **Traffic intensity** — how busy the system is.

How Do We Know μ?

To estimate μ:

- If 1 doctor takes 10 minutes per patient:

$$\mu = \frac{60}{10} = 6 \text{ patients/hour}$$

- More generally:

$$\mu = \frac{1}{\text{Average service time in hours}}$$

  Example: If average service time is 15 minutes:

$$\mu = \frac{1}{15/60} = 4 \text{ patients/hour}$$

M/M/1 Queue: One Doctor

Assumptions

- Patients arrive randomly (Poisson process).
- Each service time is exponentially distributed.
- Only 1 doctor.
- Infinite queue and no patient leaves the queue (FCFS).

Key Formulas

$$\rho = \frac{\lambda}{\mu} \quad \text{(Must be } < 1 \text{ for stability)}$$

$$L = \frac{\rho}{1-\rho} \quad \text{(Avg number of patients in system)}$$

$$L_q = \frac{\rho^2}{1-\rho} \quad \text{(Avg number of patients in queue)}$$

$$W = \frac{1}{\mu - \lambda} \quad \text{(Avg time a patient spends in the system)}$$

$$W_q = \frac{\rho}{\mu - \lambda} \quad \text{(Avg waiting time in queue)}$$

$$P_{\text{wait}} = \rho \quad \text{(Probability an arrival must wait)}$$

M/M/c Queue: Multiple Doctors

Assumptions

Same as M/M/1, but now we have c doctors working in parallel.

Formulas

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

Probability that no patients are in the system (idle):

$$\pi_0 = \left[\sum_{n=0}^{c-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^c}{c!(1-\rho)} \right]^{-1}$$

Probability a patient has to wait (Erlang C):

$$P_{\text{wait}} = \frac{(\lambda/\mu)^c}{c!(1-\rho)} \cdot \pi_0$$

Expected number of patients in queue:

$$L_q = \frac{P_{\text{wait}} \cdot \rho}{1-\rho}$$

Expected waiting time in queue:

$$W_q = \frac{L_q}{\lambda}$$

Total time in system:

$$W = W_q + \frac{1}{\mu}$$

Average number of patients in system:

$$L = \lambda W$$

Summary Table

Symbol	Meaning
λ	Arrival rate (patients/hour)
μ	Service rate (patients/hour per doctor)
c	Number of servers (doctors)
ρ	Traffic intensity
L	Avg number in system (waiting + service)
L_q	Avg number waiting
W	Avg time in system
W_q	Avg waiting time
P_{wait}	Probability a patient has to wait

Important Notes for Beginners

- The system must be stable: $\rho < 1$.
- Use assumptions carefully — real ERs may have limited beds or time-varying arrival rates.
- These models are used to help make decisions about how many doctors are needed.

📊 M/M/c System Summary Table (λ = 10, μ = 5, c = 3)

Metric	Value	Interpretation
ρ (Traffic Intensity)	0.6667	System is 66.67% utilized — stable (ρ < 1), no infinite queue buildup
Pw (Prob. of Waiting)	0.4444	44.44% of patients will need to wait before seeing a doctor
Lq (Avg. in Queue)	0.8889	On average, 0.89 patients are waiting — queue is short
Wq (Avg. Wait Time)	0.0889 hrs	Wait time ≈ 5.33 minutes for those who wait
L (Avg. in System)	2.8889	Average of 2.89 patients in clinic (waiting + being treated)
W (Avg. Time in System)	0.2889 hrs	Total time per patient ≈ 17.33 minutes (wait + service)
System Stability	✅ Stable	Since ρ < 1, system can handle current patient arrival rate without collapsing

Appendix: Step-by-Step Derivations of M/M/1 Formulas

1. Traffic Intensity (ρ)

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

This means: how busy is the doctor compared to their capacity?

2. Steady-State Probabilities (π_n)

We model the number of patients as a birth-death process. Using balance equations:

$$\pi_n = \rho^n \pi_0$$

Normalization gives:

$$\pi_0 = 1 - \rho \Rightarrow \pi_n = (1-\rho)\rho^n$$

3. Average Number in System (L)

$$L = \sum_{n=0}^{\infty} n\pi_n = \frac{\rho}{1-\rho}$$

4. Average Number in Queue (L_q)

$$L_q = L - \rho = \frac{\rho^2}{1-\rho}$$

5. Waiting Times (W, W_q)

By Little's Law:

$$W = \frac{L}{\lambda} = \frac{1}{\mu - \lambda}, \quad W_q = \frac{L_q}{\lambda} = \frac{\rho}{\mu - \lambda}$$

6. Probability of Waiting

$$P_{\text{wait}} = \rho$$

- **Average time waiting in queue (W_x or Wq)** is the time a customer (or patient) spends **waiting for service to begin**—i.e., from the moment they arrive until a server (doctor) actually starts serving them.
- **Average time in system (W)** is the total time a customer spends **in the entire system**, from the moment they arrive until they depart. That includes both:
  1. The time spent **waiting in queue** (Wq), and
  2. The time spent **being served** (service time).

Mathematically:

$$W = W_q + \frac{1}{\mu}$$

where $1/\mu$ is the **mean service time** (in hours). Equivalently, if you already know the average number in system L and the arrival rate λ, then ↓

K. N. Toosi University of Technology
Department of Computer Science
Faculty of Mathematics