Queueing Theory Analysis: M/M/1 and M/M/k Systems

Executive Summary

This project derives and validates analytical formulas for queueing systems from first principles. Starting with balance equations, we derive steady-state distributions, expected wait times, and queue lengths for both single-server (M/M/1) and multi-server (M/M/k) systems. A discrete-event simulation validates the theoretical results.

Analytical Framework

M/M/1 Queue

Single server with Poisson arrivals (rate λ) and exponential service (rate μ). Stability requires $\rho = \lambda/\mu < 1$.

Steady-state distribution: $\pi_n = (1 - \rho)\rho^n$

Expected customers in system: $E[L] = \rho/(1 - \rho)$

Expected time in system: $E[W] = 1/(\mu - \lambda)$

Expected wait in queue: $E[Wq] = \rho/(\mu - \lambda)$

M/M/k Queue

k parallel servers, same arrival process. Utilization per server: $\rho = \lambda/(k\mu) < 1$.

Empty system probability: $\pi_0 = [\Sigma_{n=0}^{k-1} (k\rho)^n/n! + (k\rho)^k/(k!(1-\rho))]^{-1}$

Erlang-C formula: $P(\text{wait}) = [(k\rho)^k/(k!(1-\rho))] \cdot \pi_0$

Expected queue length: $E[Lq] = P(\text{wait}) \cdot \rho/(1 - \rho)$

Expected wait in queue: $E[Wq] = P(\text{wait})/(k\mu - \lambda)$

Expected time in system: $E[W] = E[Wq] + 1/\mu$

Key Derivation: Little's Law

Little's Law connects queue length to wait time without requiring specific distributional assumptions:

$$\mathbf{L = \lambda W}$$

where L is expected number in system, λ is arrival rate, and W is expected time in system.

Simulation Results

Discrete-event simulation with 50,000 customers per test case (5,000 warmup discarded). Results compared against analytical formulas.

Test 1: M/M/1 Queue (Moderate Load)

$\lambda = 0.8$, $\mu = 1.0$, $k = 1$, $\rho = 0.80$

Metric	Simulation	Analytical	Error %
E[Wq]	4.322459	4.000000	8.06%
E[W]	5.328370	5.000000	6.57%
E[L]	4.199331	4.000000	4.98%

Test 2: M/M/1 Queue (High Load)

$\lambda = 0.95$, $\mu = 1.0$, $k = 1$, $\rho = 0.95$

Metric	Simulation	Analytical	Error %
E[Wq]	24.866331	19.000000	30.88%
E[W]	25.872242	20.000000	29.36%
E[L]	23.512980	19.000000	23.75%

Test 3: M/M/3 Queue (Multi-Server)

$\lambda = 2.0$, $\mu = 1.0$, $k = 3$, $\rho = 0.67$

Metric	Simulation	Analytical	Error %
E[Wq]	0.477445	0.444444	7.43%
E[W]	1.483353	1.444444	2.69%
E[L]	2.952623	2.888889	2.21%
P(wait)	0.457545	0.444444	2.95%

Test 4: M/M/5 Queue (Multi-Server)

$\lambda = 4.0$, $\mu = 1.0$, $k = 5$, $\rho = 0.80$

Metric	Simulation	Analytical	Error %
E[Wq]	0.603261	0.554113	8.87%
E[W]	1.609160	1.554113	3.54%
E[L]	6.405527	6.216450	3.04%
P(wait)	0.574965	0.554113	3.76%

Analysis

Convergence Behavior

Tests 1, 3, and 4 show errors in the 2-9% range, confirming the analytical formulas. Test 2 ($\rho = 0.95$) shows ~30% deviation. This is expected behavior, not an error.

Why High-ρ Systems Converge Slowly

Near $\rho = 1$, the variance of queue length is:

$$\mathbf{Var[L] = \rho/(1 - \rho)^2}$$

At $\rho = 0.95$, this equals 380. A single long busy period can skew the entire sample mean. Reliable estimates would require 500,000+ customers.

Key Insight

The simulation consistently overestimates wait times at high utilization. Understanding why models deviate from simulation—not just reporting numbers—is the core skill this project demonstrates.

Conclusions

The M/M/1 and M/M/k analytical formulas are validated by simulation. The project demonstrates: (1) first-principles derivation of steady-state distributions and performance metrics, (2) implementation of discrete-event simulation, (3) understanding of convergence behavior and model limitations.