Restaurant Wait Time Modeling

Queueing Theory: From First Principles to Assumption Failures

Complete Derivations, Simulation Results, and Analysis

Quantitative Finance Project Portfolio

Abstract

This project derives and validates queueing theory formulas from first principles, then demonstrates when standard assumptions fail. We develop analytical solutions for M/M/1 and M/M/k queues, implement discrete-event simulations to validate theory, and quantify the impact of non-Poisson arrivals (batch and Hawkes processes). Key finding: bursty arrivals increase wait times by 60-90% compared to Poisson predictions, with multiple servers partially mitigating this effect.

Contents

Part I
Theoretical Framework

1 M/M/1 Queue: Single Server Analysis

1.1 Model Assumptions

- **Arrivals:** Poisson process with rate λ (interarrival times $\sim \text{Exp}(\lambda)$)
- **Service times:** Exponential with rate μ (mean $E[S] = 1/\mu$)
- **Servers:** Single server
- **Queue capacity:** Infinite
- **Discipline:** First-Come-First-Served (FCFS)
- **Stability:** Utilization $\rho = \lambda/\mu < 1$

1.2 Steady-State Distribution Derivation

Let $\pi_n = P(\text{system has } n \text{ customers in steady state})$.

Balance equations: Flow into state n = Flow out of state n.

For $n = 0$:

$$\mu\pi_1 = \lambda\pi_0 \implies \pi_1 = \rho\pi_0 \tag{1}$$

For $n \geq 1$:

$$\lambda\pi_{n-1} + \mu\pi_{n+1} = (\lambda + \mu)\pi_n \tag{2}$$

Solution: Guess $\pi_n = \rho^n\pi_0$ and verify by substitution.

Normalization:

$$\sum_{n=0}^{\infty} \pi_n = \pi_0 \sum_{n=0}^{\infty} \rho^n = \frac{\pi_0}{1-\rho} = 1 \implies \pi_0 = 1 - \rho \tag{3}$$

Result

Steady-State Distribution:

$$\pi_n = (1-\rho)\rho^n, \quad n = 0, 1, 2, \ldots \tag{4}$$

1.3 Performance Metrics

Expected number in system:

$$E[L] = \sum_{n=0}^{\infty} n\pi_n = (1-\rho)\sum_{n=0}^{\infty} n\rho^n = (1-\rho)\cdot\frac{\rho}{(1-\rho)^2} = \frac{\rho}{1-\rho} \tag{5}$$

Little's Law: $L = \lambda W$ (remarkably general — no distributional assumptions needed)

Result

M/M/1 Performance Formulas:

$$E[L] = \frac{\rho}{1-\rho} \quad \text{(customers in system)} \tag{6}$$

$$E[W] = \frac{1}{\mu-\lambda} \quad \text{(time in system)} \tag{7}$$

$$E[W_q] = \frac{\rho}{\mu-\lambda} \quad \text{(wait in queue)} \tag{8}$$

$$E[L_q] = \frac{\rho^2}{1-\rho} \quad \text{(customers in queue)} \tag{9}$$

2 M/M/k Queue: Multiple Servers

2.1 Model Setup

- k parallel servers, each with service rate μ
- Total service rate: $\min(n, k) \cdot \mu$ when n customers present
- Utilization per server: $\rho = \lambda/(k\mu) < 1$
- Offered load: $a = \lambda/\mu = k\rho$

2.2 Steady-State Distribution

For $n \leq k$ (not all servers busy):

$$\pi_n = \frac{a^n}{n!}\pi_0 \tag{10}$$

The $n!$ arises because service rate increases as $\mu, 2\mu, \ldots, n\mu$.

For $n \geq k$ (all servers busy):

$$\pi_n = \frac{a^k}{k!} \cdot \rho^{n-k}\pi_0 \tag{11}$$

Geometric in ρ because service rate is constant at $k\mu$.

Normalization:

$$\pi_0 = \left[\sum_{n=0}^{k-1} \frac{a^n}{n!} + \frac{a^k}{k!(1-\rho)}\right]^{-1} \tag{12}$$

2.3 Erlang-C Formula

Probability of waiting (all servers busy at arrival):

Result

Erlang-C Formula:

$$P(\text{wait}) = C(k,a) = \frac{\frac{a^k}{k!} \cdot \frac{1}{1-\rho}}{\sum_{n=0}^{k-1} \frac{a^n}{n!} + \frac{a^k}{k!} \cdot \frac{1}{1-\rho}} \tag{13}$$

Performance Metrics:

$$E[W_q] = \frac{P(\text{wait})}{k\mu - \lambda} \tag{14}$$

$$E[W] = E[W_q] + \frac{1}{\mu} \tag{15}$$

3 M/G/1 Queue: General Service Distribution

3.1 Pollaczek-Khinchin Formula

For Poisson arrivals with *general* service distribution:

$$E[W_q] = \frac{\lambda E[S^2]}{2(1-\rho)} \tag{16}$$

Using $E[S^2] = E[S]^2(1 + C_s^2)$ where $C_s^2 = \text{Var}[S]/E[S]^2$:

Result

Pollaczek-Khinchin (Alternative Form):

$$E[W_q] = \frac{\rho}{1-\rho} \cdot \frac{1 + C_s^2}{2\mu} \tag{17}$$

Key Insight

The factor $(1 + C_s^2)/2$ is the **variance multiplier**:

- Deterministic service ($C_s^2 = 0$): Wait is **halved**
- Exponential service ($C_s^2 = 1$): Baseline M/M/1
- Heavy-tailed service ($C_s^2 > 1$): Wait **increases**

Part II
Assumption Failures: Non-Poisson Arrivals

4 Batch Arrivals: $M^X/M/1$

4.1 Model Setup

- Batch arrival rate: λ_b
- Batch size: B (random variable)
- Effective arrival rate: $\lambda = \lambda_b \cdot E[B]$

4.2 Internal Batch Delay

Customer i in a batch waits for customers $1, \ldots, i-1$:

$$\text{Internal wait for customer } i = (i-1) \cdot E[S] = \frac{i-1}{\mu} \tag{18}$$

Total internal waiting across batch:

$$\sum_{i=1}^{B} (i-1)E[S] = \frac{B(B-1)}{2\mu} \tag{19}$$

Expected internal delay per customer:

$$E[\text{internal delay}] = \frac{E[B^2] - E[B]}{2\mu} \tag{20}$$

4.3 Key Result: Batch Penalty

Result

Batch Arrival Wait Time:

$$E[W_q]_{\text{batch}} = E[W_q]_{M/M/1} + \frac{E[B^2] - E[B]}{2\mu} \tag{21}$$

Heavy Traffic Ratio:

$$\frac{E[W_q]_{\text{batch}}}{E[W_q]_{\text{Poisson}}} \approx \frac{E[B^2]}{2E[B]} \quad \text{as } \rho \to 1 \tag{22}$$

5 Hawkes Process Arrivals

5.1 Self-Exciting Intensity

$$\lambda(t) = \lambda_0 + \sum_{t_i < t} \alpha e^{-\beta(t-t_i)} \tag{23}$$

- λ_0 = baseline rate
- α = jump size per arrival
- β = decay rate
- Stability requires $\alpha/\beta < 1$

5.2 Mean and Variance

Result

Hawkes Process Properties:

$$\bar{\lambda} = \frac{\lambda_0}{1 - \alpha/\beta} \qquad \text{(mean rate)} \tag{24}$$

$$\text{Var}[N(T)] = \bar{\lambda} T \cdot \frac{1}{(1 - \alpha/\beta)^2} \qquad \text{(overdispersed)} \tag{25}$$

Key Insight

Hawkes arrivals have $\mathrm{Var}[N(T)] > \bar{\lambda}T$ (overdispersed compared to Poisson). This is because arrivals are positively correlated: $\mathrm{Cov}[X_i, X_j] > 0$.

Part III
Simulation Results

6 Validation: M/M/1 and M/M/k

Discrete-event simulation with 50,000 customers per test (5,000 warmup discarded).

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

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

Metric	Simulation	Analytical	Error %
$E[W_q]$	4.3225	4.0000	8.06%
$E[W]$	5.3284	5.0000	6.57%
$E[L]$	4.1993	4.0000	4.98%

Verdict: Good agreement (5-8% error typical for 45k samples).

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

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

Metric	Simulation	Analytical	Error %
$E[W_q]$	24.8663	19.0000	30.88%
$E[W]$	25.8722	20.0000	29.36%
$E[L]$	23.5130	19.0000	23.75%

Key Insight

The 30% error is **expected behavior**, not a bug. Near $\rho = 1$:

$$\mathrm{Var}[L] = \frac{\rho}{(1-\rho)^2} = \frac{0.95}{0.0025} = 380 \tag{26}$$

High variance means slow convergence. Would need 500k+ samples for tight estimates.

6.3 Test 3: M/M/3 Queue

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

Metric	Simulation	Analytical	Error %
$E[W_q]$	0.4774	0.4444	7.43%
$E[W]$	1.4834	1.4444	2.69%
$E[L]$	2.9526	2.8889	2.21%
$P(\text{wait})$	0.4575	0.4444	2.95%

Verdict: Excellent agreement (2-7% error).

6.4 Test 4: M/M/5 Queue

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

Metric	Simulation	Analytical	Error %
$E[W_q]$	0.6033	0.5541	8.87%
$E[W]$	1.6092	1.5541	3.54%
$E[L]$	6.4055	6.2165	3.04%
$P(\text{wait})$	0.5750	0.5541	3.76%

Verdict: Good agreement (3-9% error).

7 Assumption Failure Analysis

7.1 M/M/1 with Bursty Arrivals

Base Parameters: $\lambda = 0.8$, $\mu = 1.0$, $\rho = 0.80$
All scenarios calibrated to same effective arrival rate.

Arrival Type	$E[W_q]$	$E[W]$	$E[L]$	vs M/M/1
M/M/1 Theory	4.0000	5.0000	4.0000	baseline
Poisson (sim)	4.3225	5.3284	4.1993	+6.6%
Batch ($\mu_B = 2$)	8.6223	9.6218	7.9088	**+92.4%**
Hawkes	7.9380	8.9384	6.9977	**+78.8%**

7.2 M/M/4 with Bursty Arrivals

Parameters: $\lambda = 3.2$, $\mu = 1.0$, $k = 4$, $\rho = 0.80$

Arrival Type	$E[W_q]$	$E[W]$	$P(\text{wait})$	vs Theory
M/M/k Theory	0.7455	1.7455	0.5964	baseline
Poisson (sim)	0.8123	1.8182	0.6158	+4.2%
Batch ($\mu_B = 2$)	1.7725	2.7720	0.7524	**+58.8%**
Hawkes	1.5077	2.5050	0.6752	**+43.5%**

7.3 Key Comparison: M/M/1 vs M/M/k Under Burstiness

Both systems at $\rho = 0.8$ utilization

System	Batch Penalty	Hawkes Penalty
M/M/1	+92.4%	+78.8%
M/M/4	+58.8%	+43.5%

Key Insight

Multiple servers reduce **both** absolute wait times **and** the relative penalty from burstiness. This is because k servers can absorb short bursts before queue builds — a burst of size $B < k$ can be served immediately if servers are available.

Part IV
Conclusions and Implications

8 Summary of Results

8.1 Theoretical Contributions

1. **M/M/1 derivation:** Balance equations → geometric steady-state → Little's Law
2. **M/M/k derivation:** State-dependent service rates → Erlang-C formula
3. **M/G/1:** Pollaczek-Khinchin shows variance multiplier effect
4. **Batch arrivals:** Internal delay formula, heavy-traffic ratio $E[B^2]/(2E[B])$
5. **Hawkes process:** Mean rate $\lambda_0/(1-\alpha/\beta)$, overdispersion formula

8.2 Simulation Validation

Test Case	Utilization	Error Range	Verdict
M/M/1 ($\rho = 0.8$)	80%	5-8%	✓ Good
M/M/1 ($\rho = 0.95$)	95%	24-31%	✓ Expected (high variance)
M/M/3 ($\rho = 0.67$)	67%	2-7%	✓ Excellent
M/M/5 ($\rho = 0.8$)	80%	3-9%	✓ Good

8.3 Assumption Failure Findings

1. **Batch arrivals** increase wait time by ∼90% vs Poisson prediction (M/M/1)
2. **Hawkes arrivals** increase wait time by ∼80% vs Poisson prediction (M/M/1)
3. **Multiple servers** reduce the burstiness penalty from ∼90% to ∼60%
4. **M/M/1 formulas underestimate** congestion when arrivals are bursty

9 The Variance Principle

Theorem 1 (Unifying Principle)**.** *For fixed mean arrival rate λ and mean service rate μ:*

$$E[W_q] \propto Var[arrivals] + Var[service] \tag{27}$$

Higher variance in either process increases congestion, even at constant utilization.

10 Practical Implications

Domain	**Arrival Pattern**	**Implication**
Restaurant operations	Batch (families, groups)	M/M/k underestimates wait by 60-90%
Trading order flow	Hawkes (self-exciting)	Congestion clusters; queue position has information content
Call centers	Time-varying, bursty	Staffing models must account for arrival variance
Network traffic	Packet bursts	Poisson models inadequate for QoS guarantees

11 When Models Break: Summary

Assumption	**Violation**	**Effect on** $E[W]$
Poisson arrivals	Batch/Hawkes	Underestimates by 60-90%
Exponential service	Heavy-tailed	Increases with C_s^2
$\rho < 1$	$\rho \to 1$	Variance explodes
Steady state	Time-varying λ	Transient analysis needed
Infinite queue	Finite capacity	Some customers lost