Queueing Theory: Complete Derivations

From First Principles to Assumption Failures

Quantitative Finance Project

Contents

1 Baseline: M/M/1 Waiting Time

1.1 Assumptions

- **Arrivals:** Poisson process with rate λ
- **Service times:** Exponential with rate μ (mean service time $E[S] = 1/\mu$)
- **Utilization:** $\rho = \lambda/\mu < 1$ (stability condition)
- **Queue discipline:** First-Come-First-Served (FCFS)
- **Capacity:** Infinite queue

1.2 Derivation of Steady-State Distribution

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

Balance equations: In steady state, flow into state n = flow out of state n.

For $n = 0$:

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

For $n \geq 1$:

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

Solution by substitution: From the $n = 0$ equation:

$$\pi_1 = \frac{\lambda}{\mu}\pi_0 = \rho\pi_0 \tag{3}$$

Guess the pattern $\pi_n = \rho^n\pi_0$ and verify in the balance equation:

$$\lambda\rho^{n-1}\pi_0 + \mu\rho^{n+1}\pi_0 = (\lambda + \mu)\rho^n\pi_0 \tag{4}$$

$$\lambda + \mu\rho^2 = (\lambda + \mu)\rho \tag{5}$$

$$\lambda + \mu \cdot \frac{\lambda^2}{\mu^2} = \lambda + \mu \cdot \frac{\lambda}{\mu} \quad \checkmark \tag{6}$$

Normalization:

$$\sum_{n=0}^{\infty} \pi_n = \pi_0 \sum_{n=0}^{\infty} \rho^n = \frac{\pi_0}{1 - \rho} = 1 \tag{7}$$

This converges if and only if $\rho < 1$, giving:

$$\boxed{\pi_0 = 1 - \rho, \quad \pi_n = (1 - \rho)\rho^n} \tag{8}$$

1.3 Performance Metrics

Expected number in system:

$$E[L] = \sum_{n=0}^{\infty} n\pi_n = (1 - \rho)\sum_{n=0}^{\infty} n\rho^n \tag{9}$$

Using the identity $\sum_{n=0}^{\infty} nx^n = \frac{x}{(1-x)^2}$ (derived by differentiating the geometric series):

$$\boxed{E[L] = \frac{\rho}{1 - \rho}} \tag{10}$$

Expected time in system (Little's Law):

$$L = \lambda W \implies \boxed{E[W] = \frac{E[L]}{\lambda} = \frac{1}{\mu - \lambda}} \tag{11}$$

Expected wait in queue:

$$E[W] = E[W_q] + E[S] \implies \boxed{E[W_q] = \frac{1}{\mu - \lambda} - \frac{1}{\mu} = \frac{\rho}{\mu(1-\rho)} = \frac{\rho}{\mu - \lambda}} \tag{12}$$

2 General Single Server Queue (M/G/1)

2.1 Pollaczek-Khinchin Formula

For M/G/1 (Poisson arrivals, *general* service distribution):

$$\boxed{E[W_q] = \frac{\lambda E[S^2]}{2(1-\rho)}} \tag{13}$$

2.2 Derivation

The key insight is that $E[S^2]$ can be decomposed:

$$E[S^2] = \mathrm{Var}[S] + (E[S])^2 = E[S]^2(1 + C_s^2) \tag{14}$$

where $C_s^2 = \frac{\mathrm{Var}[S]}{E[S]^2}$ is the **squared coefficient of variation**.
Substituting $E[S] = 1/\mu$ and $\rho = \lambda/\mu$:

$$E[W_q] = \frac{\lambda \cdot \frac{1}{\mu^2}(1 + C_s^2)}{2(1-\rho)} \tag{15}$$

$$= \frac{\rho}{\mu(1-\rho)} \cdot \frac{1 + C_s^2}{2} \tag{16}$$

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

2.3 Interpretation: The Variance Multiplier

Remark 1. *For exponential service ($C_s^2 = 1$), this reduces to M/M/1.*
For deterministic service ($C_s^2 = 0$), wait time is ***halved****:*

$$E[W_q]_{M/D/1} = \frac{1}{2} E[W_q]_{M/M/1} \tag{18}$$

For heavy-tailed service ($C_s^2 > 1$), wait time ***increases****.*

Distribution	C_s^2	**Wait vs M/M/1**
Deterministic	0	0.5×
Exponential	1	1× (baseline)
Hyperexponential	> 1	$> 1\times$
Pareto ($\alpha = 2.5$)	~ 2.7	$\sim 1.85\times$

Key insight: If $C_s^2 \uparrow$ then $E[W_q] \uparrow$ even when ρ is fixed. Variance matters.

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

3.1 Setup

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

3.2 Internal Waiting Within a Batch

When a batch of size B arrives, the i-th customer in the batch must wait for customers $1, 2, \ldots, i-1$ to be served (in addition to any queue delay).

Expected service time per customer: $E[S] = 1/\mu$

Wait for customer i due to batch-mates:

$$(i-1) \cdot E[S] = \frac{i-1}{\mu} \tag{19}$$

Total internal waiting (summed over all customers in batch):

$$\sum_{i=1}^{B}(i-1)E[S] = E[S]\sum_{i=1}^{B}(i-1) \tag{20}$$

$$= E[S] \cdot \frac{B(B-1)}{2} \tag{21}$$

$$= \frac{B(B-1)}{2\mu} \tag{22}$$

3.3 Expected Internal Delay per Customer

Taking expectations over batch size B:

$$E[\text{internal delay}] = \frac{E[B(B-1)]}{2\mu} = \frac{E[B^2] - E[B]}{2\mu} \tag{23}$$

3.4 Total Expected Wait

The total wait combines the M/M/1 queue wait plus internal batch delay:

$$\boxed{E[W_q]_{\text{batch}} = E[W_q]_{M/M/1} + \frac{E[B^2] - E[B]}{2\mu}} \tag{24}$$

3.5 Ratio to Poisson (Key Result)

$$\frac{E[W_q]_{\text{batch}}}{E[W_q]_{\text{Poisson}}} = 1 + \frac{E[B^2] - E[B]}{2\mu \cdot E[W_q]_{M/M/1}} \tag{25}$$

$$= 1 + \frac{(E[B^2] - E[B])(1-\rho)}{2\rho} \tag{26}$$

In heavy traffic ($\rho \to 1$), this simplifies to:

$$\boxed{\frac{E[W_q]_{\text{batch}}}{E[W_q]_{\text{Poisson}}} \approx \frac{E[B^2]}{2E[B]}} \tag{27}$$

3.6 Example: Geometric Batch Size

For geometric distribution with mean $E[B] = m$:

- $\text{Var}[B] = m(m-1)$ (for geometric starting at 1)
- $E[B^2] = \text{Var}[B] + E[B]^2 = m(m-1) + m^2 = 2m^2 - m$

Ratio:

$$\frac{E[B^2]}{2E[B]} = \frac{2m^2 - m}{2m} = m - \frac{1}{2} \tag{28}$$

For $m = 2$: ratio ≈ 1.5, so batch arrivals increase wait by $\sim 50\%$.

Remark 2. *Our simulation showed* $\sim 92\%$ *increase. The discrepancy arises because:*

1. *The heavy-traffic approximation isn't exact at* $\rho = 0.8$
2. *Geometric distribution has higher variance than assumed*

4 Hawkes Process Arrivals

4.1 Definition

A Hawkes process is a **self-exciting** point process with intensity:

$$\boxed{\lambda(t) = \lambda_0 + \sum_{t_i < t} \alpha e^{-\beta(t - t_i)}} \tag{29}$$

- λ_0 = baseline arrival rate
- α = jump size (temporary intensity boost per arrival)
- β = decay rate (exponential decay of excitation)

4.2 Mean Arrival Rate

Each arrival contributes expected future arrivals:

$$\int_0^\infty \alpha e^{-\beta s}\, ds = \frac{\alpha}{\beta} \tag{30}$$

This is the **branching ratio**. For stability, we need $\alpha/\beta < 1$.
The total rate satisfies:

$$\bar{\lambda} = \lambda_0 + \bar{\lambda} \cdot \frac{\alpha}{\beta} \tag{31}$$

Solving:

$$\boxed{\bar{\lambda} = \frac{\lambda_0}{1 - \alpha/\beta}} \tag{32}$$

4.3 Variance of Arrival Count

For Poisson, $\text{Var}[N(T)] = \bar{\lambda}T$.
For Hawkes:

$$\boxed{\text{Var}[N(T)] = \bar{\lambda}T \cdot \frac{1}{(1 - \alpha/\beta)^2} > \bar{\lambda}T} \tag{33}$$

4.4 Interpretation

The variance formula decomposes as:

$$\text{Var}[N(T)] \approx (\text{mean arrivals}) \times (\text{mean cluster size})^2 \tag{34}$$

where mean cluster size $= 1/(1 - \alpha/\beta)$.
Why arrivals are overdispersed:
For Poisson: $\text{Cov}[X_i, X_j] = 0$ (arrivals independent)
For Hawkes: $\text{Cov}[X_i, X_j] > 0$ (arrivals *cause* more arrivals)
Using $\text{Var}[\sum X_i] = \sum \text{Var}[X_i] + 2\sum_{i<j} \text{Cov}[X_i, X_j]$:

$$\text{Var}[N(T)]_{\text{Hawkes}} > \text{Var}[N(T)]_{\text{Poisson}} \tag{35}$$

4.5 Stability Condition

$$\frac{\alpha}{\beta} < 1 \implies \text{arrivals die down (stable)} \tag{36}$$

$$\frac{\alpha}{\beta} \geq 1 \implies \text{explosion (unstable)} \tag{37}$$

4.6 Impact on Queueing

Little's Law still holds: $L = \bar{\lambda} W$
But since Hawkes has:

- Long silent periods (low intensity)
- Violent bursts (high intensity after arrivals)

The queue builds up during bursts, increasing average wait:

$$\boxed{E[W]_{\text{Hawkes}} > E[W]_{\text{Poisson}} \quad \text{for the same } \bar{\lambda}} \tag{38}$$

There is no simple closed form — simulation is required.

5 Multi-Server Queue: M/M/k

5.1 Setup

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

5.2 Steady-State Distribution

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

$$\pi_n = \frac{(k\rho)^n}{n!}\pi_0 = \frac{a^n}{n!}\pi_0 \tag{39}$$

where $a = \lambda/\mu = k\rho$ is the offered load.
The $n!$ appears because service rate increases as $\mu, 2\mu, \ldots, n\mu$.
For $n \geq k$ (all servers busy):

$$\pi_n = \frac{a^k}{k!} \cdot \rho^{n-k}\pi_0 \tag{40}$$

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

5.3 Normalization

$$\pi_0 = \left[\sum_{n=0}^{k-1} \frac{a^n}{n!} + \frac{a^k}{k!(1-\rho)} \right]^{-1} \tag{41}$$

5.4 Erlang-C Formula

The probability of waiting (all servers busy at arrival):

$$\boxed{P(\text{wait}) = C(k,a) = \frac{\frac{a^k}{k!} \cdot \frac{1}{1-\rho}}{\sum_{n=0}^{k-1} \frac{a^n}{n!} + \frac{a^k}{k!} \cdot \frac{1}{1-\rho}}} \tag{42}$$

5.5 Performance Metrics

$$E[W_q] = \frac{P(\text{wait})}{k\mu - \lambda} \tag{43}$$

$$E[W] = E[W_q] + \frac{1}{\mu} \tag{44}$$

6 Batch Arrivals with M/M/k

6.1 Key Result

For batch arrivals in M/M/k:

$$E[W_q] \propto \frac{\text{arrival variance} + \text{service variance}}{2(1-\rho)} \tag{45}$$

The arrival variance is proportional to $E[B^2]$.

6.2 Heavy Traffic Ratio

$$\frac{E[W_q]_{\text{batch}}}{E[W_q]_{\text{Poisson}}} \to \frac{E[B^2]}{E[B]} \quad \text{as } \rho \to 1 \tag{46}$$

6.3 Why M/M/k Has Smaller Penalty Than M/M/1

For queueing to occur, we need $N \geq k$ (all servers busy).

$$P(\text{queue forms}) = P(N \geq k) \tag{47}$$

With multiple servers:

- A burst of size $B < k$ can be absorbed without queueing
- Only bursts exceeding available capacity create delays

Simulation results:

System	Batch Penalty	Hawkes Penalty
M/M/1 ($\rho = 0.8$)	+92%	+79%
M/M/4 ($\rho = 0.8$)	+59%	+44%

6.4 Key Insight

$$\boxed{E[W_q]^{\text{bursty}}_{M/M/k} < E[W_q]^{\text{bursty}}_{M/M/1}} \tag{48}$$

$E[W]$ is controlled by:

1. Variance of arrivals
2. The factor $1/(1-\rho)$

Not by λ alone.
Multiple servers reduce both absolute wait times and the relative penalty from burstiness.

7 Summary: When Models Break

7.1 Assumption Violations and Effects

Assumption	**Violation**	**Effect**
Poisson arrivals	Batch/Hawkes	Underestimates wait
Exponential service	Heavy-tailed	Underestimates wait
$\rho < 1$	$\rho \to 1$	Variance explodes
Infinite queue	Finite capacity	Overestimates wait
Steady state	Time-varying λ	Transient needed

7.2 The Variance Principle

Across all models, a unifying theme:

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

$$E[W_q] \propto Var[arrivals] + Var[service] \tag{49}$$

Higher variance in either process increases congestion.

7.3 Practical Implications

1. **Restaurant rushes:** Batch arrivals (families) increase wait beyond M/M/k prediction
2. **Trading order flow:** Hawkes-like self-excitation means congestion clusters
3. **Call centers:** Heavy-tailed service (complex queries) dominates simple metrics
4. **Staffing decisions:** M/M/k formulas are optimistic when arrivals are bursty