Queueing Theory I
Summary

- Little’s Law
- Queueing System Notation
- Stationary Analysis of Elementary Queueing Systems
  - M/M/1
  - M/M/m
  - M/M/1/K
  - ...

Little’s Law

- \( a(t) \): the process that counts the number of arrivals up to \( t \).
- \( d(t) \): the process that counts the number of departures up to \( t \).
- \( N(t) = a(t) - d(t) \)

- Average arrival rate (up to \( t \)) \( \lambda_t = \frac{a(t)}{t} \)
- Average time each customer spends in the system \( T_t = \frac{\gamma(t)}{a(t)} \)
- Average number in the system \( N_t = \frac{\gamma(t)}{t} \)
Little’s Law

\[ N_t = \lambda_t T_t \]

- Taking the limit as \( t \) goes to infinity

\[ E[N] = \lambda E[T] \]

- Expected number of customers in the system
- Expected time in the system
- Arrival rate \textbf{IN} the system
Generality of Little’s Law

\[ E[N] = \lambda E[T] \]

- Little’s Law is a pretty general result
- It does not depend on the arrival process distribution
- It does not depend on the service process distribution
- It does not depend on the number of servers and buffers in the system.
Specification of Queueing Systems

- Customer *arrival* and *service* stochastic models

- Structural Parameters
  - Number of servers
  - Storage capacity

- Operating policies
  - Customer class differentiation (are all customers treated the same or do some have priority over others?)
  - Scheduling/Queueing policies (which customer is served next)
  - Admission policies (which/when customers are admitted)
Queueing System Notation

Arrival Process
- **M**: Markovian
- **D**: Deterministic
- **Er**: Erlang
- **G**: General

Service Process
- **M**: Markovian
- **D**: Deterministic
- **Er**: Erlang
- **G**: General

A/B/m/K/N

- Number of servers $m = 1, 2, \ldots$
- Storage Capacity $K = 1, 2, \ldots$ (if $\infty$ then it is omitted)
- Number of customers $N = 1, 2, \ldots$ (for closed networks otherwise it is omitted)
Performance Measures of Interest

- We are interested in steady state behavior
  - Even though it is possible to pursue transient results, it is a significantly more difficult task.

- $E[S]$ average **system time** (average time spent in the system)
- $E[W]$ average **waiting time** (average time spent waiting in queue(s))
- $E[X]$ average **queue length**
- $E[U]$ average **utilization** (fraction of time that the resources are being used)
- $E[R]$ average **throughput** (rate that customers leave the system)
- $E[L]$ average customer **loss** (rate that customers are lost or probability that a customer is lost)
Recall the Birth-Death Chain Example

At steady state, we obtain

\[-\lambda_0 \pi_0 + \mu_1 \pi_1 = 0 \Rightarrow \pi_1 = \frac{\lambda_0}{\mu_1} \pi_0\]

In general

\[\lambda_{j-1} \pi_{j-1} - \left( \lambda_j + \mu_j \right) \pi_j + \mu_{j+1} \pi_{j+1} = 0 \Rightarrow \pi_{j+1} = \left( \frac{\lambda_0 \ldots \lambda_j}{\mu_1 \ldots \mu_{j+1}} \right) \pi_0\]

Making the sum equal to 1

\[\pi_0 \left( 1 + \sum_{j=1}^{\infty} \left( \frac{\lambda_0 \ldots \lambda_{j-1}}{\mu_1 \ldots \mu_j} \right) \right) = 1\]

Solution exists if

\[S = 1 + \sum_{j=1}^{\infty} \left( \frac{\lambda_0 \ldots \lambda_{j-1}}{\mu_1 \ldots \mu_j} \right) < \infty\]
M/M/1 Example

**Meaning:** Poisson Arrivals, exponentially distributed service times, one server and infinite capacity buffer.

Using the birth-death result $\lambda_j = \lambda$ and $\mu_j = \mu$, we obtain

$$
\pi_j = \left( \frac{\lambda}{\mu} \right)^j \pi_0, \quad j = 0, 1, 2, \ldots
$$

Therefore

$$
\pi_0 \left( 1 + \sum_{j=1}^{\infty} \left( \frac{\lambda}{\mu} \right)^j \right) = 1 \quad \text{for } \lambda/\mu = \rho < 1
$$

$$
\pi_0 = 1 - \rho
$$

$$
\pi_j = (1 - \rho) \rho^j, \quad j = 1, 2, \ldots
$$
M/M/1 Performance Metrics

- **Server Utilization**
  \[ E[U] = \sum_{j=1}^{\infty} \pi_j = 1 - \pi_0 = 1 - (1 - \rho) = \rho \]

- **Throughput**
  \[ E[R] = \mu \sum_{j=1}^{\infty} \pi_j = \mu (1 - \pi_0) = \mu \rho = \lambda \]

- **Expected Queue Length**
  \[ E[X] = \sum_{j=0}^{\infty} j \pi_j = (1 - \rho) \sum_{j=0}^{\infty} j \rho^j = \rho (1 - \rho) \sum_{j=0}^{\infty} \frac{d\{\rho^j\}}{d\rho} = \]
  \[ = \rho (1 - \rho) \frac{d}{d\rho} \left\{ \sum_{j=0}^{\infty} \rho^j \right\} = \rho (1 - \rho) \frac{d}{d\rho} \left\{ \frac{1}{(1 - \rho)} \right\} = \frac{\rho}{(1 - \rho)} \]
M/M/1 Performance Metrics

- **Average System Time**

\[
E[X] = \lambda E[S] \Rightarrow E[S] = \frac{1}{\lambda} E[X]
\]

\[
E[S] = \frac{1}{\lambda} \frac{\rho}{1 - \rho} = \frac{1}{\mu(1 - \rho)}
\]

- **Average waiting time in queue**

\[
\]

\[
E[W] = \frac{1}{\mu(1 - \rho)} - \frac{1}{\mu} = \frac{\rho}{\mu(1 - \rho)}
\]
M/M/1 Performance Metrics Examples

$\mu = 0.5$

![Graph showing performance metrics $E[S]$, $E[W]$, and $E[X]$ as functions of $\rho$. The x-axis represents $\rho$ and the y-axis represents delay (time units) / number of customers. The graph illustrates how these metrics change with varying $\rho$.](image)
PASTA Property

- PASTA: Poisson Arrivals See Time Averages
- Let $\pi_j(t) = \Pr\{\text{System state } X(t) = j\}$
- Let $a_j(t) = \Pr\{\text{Arriving customer at } t \text{ finds } X(t) = j\}$
- In general $\pi_j(t) \neq a_j(t)$!
  - Suppose a D/D/1 system with interarrival times equal to 1 and service times equal to 0.5

Thus $\pi_0(t) = 0.5$ and $\pi_1(t) = 0.5$ while $a_0(t) = 1$ and $a_1(t) = 0$!
Theorem

For a queueing system, when the arrival process is Poisson and independent of the service process then, the probability that an arriving customer finds $j$ customers in the system is equal to the probability that the system is at state $j$. In other words,

$$a_j(t) = \pi_j(t) = \Pr\{X(t) = j\}, \quad j = 0, 1, \ldots$$

Proof:

$$a_j(t) \equiv \lim_{\Delta t \to 0} \Pr\{X(t) = j \mid a(t, t + \Delta t)\}$$

$$= \lim_{\Delta t \to 0} \frac{\Pr\{X(t) = j, a(t, t + \Delta t)\}}{\Pr\{a(t, t + \Delta t)\}}$$

$$= \lim_{\Delta t \to 0} \frac{\Pr\{X(t) = j\} \Pr\{a(t, t + \Delta t)\}}{\Pr\{a(t, t + \Delta t)\}} = \Pr\{X(t) = j\} = \pi_j(t)$$
**M/M/m Queueing System**

**Meaning:** Poisson Arrivals, exponentially distributed service times, *m* identical servers and infinite capacity buffer.

\[
\lambda_j = \lambda \quad \text{and} \quad \mu_j = \begin{cases} 
  j\mu & \text{if } 0 \leq j \leq m \\
  m\mu & \text{if } j \geq m 
\end{cases}
\]
M/M/m Queueing System

- Using the general birth-death result

\[
\pi_j = \frac{1}{j!} \left( \frac{\lambda}{\mu} \right)^j \pi_0, \quad \text{if } j < m \\
\pi_j = \frac{m^m}{m!} \left( \frac{\lambda}{m \mu} \right)^j \pi_0, \quad \text{if } j \geq m
\]

- Letting \( \rho = \frac{\lambda}{m \mu} \) we get

\[
\pi_j = \begin{cases} 
\frac{(m \rho)^j}{j!} \pi_0 & \text{if } j < m \\
\frac{m^m \rho^j}{m!} \pi_0 & \text{if } j \geq m
\end{cases}
\]

- To find \( \pi_0 \)

\[
\pi_0 \left( 1 + \sum_{j=1}^{m-1} \frac{(m \rho)^j}{j!} + \sum_{j=m}^{\infty} \frac{m^m \rho^j}{m!} \right) = 1 \Rightarrow \pi_0 = \left( 1 + \sum_{j=1}^{m-1} \frac{(m \rho)^j}{j!} + \frac{(m \rho)^m}{m! (1 - \rho)} \right)^{-1}
\]
M/M/m Performance Metrics

Server Utilization

$$E[U] = \sum_{j=1}^{m-1} j \pi_j + m \Pr \{ X \geq m \} = \pi_0 \left( \sum_{j=1}^{m-1} \frac{(m \rho)^j}{j!} + m \sum_{j=m}^{\infty} \frac{m^j \rho^j}{m!} \right)$$

$$= \pi_0 \left( (m \rho) + \sum_{j=2}^{m-1} \frac{(m \rho)^j}{(j-1)!} + m \frac{(m \rho)^m}{m!(1-\rho)} \right)$$

$$= \pi_0 m \rho \left( 1 + \sum_{j=2}^{m-1} \frac{(m \rho)^{j-1}}{(j-1)!} + \frac{(m \rho)^{m-1}}{(m-1)!} - \frac{(m \rho)^{m-1}}{(m-1)!} + \frac{m(m \rho)^{m-1}}{m!(1-\rho)} \right)$$

$$= \pi_0 m \rho \left( 1 + \sum_{j=1}^{m-1} \frac{(m \rho)^j}{j!} + \frac{(m \rho)^m}{m!(1-\rho)} \right)$$

$$= \pi_0 m \rho \frac{1}{\pi_0} = m \rho = \frac{\lambda}{\mu}$$
M/M/m Performance Metrics

- **Throughput**
  \[ E[R] = \mu \sum_{j=1}^{m-1} j \pi_j + m \mu \sum_{j=m}^{\infty} \pi_j = \lambda \]

- **Expected Queue Length**
  \[ E[X] = \sum_{j=0}^{\infty} j \pi_j = \pi_0 \left( \sum_{j=1}^{m-1} j \frac{(m\rho)^j}{j!} + \frac{m^m}{m!} \sum_{j=m}^{\infty} j \rho^j \right) = ... \]
  \[ E[X] = m\rho + \frac{(m\rho)^m}{m!} \frac{\rho}{(1-\rho)^2} \pi_0 \]

- **Using Little’s Law**
  \[ E[S] = \frac{1}{\lambda} E[X] = \frac{1}{\lambda} \left( m\rho + \frac{(m\rho)^m}{m!} \frac{\rho}{(1-\rho)^2} \pi_0 \right) \]

- **Average Waiting time in queue**
  \[ E[W] = E[S] - \frac{1}{\mu} \]
M/M/m Performance Metrics

Queueing Probability

\[ P_Q = \Pr \{ X \geq m \} = \sum_{j=m}^{\infty} \pi_j = \pi_0 \sum_{j=m}^{\infty} \frac{m^m \rho^j}{m!} = \frac{\pi_0 (m\rho)^m}{m!(1-\rho)} \]

Erlang C Formula
Example

Suppose that customers arrive according to a Poisson process with rate $\lambda = 1$. You are given the following two options,

- Install a single server with processing capacity $\mu_1 = 1.5$
- Install two identical servers with processing capacity $\mu_2 = 0.75$ and $\mu_3 = 0.75$
- Split the incoming traffic to two queues each with probability 0.5 and have $\mu_2 = 0.75$ and $\mu_3 = 0.75$ serve each queue.
Example

- Throughput
  - It is easy to see that all three systems have the same throughput $E[R_A] = E[R_B] = E[R_C] = \lambda$

- Server Utilization

  $E[U_A] = \frac{\lambda}{\mu_1} = \frac{1}{1.5} = \frac{2}{3}$
  $E[U_B] = \frac{\lambda}{\mu_2} = \frac{1}{0.75} = \frac{4}{3}$  Therefore, each server is $2/3$ utilized
  $E[U_C] = \frac{0.5\lambda}{\mu_2} = \frac{1}{2 \times 0.75} = \frac{2}{3}$

- Therefore, all servers are similarly loaded.
Example

- Probability of being idle

\[ \pi_{0A} = 1 - \frac{\lambda}{\mu_1} = \frac{1}{3} \]

\[ \pi_{0B} = \left( 1 + \sum_{j=1}^{m-1} \frac{(m\rho)^j}{j!} + \frac{(m\rho)^m}{m!(1-\rho)} \right)^{-1} = \left( 1 + \frac{4}{3} + \frac{\left( \frac{4}{3} \right)^2}{2 \left( 1 - \frac{2}{3} \right)} \right)^{-1} = \frac{1}{5} \]

\[ \pi_{0C} = 1 - \frac{\lambda}{2\mu_2} = \frac{1}{3} \quad \text{For each server} \]
Example

- Queue length and delay

\[
E[X_A] = \frac{\lambda}{\mu_1 - \lambda} = \frac{1}{1.5 - 1} = 2
\]

\[
E[S_A] = \frac{1}{\lambda} E[X_A] = 2
\]

\[
E[X_B] = m\rho + \frac{(m\rho)^m}{m!} \frac{\rho}{(1-\rho)^2} \pi_0 = \frac{12}{5}
\]

\[
E[S_B] = \frac{1}{\lambda} E[X_B] = \frac{12}{5}
\]

\[
E[X_{1c}] = \frac{\lambda / 2}{\mu_2 - \lambda / 2} = \frac{0.5}{0.75 - 0.5} = 2
\]

\[
\Rightarrow E[X_C] = 2 \times E[X_{1c}] = 4
\]

\[
E[X_C] = \frac{1}{\lambda} E[X_C] = 4
\]

For each queue!
M/M/∞ Queueing System

- Special case of the M/M/m system with m going to ∞

\[ \lambda_j = \lambda \quad \text{and} \quad \mu_j = j\mu \quad \text{for all} \quad j \]

- Let \( \rho = \lambda / \mu \) then, the state probabilities are given by

\[
\pi_j = \frac{\rho^j}{j!} \pi_0 \quad \pi_0 \left(1 + \sum_{j=1}^{\infty} \frac{\rho^j}{j!}\right) = 1 \Rightarrow \pi_0 = e^{-\rho} \quad \Rightarrow \pi_j = \frac{\rho^j e^{-\rho}}{j!}
\]

- System Utilization and Throughput

\[
E[U] = 1 - \pi_0 = 1 - e^{-\rho} \quad \quad \quad E[R] = \lambda
\]
M/M/$\infty$ Performance Metrics

- **Expected Number in the System**
  \[
  E[X] = \sum_{j=0}^{\infty} j \pi_j = \sum_{j=0}^{\infty} j \frac{\rho^j}{j!} e^{-\rho} = \rho e^{-\rho} \sum_{j=1}^{\infty} \frac{\rho^{j-1}}{(j-1)!} = \rho
  \]

  Number of busy servers

- **Using Little’s Law**
  \[
  E[S] = \frac{1}{\lambda} E[X] = \frac{1}{\lambda} \frac{\lambda}{\mu} = \frac{1}{\mu}
  \]

  No queueing!
M/M/1/K – Finite Buffer Capacity

**Meaning:** Poisson Arrivals, exponentially distributed service times, one server and finite capacity buffer $K$.

Using the birth-death result $\lambda_j = \lambda$ and $\mu_j = \mu$, we obtain

$$\pi_j = \left(\frac{\lambda}{\mu}\right)^j \pi_0, \quad j = 0, 1, 2, \ldots K$$

Therefore

$$\pi_0 \left(1 + \sum_{j=1}^{K} \left(\frac{\lambda}{\mu}\right)^j\right) = 1 \quad \text{for } \frac{\lambda}{\mu} = \rho$$

$$\pi_0 = \frac{1 - \rho}{1 - \rho^{K+1}}$$

$$\pi_j = \frac{(1 - \rho) \rho^j}{1 - \rho^{K+1}}, \quad j = 1, 2, \ldots K$$
M/M/1/K Performance Metrics

- **Server Utilization**
  
  \[ E[U] = 1 - \pi_0 = 1 - \frac{(1 - \rho)}{1 - \rho^{K+1}} = \frac{\rho(1 - \rho^K)}{1 - \rho^{K+1}} \]

- **Throughput**
  
  \[ E[R] = \mu (1 - \pi_0) = \lambda \frac{1 - \rho^K}{1 - \rho^{K+1}} < \lambda \]

- **Blocking Probability**
  
  \[ P_B = \pi_K = \frac{(1 - \rho) \rho^K}{1 - \rho^{K+1}} \]

Probability that an arriving customer finds the queue full (at state \( K \))
M/M/1/K Performance Metrics

- Expected Queue Length

\[ E[X] = \sum_{j=0}^{K} j \pi_j = \frac{(1-\rho)}{1-\rho^{K+1}} \sum_{j=0}^{K} j \rho^j = \frac{(1-\rho)\rho}{1-\rho^{K+1}} \sum_{j=0}^{K} \frac{d\{\rho^j\}}{d\rho} = \]

\[ = \frac{(1-\rho)\rho}{1-\rho^{K+1}} \frac{d}{d\rho} \left\{ \sum_{j=0}^{K} \rho^j \right\} = \frac{(1-\rho)\rho}{1-\rho^{K+1}} \frac{d}{d\rho} \left\{ \frac{1-\rho^{K+1}}{(1-\rho)} \right\} = \]

\[ = \frac{(1-\rho)\rho}{1-\rho^{K+1}} \left( \frac{(1-\rho^{K+1})-(1-\rho)(K+1)\rho^K}{(1-\rho)^2} \right) = \]

\[ = \frac{\rho}{1-\rho^{K+1}} \left( \frac{1-\rho^K}{1-\rho} - K\rho^K \right) \]

- System time

\[ E[X] = \lambda (1-\pi_K) E[S] \]
M/M/m/m – Queueing System

- **Meaning**: Poisson Arrivals, exponentially distributed service times, $m$ servers and no storage capacity.

- Using the birth-death result $\lambda_j=\lambda$ and $\mu_j=\mu$, we obtain
  \[
  \pi_j = \frac{1}{j!} \left( \frac{\lambda}{\mu} \right)^j \pi_0, \quad j = 0, 1, 2, \ldots m
  \]

- Therefore
  \[
  \pi_0 \left( \sum_{j=0}^{m} \frac{1}{j!} \left( \frac{\lambda}{\mu} \right)^j \right) = 1 \quad \text{for } \lambda/\mu = \rho
  \]
  \[
  \pi_0 = \left( \sum_{j=0}^{m} \frac{\rho^j}{j!} \right)^{-1}
  \]
  \[
  \pi_j = \frac{\rho^j}{j!} \pi_0, \quad j = 1, 2, \ldots m
  \]
M/M/m/m Performance Metrics

- **Blocking Probability**

  \[ P_B = \pi_m = \frac{\rho^m / m!}{\sum_{j=0}^{m} \frac{\rho^j}{j!}} \]

  Erlang B Formula

  Probability that an arriving customer finds all servers busy (at state \( m \))

- **Throughput**

  \[ E[R] = \lambda \left( 1 - \pi_m \right) = \lambda \left( 1 - \frac{\rho^m / m!}{\sum_{j=0}^{m} \frac{\rho^j}{j!}} \right) < \lambda \]
M/M/1//N – Closed Queueing System

**Meaning:** Poisson Arrivals, exponentially distributed service times, one server and the number of customers are fixed to $N$.

Using the birth-death result, we obtain

$$\pi_j = \frac{N!}{(N-j)!} \rho^j \pi_0, \quad j = 1, 2, \ldots, N$$

$$\pi_0 = \left[ \sum_{j=0}^{N} \frac{N!}{(N-j)!} \rho^j \right]^{-1}$$
M/M/1//N – Closed Queueing System

- **Response Time**
  - Time from the moment the customer entered the queue until it received service.
  - For the queue, using Little’s law we get,
    \[ E[X] = \mu (1 - \pi_0) E[S] \]

- In the “thinking” part,
  \[ E[N - X] = \mu (1 - \pi_0) \frac{1}{\lambda} \]

- Therefore
  \[ E[S] = \frac{N - \mu (1 - \pi_0) \frac{1}{\lambda}}{\mu (1 - \pi_0)} = \frac{N}{\mu (1 - \pi_0)} - \frac{1}{\lambda} \]