Queueing (Waiting Line) Theory

Queue or Waiting Line: An Example at a check-out counter

Queueing Examples In Real Life

Why do we need to study queueing and queueing theory?
- From a customer’s prospective:
  - Line is too long
  - Perceived time to be served is too long
  - Someone cut line in front of you
- From a business prospective:
  - Multiple tasking, Productivity, Fairness
  - Put queueing theory in use
  - Combine intuition, common sense & queueing theory

A 30 minutes observation at a 911 call center (starts at 8:00am at time 0):

<table>
<thead>
<tr>
<th>id</th>
<th>Time</th>
<th>Server busy min</th>
<th>Server idle min</th>
<th>no in Queue</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>12</td>
<td>2</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>17</td>
<td>3</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>18</td>
<td>8</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>26</td>
<td>8</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>28</td>
<td>8</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>30</td>
<td>5</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Characteristics of Queuing

- Arrival process: \( \lambda \) mean arrival rate per time unit (hour) versus mean inter-arrival time \( 1/\lambda \). If 160 customers arrive for service at a bar in an eight hour day,
  - \( \lambda \) Arrival rate per hour?
- $\frac{1}{\lambda}$ inter-arrival time (hour)?

Please note: the mean arrival rate $\lambda$ and the inter-arrival time $1/\lambda$ should initially have the same time units, an hour, for example.

- $\lambda$ Arrival rate per 15 minutes?

- $\frac{1}{\lambda}$ inter-arrival time?

• Service process: $(\mu)$ mean service rate per time unit (hour) versus mean service time $(1/\mu)$. If the bar can serve 240 customers in an eight hour day,

  - $\mu$ Service rate per hour?

  - $\frac{1}{\mu}$ Mean service time (hour)?

Please note: the mean service rate $\mu$ and the mean service time $1/\mu$ should initially have the same time units, an hour, for example.

• The arrival rate $\lambda$ and the service rate $\mu$:

  $\begin{array}{ccccccc}
  a_9 & a_8 & a_7 & a_6 & a_5 & a_4 & a_3 & a_2 & a_1 & \text{Server}
  \end{array}$

  - $\lambda < \mu$ in steady state
  - $\lambda \approx \mu$ or $\lambda$ is close to $\mu$, what would happen if the server has to wait for an arrival to come? A server can’t get any lost time back.

• Number of servers: single versus multiple

• Number of queue positions: infinite versus finite

• Source population: infinite versus finite (machine repairs)
• Queueing rules (disciplines / job sequencing):
  o FIFS or First In and First Serve: Bank teller
  o LIFS or Last In and First Serve: Elevator
  o Shortest Processing Time: Express lane at Grocery store
  o Earliest Due Date: Order processing to reduce customer complaints
  o Balking, reneging, jockeying queue positions: Special permit / ticket

What are Operating Characteristics of Queuing Theory?

$\lambda$ = mean arrival rate (mean number of arrivals per time unit)

$1/\lambda$ = mean inter-arrival time for arrivals

$\mu$ = mean service rate (mean number of services per time unit)

$1/\mu$ = mean service time per customer or job

$L_q$ = average queue length or number of units in line waiting for service

$W_q$ = average waiting time a unit spent in queue before being served

$L_q = \lambda W_q$

 ✓ The average queue length is the arrival rate multiplies by the average time spent waiting in the queue.

 ✓ Jobs blocked and refused entry to the system are not counted in $\lambda$.

$L = $ average number of units in the system ($L_q$ in queue plus being served)

$W = $ average time a unit spent in the system (in queue plus being served)

$L = \lambda W$

 ✓ The average queue length plus the one being served is the arrival rate multiplies by the average time spent waiting in the queue plus the time being served.

 ✓ Jobs blocked and refused entry to the system are not counted in $\lambda$.

$s$ = number of parallel or equivalent servers in the system

$\rho$ (Rho) or $U$ = server utilization factor = the proportion of time the server is busy

$P_w$ = Probability of an arriving unit to wait in the queue before being served

$P_0$ = Probability of no unit in the system (empty) (neither in queue nor being served)

$P_n$ = Probability of having $n$ units in the system (in queue plus being served)
Operating Characteristics of Basic Single-Server M/M/1 Queueing Model with FCFS, Infinite queue and source:

\[\lambda \quad \frac{1}{\lambda} \quad \mu \quad \frac{1}{\mu} \quad \rho = U = P_w = \frac{\lambda}{\mu} = P(n \geq 1) = 1 - P_0\]

\[P_0 = 1 - \frac{\lambda}{\mu} = 1 - \rho \quad P_n = P_0 \left[\frac{\lambda}{\mu}\right]^n = P_0 \rho^n \quad W = \frac{1}{\mu - \lambda} = W_q + \frac{1}{\mu} = \frac{\lambda}{\mu(1-\rho)}\]

\[L = \lambda W = \frac{\lambda}{\mu - \lambda} = L_q + \frac{\lambda}{\mu} = \frac{\rho}{1-\rho} \quad W_q = \frac{\lambda}{\mu(\mu - \lambda)} = W - \frac{1}{\mu} = \frac{l_q}{\lambda} = \frac{\rho}{\mu(1-\rho)} \quad L_q = \lambda W_q = \frac{\lambda^2}{\mu(\mu - \lambda)} = \frac{\rho^2}{1-\rho}\]

Operating Characteristics of Basic Double-Server M/M/2 Queueing Model with FCFS, Infinite queue and source:

\[\lambda \quad \frac{1}{\lambda} \quad \mu \quad \frac{1}{\mu} \quad \rho = U = P_w = \frac{\lambda}{2\mu} = P(n \geq 1) = 1 - P_0\]

\[P_n = \begin{cases} \left(\frac{\lambda/\mu}{n!}\right) P_0, & f\ for\ n \leq 2 \\ \left(\frac{(\lambda/\mu)^n}{2(n-1)!}\right) P_0, & f\ for\ n > 2 \end{cases} \]

\[L_q = \lambda W_q = 2\rho P_0 \left(\frac{\rho}{1-\rho}\right)^2 = \frac{\rho(\lambda/\mu)^3}{(2-\lambda/\mu)^2} \quad W_q = W - \frac{1}{\mu} = \frac{l_q}{\lambda} \quad L = \lambda W = L_q + \frac{\lambda}{\mu} = L_q + 2\rho \]

\[W = \frac{\lambda}{\mu} = W_q + \frac{1}{\mu} \quad P_w = 2\rho^2 \left(\frac{1}{1-\rho}\right) P_0 = \left(\frac{\lambda^2}{\mu(2\mu - \lambda)}\right) P_0\]

Operating Characteristics of Basic Multiple-Server M/M/s Queueing Model with FCFS, Infinite queue and source:

\[\lambda \quad \frac{1}{\lambda} \quad \mu \quad \frac{1}{s\mu} \quad \rho = U = \frac{\lambda}{s\mu} \quad \rho = U = \frac{\lambda}{s\mu}\]

\[P_0 = \sum_{n=0}^{s-1} \left(\frac{n!}{n!}\right) P_0, f\ for\ n \leq s \\ \left(\frac{(\lambda/\mu)^n}{2^{n-1}}\right) P_0, f\ for\ n > s \quad P_n = \begin{cases} \left(\frac{(\lambda/\mu)^n}{n!}\right) P_0, & f\ for\ n \leq s \\ \left(\frac{(\lambda/\mu)^n}{2^{n-s}}\right) P_0, & f\ for\ n > s \end{cases} \quad L_q = \lambda W_q = \frac{(\lambda/\mu)^{s+1}}{(s-1)(s-\lambda/\mu)^2} P_0 \]

\[W_q = W - \frac{1}{\mu} = \frac{l_q}{\lambda} \quad L = \lambda W = L_q + \frac{\lambda}{\mu} \quad W = \frac{\lambda}{\mu} = W_q + \frac{1}{\mu} \quad P_w = \frac{1}{s!} \left(\frac{\lambda}{\mu}\right)^s \left(\frac{s\mu}{s\mu - \lambda}\right) P_0\]

Operating Characteristics of Basic Single-Server M/G/1 Queueing Model with FCFS, Infinite queue and source:

\[\lambda \quad \frac{1}{\lambda} \quad \mu \quad \frac{1}{\mu} \quad \rho = U = P_w = \frac{\lambda}{\mu} \quad P_0 = 1 - \frac{\lambda}{\mu} = 1 - \rho \quad P_n = P_0 \left[\left(\frac{1}{\mu}\right)^n\right] = P_0 \rho^n \]

\[W = W_q + \frac{1}{\lambda} = \frac{l_q}{\lambda} \quad L = \lambda W = L_q + \frac{\lambda}{\mu} \quad W_q = W - \frac{1}{\mu} = \frac{l_q}{\lambda} \quad L_q = \lambda W_q = \frac{\lambda^2 + (\lambda/\mu)^2}{2(1-\lambda/\mu)}\]