Steady-State Behavior of Infinite-Population Markovian Models

- Markovian models: exponential-distribution arrival process (mean arrival rate = $\lambda$)
- Service times may be exponentially distributed as well ($M$) or arbitrary ($G$)
- A queueing system is in statistical equilibrium if the probability that the system is in a given state is not time dependent:
  \[ P(L(t) = n) = P_n(t) = P_n \]
- Mathematical models in this chapter can be used to obtain approximate results even when the model assumptions do not strictly hold (as a rough guide)
- Simulation can be used for more refined analysis (more faithful representation for complex systems)

For the simple model studied in this chapter, the steady-state parameter, $L$, the time-average number of customers in the system is:

\[ L = \sum_{n=0}^{\infty} nP_n \]

Apply Little's equation to the whole system and to the queue alone:

\[ w = \frac{L}{\lambda} \]
\[ w_Q = w - \frac{1}{\mu} \]
\[ L_Q = \lambda w_Q \]

$G/G/c/\infty/\infty$ example:

Necessary and sufficient condition to have statistical equilibrium: $\lambda/(c\mu) < 1$
M/G/1 Queues

- Single-server queues, Poisson arrivals & unlimited capacity.
- Service times with mean $1/\mu$, variance $\sigma^2$, and $\rho = \lambda/\mu < 1$
- Steady-state parameters of $M/G/1$ queue:
  $\rho = \lambda/\mu$, $P_0 = 1 - \rho$

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho$</td>
<td>$\frac{\lambda}{\mu}$</td>
</tr>
<tr>
<td>$L$</td>
<td>$\rho + \frac{\lambda^2(1/\mu^2 + \sigma^2)}{2(1 - \rho)} = \rho + \frac{\rho^2(1 + \sigma^2\mu^2)}{2(1 - \rho)}$</td>
</tr>
<tr>
<td>$w$</td>
<td>$\frac{1}{\mu} + \frac{\lambda(1/\mu^2 + \sigma^2)}{2(1 - \rho)}$</td>
</tr>
<tr>
<td>$w_Q$</td>
<td>$\frac{\lambda(1/\mu^2 + \sigma^2)}{2(1 - \rho)}$</td>
</tr>
<tr>
<td>$L_Q$</td>
<td>$\frac{\lambda^2(1/\mu^2 + \sigma^2)}{2(1 - \rho)} = \frac{\rho^2(1 + \sigma^2\mu^2)}{2(1 - \rho)}$</td>
</tr>
<tr>
<td>$P_0$</td>
<td>$1 - \rho$</td>
</tr>
</tbody>
</table>

M/G/1 Queues

No simple expression for the steady-state probabilities $P_0$, $P_1$, …
$L - L_Q = \rho$ is the time-average number of customers being served.

Average length of queue, $L_Q$, can be rewritten as:

$$L_Q = \frac{\rho^2}{2(1 - \rho)} + \frac{\lambda^2\sigma^2}{2(1 - \rho)}$$

If $\lambda$ and $\mu$ are held constant, $L_Q$ depends on the variability, $\sigma^2$, of the service times.
M/G/1 Queues

Example: Two workers competing for a job, Able claims to be faster than Baker on average, but Baker claims to be more consistent,

Poisson arrivals at rate $\lambda = 2$ per hour ($1/30$ per minute)
Able: $1/\mu = 24$ minutes and $\sigma^2 = 20^2 = 400$ minutes$^2$:

\[ \rho = \frac{\lambda}{\mu} = \frac{24}{30} = \frac{4}{5} \]

\[ L_Q = \frac{(1/30)^2[24^2 + 400]}{2(1 - 4/5)} = 2.711 \text{ customers} \]

The proportion of arrivals who find Able idle and thus experience no delay is $P_0 = 1 - \rho = 1/5 = 20\%$.

M/G/1 Queues

Baker: $1/\mu = 25$ minutes and $\sigma^2 = 2^2 = 4$ minutes$^2$:

\[ L_Q = \frac{(1/30)^2[25^2 + 4]}{2(1 - 5/6)} = 2.097 \text{ customers} \]

The proportion of arrivals who find Baker idle and thus experience no delay is $P_0 = 1 - \rho = 1/6 = 16.7\%$

Although working faster on average, Able’s greater service variability results in an average queue length about $30\%$ greater than Baker’s
M/M/1 Queues

Suppose the service times in an $M/G/1$ queue are exponentially distributed with mean $1/\mu$, then the variance is $\sigma^2 = 1/\mu^2$. $M/M/1$ queue is a useful approximate model when service times have standard deviation approximately equal to their means. The steady-state parameters ($\rho = \lambda/\mu$):

\[
\begin{align*}
L &= \frac{\lambda}{\mu - \lambda} = \frac{\rho}{1 - \rho} \\
w &= \frac{1}{\mu - \lambda} = \frac{1}{\mu(1 - \rho)} \\
w_Q &= \frac{\lambda}{\mu(\mu - \lambda)} = \frac{\rho}{\mu(1 - \rho)} \\
L_Q &= \frac{\lambda^2}{\mu(\mu - \lambda)} = \frac{\rho^2}{1 - \rho} \\
P_n &= \left(1 - \frac{\lambda}{\mu}\right) \left(\frac{\lambda}{\mu}\right)^n = (1 - \rho)\rho^n
\end{align*}
\]

M/M/1 Queues

Example: $M/M/1$ queue
Service rate $\mu=10$ customers/hour
Consider how $L$ and $w$ increase as arrival rate, $\lambda$, increases from 5 to 8.64 by increments of 20%:

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>5.0</th>
<th>6.0</th>
<th>7.2</th>
<th>8.64</th>
<th>10.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho$</td>
<td>0.500</td>
<td>0.600</td>
<td>0.720</td>
<td>0.864</td>
<td>1.0</td>
</tr>
<tr>
<td>$L$</td>
<td>1.00</td>
<td>1.50</td>
<td>2.57</td>
<td>6.35</td>
<td>$\infty$</td>
</tr>
<tr>
<td>$w$</td>
<td>0.20</td>
<td>0.25</td>
<td>0.36</td>
<td>0.73</td>
<td>$\infty$</td>
</tr>
</tbody>
</table>

If $\lambda/\mu \geq 1$, waiting lines tend to continually grow
Increase in average system time ($w$) and average number in system ($L$) is highly nonlinear as a function of $\rho$
Effect of Utilization and Service Variability

- For almost all queues, if lines are too long, they can be reduced by decreasing server utilization ($\rho$) or by decreasing the service time variability ($\sigma^2$).

- A measure of the variability of a distribution, coefficient of variation (cv):
  \[(cv)^2 = \frac{V(X)}{[E(X)]^2}\]

- The larger cv is, the more variable the distribution relative to its expected value is.

Effect of Utilization and Service Variability

Consider $L_Q$ for any $M/G/1$ queue:
\[(cv)^2 = \sigma^2/(1/\mu)^2 = \sigma^2\mu^2\]

\[L_Q = \frac{\rho^2(1 + \sigma^2\mu^2)}{2(1 - \rho)}\]
\[= \frac{\rho^2(1 + (cv)^2)}{2(1 - \rho)}\]
\[= \left(\frac{\rho^2}{1 - \rho}\right) \left(\frac{1 + (cv)^2}{2}\right)\]

$L_Q$ for $M/M/1$ queue

Account for non-exponential service time dist'n
Multiserver Queue

$M/M/c/\infty/\infty$ queue

- $c$ channels operating in parallel
- Each channel has an independent and identical exponential service-time distribution, with mean $1/\mu$
- To achieve statistical equilibrium, the offered load $(\lambda/\mu)$ must satisfy $\lambda/\mu < c$
  - $\lambda/(c\mu) = \rho$ – server utilization

### Multiserver Queue

#### Steady-state parameters:

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>$\frac{\lambda}{c\mu}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_0$</td>
<td>$\frac{\left[\sum_{n=0}^{c-1} \frac{\lambda \lambda^n}{n!}\right] + \left[\frac{\lambda}{\mu} - 1\right] \left[\frac{c\mu}{c\mu - \lambda}\right]}{\left[\sum_{n=0}^{c-1} \frac{(c\rho)^n}{n!}\right] + \left[\frac{1}{c!} - 1\right] \left[\frac{1}{1 - \rho}\right]}^{-1}$</td>
</tr>
<tr>
<td>$P(L(\infty) \geq c)$</td>
<td>$\frac{(\lambda/\mu)^c P_0}{c! (1 - \lambda/c\mu)} = \frac{(c\rho)^c P_0}{c! (1 - \rho)}$</td>
</tr>
<tr>
<td>$L$</td>
<td>$c\rho + \frac{(c\rho)^c+1 P_0}{c! (1 - \rho)^2} = c\rho + \frac{\rho P(L(\infty) \geq c)}{1 - \rho}$</td>
</tr>
<tr>
<td>$w$</td>
<td>$L \frac{\lambda}{c\mu}$</td>
</tr>
<tr>
<td>$w_Q$</td>
<td>$w - \frac{1}{\mu}$</td>
</tr>
<tr>
<td>$L_Q$</td>
<td>$\frac{\lambda w_Q}{c! (1 - \rho)^2} = \frac{\rho P(L(\infty) \geq c)}{1 - \rho}$</td>
</tr>
<tr>
<td>$L - L_Q$</td>
<td>$\frac{\lambda}{c\mu} = c\rho$</td>
</tr>
</tbody>
</table>
Steady-State Behavior of Finite-Population Models

- When the calling population is small, the presence of one or more customers in the system has a strong effect on the distribution of future arrivals.

- Consider a finite-calling population model with $K$ customers ($M/M/c/K/K$).

- The time between the end of one service visit and the next call for service is exponentially distributed, (mean = $1/\lambda$).

- Service times are also exponentially distributed.

- $c$ parallel servers and system capacity $K$.

Steady-State Behavior of Finite-Population Models

$$
\begin{align*}
P_0 &= \sum_{n=0}^{c-1} \binom{K}{n} \left( \frac{\lambda}{\mu} \right)^n + \sum_{n=c}^{K} \frac{K!}{(K-n)!c!\mu^{n-c}} \left( \frac{\lambda}{\mu} \right)^n \times P_0^{-1} \\
P_n &= \begin{cases} \\
\binom{K}{n} \left( \frac{\lambda}{\mu} \right)^n P_0, & n = 0, 1, \ldots, c - 1 \\
\binom{K}{n} \frac{K!}{(K-n)!c!\mu^{n-c}} \left( \frac{\lambda}{\mu} \right)^n P_0, & n = c, c + 1, \ldots, K \\
\end{cases} \\
L &= \sum_{n=0}^{K} nP_n \\
L_Q &= \sum_{n=c}^{K} (n-c)P_n \\
\lambda_e &= \sum_{n=0}^{K} (K-n) \lambda P_n \\
w &= \frac{L}{\lambda_e} \\
w_Q &= \frac{L_Q}{\lambda_e} \\
\rho &= \frac{L - L_Q}{c} = \frac{\lambda_e}{c\mu} \\
\end{align*}
$$

Steady-state parameters:
Steady-State Behavior of Finite-Population Models

Example:
- two workers who are responsible for 10 milling machines
- run on the average for 20 minutes
- require an average 5-minute service period
- both times exponentially distributed: $\lambda = 1/20$ and $\mu = 1/5$
- Compute the various measures of performance for this system

- All measures depend on $P_n$, which is:

$$
\left[ \sum_{n=0}^{2-1} \binom{10}{n} \left( \frac{5}{20} \right)^n + \sum_{n=2}^{10} \frac{10!}{(10-n)!2!2^{n-2}} \left( \frac{5}{20} \right)^n \right]^{-1} = 0.065
$$

Steady-State Behavior of Finite-Population Models

Then, we can obtain the $P_n$, and the expected number of machines in system:

$$
L_Q = \sum_{n=3}^{10} (n-2)P_n = 1.46 \text{ machines}
$$

Effective arrival rate:

$$
\lambda_e = \sum_{n=0}^{10} (10-n) \left( \frac{1}{20} \right) P_n = 0.342 \text{ machines/minute}
$$

Average waiting time in queue:

$$
w_Q = \frac{L_Q}{\lambda_e} = 4.27 \text{ minutes}
$$
Steady-State Behavior of Finite-Population Models

Expected no. of machines being served:

\[ L = \sum_{n=0}^{10} nP_n = 3.17 \text{ machines} \]

Average no. of machines being served:

\[ L = L_Q = 3.17 - 1.46 = 1.71 \text{ machines} \]

Average no. of running machines:

\[ K - L = 10 - 3.17 = 6.83 \text{ machines} \]

What if the no. of servers \( c = 3 \)?

\[ K - L = 7.74 \text{ machines} \]

What if the no. of servers \( c = 1 \)?

\[ K - L = 3.98 \text{ machines} \]