Chapter 6
Queueing Models (2)

Banks, Carson, Nelson & Nicol
Discrete-Event System Simulation
Outline

- Server Utilization
- System Performance
- Steady-State Behavior of Infinite-Population Models
- Steady-State Behavior of Finite-Population Models
- Networks of Queues
Definition: the proportion of time that a server is busy.

- Observed server utilization, $\hat{\rho}$, is defined over a specified time interval $[0,T]$.
- Long-run server utilization is $\rho$.
- For systems with long-run stability: $\hat{\rho} \rightarrow \rho$ as $T \rightarrow \infty$
Server Utilization

[Characteristics of Queueing System]

- For $G/G/1/\infty/\infty$ queues:

- In general, for a single-server queue:

$$\rho = \frac{\lambda E(s)}{\mu} = \frac{\lambda}{\mu} \quad \rho = \frac{\lambda}{\mu} < 1$$

- For a single-server stable queue:

- For an unstable queue ($\lambda > \mu$), long-run server utilization is 1.
Server Utilization

[Characteristics of Queueing System]

- For $G/G/c/\infty/\infty$ queues:
  - A system with $c$ identical servers in parallel.
  - If an arriving customer finds more than one server idle, the customer chooses a server without favoring any particular server.
  - The long-run average server utilization is:

\[
\rho = \frac{\lambda}{c \mu}, \quad \text{where} \quad \lambda < c \mu \quad \text{for stable systems}
\]
Server Utilization and System Performance

[Characteristics of Queueing System]

- System performance varies widely for a given utilization $\rho$.
  - For example, a $D/D/1$ queue where $E(A) = 1/\lambda$ and $E(S) = 1/\mu$, where:
    \[ L = \rho = \frac{\lambda}{\mu}, \quad w = E(S) = \frac{1}{\mu}, \quad L_Q = W_Q = 0. \]
  - By varying $\lambda$ and $\mu$, server utilization can assume any value between 0 and 1.
  - Yet there is never any line.
  - In general, variability of interarrival and service times causes lines to fluctuate in length.
Server Utilization and System Performance

[Characteristics of Queueing System]

- Example: A physician who schedules patients every 10 minutes and spends $S_i$ minutes with the $i^{th}$ patient:

  $S_i = \begin{cases} 9 \text{ minutes with probability } 0.9 \\ 12 \text{ minutes with probability } 0.1 \end{cases}$

- Arrivals are deterministic, $A_1 = A_2 = \ldots = \lambda^{-1} = 10$.
- Services are stochastic, $E(S_i) = 9.3$ min and $V(S_i) = 0.81$ min$^2$.
- On average, the physician's utilization $\rho = \lambda/\mu = 0.93 < 1$.
- Consider the system is simulated with service times: $S_1 = 9$, $S_2 = 12$, $S_3 = 9$, $S_4 = 9$, $S_5 = 9$, $\ldots$. The system becomes:

- The occurrence of a relatively long service time ($S_2 = 12$) causes a waiting line to form temporarily.
Costs in Queueing Problems

[Characteristics of Queueing System]

- Costs can be associated with various aspects of the waiting line or servers:
  - System incurs a cost for each customer in the queue, say at a rate of $10 per hour per customer.
    - The average cost per customer is:
      \[ \sum_{j=1}^{N} \frac{10 \cdot W_j^Q}{N} = 10 \cdot \hat{W}_Q \]
    - If \( \hat{\lambda} \) customers per hour arrive (on average), the average cost per hour is:
      \[ \left( \frac{\hat{\lambda}}{\text{hour}} \right) \left( \frac{10 \cdot \hat{W}_Q}{\text{customer}} \right) = 10 \cdot \hat{\lambda} \hat{W}_Q = 10 \cdot \hat{L}_Q \text{ / hour} \]
  - Server may also impose costs on the system, if a group of \( c \) parallel servers \( (1 \leq c \leq \infty) \) have utilization \( \rho \), each server imposes a cost of $5 per hour while busy.
    - The total server cost is: \( 5 \cdot c \rho \).
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).
Steady-State Behavior of Infinite-Population Markovian Models

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:

$$w = \frac{L}{\lambda}, \quad w_Q = w - \frac{1}{\mu}$$

$$L_Q = \lambda w_Q$$

- $G/G/c/\infty/\infty$ example: to have a statistical equilibrium, a necessary and sufficient condition is $\lambda/(c\mu) < 1$. 
M/G/1 Queues

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

\[
\begin{align*}
\rho &= \frac{\lambda}{\mu}, \quad P_0 = 1 - \rho \\
L &= \rho + \frac{\rho^2(1 + \sigma^2/\mu^2)}{2(1 - \rho)}, \quad L_Q = \frac{\rho^2(1 + \sigma^2/\mu^2)}{2(1 - \rho)} \\
w &= \frac{1}{\mu} + \frac{\lambda(1/\mu^2 + \sigma^2)}{2(1 - \rho)}, \quad w_Q = \frac{\lambda(1/\mu^2 + \sigma^2)}{2(1 - \rho)}
\end{align*}
\]
M/G/1 Queues

No simple expression for the steady-state probabilities $P_0, P_1, \ldots$

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

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

\[
\rho = \lambda / \mu, \quad P_n = (1 - \rho)\rho^n
\]

\[
L = \frac{\lambda}{\mu - \lambda} = \frac{\rho}{1 - \rho}, \quad L_Q = \frac{\lambda^2}{\mu(\mu - \lambda)} = \frac{\rho^2}{1 - \rho}
\]

\[
w = \frac{1}{\mu - \lambda} = \frac{1}{\mu(1 - \rho)}, \quad w_Q = \frac{\lambda}{\mu(\mu - \lambda)} = \frac{\rho}{\mu(1 - \rho)}
\]
M/M/1 Queues

Example: $M/M/1$ queue with service rate $\mu = 10$ customers per 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>1.000</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 in length.
- 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

[Steady-State of Markovian Model]

- 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 is the distribution relative to its expected value
Effect of Utilization and Service Variability

[Steady-State of Markovian Model]

Consider $L_Q$ for any $M/G/1$ queue:

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

$$= \left( \frac{\rho^2}{1 - \rho} \right) \left( \frac{1 + (cv)^2}{2} \right)$$

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

Corrects the $M/M/1$ formula to account for a 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 $\lambda/\mu < c$, where $\lambda/(c\mu) = \rho$ is the server utilization.
  - Some of the steady-state probabilities:
    \[ \rho = \frac{\lambda}{c\mu} \]
    \[ P_0 = \left\{ \sum_{n=0}^{c-1} \frac{(\lambda/\mu)^n}{n!} \right\} + \left[ \left( \frac{\lambda}{\mu} \right)^c \left( \frac{1}{c!} \right) \left( \frac{c\mu}{c\mu - \lambda} \right) \right]^{-1} \]
    \[ L = c\rho + \frac{(c\rho)^{c+1} P_0}{c(c!)(1 - \rho)^2} = c\rho + \frac{\rho P(L(\infty) \geq c)}{1 - \rho} \]
    \[ w = \frac{L}{\lambda} \]
Other common multiserver queueing models:

- **$M/G/c/\infty$**: general service times and $c$ parallel server. The parameters can be approximated from those of the $M/M/c/\infty/\infty$ model.

- **$M/G/\infty$**: general service times and infinite number of servers, e.g., customer is its own system, service capacity far exceeds service demand.

- **$M/M/C/N/\infty$**: service times are exponentially distributed at rate $m$ and $c$ servers where the total system capacity is $N \geq c$ customer (when an arrival occurs and the system is full, that arrival is turned away).
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 is $K$. 
Steady-State Behavior of Finite-Population Models

Some of the steady-state probabilities:

\[ P_0 = \left\{ \sum_{n=0}^{c-1} \binom{K}{n} \left( \frac{\lambda}{\mu} \right)^n + \sum_{n=c}^{K} \frac{K!}{(K-n)!c!c^{n-c}} \left( \frac{\lambda}{\mu} \right)^n \right\}^{-1} \]

\[ P_n = \left\{ \binom{K}{n} \left( \frac{\lambda}{\mu} \right)^n P_0, \quad n = 0, 1, \ldots, c - 1 \right\} \]

\[ P_n = \frac{K!}{(K-n)!c!c^{n-c}} \left( \frac{\lambda}{\mu} \right)^n, \quad n = c, c + 1, \ldots, K \]

\[ L = \sum_{n=0}^{K} nP_n, \quad w = L / \lambda_e, \quad \rho = \lambda_e / c\mu \]

where \( \lambda_e \) is the long run effective arrival rate of customers to queue (or entering/exiting service)

\[ \lambda_e = \sum_{n=0}^{K} (K-n) \lambda P_n \]
Steady-State Behavior of Finite-Population Models

- Example: two workers who are responsible for 10 milling machines.
  - Machines run on the average for 20 minutes, then require an average 5-minute service period, both times exponentially distributed: $\lambda = \frac{1}{20}$ and $\mu = \frac{1}{5}$.
  - All of the performance measures depend on $P_0$:
    \[
    P_0 = \left\{2^{-1}\sum_{n=0}^{10} \binom{10}{n} \left(\frac{5}{20}\right)^n + \sum_{n=2}^{10} \frac{10!}{(10-n)!} \frac{5^n}{2^{n-2}}\right\}^{-1} = 0.065
    \]
    - Then, we can obtain the other $P_n$.
    - Expected number of machines in system:
      \[
      L = \sum_{n=0}^{10} nP_n = 3.17 \text{ machines}
      \]
    - The average number of running machines:
      \[
      K - L = 10 - 3.17 = 6.83 \text{ machines}
      \]
Networks of Queues

- Many systems are naturally modeled as networks of single queues: customers departing from one queue may be routed to another.

- The following results assume a stable system with infinite calling population and no limit on system capacity:
  - Provided that no customers are created or destroyed in the queue, then the departure rate out of a queue is the same as the arrival rate into the queue (over the long run).
  - If customers arrive to queue $i$ at rate $\lambda_i$, and a fraction $0 \leq p_{ij} \leq 1$ of them are routed to queue $j$ upon departure, then the arrival rate from queue $i$ to queue $j$ is $\lambda_i p_{ij}$ (over the long run).
Networks of Queues

- The overall arrival rate into queue $j$:

$$\lambda_j = a_j + \sum_{all\ i} \lambda_i p_{ij}$$

Arrival rate from outside the network

Sum of arrival rates from other queues in network

- If queue $j$ has $c_j < \infty$ parallel servers, each working at rate $\mu_j$, then the long-run utilization of each server is $\rho_j = \lambda_j/(c_j \mu_j)$ (where $\rho_j < 1$ for stable queue).

- If arrivals from outside the network form a Poisson process with rate $a_j$ for each queue $j$, and if there are $c_j$ identical servers delivering exponentially distributed service times with mean $1/\mu_j$, then, in steady state, queue $j$ behaves like an $M/M/c_j$ queue with arrival rate

$$\lambda_j = a_j + \sum_{all\ i} \lambda_i p_{ij}$$
Network of Queues

- Discount store example:
  - Suppose customers arrive at the rate 80 per hour and 40% choose self-service. Hence:
    - Arrival rate to service center 1 is \( \lambda_1 = 80(0.4) = 32 \) per hour
    - Arrival rate to service center 2 is \( \lambda_2 = 80(0.6) = 48 \) per hour.
  - \( c_2 = 3 \) clerks and \( \mu_2 = 20 \) customers per hour.
  - The long-run utilization of the clerks is:
    \[
    \rho_2 = \frac{48}{(3 \times 20)} = 0.8
    \]
  - All customers must see the cashier at service center 3, the overall rate to service center 3 is \( \lambda_3 = \lambda_1 + \lambda_2 = 80 \) per hour.
    - If \( \mu_3 = 90 \) per hour, then the utilization of the cashier is:
      \[
      \rho_3 = \frac{80}{90} = 0.89
      \]
Summary

- Introduced basic concepts of queueing models.
- Show how simulation, and sometimes mathematical analysis, can be used to estimate the performance measures of a system.
- Commonly used performance measures: $L$, $L_Q$, $w$, $w_Q$, $\rho$, and $\lambda_e$.
- When simulating any system that evolves over time, analysts must decide whether to study transient behavior or steady-state behavior.
  - Simple formulas exist for the steady-state behavior of some queues.
- Simple models can be solved mathematically, and can be useful in providing a rough estimate of a performance measure.