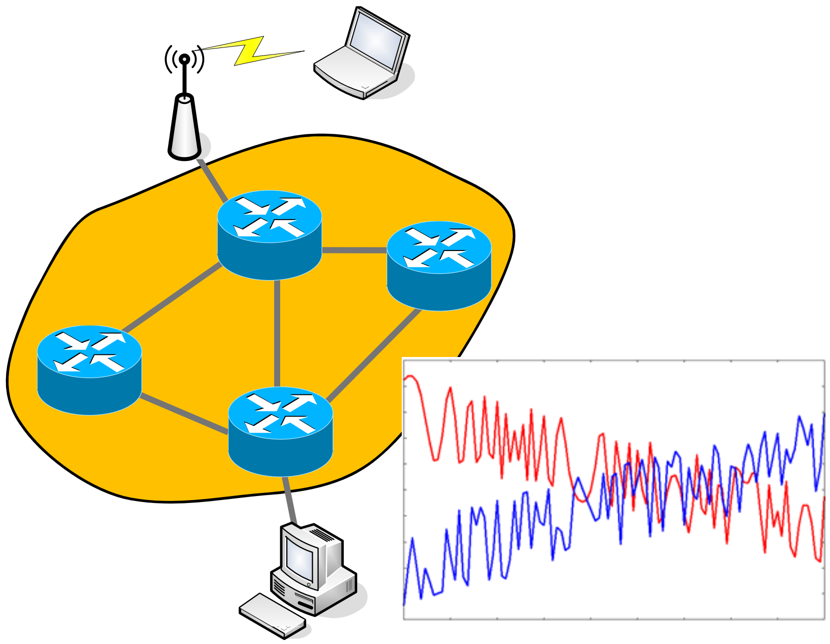
Chapter 8

Queueing Models

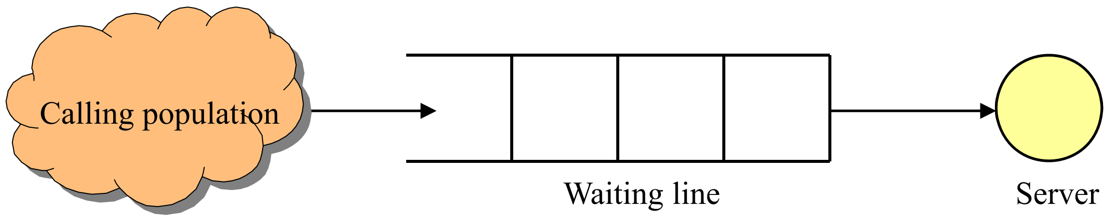


Contents

- Characteristics of Queueing Systems
- Queueing Notation – Kendall Notation
- Long-run Measures of Performance of Queueing Systems
- Steady-state Behavior of Infinite-Population Markovian Models
- Steady-state Behavior of Finite-Population Models
- Networks of Queues

Purpose

- Simulation is often used in the analysis of queueing models.
- A simple but typical queueing model



- Queueing models provide the analyst with a powerful tool for designing and evaluating the performance of queueing systems.
- Typical measures of system performance
  - Server utilization, length of waiting lines, and delays of customers
  - For relatively simple systems: compute mathematically
  - For realistic models of complex systems: simulation is usually required

Outline

- Discuss some well-known models
  - Not development of queueing theory, for this see other class!

- We will deal with
  - General characteristics of queues
  - Meanings and relationships of important performance measures
  - Estimation of mean measures of performance
  - Effect of varying input parameters
  - Mathematical solutions of some basic queueing models

Characteristics of Queueing Systems

Characteristics of Queueing Systems

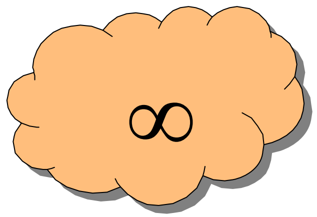
- Key elements of queueing systems
  - Customer: refers to anything that arrives at a facility and requires service, e.g., people, machines, trucks, emails, packets, frames.
  - Server: refers to any resource that provides the requested service, e.g., repairpersons, machines, runways at airport, host, switch, router, disk drive, algorithm.

System	Customers	Server
Reception desk	People	Receptionist
Hospital	Patients	Nurses
Airport	Airplanes	Runway
Production line	Cases	Case-packer
Road network	Cars	Traffic light
Grocery	Shoppers	Checkout station
Computer	Jobs	CPU, disk, CD
Network	Packets	Router

Calling Population

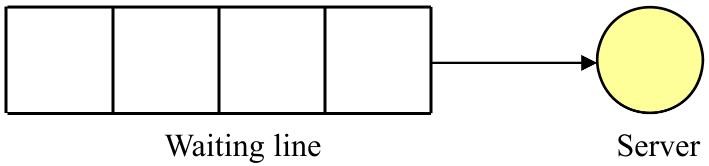
- Calling population: the population of potential customers, may be assumed to be **finite** or **infinite**.
  - Finite population model: if arrival rate depends on the number of customers being served and waiting, e.g., model of one corporate jet, if it is being repaired, the repair arrival rate becomes zero.

  - Infinite population model: if arrival rate is not affected by the number of customers being served and waiting, e.g., systems with large population of potential customers.

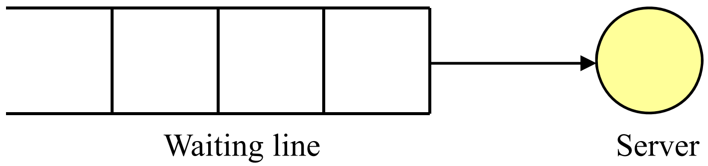


System Capacity

- System Capacity: a limit on the number of customers that may be in the waiting line or system.
  - Limited capacity, e.g., an automatic car wash only has room for 10 cars to wait in line to enter the mechanism.
  - If system is full no customers are accepted anymore



  - Unlimited capacity, e.g., concert ticket sales with no limit on the number of people allowed to wait to purchase tickets.

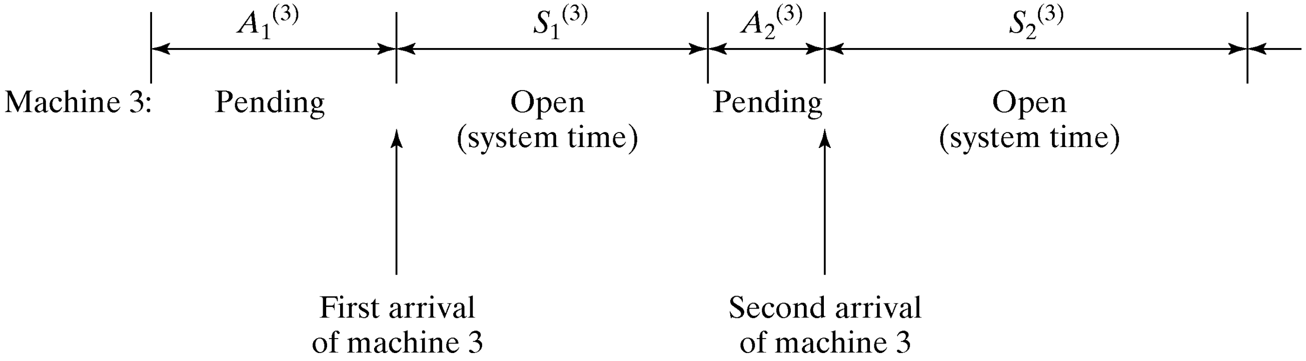


Arrival Process

- For infinite-population models:
  - In terms of interarrival times of successive customers.
- Arrival types:
  - Random arrivals: interarrival times usually characterized by a probability distribution.
    - Most important model: Poisson arrival process (with rate λ), where a time represents the interarrival time between customer n-1 and customer n, and is exponentially distributed (with mean $1/\lambda$).
  - Scheduled arrivals: interarrival times can be constant or constant plus or minus a small random amount to represent early or late arrivals.
    - Example: patients to a physician or scheduled airline flight arrivals to an airport
- At least one customer is assumed to always be present, so the server is never idle, e.g., sufficient raw material for a machine.

Arrival Process

- For finite-population models:
  - Customer is pending when the customer is outside the queueing system, e.g., machine-repair problem: a machine is "pending" when it is operating, it becomes "not pending" the instant it demands service from the repairman.
  - Runtime of a customer is the length of time from departure from the queueing system until that customer's next arrival to the queue, e.g., machine-repair problem, machines are customers and a runtime is time to failure (TTF).
  - Let $A_1^{(i)}, A_2^{(i)}, \ldots$ be the successive runtimes of customer i, and $S_1^{(i)}$, $S_2^{(i)}$ be the corresponding successive system times:



Queue Behavior and Queue Discipline

- Queue behavior: the actions of customers while in a queue waiting for service to begin, for example:
  - Balk: leave when they see that the line is too long
  - Renege: leave after being in the line when its moving too slowly
  - Jockey: move from one line to a shorter line

- Queue discipline: the logical ordering of customers in a queue that determines which customer is chosen for service when a server becomes free, for example:
  - First-in-first-out (FIFO)
  - Last-in-first-out (LIFO)
  - Service in random order (SIRO)
  - Shortest processing time first (SPT)
  - Service according to priority (PR)

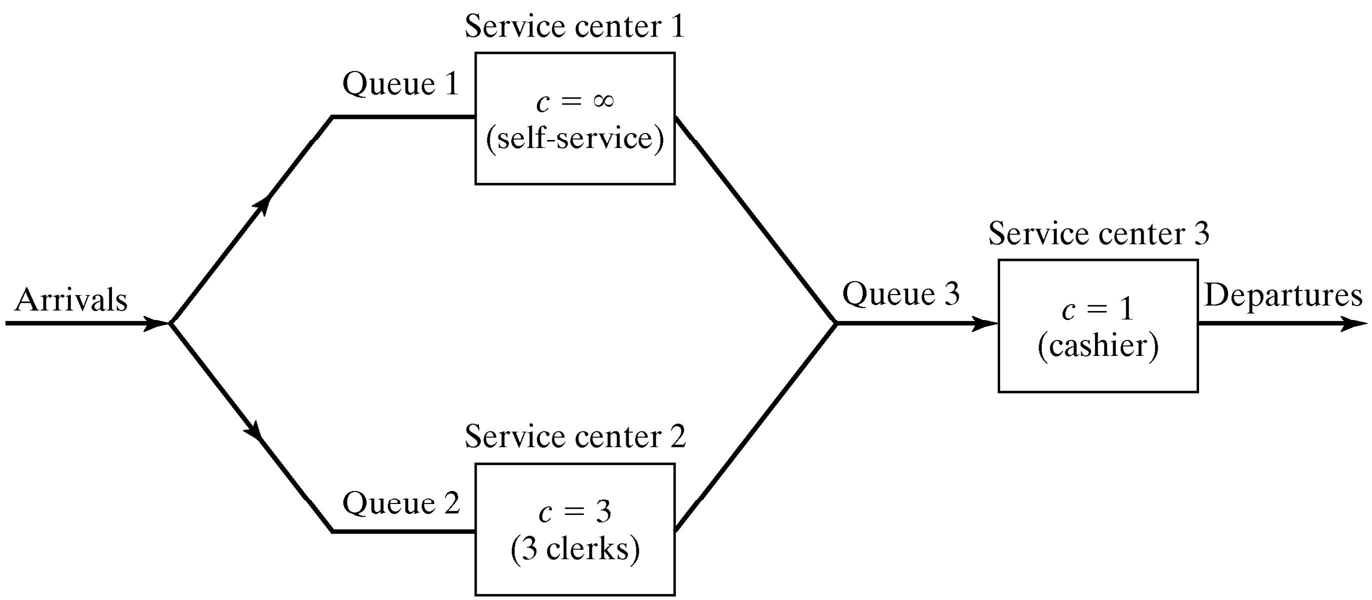
Service Times and Service Mechanism

- Service times of successive arrivals are denoted by S_1, S_2, S_3.
  - May be constant or random.
  - $\{S_1, S_2, S_3, \ldots\}$ is usually characterized as a sequence of independent and identically distributed (IID) random variables, e.g.,
    - Exponential, Weibull, Gamma, Lognormal, and Truncated normal distribution.

- A queueing system consists of a number of service centers and interconnected queues.
  - Each service center consists of some number of servers (c) working in parallel, upon getting to the head of the line, a customer takes the 1^{st} available server.

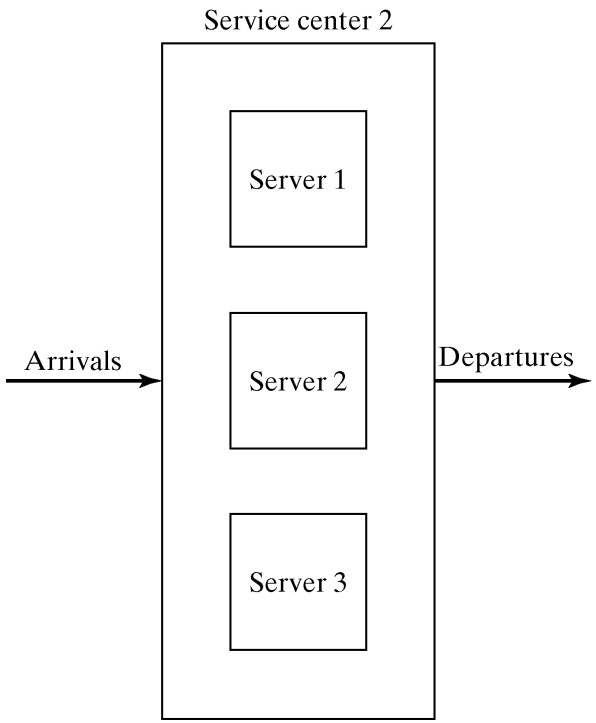
Queuing System: Example 1

- Example: consider a discount warehouse where customers may
  - serve themselves before paying at the cashier (service center 1) or
  - served by a clerk (service center 2)



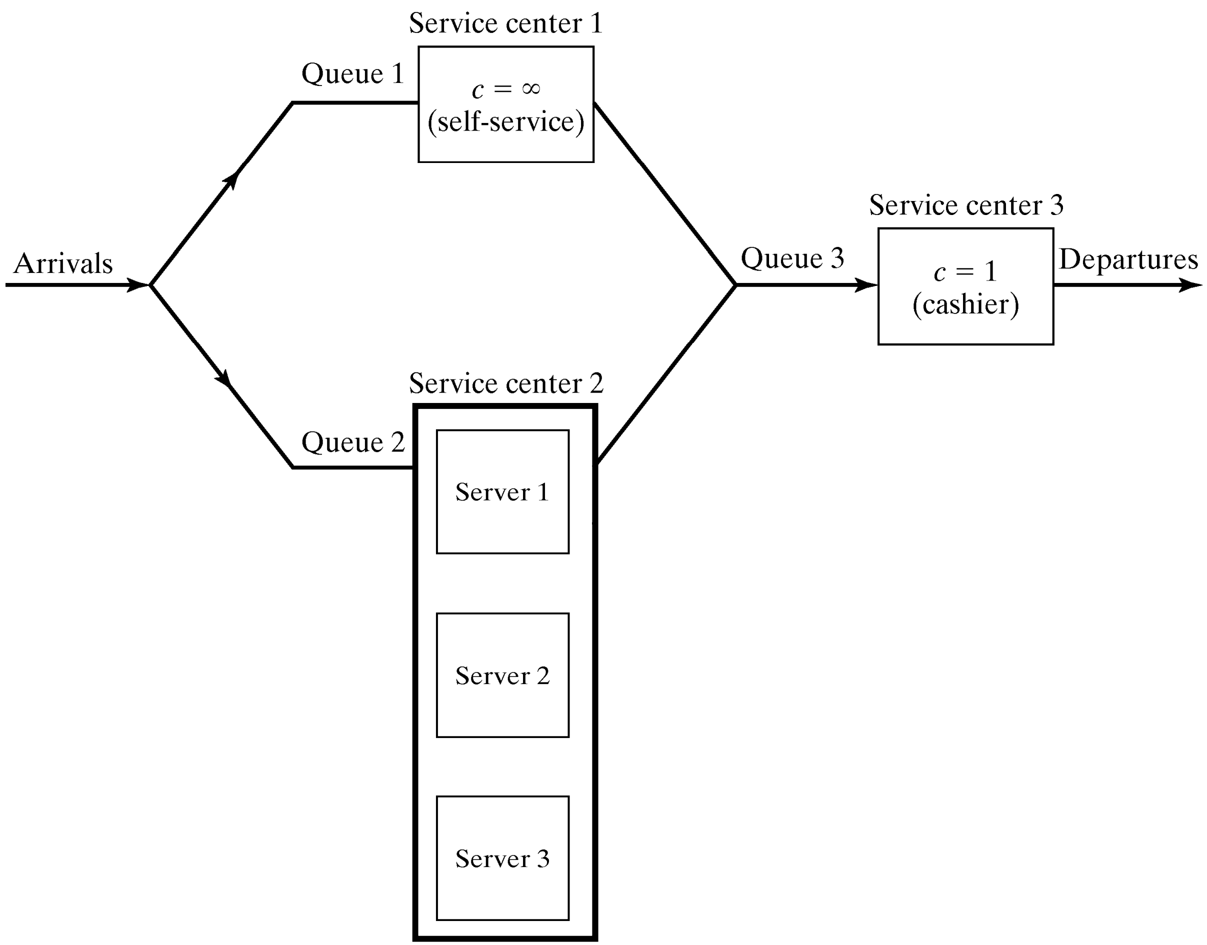
Queuing System: Example 1

- Wait for one of the three clerks:



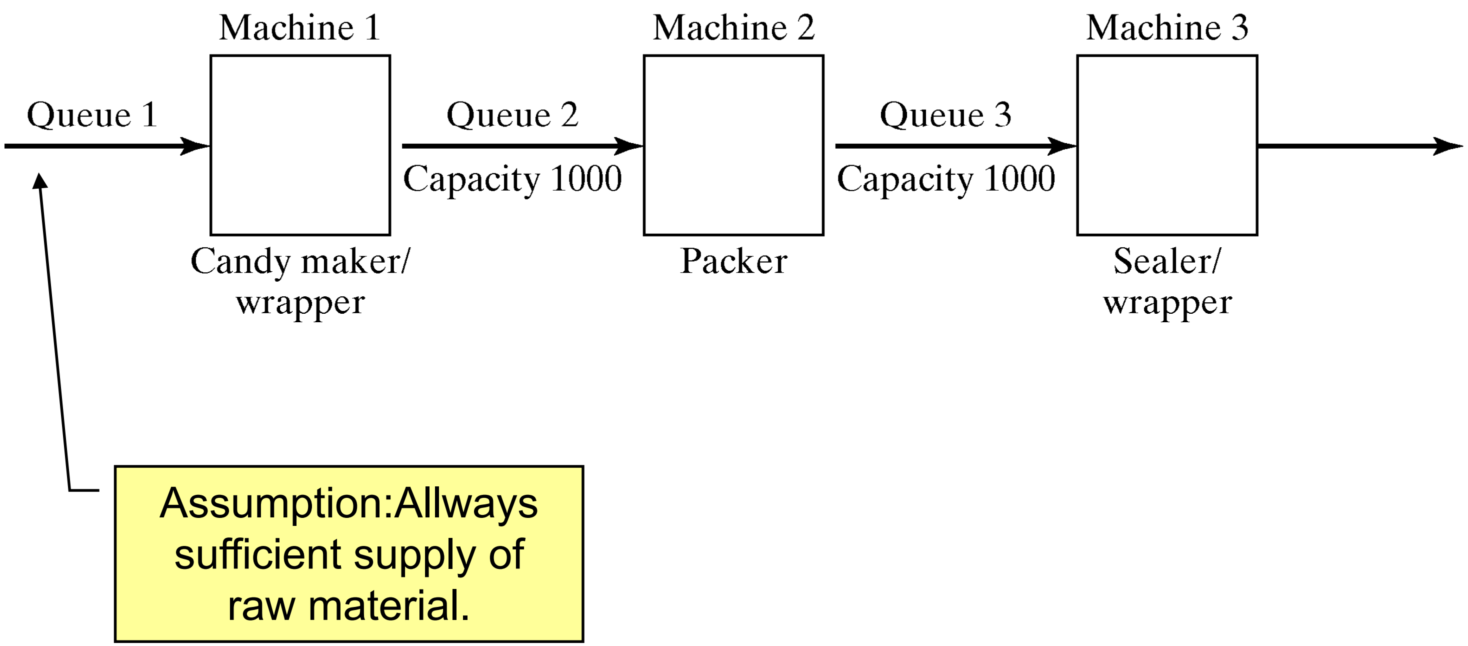
- Batch service (a server serving several customers simultaneously), or customer requires several servers simultaneously.

Queuing System: Example 1



Queuing System: Example 2

- Candy production line
  - Three machines separated by buffers
  - Buffers have capacity of 1000 candies



Queueing Notation

The Kendall Notation

Queueing Notation: Kendall Notation

- A notation system for parallel server queues: $A/B/c/N/K$
  - A represents the interarrival-time distribution
  - B represents the service-time distribution
  - c represents the number of parallel servers
  - N represents the system capacity
  - K represents the size of the calling population
  - N, K are usually dropped, if they are infinity
- Common symbols for A and B
  - M Markov, exponential distribution
  - D Constant, deterministic
  - E_k Erlang distribution of order k
  - H Hyperexponential distribution
  - G General, arbitrary
- Examples
  - $M/M/1/\infty/\infty$ same as $M/M/1$: Single-server with unlimited capacity and call-population. Interarrival and service times are exponentially distributed
  - $G/G/1/5/5$: Single-server with capacity 5 and call-population 5.
  - $M/M/5/20/1500/\mathrm{FIFO}$: Five parallel server with capacity 20, call-population 1500, and service discipline FIFO

Queueing Notation

- General performance measures of queueing systems:
  - P_n steady-state probability of having n customers in system
  - $P_n(t)$ probability of n customers in system at time t
  - λ arrival rate
  - λ_e effective arrival rate
  - μ service rate of one server
  - ρ server utilization
  - A_n interarrival time between customers n-1 and n
  - S_n service time of the n-th arriving customer
  - W_n total time spent in system by the n-th customer
  - W_n^Q total time spent in the waiting line by customer n
  - $L(t)$ the number of customers in system at time t
  - $L_Q(t)$ the number of customers in queue at time t
  - L long-run time-average number of customers in system
  - L_Q long-run time-average number of customers in queue
  - W long-run average time spent in system per customer
  - w_Q long-run average time spent in queue per customer

Long-run Measures of Performance of Queueing Systems

Long-run Measures of Performance of Queueing Systems

- Primary long-run measures of performance are
  - L long-run time-average number of customers in system
  - L_Q long-run time-average number of customers in queue
  - W long-run average time spent in system per customer
  - w_Q long-run average time spent in queue per customer
  - ρ server utilization

- Other measures of interest are
  - Long-run proportion of customers who are delayed longer than t_0 time units
  - Long-run proportion of customers turned away because of capacity constraints
  - Long-run proportion of time the waiting line contains more than k_0 customers

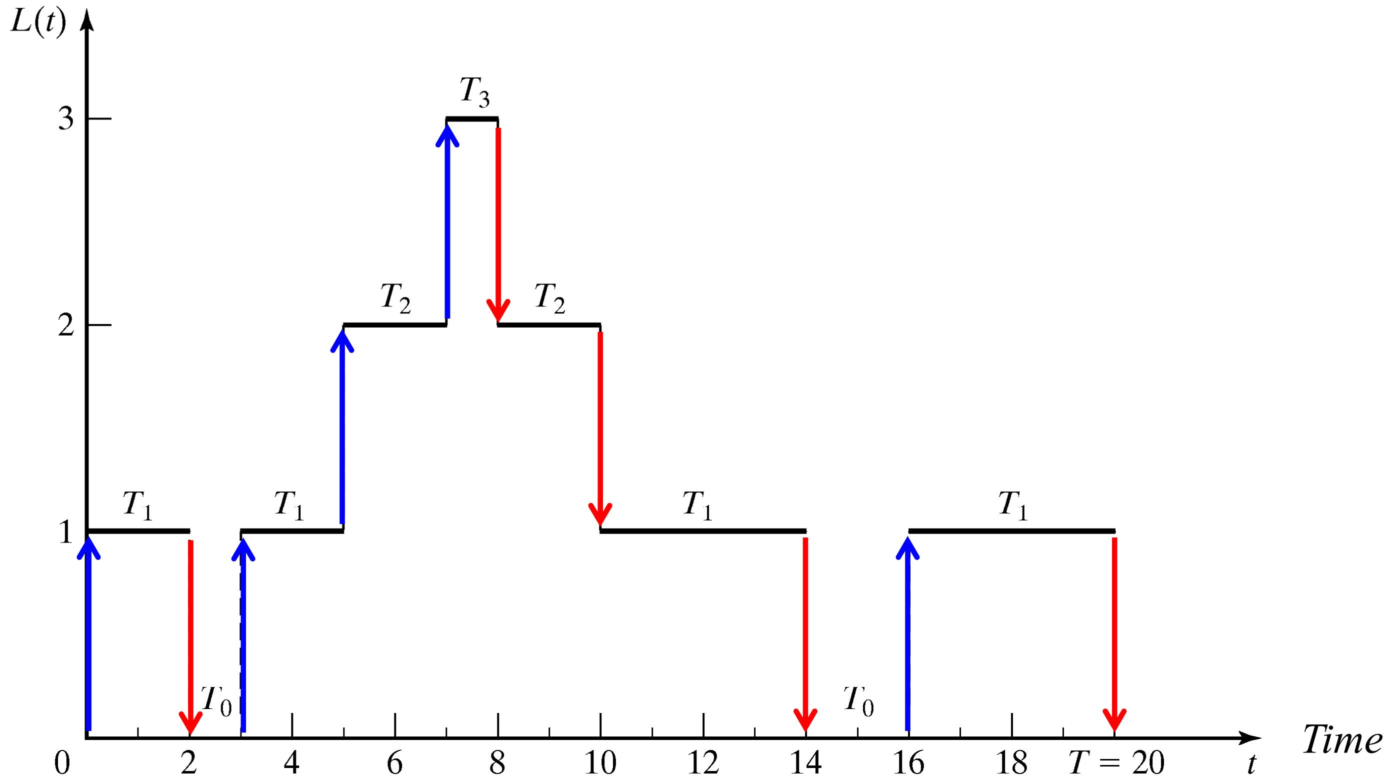
Long-run Measures of Performance of Queueing Systems

- Goal of this section
  - Major measures of performance for a general $G/G/c/N/K$ queueing system
  - How these measures can be estimated from simulation runs

- Two types of estimators
  - Sample average
  - Time-integrated sample average

Time-Average Number in System L

Number of customers in the system



Time-Average Number in System L

- Consider a queueing system over a period of time T
  - Let T_i denote the total time during $[0,T]$ in which the system contained exactly i customers, the time-weighted-average number in the system is defined by:

$$\hat{L} = \frac{1}{T}\sum_{i=0}^{\infty} iT_i = \sum_{i=0}^{\infty} i\left(\frac{T_i}{T}\right)$$

  - Consider the total area under the function is $L(t)$, then,

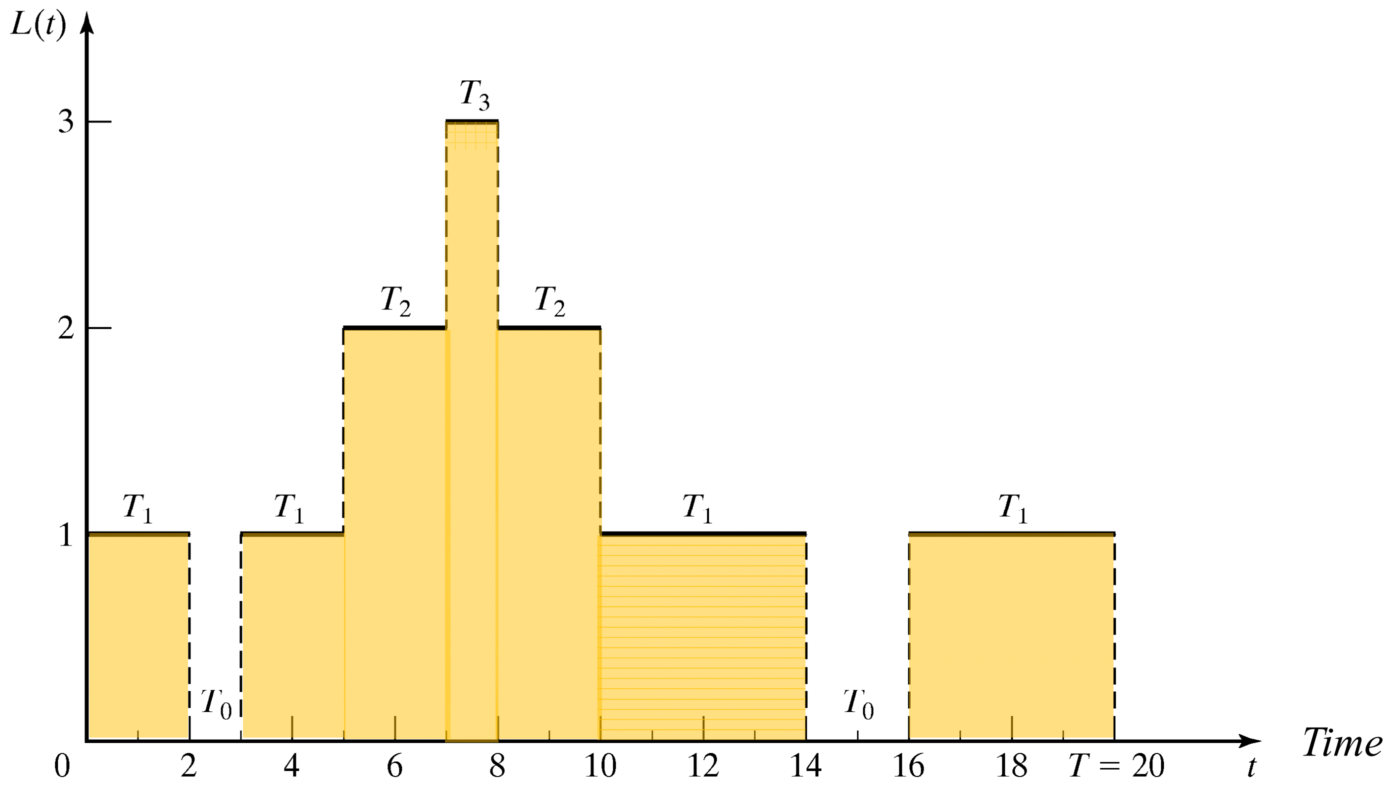
$$\hat{L} = \frac{1}{T}\sum_{i=0}^{\infty} iT_i = \frac{1}{T}\int_0^T L(t)\,dt$$

  - The long-run time-average number of customers in system, with probability 1:

$$\hat{L} = \frac{1}{T}\int_0^T L(t)\,dt \xrightarrow[T\to\infty]{} L$$

Time-Average Number in System L

Number of customers in the system

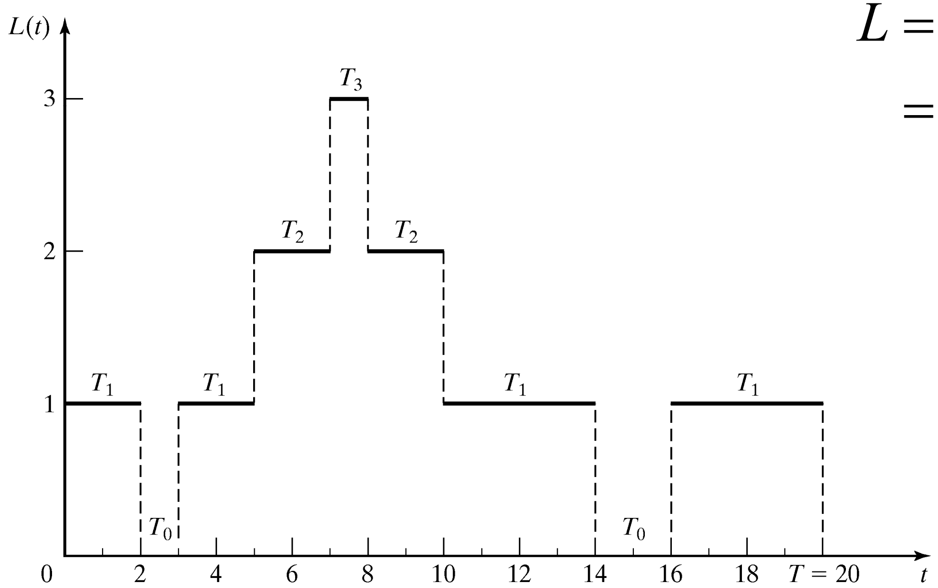


Time-Average Number in System L

- The time-weighted-average number in queue is:

$$\hat{L}_Q = \frac{1}{T}\sum_{i=0}^{\infty} iT_i^Q = \frac{1}{T}\int_0^T L_Q(t)\,dt \xrightarrow[T\to\infty]{} L_Q$$

  - $G/G/1/N/K$ example: consider the results from the queueing system ($N \geq 4,\ K \geq 3$).

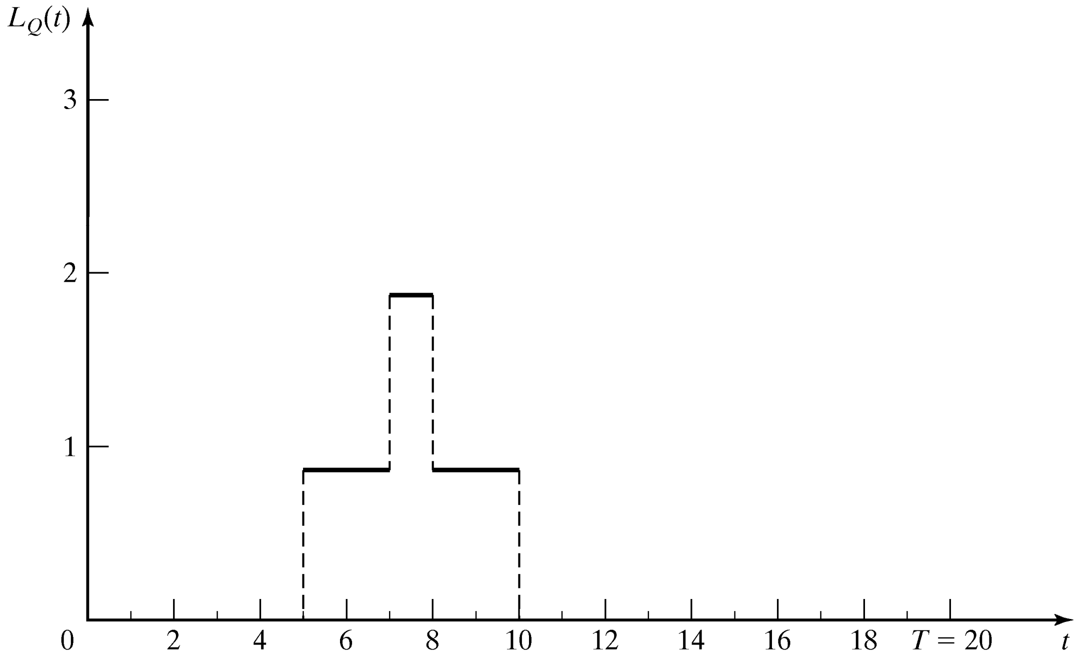


$$\hat{L} = [0(3) + 1(12) + 2(4) + 3(1)]/20$$
$$= 23/20 = 1.15 \text{ customers}$$

Time-Average Number in System L

$$L_Q(t) = \begin{cases} 0, & \text{if } L(t) = 0 \\ L(t) - 1, & \text{if } L(t) \geq 1 \end{cases}$$

$$\hat{L}_Q = \frac{0(15) + 1(4) + 2(1)}{20} = 0.3 \text{ customers}$$



Average Time Spent in System Per Customer w

- The average time spent in system per customer, called the average system time, is:

$$\hat{w} = \frac{1}{N}\sum_{i=1}^{N} W_i$$

where $W_1, W_2, \ldots, W_N$ are the individual times that each of the N customers spend in the system during $[0,T]$.

- For stable systems: $\hat{w} \to w \text{ as } N \to \infty$

- If the system under consideration is the queue alone:

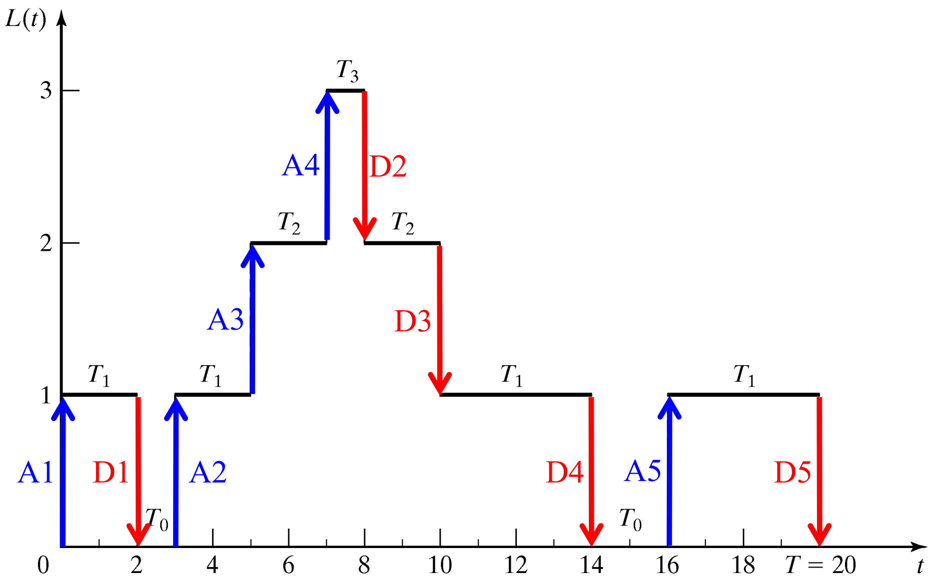
$$\hat{w}_Q = \frac{1}{N}\sum_{i=1}^{N} W_i^Q \xrightarrow[N\to\infty]{} w_Q$$

Average Time Spent in System Per Customer w

- $G/G/1/N/K$ example (cont.):
  - The average system time is ($W_i = D_i - A_i$)

$$\hat{w} = \frac{W_1 + W_2 + ... + W_5}{5} = \frac{2 + (8-3) + (10-5) + (14-7) + (20-16)}{5} = 4.6 \text{ time units}$$

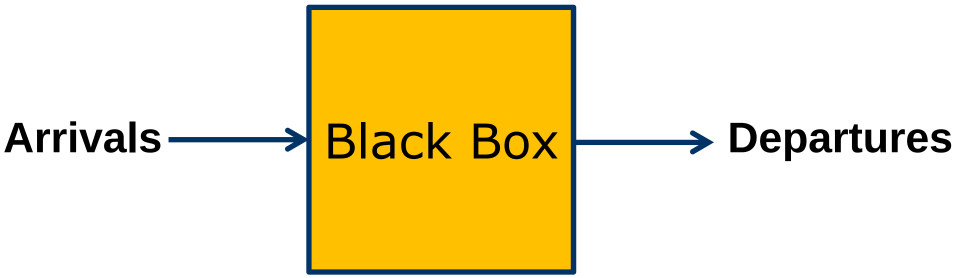
  - The average queuing time is $$\hat{w}_Q = \frac{0+0+3+3+0}{5} = 1.2 \text{ time units}$$



The Conservation Equation or Little's Law

The Conservation Equation: Little's Law

- One of the most common theorems in queueing theory
- Mean number of customers in system
- Conservation equation (a.k.a. Little's law)



$$\text{average number in system} = \text{arrival rate} \times \text{average system time}$$

The Conservation Equation: Little's Law

- Conservation equation (a.k.a. Little's law)

$$\hat{L} = \hat{\lambda}\hat{w}$$

Average # in system

Arrival rate

Average System time

$$L = \lambda w \quad \text{as} \quad T \to \infty \ \text{ and } \ N \to \infty$$

- Holds for almost all queueing systems or subsystems (regardless of the number of servers, the queue discipline, or other special circumstances).
- *G/G/1/N/K* example (cont.): On average, one arrival every 4 time units and each arrival spends 4.6 time units in the system. Hence, at an arbitrary point in time, there are (1/4)(4.6) = 1.15 customers present on average.

Server Utilization

- 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 ρ.
  - For systems with long-run stability: $\hat{\rho} \rightarrow \rho \quad \text{as} \quad T \rightarrow \infty$

Server Utilization

- For $G/G/1/\infty/\infty$ queues:
  - Any single-server queueing system with
    - average arrival rate λ customers per time unit,
    - average service time $E(S) = 1/\mu$ time units, and
    - infinite queue capacity and calling population.
  - Conservation equation, $L = \lambda w$, can be applied.
  - For a stable system, the average arrival rate to the server, λ_s, must be identical to λ.
  - The average number of customers in the server is:

$$\hat{L}_s = \frac{1}{T}\int_0^T \left(L(t) - L_Q(t)\right)dt = \frac{T - T_0}{T}$$

Server Utilization

- In general, for a single-server queue:

$$\hat{L}_s = \hat{\rho} \xrightarrow[T\to\infty]{} L_s = \rho$$

$$\text{and} \qquad \rho = \lambda \cdot E(s) = \frac{\lambda}{\mu}$$

  - For a single-server stable queue: $\rho = \dfrac{\lambda}{\mu} < 1$
  - For an unstable queue ($\lambda > \mu$), long-run server utilization is 1.

Server Utilization

- 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.
  - For systems in statistical equilibrium, the average number of busy servers, L_s, is:

$$L_S = \lambda E(S) = \frac{\lambda}{\mu}$$

  - Clearly $0 \le L_S \le c$
  - The long-run average server utilization is:

$$\rho = \frac{L_s}{c} = \frac{\lambda}{c\mu}, \quad \text{where } \lambda < c\mu \text{ for stable systems}$$

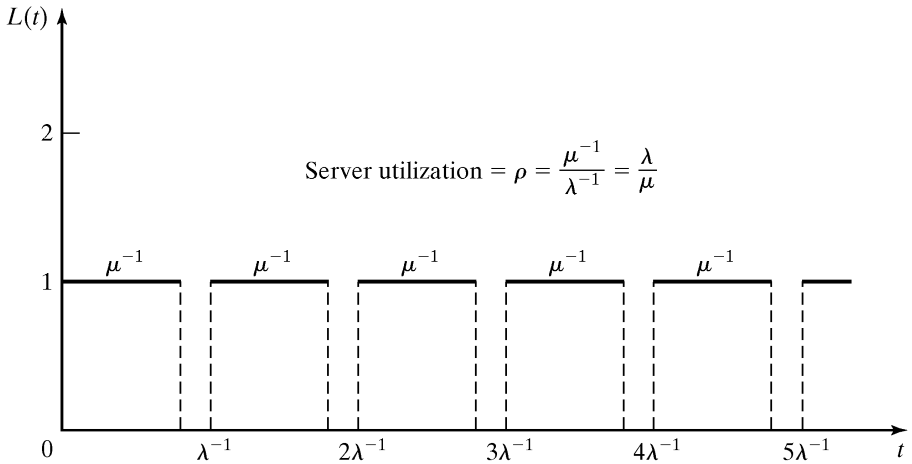
Server Utilization and System Performance

- System performance varies widely for a given utilization ρ.
  - For example, a $D/D/1$ queue where $E(A) = 1/\lambda$ and $E(S) = 1/\mu$, where:

$$L = \rho = \lambda/\mu, \quad w = E(S) = 1/\mu, \quad L_Q = W_Q = 0$$

    - By varying λ and μ, server utilization can assume any value between 0 and 1.

  - In general, variability of interarrival and service times causes lines to fluctuate in length.



Server Utilization and System Performance

- 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:

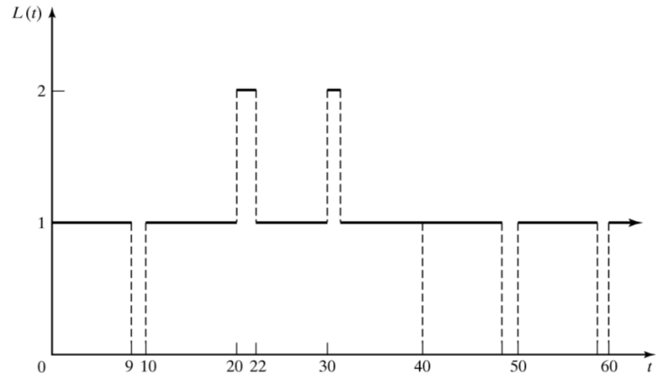
$$A_1 = A_2 = \ldots = \lambda^{-1} = 10$$

- Services are stochastic
  - $E(S_i) = 9.3$ min
  - $V(S_0) = 0.81$ min^2
  - $\sigma = 0.9$ min

- On average, the physician's utilization is
$\rho = \lambda/\mu = 0.93 < 1$

- 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

- 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:

W_j^Q is the time customer j spends in queue

$$\sum_{j=1}^{N} \frac{\$10 \cdot W_j^Q}{N} = \$10 \cdot \hat{w}_Q$$

    - If $\hat{\lambda}$ customers per hour arrive (on average), the average cost per hour is:

$$\left(\hat{\lambda}\frac{\text{customer}}{\text{hour}}\right)\left(\frac{\$10 \cdot \hat{w}_Q}{\text{customer}}\right) = \$10 \cdot \hat{\lambda} \cdot \hat{w}_Q = \frac{\$10 \cdot \hat{L}_Q}{\text{hour}}$$

  - Server may also impose costs on the system, if a group of c parallel servers ($1 \leq c \leq \infty$) have utilization ρ, each server imposes a cost of \$5 per hour while busy.
    - The total server cost is: $\$5 \cdot c \cdot \rho$

Steady-state Behavior of Infinite-Population Markovian Models

Steady-State Behavior of Markovian Models

- Markovian models:
  - Exponential-distributed arrival process (mean arrival rate $= 1/\lambda$).
  - Service times may be exponentially (M) or arbitrary (G) distributed.
  - Queue discipline is FIFO.
  - 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 Markovian Models

- Properties of processes with statistical equilibrium
  - The state of statistical equilibrium is reached from any starting state.
  - The process remains in statistical equilibrium once it has reached it.

Steady-State Behavior of 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, $L=\lambda w$, to the whole system and to the queue alone:

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

- $G/G/c/\infty/\infty$ example: to have a statistical equilibrium, a necessary and sufficient condition is:

$$\rho = \frac{\lambda}{c\mu} < 1$$

M/G/1 Queues

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

$$\rho = \frac{\lambda}{\mu}$$

$$P_0 = 1 - \rho$$

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

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

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

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

The particular distribution is not known!

ρ	server utilization
P_0	probability of empty system
L	long-run time-average number of customers in system
w	long-run average time spent in system per customer
L_Q	long-run time-average number of customers in queue
w_Q	long-run average time spent in queue per customer

M/G/1 Queues

- There are no simple expressions for the steady-state probabilities $P_0, P_1, P_2, \ldots$
- $L - L_Q = \rho$ 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 λ and μ are held constant, L_Q depends on the variability, σ^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:

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

    - The proportion of arrivals who find Able idle and thus experience no delay is $P_0 = 1\text{-}\rho = 1/5 = 20\%$.
  - 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 $P_0 = 1\text{-}\rho = 1/6 = 16.7\%$.
  - 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

$$\rho = \frac{\lambda}{\mu}$$

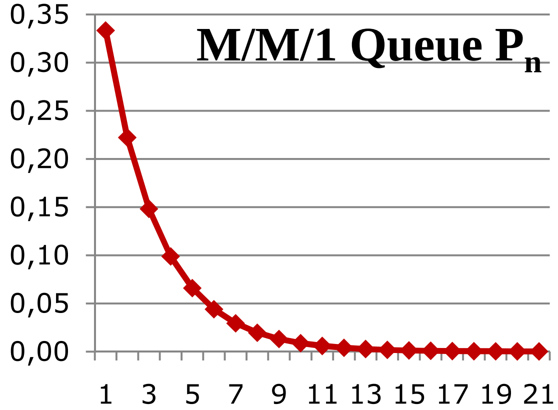
$$P_n = (1-\rho)\rho^n \quad \longrightarrow \quad P_0 = 1-\rho$$

$$L = \frac{\lambda}{\mu-\lambda} = \frac{\rho}{1-\rho}$$

$$w = \frac{1}{\mu-\lambda} = \frac{1}{\mu(1-\rho)}$$

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

$$w_Q = \frac{\lambda}{\mu(\mu-\lambda)} = \frac{\rho}{\mu(1-\rho)}$$



ρ	server utilization
P_0	probability of empty system
L	long-run time-average number of customers in system
w	long-run average time spent in system per customer
L_Q	long-run time-average number of customers in queue
w_Q	long-run average time spent in queue per customer

M/M/1 Queues

- Single-chair unisex hair-styling shop
  - Interarrival and service times are exponentially distributed
  - $\lambda = 2$ customers/hour and $\mu = 3$ customers/hour

$$\rho = \frac{\lambda}{\mu} = \frac{2}{3}$$

$$P_0 = 1 - \rho = \frac{1}{3}$$

$$P_1 = \frac{1}{3} \cdot \left(\frac{2}{3}\right)^1 = \frac{2}{9}$$

$$P_2 = \frac{1}{3} \cdot \left(\frac{2}{3}\right)^2 = \frac{4}{27}$$

$$P_{\geq 4} = 1 - \sum_{n=0}^{3} P_n = \frac{16}{81}$$

$$L = \frac{\lambda}{\mu - \lambda} = \frac{2}{3-2} = 2 \text{ Customers}$$

$$w = \frac{L}{\lambda} = \frac{2}{2} = 1 \text{ hour}$$

$$w_Q = w - \frac{1}{\mu} = 1 - \frac{1}{3} = \frac{2}{3} \text{ hour}$$

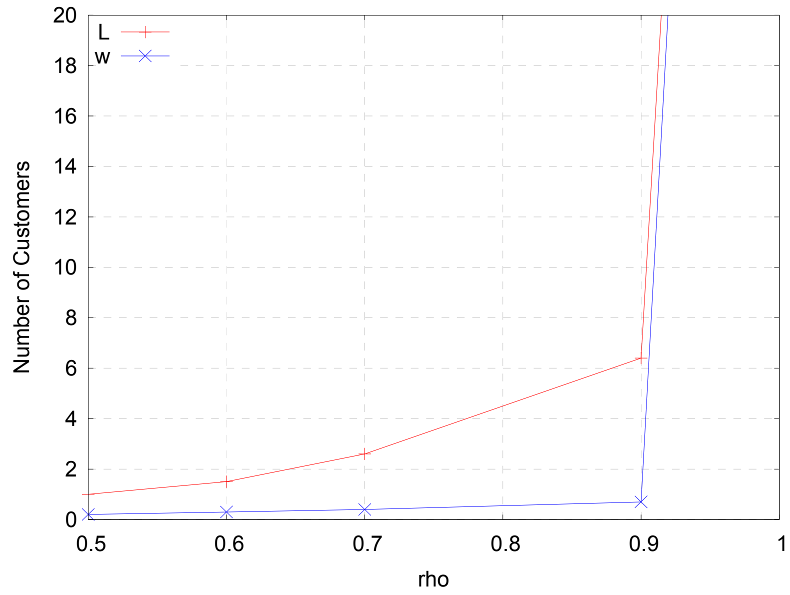
$$L_Q = \frac{\lambda^2}{\mu(\mu - \lambda)} = \frac{4}{3(3-2)} = \frac{4}{3} \text{ Customers}$$

$$L = L_Q + \frac{\lambda}{\mu} = \frac{4}{3} + \frac{2}{3} = 2 \text{ Customers}$$

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, λ, increases from 5 to 8.64 by increments of 20%
  - 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 ρ.

λ	5	6	7.2	8.64	10
ρ	0.5	0.60	0.72	0.864	1
L	1.0	1.50	2.57	6.350	∞
w	0.2	0.25	0.36	0.730	∞



Effect of Utilization and Service Variability

- For almost all queues, if lines are too long, they can be reduced by decreasing server utilization (ρ) or by decreasing the service time variability (σ^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
  - For exponential service times with rate μ
    - $E(X) = 1/\mu$
    - $V(X) = 1/\mu^2$

    ➡ $cv = 1$

Effect of Utilization and Service Variability

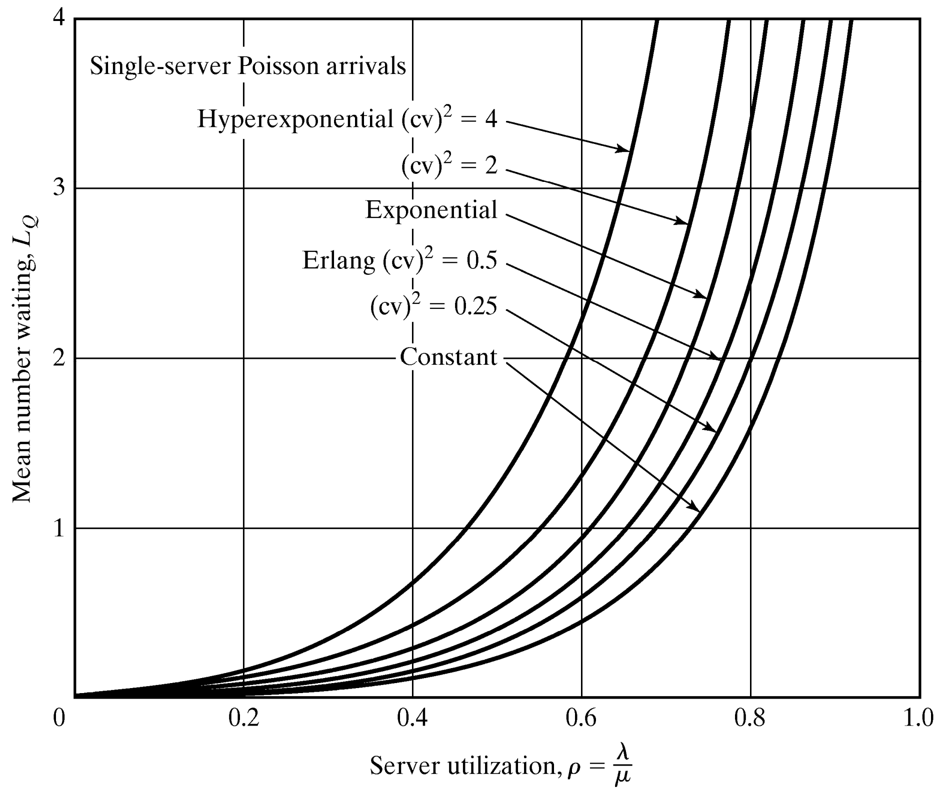
- Consider L_Q for any $M/G/1$ queue:

For any M/G/1
$(cv)^2 = \sigma^2/(1/\mu)^2 = \sigma^2\mu^2$

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

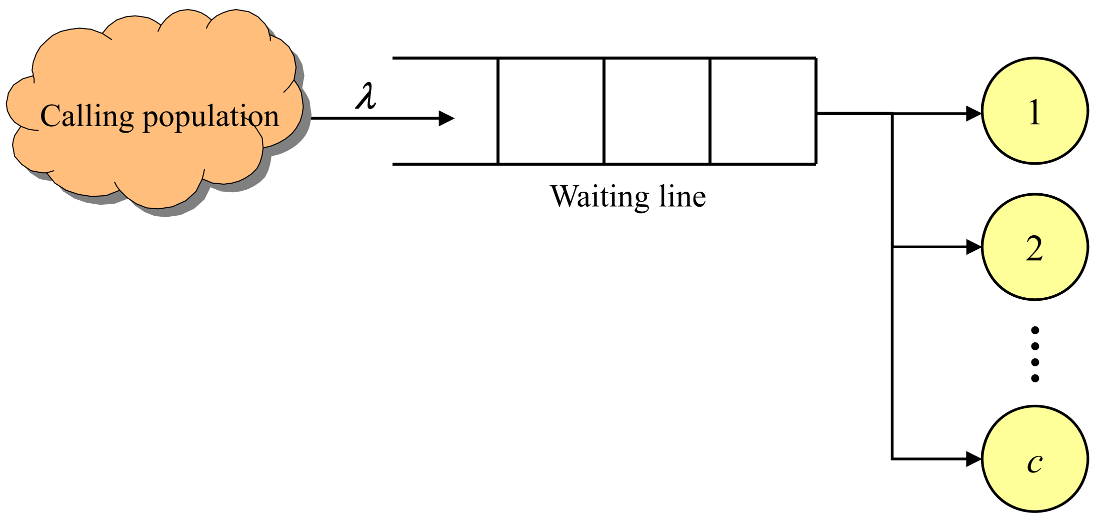
L_Q for $M/M/1$ queue

Corrects the $M/M/1$ formula to account for a non-exponential service time dist'n



Multiserver Queue: $M/M/c$

- $M/M/c/\infty/\infty$ queue: c servers operating in parallel
  - Arrival process is poisson with rate λ
  - Each server has an independent and identical exponential service-time distribution, with mean $1/\mu$.
  - To achieve statistical equilibrium, the offered load (λ/μ) must satisfy $\lambda/\mu < c$, where $\lambda/(c\mu) = \rho$ is the server utilization.



Multiserver Queue: *M/M/c*

- The steady-state parameters for *M/M/c*

Probability that all servers are busy

$$\rho = \frac{\lambda}{c\mu}$$

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

$$P\big(L(\infty) \ge c\big) = \frac{(c\rho)^c P_0}{c!(1-\rho)}$$

$$L = c\rho + \frac{(c\rho)^{c+1} P_0}{c(c!)(1-\rho)^2} = c\rho + \frac{\rho \cdot P\big(L(\infty) \ge c\big)}{1-\rho}$$

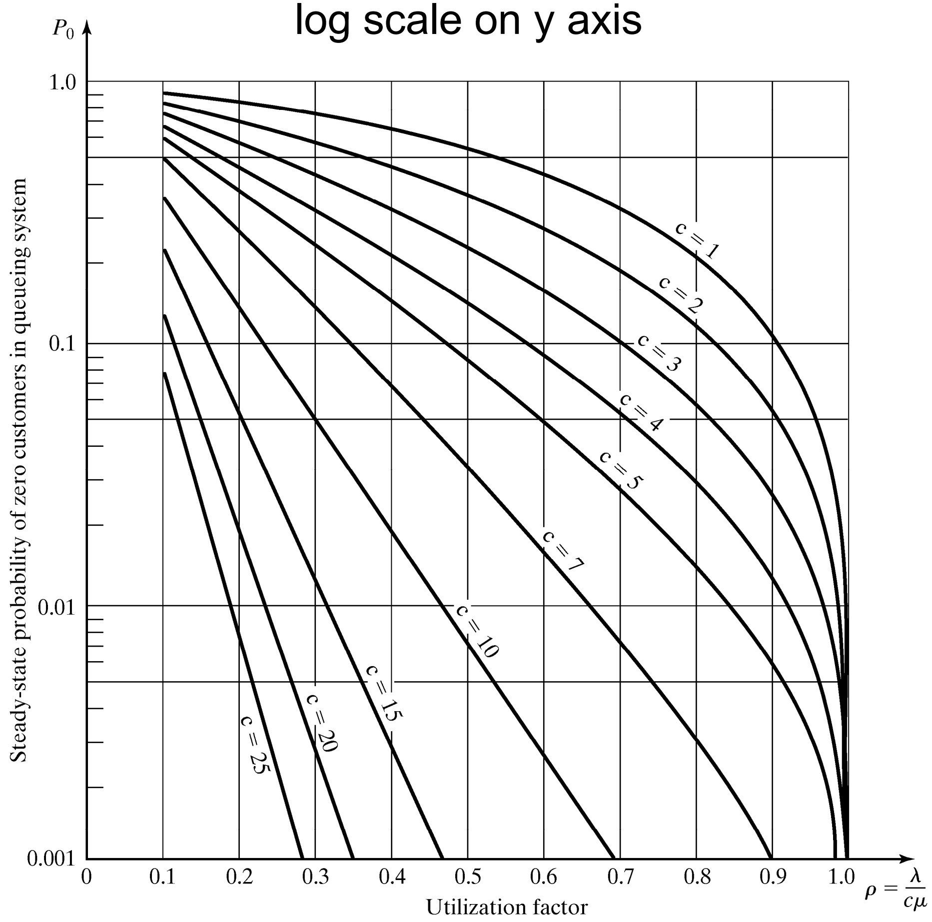
$$w = \frac{L}{\lambda}$$

$$L_Q = \frac{\rho \cdot P\big(L(\infty) \ge c\big)}{1-\rho}$$

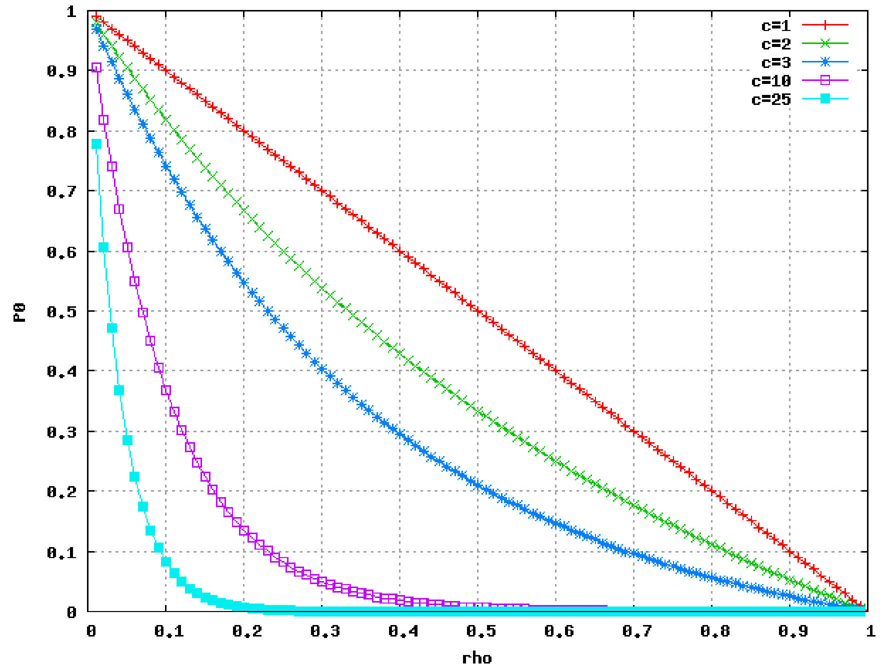
$$L - L_Q = c\rho$$

Multiserver Queue: *M/M/c*

Probability of empty system
log scale on y axis

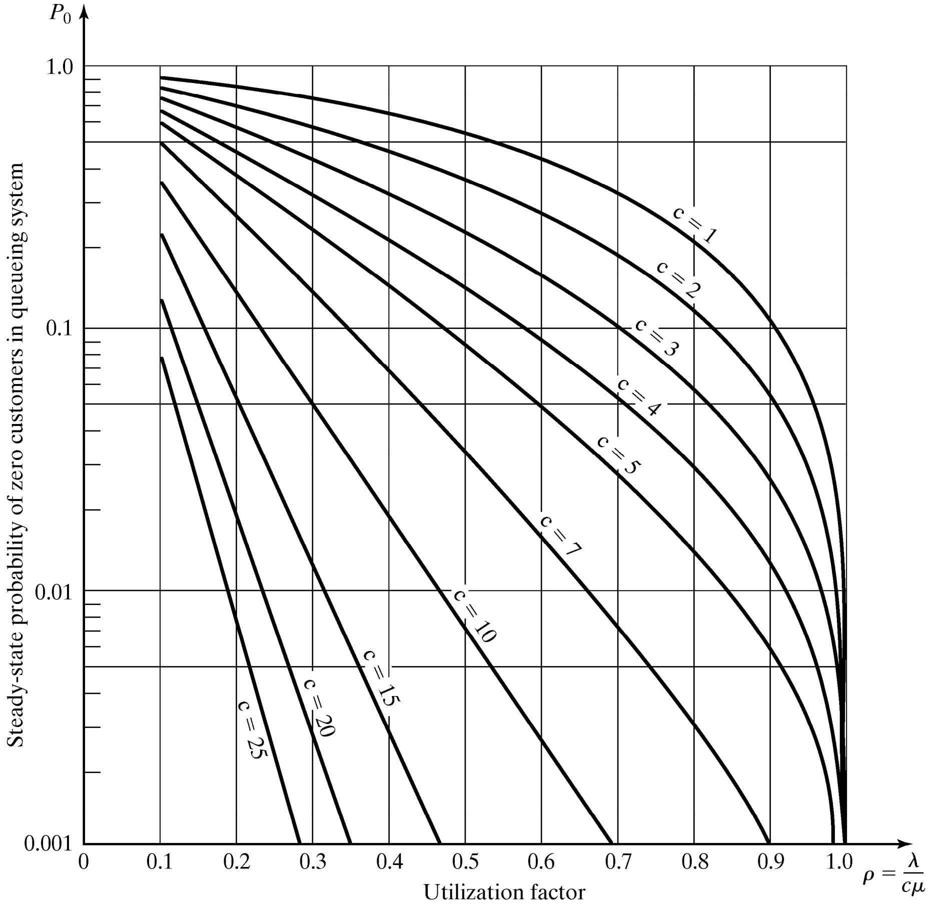


Probability of empty system

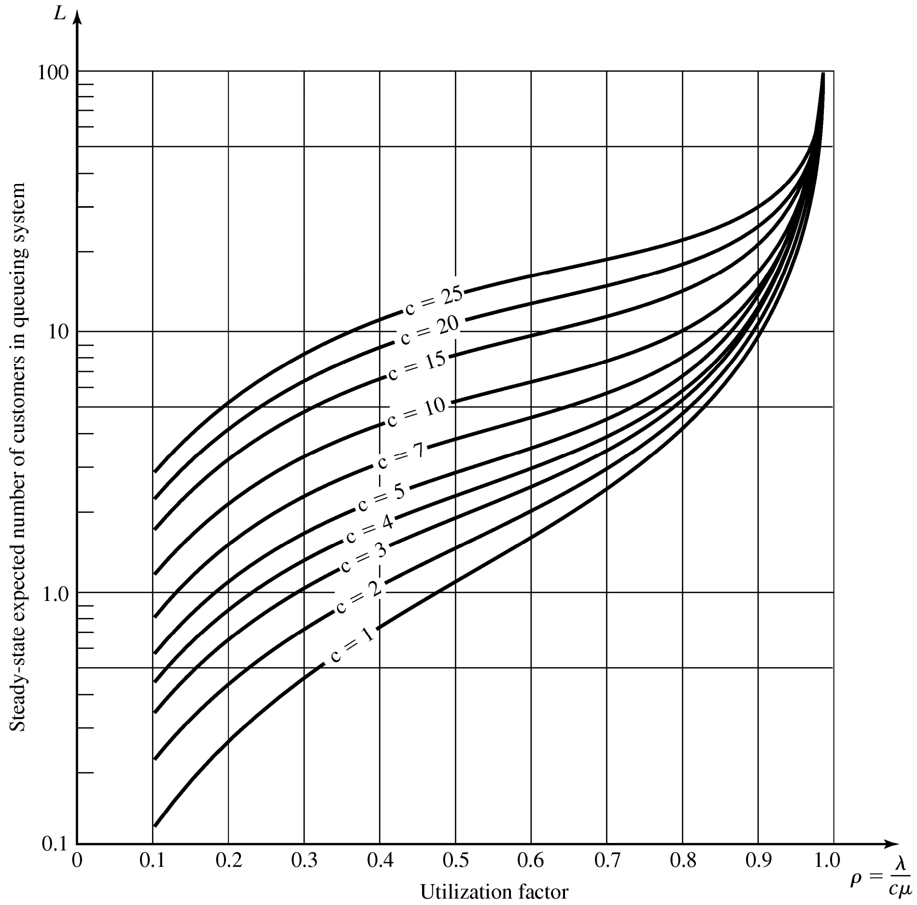


Multiserver Queue: $M/M/c$

Probability of empty system



Number of customers in system



Multiserver Queue: Common Models

- Other common multiserver queueing models

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

L_Q for $M/M/1$ queue

Corrects the $M/M/1$ formula

- $M/G/c/\infty$: 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.
- $M/M/c/N/\infty$: service times are exponentially distributed at rate μ 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.

Multiserver Queue: $M/G/\infty$

- $M/G/\infty$: general service times and infinite number of servers
  - customer is its own server
  - service capacity far exceeds service demand
  - when we want to know how many servers are required so that customers are rarely delayed

$$P_n = e^{-\frac{\lambda}{\mu}} \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!}, n = 0,1,\ldots$$

$$P_0 = e^{-\frac{\lambda}{\mu}}$$

$$w = \frac{1}{\mu}$$

$$w_Q = 0$$

$$L = \frac{\lambda}{\mu}$$

$$L_Q = 0$$

Multiserver Queue: $M/G/\infty$

- How many users can be logged in simultaneously in a computer system
  - Customers log on with rate $\lambda = 500$ per hour
  - Stay connected in average for $1/\mu = 180$ minutes $= 3$ hours
  - For planning purposes it is pretended that the simultaneous logged in users is infinite
  - Expected number of simultaneous users L

$$L = \frac{\lambda}{\mu} = 500 \cdot 3 = 1500$$

  - To ensure providing adequate capacity 95% of the time, the number of parallel users c has to be restricted

$$P(L(\infty) \le c) = \sum_{n=0}^{c} P_n = \sum_{n=0}^{c} \frac{e^{-1500}(1500)^n}{n!} \ge 0.95$$

  - The capacity $c = 1564$ simultaneous users satisfies this requirement

Multiserver Queue with Limited Capacity

- $M/M/c/N/\infty$: service times are exponentially distributed at rate μ 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
  - Effective arrival rate λ_e is defined as the mean number of arrivals per time unit who enter and remain in the system

$$a = \frac{\lambda}{\mu}$$

$$\rho = \frac{\lambda}{c\mu}$$

$$P_0 = \left[1 + \sum_{n=1}^{c} \frac{a^n}{n!} + \frac{a^c}{c!} \sum_{n=c+1}^{N} \rho^{n-c} \right]^{-1}$$

$$P_N = \frac{a^N}{c!c^{N-c}} P_0$$

$$L_Q = \frac{P_0 a^c \rho}{c!(1-\rho)} \left(1 - \rho^{N-c} - (N-c)\rho^{N-c}(1-\rho) \right)$$

$$\lambda_e = \lambda(1 - P_N)$$

$$w_Q = \frac{L_Q}{\lambda_e}$$

$$w = w_Q + \frac{1}{\mu}$$

$$L = \lambda_e w$$

($1 - P_N$) probability that a customer will find a space and be able to enter the system

Multiserver Queue with Limited Capacity

Single-chair unisex hair-styling shop (again!)

- Space only for 3 customers: one in service and two waiting
- First compute P_0

$$P_0 = \frac{1}{\left[1 + \frac{2}{3} + \frac{2}{3}\sum_{n=2}^{3}\left(\frac{2}{3}\right)^{n-1}\right]} = 0.415$$

- P(system is full)

$$P_N = P_3 = \frac{\left(\frac{2}{3}\right)^3}{1!1^2} P_0 = \frac{8}{65} = 0.123$$

- Average of the queue

$$L_Q = 0.431$$

- Effective arrival rate

$$\lambda_e = 2\left(1 - \frac{8}{65}\right) = \frac{114}{65} = 1.754$$

- Queue time

$$w_Q = \frac{L_Q}{\lambda_e} = \frac{28}{114} = 0.246$$

- System time, time in shop

$$w = w_Q + \frac{1}{\mu} = \frac{66}{114} = 0.579$$

- Expected number of customers in shop

$$L = \lambda_e w = \frac{66}{65} = 1.015$$

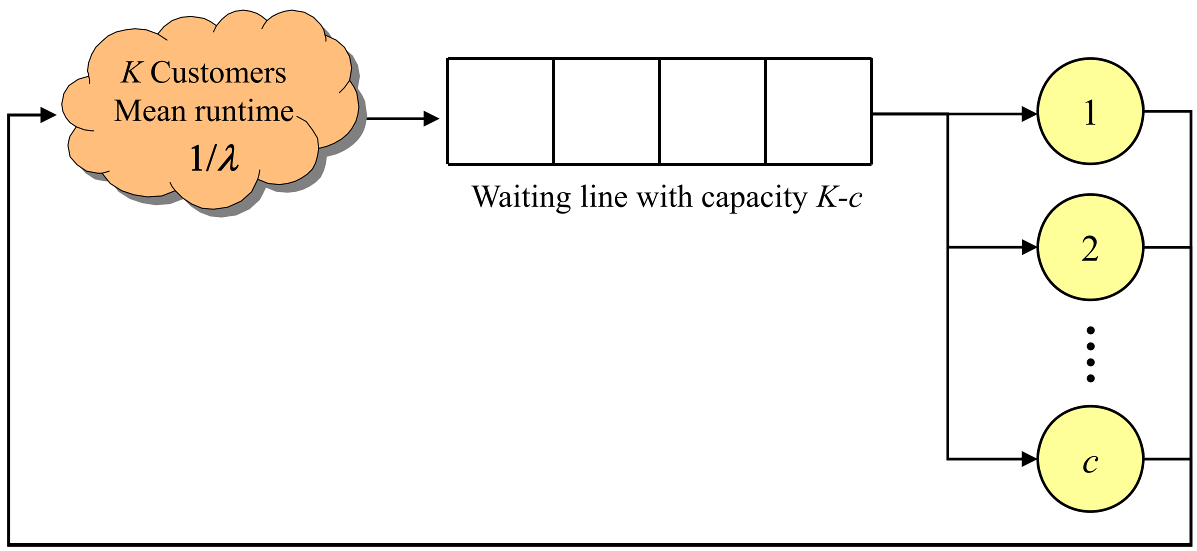
- Probability of busy shop

$$1 - P_0 = \frac{\lambda_e}{\mu} = 0.585$$

Steady-state Behavior of Finite-Population Models

Steady-State Behavior of Finite-Population Models

- In practical problems calling population is finite
  - 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 with mean $= 1/\lambda$.
  - Service times are also exponentially distributed with mean $1/\mu$.
  - c parallel servers and system capacity is K.



Steady-State Behavior of Finite-Population Models

- Some of the steady-state probabilities of $M/M/c/K/K$:

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

$$P_n = \begin{cases} \binom{K}{n} \left(\frac{\lambda}{\mu}\right)^n P_0, & n = 0,1,\ldots,c-1 \\ \frac{K!}{(K-n)!c!c^{n-c}} \left(\frac{\lambda}{\mu}\right)^n, & n = c, c+1, \ldots K \end{cases}$$

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

where λ_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: $\lambda = 1/20$ and $\mu = 1/5$.
  - All of the performance measures depend on P_0:

$$P_0 = \left[\sum_{n=0}^{2-1} \binom{10}{n} \left(\frac{5}{20} \right)^n + \sum_{n=2}^{10} \frac{10!}{(10-n)!2!2^{n-2}} \left(\frac{5}{20} \right)^n \right]^{-1} = 0.065$$

  - Then, we can obtain the other P_n, and can compute the expected number of machines in system:

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

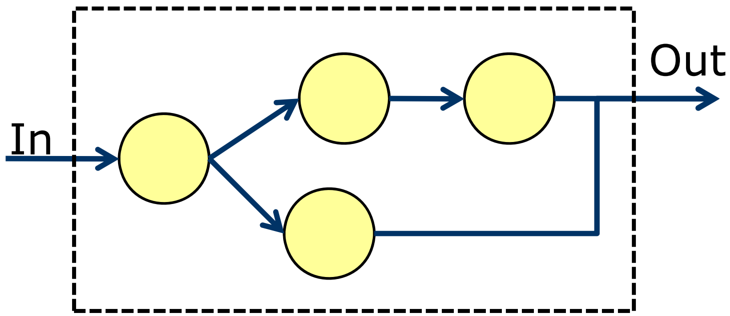
  - The average number of running machines:

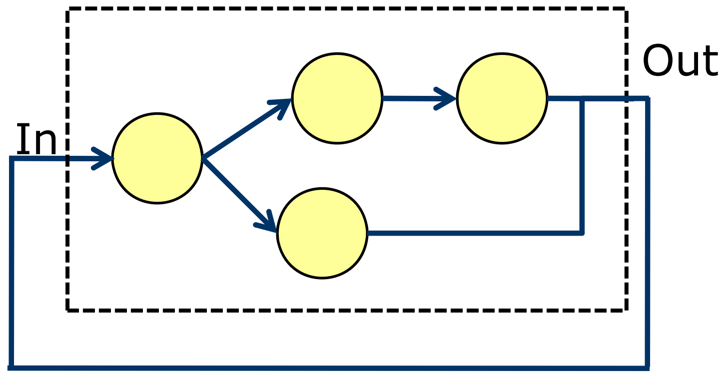
$$K - L = 10 - 3.17 = 6.83 \text{ machines}$$

Networks of Queues

Networks of Queues

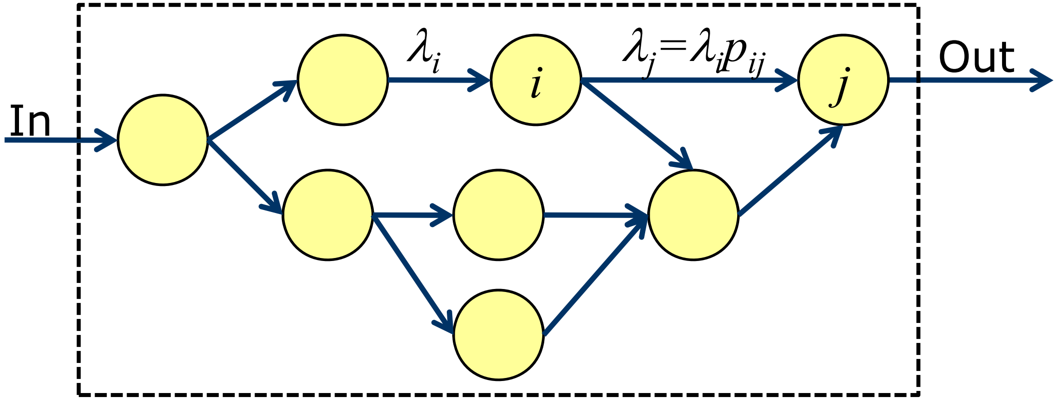
- No simple notation for networks of queues
- Two types of networks of queues
  - Open queueing network
    - External arrivals and departures
    - Number of customers in system varies over time
  - Closed queueing network
    - No external arrivals and departures
    - Number of customers in system is constant





Networks of Queues

- Many systems are 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 λ_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 $\lambda_j = \lambda_i p_{ij}$ over the long run.

Networks of Queues

- The overall arrival rate into queue j:

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

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 μ_j, then the long-run utilization of each server is: (where $\rho_j < 1$ for stable queue).

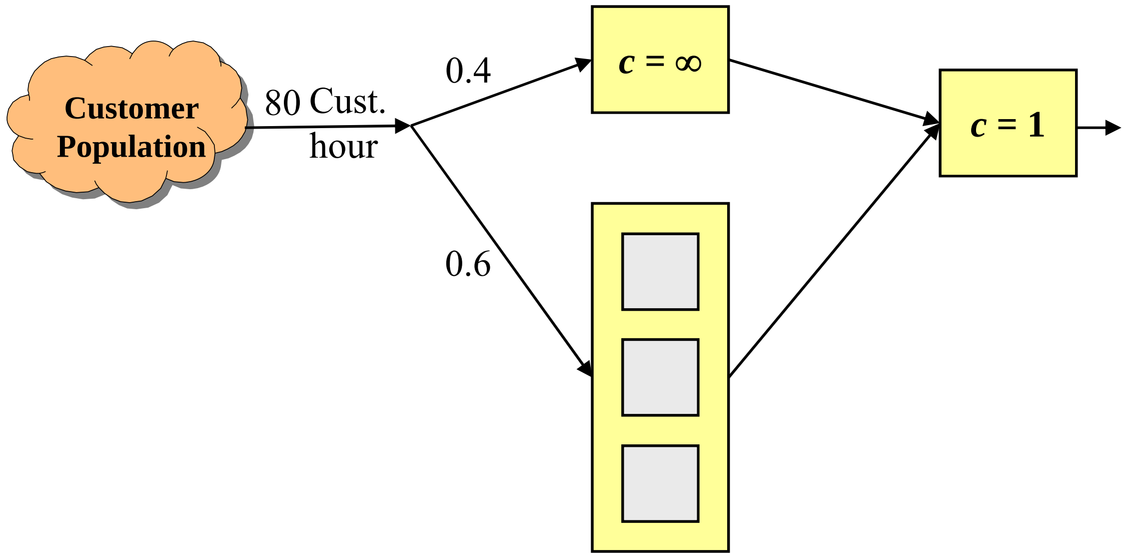
$$\rho_j = \frac{\lambda_j}{c_j \mu_j}$$

- 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 likes an $M/M/c_j$ queue with arrival rate

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

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:

$$\rho_2 = 48/(3\times 20) = 0.8$$

  - 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:

$$\rho_3 = 80/90 = 0.89$$

Summary

- Introduced basic concepts of queueing models.
- Showed 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, ρ, and λ_e.
- When simulating any system that evolves over time, analyst must decide whether to study **transient** 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.