Queueing Models
to be used in Simulation

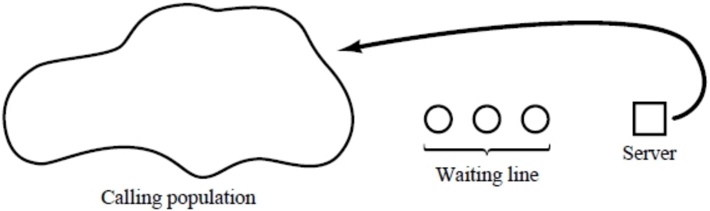
Radu Trîmbițaș

UBB

1st Semester 2011-2012

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 theories):

- General characteristics of queues,

Outline

Discuss some well-known models (not development of queueing theories):

- General characteristics of queues,
- Meanings and relationships of important performance measures,

Outline

Discuss some well-known models (not development of queueing theories):

- General characteristics of queues,
- Meanings and relationships of important performance measures,
- Estimation of mean measures of performance.

Outline

Discuss some well-known models (not development of queueing theories):

- General characteristics of queues,
- Meanings and relationships of important performance measures,
- Estimation of mean measures of performance.
- Effect of varying input parameters,

Outline

Discuss some well-known models (not development of queueing theories):

- General characteristics of queues,
- Meanings and relationships of important performance measures,
- Estimation of mean measures of performance.
- Effect of varying input parameters,
- Mathematical solution of some basic queueing models.

Key elements 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.
  - Server: refers to any resource that provides the requested service, e.g., repairpersons, retrieval machines, runways at airport.

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

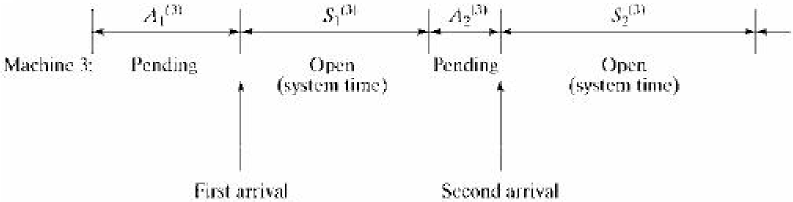
- **System Capacity**: a limit on the number of customers that may be in the waiting line or system.
  - Limited capacity, e.g., an automatic car wash only has room for 10 cars to wait in line to enter the mechanism.
  - Unlimited capacity, e.g., concert ticket sales with no limit on the number of people allowed to wait to purchase tickets.

Arrival Processes - Infinite population models

- In terms of interarrival times of successive customers.
- Random arrivals: interarrival times usually characterized by a probability distribution.
    - Most important model: Poisson arrival process (with rate λ), where A_n 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.
    - e.g., 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 Processes - 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 form 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.
- Let $A_1^{(i)}$, $A_2^{(i)}$, ... 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 it's 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 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.

Service Times and Service Mechanism II

- Example: consider a discount warehouse where customers may:
- Serve themselves before paying at the cashier:

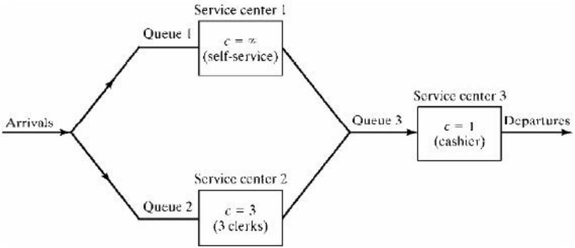
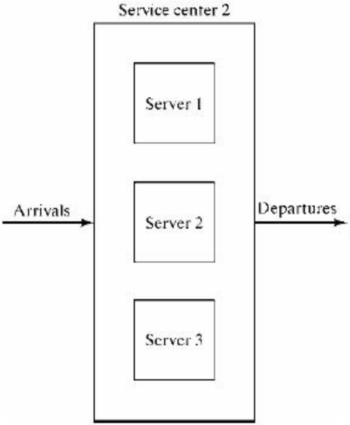


Figure: Warehouse example

Service Times and Service Mechanism III

- Wait for one of the three clerks:



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

Queueing notation I

- A notation system for parallel server queues: $A/B/c/N/K$, (due to Kendall) where
- 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.

Queueing notation II

- Primary 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 nth arriving customer,
  - W_n: total time spent in system by the nth arriving 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.

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 a system is defined by:

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

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

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

- The long-run time-average # in system, with probability 1:

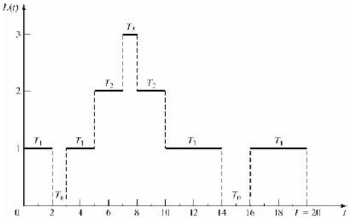
$$\widehat{L} = \frac{1}{T}\int_0^T L(t)dt \to L \text{ as } T \to \infty$$

Time-Average Number in Queue

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

$$\widehat{L}_Q = \frac{1}{T}\sum_{i=0}^{\infty} iT_i^Q = \frac{1}{T}\int_0^T L_Q(t)dt \to L_Q \qquad \text{as } T \to \infty$$

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



$$\widehat{L} = (0\cdot 3 + 1\cdot 12 + 2\cdot 4 + 3\cdot 1)/20$$
$$= 1.15 \text{ customers}$$

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

$$L_Q(t) = \frac{0\cdot 15 + 1\cdot 4 + 2\cdot 1}{20} = 0.3 \text{ customers}$$

Average Time Spent in System Per Customer

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

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

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

- for stable systems $\widehat{w} \to w$ as $N \to \infty$
- If the system under consideration is the queue alone:

$$\widehat{w}_Q = \frac{1}{N} \sum_{i=1}^{N} W_i^Q \to w_Q \text{ as } N \to \infty$$

- G/G/1/N/K example (cont.): the average system time is

$$\widehat{w} = \frac{\sum_{i=1}^{5} W_i}{5} = \frac{2 + (8-3) + \cdots + (20-16)}{5} = 4.6 \text{ time units}$$

The Conservation Equation

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

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

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

where $\widehat{L}$ average # in system, $\widehat{\lambda}$ arrival rate, $\widehat{w}$ average system time

- 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 is $(1/4)(4.6) = 1.15$ customers present on average.

Server Utilization I

- Definition: the proportion of time that a server is busy.
  - Observed server utilization, $\widehat{\rho}$, is defined over a specified time interval $[0, T]$.
  - Long-run server utilization is ρ.
  - For systems with long-run stability: $\widehat{\rho} \to \rho$ as $T \to \infty$
- For $G/G/1/\infty/\infty$ queues:
  - Any single-server queueing system with average arrival rate λ customers per time unit, where average service time $E(S) = 1/\mu$ time units, 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:

$$\widehat{L}_s = \frac{1}{T}\int_0^T (L(t) - L_Q(t))\, dt = \frac{T - T_0}{T}$$

Server Utilization II

- In general, for a single-server queue:

$$\widehat{L}_s = \widehat{\rho} \to L_s = \rho \text{ as } T \to \infty$$

and

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

  - For a single server stable queue

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

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

Server Utilization III

- 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: $L_s = \lambda E(s) = \lambda/\mu$.
  - The long-run average server utilization is:

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

Server Utilization and System Performance

- System performance varies widely for a given utilization ρ.
  - For example, a D/D/1 queue with

$$E(A) = 1/\lambda$$
$$E(S) = 1/\mu,$$

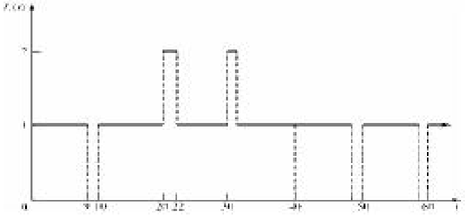
    where:

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

  - By varying λ and μ, 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

- Example: A physician who schedules patients every 10 minutes and spends S_i minutes with the ith patient: $S_i = \begin{pmatrix} 9 & 12 \\ 0.9 & 0.1 \end{pmatrix}$
- Arrivals are deterministic, $A_1 = A_2 = \ldots = \lambda^{-1} = 10$.
- Services are stochastic, $E(S_i) = 9.3$ min and $V(S_0) = 0.81 \text{ min}^2$.
- On average, the physician's utilization $= \rho = \lambda/\mu = 0.93 < 1$.
- Consider the system is simulated with service times: $S1 = 9$, $S2 = 12$, $S3 = 9$, $S4 = 9$, $S5 = 9$, $\ldots$. 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):

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

  - If λ customers per hour arrive (on average), the average cost per hour is:

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

- Server may also impose costs on the system, if a group of c parallel servers ($1 \leq c \leq \infty$) have utilization r, 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 I

- 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 II

- 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:

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

- $G/G/c/\infty/\infty$ 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 σ^2 and $\rho = \lambda/\mu < 1$, the steady-state parameters of $M/G/1$ queue:

$$\rho = \frac{\lambda}{\mu}, \qquad P_0 = 1 - \rho$$

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

$$w = \frac{1}{\mu} + \frac{\lambda \left(\frac{1}{\mu^2} + \sigma^2\right)}{2\left(1-\rho\right)}, \qquad w_Q = \frac{\lambda \left(\frac{1}{\mu^2} + \sigma^2\right)}{2\left(1-\rho\right)}$$

M/G/1 Queues II

- 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 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:

$$L_Q = \frac{(1/30)^2\left(24^2 + 400\right)}{2\left(1 - \frac{4}{5}\right)} = 2.7111 \text{ customers}$$

  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:

$$L_Q = \frac{(1/30)^2\left(25^2 + 4\right)}{2\left(1 - \frac{5}{6}\right)} = 2.0967 \text{ customers}$$

  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 I

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

$$\begin{aligned}
\rho &= \frac{\lambda}{\mu}, \\
P_n &= \rho\left(1-\rho\right)^n \\
L &= \frac{\lambda}{\mu-\lambda} = \frac{\rho}{1-\rho}, \qquad L_Q = \frac{\lambda^2}{\mu\left(\mu-\lambda\right)} = \frac{\rho^2}{1-\rho} \\
w &= \frac{1}{\mu-\lambda} = \frac{1}{\mu\left(1-\rho\right)}, \qquad w_Q = \frac{\lambda}{\mu\left(\mu-\lambda\right)} = \frac{\rho}{\mu\left(1-\rho\right)}
\end{aligned}$$

M/M/1 Queues II

- **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.

λ	5	6	7.2	8.64	10
ρ	0.5	0.6	0.72	0.864	1.000
L	1	1.5	2.57	6.35	∞
w	0.2	0.25	0.36	0.73	∞

- Increase in average system time, w, and average number in system, L, is highly nonlinear as a function of ρ

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):

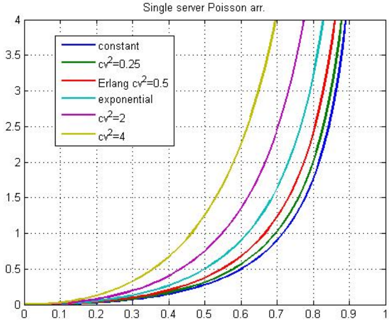
$$cv = \frac{V(X)}{E(X)^2}$$

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

Effect of Utilization and Service Variability II

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

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



Multiserver Queue

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 I

- 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 II

- Some of the steady-state probabilities:

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

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

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

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 III

- **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:

$$P_0 = \left[\sum_{n=0}^{2-1} \binom{10}{n} \left(\frac{1}{4}\right)^n + \sum_{n=2}^{10} \frac{10!}{(10-n)!2!2^{n-2}} \left(\frac{1}{4}\right)^n\right]^{-1} = 0.0647$$

  - 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 I

- 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 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 $\lambda_i p_{ij}$ (over the long run).
- The overall arrival rate into queue j:

$$\lambda_j = a_j + \sum_i \lambda_i p_{ij}$$

Networks of Queues II

- If queue j has $c_j < \infty$ parallel servers, each working at rate μ_j, then the long-run utilization of each server is $\rho_j = \frac{\lambda_j}{c\mu_j}$ (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 likes an $M/M/c_j$ queue with arrival rate

$$\lambda_j = a_j + \sum_i \lambda_i p_{ij}$$

- **Discount store example:** see Figure 1
- 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.

Networks of Queues III

- $c_2 = 3$ clerks and $\mu_2 = 20$ customers per hour.
- The long-run utilization of the clerks is:

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

- 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.
- Show how simulation, and some times 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 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.