Overview
of
Queuing Models

Queuing Models

- Whether solved mathematically or analyzed through simulation, QM provide useful info for designing & evaluating the performance of queuing systems.
- Typical measures:
  - Server utilization
  - Length of waiting lines
  - Delays of customers

  as a function of input parameters such as arrival rate, service demands, service rate, # of servers.
- Involves trade-offs analysis between server utilization, customer satisfaction (line lengths & delays)

Queuing Models

- For simple systems, measures can be computed mathematically or analytically.
  - Output measures are related to input parameters by a set of mathematical formulae.
  - Easier to solve
  - More restrictive assumptions
  - Rough-cut estimates
- For more complex systems, simulation is usually required.

Characteristics of Queuing System

- Two key elements: customers and servers.
- Calling or customer population: population of potential customers
  - Finite (closed system) or infinite (open system) depending on the arrival rate.
- System capacity
  - Limited: arrival rate vs. effective arrival rate
- Arrival Process / interarrival time distribution
  - Scheduled or random (Poisson process)
- Queue behavior
  - Leave, stay, or move
- Queue discipline
  - FCFS/FIFO, RR (round robin), Priority, …
- Service time distribution / mechanism

Queuing Notation

- Kendall's notation

  A/B/c/N/K
  - A: interarrival time distribution
    - M (Markovian, memoryless, exponential)
    - G (general or arbitrary)
    - D (constant or deterministic)
    - Ek (Erlang of order k)
  - B: service time distribution, same as A
  - c: number of servers
    - 1, multiple, infinite
  - N: system capacity or queue size
    - Finite, infinite
  - K: size of calling population
  - Table 6.2

Long-Run Measures of Performance of Queuing Systems

- Main long-run measures:
  - L: long-run time-average number of customers in the system
  - L_Q: in the queue
  - w: long-run time average time spent in the system
  - w_Q: in the queue

  ρ: server utilization
- Other measures:
  - long-run proportion of customers who are delayed in queue longer than t_0 time units
  - long-run proportion of customers turned away
  - long-run proportion of time the waiting line contains more than k_0 customers

Performance Measures of Queuing Systems

- Time-average number is system L: influenced by initial conditions at time 0 and the run length T.
  - L(t): number of customers in system at time t.
  - T_i: total time during [0,T] in which the system contained exactly i customers.
  - L hat: time-weighted-average number is system:
  - As T → ∞, L hat → L (long-run time-average number is system)
  - $L_Q(t)$: number of customers waiting in line
  - L_Q: long-run time-average number waiting in line.
  - L_Q hat: observed time-average number of customers in line from time 0 to time T.
  - T_i^Q : total time during [0,T] in which exactly i customers are waiting in line.

More on Performance Measures of QS

- Average time spent in system per customer, w
  - similar to L
- Conservation equation: $L = \lambda\omega$

  λ : long-run average arrival rate;

  λ hat: observed average arrival rate;

  λ hat -> λ as T → ∞ and N → ∞
  - The average number of customers in the system at an arbitrary point in time = the average number of arrivals per time unit * average time spent in the system.

More on Performance Measures of QS

- Server utilization ?: proportion of time that the server is busy.
- Server utilization in G/G/1/¥/¥
  - arrival rate: λ (customers per time unit)
  - service rate: μ (customers per time unit)
  - server can be considered as a queuing system in itself, so $L = \lambda\omega$ can be applied.
  - What is ω for the server subsystem, i.e., average server time? $\omega = \mu^{-1}$
  - L hat: observed average number is server subsystem
  - L_s: average number in server subsystem or busy servers
  - In general, for a single-server queue, $L_s = ? = \lambda\omega = \lambda/\mu$
    - also called the offered load; a measure of workload

More on Performance Measures of QS

- Server utilization in G/G/c/¥/¥
  - c identical servers in parallel: the choice of server might be made at random.
  - Maximum service rate for is $c\mu$: all servers are busy.
  - $L_s = \lambda E(S) = \lambda / \mu$ (average # of busy servers = c)
  - $? = L_s / c = \lambda / c\mu$ (long-run average server utilization = 1: proportion of time an arbitrary server is busy in the long run.)
  - For the system to be stable, $c > \lambda / \mu$
  - A stable queue can still have long lines.
  - Trade-off analysis: server utilization vs. customer satisfaction

More on Performance Measures of QS

- Costs in Queuing Problems
  - Cost can be associated with various aspects of the queue or servers.
  - $x: cost per hour per customer; $y per hour while busy
  - Average cost per customer: $x * w_Q hat
  - Average cost per hour: $x * L_Q hat / hour
  - Cost for a set of c parallel servers
    - Server is busy: $y * c?
    - Server is idle: $y * c(1 – ?)
  - Objective: minimize total costs by varying above parameters (# of servers, service rate, system capacity)

Steady-State of Infinite-Population Markovian Models

- Arrival: Poisson process with λ arrivals.
- Assumptions:
  - Arrivals occur one at a time with FIFO discipline.
  - The distribution of the #s of arrivals between t and t+s depends only on the length of the interval s and not on the starting point t.
  - The #s of arrivals during nonoverlapping time intervals are independent random variables. Or future arrivals occur completely at random, independent of the #s of arrivals in the past time intervals.
  - It has been shown that if interarrival times are exponentially & independently distributed, then the #s of arrivals is a Poisson process. (see section 5.5)
- These models are called Markovian models because of the exponential distribution assumptions

Steady-State of Infinite-Population Markovian Models

- $P_n(t)$: probability of n customers in system at time t.
- P_n: steady-state probability of having n customers in system
- A queuing system is in steady state if the system in a given state is independent of time t, i.e., $P_n(t) = P_n$.
- Sections deal with math models to get a rough guide of the system behaviors, e.g. L, instead of simulation models which delivers a statistical estimate, e.g. L hat.
- Two properties are important to consider for steady state: starting state and remaining in steady state once it is reached.

Steady-State of Infinite-Population Markovian Models

- M/G/1 (when N and K are infinite, they may be dropped from notation)
- Assumptions:
  - mean service times $1/\mu$ and variance σ^2 & one server
  - ? = $\lambda / \mu < 1$ → M/G/1 has a steady-state probability distribution.
  - Steady-state characteristics: Table 6.3.
  - What is P_0?
  - What is $1 - P_0$?
  - What is $L - L_Q$?

Steady-State of Infinite-Population Markovian Models

- M/M/1 queue
  - The service times are exponentially distributed.
  - Assumptions:
    - mean: $1/\mu$ → standard derivation = ?
    - variance: $1/\mu^2$
    - steady-state parameters shown in Table 6.4.
    - effect of ?, and L and w
      - examples 6.11 & 6.12
    - effect of utilization & service variability
      - coefficient of variation (cv):
        - » $(cv)^2 = V(x) / [E(x)]^2$
        - » Figure 6.12

Steady-State of Infinite-Population Markovian Models

- Relationship between M/G/1 and M/M/1, and M/G/c and M/M/c
  - Correction factor for L_Q and w_Q:
    - Rewrite L_Q and w_Q for M/G/1 queue (Table 6.3) in terms of the coefficient of variation and compare it with that for M/M/1 (Table 6.4)
    - Correction factor is $(1 + (cv)^2/2)$, see equation (6.19)
    - The same correction factor can also be applied to M/G/c and M/M/c to obtain L_Q and w_Q.

Networks of Queues

- In practice, many systems are modeled as networks of single queues in which customers departing one queue may be routed to another.
- Some basic principles:
  1. If no customers are created or destroyed in the queue, then departure rate = arrival rate over the long run.
  2. Arrival rate for queue i is λ_i and $0 = p_{ij} = 1$ of them are routed to queue j, then the arrival rate from queue i to j is $\lambda_i p_{ij}$.
  3. Overall rate into queue j is the sum of the arrival rate from all sources.
  4. If queue j has c_j parallel servers with service rate μ_j, then the utilization of each server is $?_j = \lambda_j / c_j \mu_j$.
  5. If, for queue j, arrivals from outside the network is Poisson process with rate a_j and there are c_j servers with exponentially distributed service times of mean $1/\mu_j$, then in steady state queue j behaves like an $M/M/c_j$ queue with arrival rate as stated in point 3.

Summary

- Queuing models are widely used to **estimate** desired performance measures of the system.
- The estimate contains random error, and thus a proper statistical analysis is required to assess the accuracy of the estimate.
- Mathematical models/solutions may not be practical, but provide a rough estimate of a performance measure.
- An important application of mathematical queuing models is determining the minimum number of servers needed at a service centre. $\lambda/c\mu < 1$ can be used to provide an initial estimate for the # of servers, c.
- Performance bottleneck or congestion may be detected using queuing modeling. Congestion may be decreased by adding more servers or by reducing the mean value and variability of service times.
- Queuing models can also be used to evaluate alternative system designs.