Computer system modeling and simulation

5. Queueing models

Sosina M.
Addis Ababa institute of technology (AAiT)
2012 E.C.

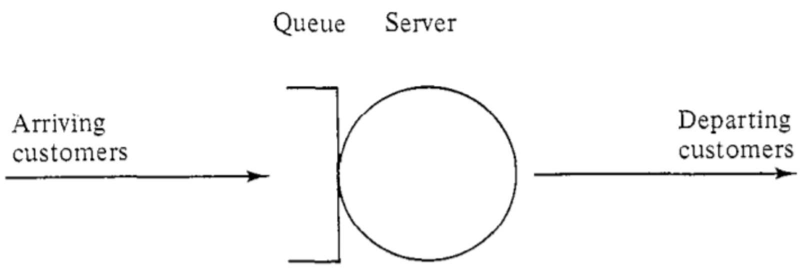
Queueing systems

- Queueing systems are models of systems providing service

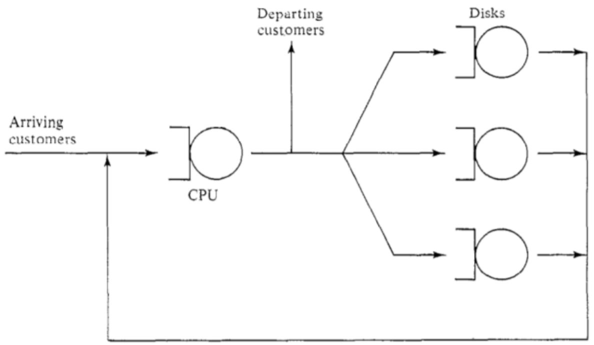
- Wide range of potential application areas
  - Vehicular traffic
    - Traffic signal, bottlenecks
  - Banking
  - Customer service
  - Communication
    - Transmission delay, medium access control, protocol evaluation
  - Computer systems
    - Parallel processing, client-server interaction , peer-to-peer
  - and so on

Queueing systems

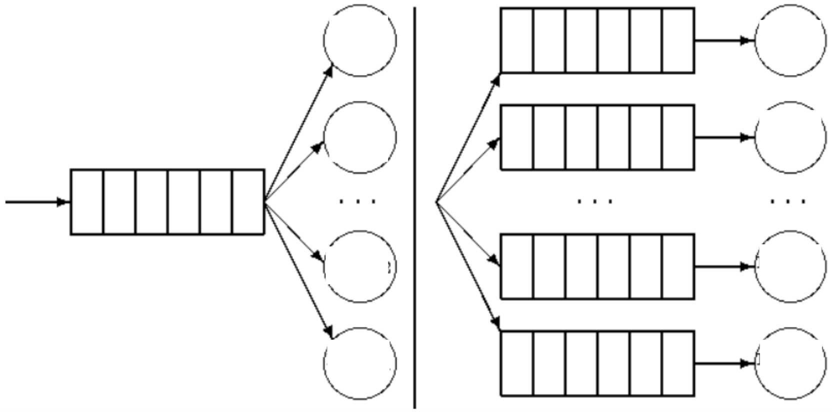
❑*A single queue system*



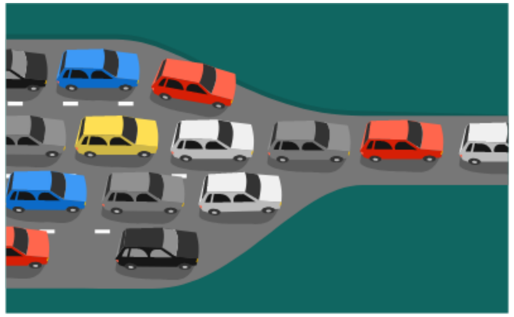
❑*A system of interconnected queues*

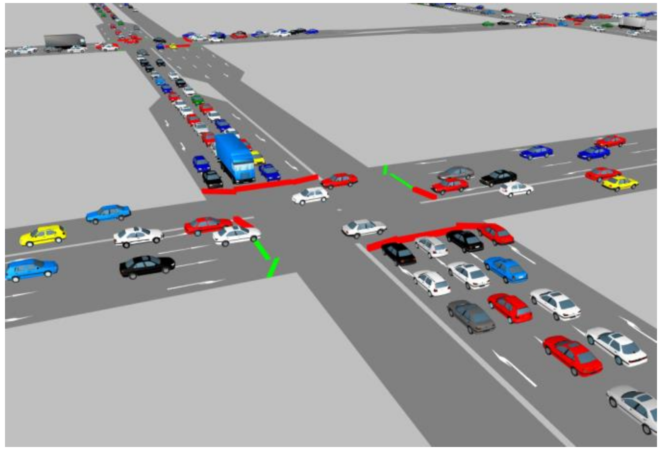


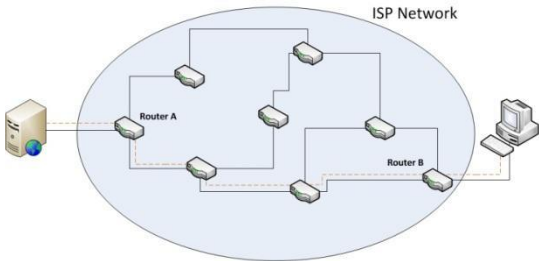
❑*A multiple queue system*



Examples







Queueing models

- Queueing models are employed for designing and evaluating the performance of queueing systems
  - Server utilization, waiting line length, waiting time, etc.
- Simple systems
  - Performance measures can be computed mathematically
- Complex systems
  - Simulation is usually required

Characteristics of queueing models

- Key elements
  - Customer – anything that arrives at a facility and requires service
  - Server – any resource that provides the requested service

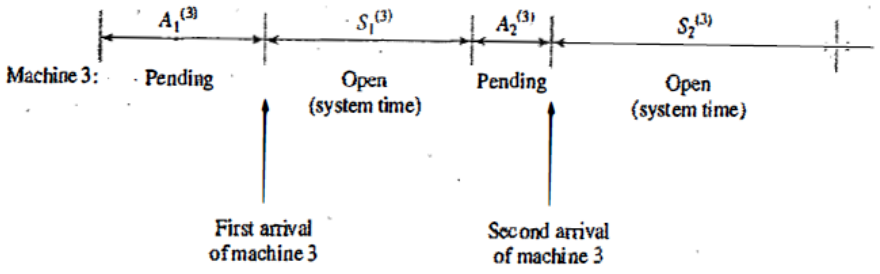
- ***The calling population***
  - The population of potential customers
  - Can be finite or infinite
  - In an ***infinite population*** model
    - The arrival rate is not affected by the number of customers being served and waiting
  - In ***finite population*** mode
    - The arrival rate to the queueing system depends on the number of customers being served and waiting

Characteristics of queueing models

- ***System capacity***
  - The number of customers that may be in the waiting line or system
- ***The arrival process***
  - ***Infinite-population models***
    - The arrival process usually characterized in terms of inter-arrival times of successive customers
    - Arrivals can be deterministic or random
    - Random arrivals
      - ✓ Interarrival times are usually characterized by a probability distribution
      - ✓ Customers may arrive one at a time or in batches
      - ✓ The batch may be of constant size or of random size
      - ✓ The most common model- Poisson model or exponential inter-arrival time

Characteristics of queueing models

- ***The arrival process (cont'd)***
  - ***Infinite-population models***
    - Scheduled (deterministic) arrivals
      - ✓ Interarrival times could be either constant or constant plus or minus a small random amount
  - ***Finite population models***
    - The arrival process is characterized in a completely different fashion
    - Pending customers- customers outside the queuing system
    - Runtime – the length of time from departure from the queueing system until that customer's next arrival to the queue
    - E.g., machine repair problem
    - Runtime – exponential, Gamma, Weibull

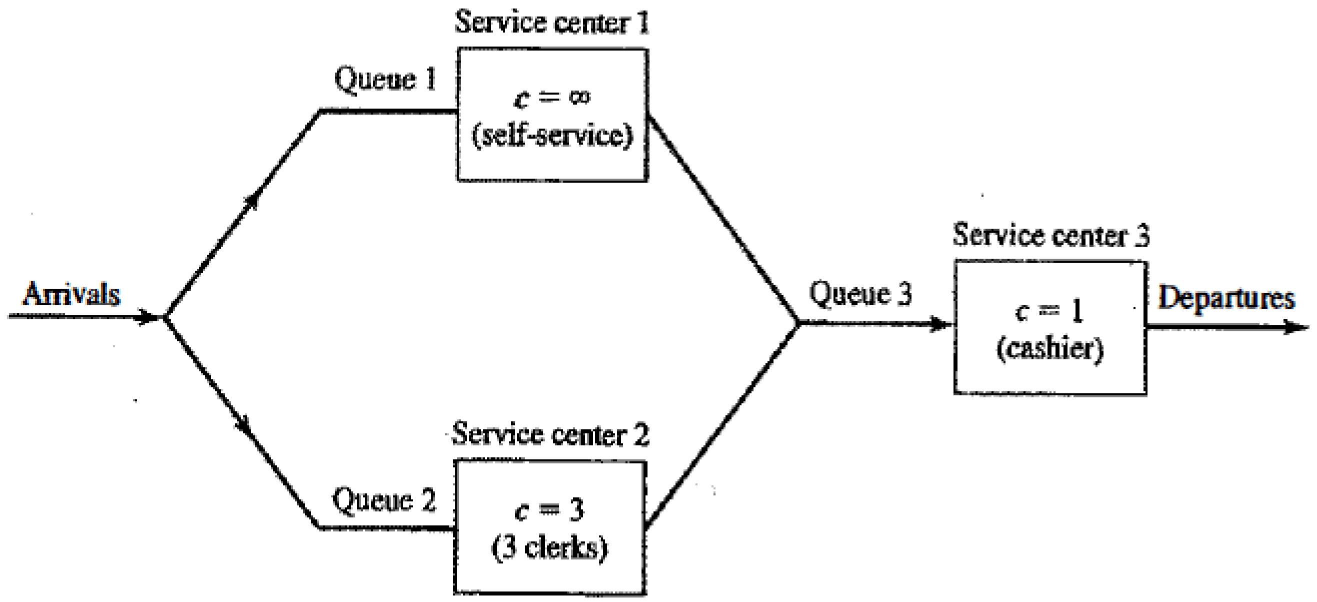


Queue behavior and queue discipline

- ***Queue discipline*** refers to the logical ordering of customers in a queue
  - First come 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 the service mechanism***
  - The service time may be constant or of random duration
  - Service times of successive arrivals {S1, S2, S3, …}are usually characterized as a sequence of IID random variables
  - Distribution used - Exponential, Weibull, gamma, lognormal and truncated normal

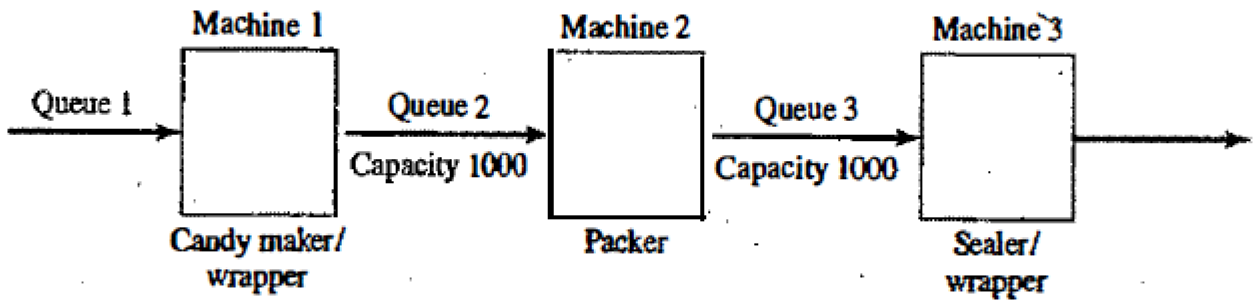
Queue behavior and queue discipline

- Example 1 – a discount warehouse
  - Customers may either serve themselves or wait for one of the three clerks
  - Finally leave after paying a single cashier



Queue behavior and queue discipline

- Example 2- a candy manufacturer
  - a production line that consists of three machines
  - The first machine makes and wraps, the second packs 50 pieces in box, the third machines seals and wraps the box



Queueing notation

- Different types of queueing systems
  - 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 calling population
  - The common symbols for A and B
    - M (exponential or markov), D (constant or deterministic), Ek (Erlang of order k), G(arbitrary or general)
  - Example $M/M/1/\infty/\infty$ (in short M/M/1)

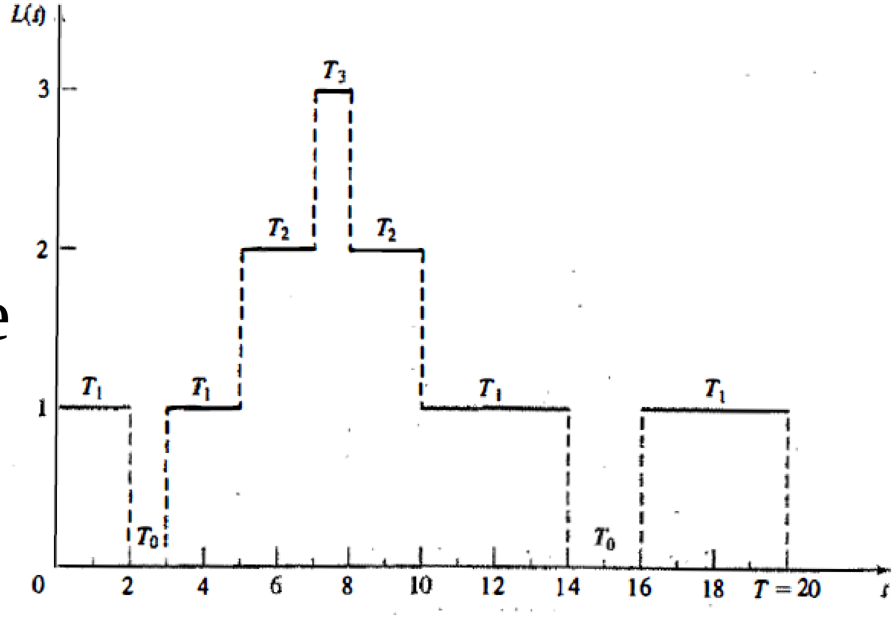
Performance of Queueing systems

- Long run measures of performance of queueing systems
  - Long-run time average number of customers in the system (L) and in the queue (L_Q)
  - The long run average time spent in system (w) and in the queue (w_Q) per customer
  - Server utilization (portion of time that a server is busy) (ρ)

Time-average number in system L

- Consider a queueing system over a period of time T
  - Let L(t) denotes the number of customers in the system at time t
  - The time-weighted average number in a system

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



- L- which is called the long run time average number in the system

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

Time-average number in system L

- If simulation run length T is sufficiently long, the estimator $\bar{L}$ becomes arbitrarily close to L
- The number of customers waiting in line

$$\overline{L_Q} = \frac{1}{T}\sum_{i=0}^{\infty} i{T_i}^{Q} = \frac{1}{T}\int_0^T L_Q(t)dt$$

Average time spent in system per customer w

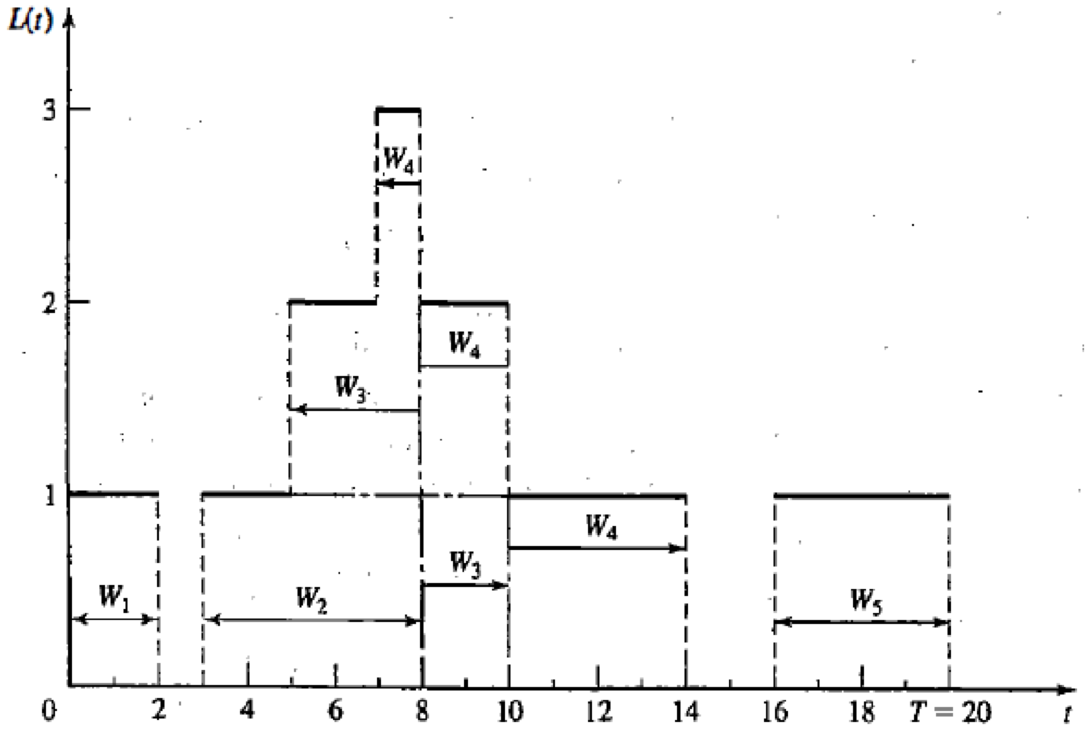
- If the queueing system is simulated for period of time T
  - Record the time each customers spends in the system (W1, W2,…,WN)
  - N = the number of arrivals during [0. T]

$$\bar{W} = \frac{1}{N}\sum_{i=0}^{\infty} W_i$$

$$\overline{W_Q} = \frac{1}{N}\sum_{i=0}^{\infty} W_i^Q$$

The conservation equation

- ***Conservation equation***
  - λ =arrival rate
  - $\overline{W}$= average waiting time
  - Then, $\overline{L} = \lambda \, \overline{W}$
- Proof
  - $\sum_{i=1}^{N} W_i = \int_0^T L(t)dt$
  - $\overline{L} = \frac{1}{T}\int_0^T L(t)dt = \frac{N}{T}\frac{1}{N}\sum_{i=1}^{N} W_i = \lambda\overline{W}$



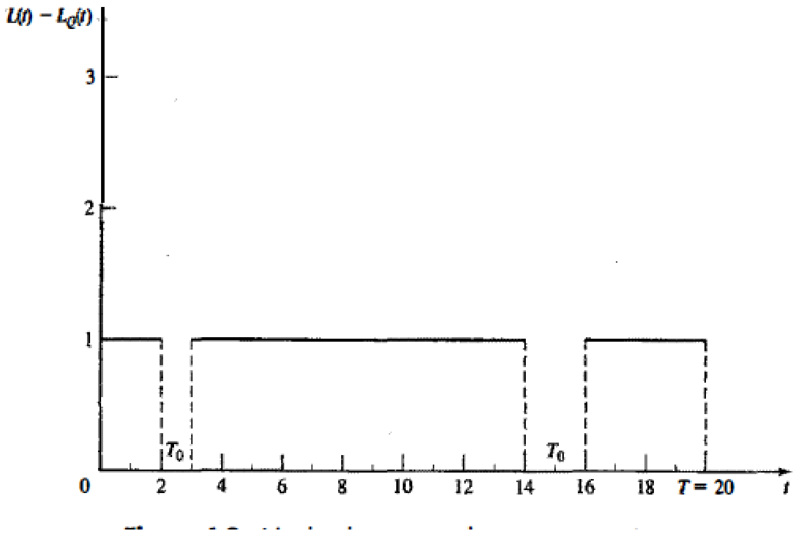
Server utilization

- Server utilization – the portion of time that a server is busy
- ***Server utilization in G/G/1/∞/∞ queues***

$$\overline{L_s} = \frac{1}{T}\int_0^T [L(t) - L_Q(t)]dt$$

$$\overline{L_s} = \frac{T - T_0}{T} \quad as\ T \to \infty, \overline{L_s} \to \rho$$

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



- ***Server utilization in G/G/c/∞/∞ queues***

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

Multiserver queues with Poisson arrivals

- M/M/c/N/∞
- If an arrival occurs when the system is full, that arrival is turned away and doesn't enter the system
- The effective arrival rate (λ_e) – the mean number of arrivals pertime unit who enter and remain in the system
  - $\lambda_e < \lambda$
  - $\lambda_e = \lambda(1 - P_N)$ $(1 - P_N)$=the probability that a customer upon arrival will find space and be able to enter the system

Network of queues

- Many systems are naturally modeled as networks of single queues
  - Customers departing from one queue may be routed to another
  - Provided that no customers are created or destroyed in the queue, ***the departure rate out of a queue is the same as the arrival rate*** into the queue, over long run
  - If customers arrive to queue i at rate λ_i, and a fraction $0 \leq P_{ij} \leq 1$ of them routed to queue j upon departure, then the arrival rate from queue I to queue j is $\boldsymbol{\lambda_i P_{ij}}$
  - The overall arrival rate into queue j, λ_j, is the sum of the arrival rate from all sources
    - $\boldsymbol{\lambda_j = a_j + \sum_{\text{all } i} \lambda_i P_{ij}}$
  - If queue j has cj parallel servers, each working at rate μ_j, the long run utilization of each server is $\boldsymbol{\rho_j = \frac{\lambda_j}{c_j \mu_j}}$

Project-2

- ***Consider a communication system with the following settings:***
  - *There are two types of packets, high and low priority packets. For each packet type, there is a separate queue and a FIFO queue discipline is applied. Packets in the low-priority queue are served only if there is no packets in the high-priority queue.*
  - *The packets are transmitted over a communication link with a capacity of 100Mb/s and the packet length distribution follows an exponential distribution with a mean 25Mb.*
  - *The packets arrive to the system according to a Poisson process at an average rate λ=2 packet/s. The probability that an arriving packet belongs to a high priority class is 0.3.*
  - ***Tasks***
    - ✓ *Show a diagrammatic representation of the queueing system*
    - ✓ *Develop a simulation model and analyze the different properties of the system – the average waiting time for each packet type, the average queue length, link utilization*
    - ✓ *Plot the CDF of inter-arrival time statistics*
  - *You can do the project in a group of 2 or 3*