Simulation and Modelling

Spring 2023
CS4056

Muhammad Shahid Ashraf

Notes

Queueing Models

Notes

Characteristics of Queueing Systems

- The Calling Population
- System Capacity
- The Arrival Process
- Queue Behavior and Queue Discipline
- Service Times and the Service Mechanism

Notes

Key elements of queueing systems

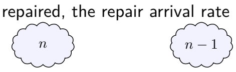
- Customer: refers to anything that arrives at a facility and requires service, e.g., people, machines, trucks, emails, packets, frames.
- Server: refers to any resource that provides the requested service, e.g., repairpersons, machines, runways at airport, host, switch, router, disk drive, algorithm.

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

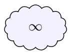
Notes

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.



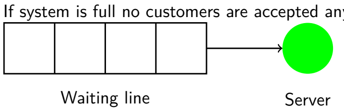
  - Infinite population model: if arrival rate is not affected by the number of customers being served and waiting, e.g., systems with large population of potential customers.



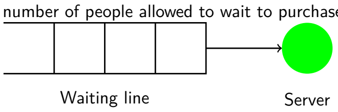
Notes

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.



Notes

Arrival Process

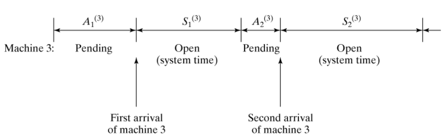
- For infinite-population models:
  - In terms of interarrival times of successive customers.
- Arrival types:
  - Random arrivals: interarrival times usually characterized by a probability distribution.
    - Most important model: Poisson arrival process (with rate λ), where a time represents the interarrival time between customer $n-1$ and customer n, and is exponentially distributed (with mean $\frac{1}{\lambda}$).
  - 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.

Notes

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.



Notes

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)

Notes

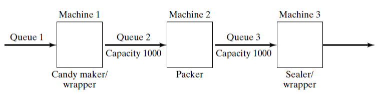
Service Times and Service Mechanism

- 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.

Notes

Examples of Queueing Systems

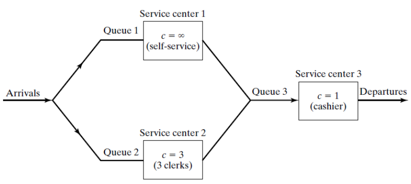
A candy manufacturer has a production line that consists of three machines separated by inventory in-process buffers. The first machine makes and wraps the individual pieces of candy, the second packs 50 pieces in a box, and the third machine seals and wraps the box. The two inventory buffers have capacities of 1000 boxes each. As illustrated by figure, the system is modeled as having three service centers, each center having c = 1 server (a machine), with queue capacity constraints between machines. It is assumed that a sufficient supply of raw material is always available at the first queue. Because of the queue capacity constraints, machine 1 shuts down whenever its inventory buffer (queue 2) fills to capacity, and machine 2 shuts down whenever its buffer empties.



Notes

Examples of Queueing Systems

Consider a discount warehouse where customers may either serve themselves or wait for one of three clerks, then finally leave after paying a single cashier. The system is represented by the flow diagram in figure. The subsystem, consisting of queue 2 and service center 2. Other variations of service mechanisms include batch service (a server serving several customers simultaneously) and a customer requiring several servers simultaneously. In the discount warehouse, a clerk might pick several small orders at the same time, but it may take two of the clerks to handle one heavy item.



Notes

Queueing Notations

A/B/c/N/K

These letters represent the following system characteristics:

- A represents the interarrival-time distribution.
- B represents the service-time distribution. Common symbols for A and B include
  - M (exponential or Markov),
  - D (constant or deterministic),
  - Ek (Erlang of order k),
  - PH (phase-type),
  - H (hyperexponential),
  - G (arbitrary or general), and
  - GI (general independent).
- c represents the number of parallel servers.
- N represents the system capacity.
- K represents the size of the calling population

Notes

Queueing Notations

$$A/B/c/N/K$$

For example,

- $M/M/1/\infty/\infty$ indicates a single-server system that has unlimited queue capacity and an infinite population of potential arrivals. The interarrival times and service times are exponentially distributed.
- When N and K are infinite, they may be dropped from the notation. For example,$M/M/1/\infty/\infty$ is often shortened to M/M/1.
- G/G/1/5/5: Single-server with capacity 5 and call-population 5.
- M/M/5/20/1500/FIFO: Five parallel server with capacity 20, call-population 1500, and service discipline FIFO

Notes

Understanding and Evaluating Queueing System Performance over Time

- Average waiting time: This is the average time a customer spends waiting in the queue before being served. It is an important measure of the quality of service provided by the system.
- Average queue length: This is the average number of customers waiting in the queue for service. It is an important measure of the efficiency of the system.
- Utilization: This is the proportion of time that the server is busy serving customers. A high utilization indicates that the system is being used efficiently, while a low utilization indicates that there is excess capacity.
- Throughput: This is the number of customers served by the system per unit of time. It is an important measure of the capacity of the system.
- Blocking probability: This is the probability that a customer is blocked (i.e., denied service) due to the system being at full capacity. It is an important measure of the effectiveness of the system in handling demand.

Notes

Average waiting time

- Primary long-run measures of performance:
- Long-run time-average number of customers in the system (L) and in the queue (L_Q)
- Long-run average time spent in system (w) and in the queue (w_Q) per customer
- Server utilization, or proportion of time that a server is busy (ρ)
- "System" refers to the waiting line plus service mechanism; "queue" refers to waiting line alone
- Other measures of performance:
- Long-run proportion of customers delayed in queue longer than t_0 time units
- Long-run proportion of customers turned away due to capacity constraints
- Long-run proportion of time waiting line contains more than k_0 customers
- Defines measures of performance for general G/G/c/N/K queueing system
- Discusses relationships and estimation using ordinary sample average or time-integrated sample average

Notes

Long-Run Measures of Performance of Queueing Systems

Consider a queueing system over a period of time T , and let $L(t)$ denote the number of customers in the system at time t. A simulation of such a system is shown in Figure .

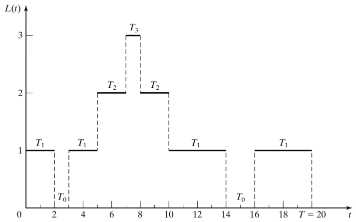
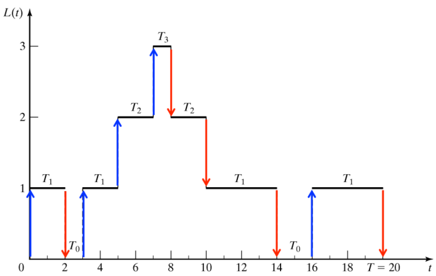


Figure: Simulation of a queueing system over a period of time T

Notes

Time-Average Number in System L



Notes

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:

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

Notes

Long-Run Measures of Performance of Queueing Systems

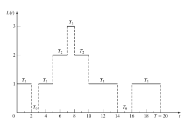


Figure: Simulation of a queueing system over a period of time T

Let T_i denote the total time during $[0, T]$ in which the system contained exactly i customers. In Figure, it is seen that $T_0 = 3$, $T_1 = 12$, $T_2 = 4$, and $T_3 = 1$. In general, the time-weighted-average number in a system is defined by:

$$L = \sum_{i=0}^{\infty} \frac{iT_i}{T}$$

Notice that T_i/T is the proportion of time the system contains exactly i customers. The estimator L is an example of a time-weighted average.

Notes

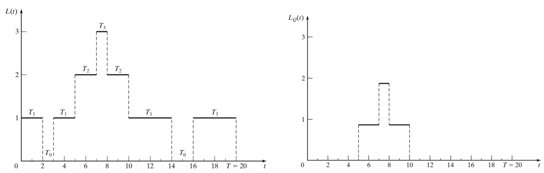
Solution

- Let T_i denote the total time during $[0, T]$ in which the system contained exactly i customers.
- In Figure 6, $T_0 = 3$, $T_1 = 12$, $T_2 = 4$, and $T_3 = 1$.
- Using Eq. (1), the time-weighted-average number of customers in the system is:

$$\begin{aligned} L &= \frac{\sum_{i=0}^{\infty} iT_i}{T} \\ &= \frac{0(3) + 1(12) + 2(4) + 3(1)}{20} \\ &= \frac{23}{20} \\ &= 1.15 \text{ customers.} \end{aligned}$$

Notes

Queueing System and Waiting Customers



Suppose that Figure represents a single-server queue—that is, a G/G/1/N/K queueing system (N $\geq$ 3, K $\geq$ 3).
Then the number of customers waiting in queue is given by $L_Q(t)$, defined by

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

and shown in Figure .

Notes

Queueing System and Waiting Customers

Suppose that Figure represents a single-server queue—that is, a G/G/1/N/K queueing system (N $\geq$ 3, K $\geq$ 3).
Then the number of customers waiting in queue is given by $L_Q(t)$, defined by

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

and shown in Figure .
Thus, $T_{0,Q} = 5 + 10 = 15$, $T_{1,Q} = 2 + 2 = 4$, and $T_{2,Q} = 1$. Therefore,

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

Notes

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: $\hat{w} \rightarrow w$ as $N \rightarrow \infty$
- If the system under consideration is the queue alone:

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

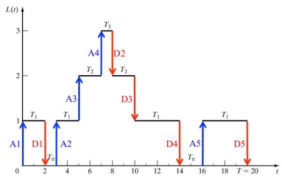
Notes

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 $\hat{w}_Q = \frac{0+0+3+3+0}{5} = 1.2$ time units



Notes

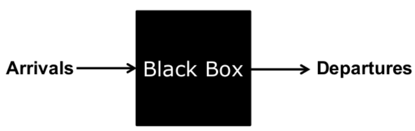
Queueing Notation

- 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

Notes

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)



average number in system $=$ arrival rate $\times$ average system time

Notes

The Conservation Equation: Little's Law

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

Average # in system

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

Average system time

Arrival rate

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

  - 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.

Notes

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: $\hat{\rho} \to \rho \quad \text{as} \quad T \to \infty$
- 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:

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

Notes

Server Utilization

- In general, for a single-server queue:

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

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

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

Notes

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:

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

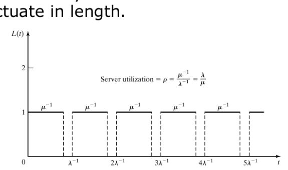
Notes

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.



Notes

Server Utilization and System Performance

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

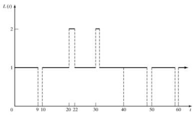
$$S_i = \begin{cases} 9 \text{ minutes with probability } 0.9 \\ 12 \text{ minutes with probability } 0.1 \end{cases}$$

  - 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: $S_1 = 9$, $S_2 = 12$, $S_3 = 9$, $S_4 = 9$, $S_5 = 9$,...
- The system becomes:



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

Notes

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:

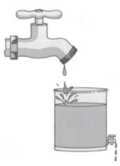
$$P(L(t) = n) = P_n(t) = P_n$$

  - 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.

Notes

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.



Notes

Formulas:	Descriptions:
$\rho = \frac{\lambda}{\mu}; P_0 = 1 - \rho$	ρ Server utilization
	P_0 Probability of empty system
$L = \rho + \frac{\rho^2(1+\sigma^2\mu^2)}{2(1-\rho)}$	L Long-run time-average number of customers in system
$w = \frac{1}{\mu} + \frac{\lambda((1/\mu^2)+\sigma^2)}{2(1-\rho)}$	w Long-run average time spent in system per customer
$L_Q = \frac{\rho^2(1+\sigma^2\mu^2)}{2(1-\rho)}$	L_Q Long-run time-average number of customers in queue
$w_Q = \frac{\lambda((1/\mu^2)+\sigma^2)}{2(1-\rho)}$	w_Q Long-run average time spent in queue per customer

Q1 Customers arrive at a walk-in shoe repair shop apparently at random. It is assumed that arrivals occur according to a Poisson process at the rate $\lambda = 1.5$ per hour. Observation over several months has found that shoe repair times by the single worker take an average time of 30 minutes, with a standard deviation of 20 minutes.

- Thus the mean service time $1/\mu =$ hour,
- the service rate is $\mu =$ per hour and $\sigma^2 =$.
- the appropriate Kendall model is the / / ,
- The proportion of time the worker is busy is $\rho =$
- the steady-state time average number of customers in the shop is L=

Q2 There are two workers competing for a job. Asif claims an average service time that is faster than Basit's, but Basit claims to be more consistent, even if not as fast. The arrivals occur according to a Poisson process at the rate of 2 per hour (1/30 per minute). Asif's service statistics are an average service time of 24 minutes with a standard deviation of 20 minutes. Baker's service statistics are an average service time of 25 minutes, but a standard deviation of only 2 minutes. If the average length of the queue is the criterion for hiring, which worker should be hired?

for Asif the average queue length is

- $\rho =$, $L_Q =$, idle time

for Basit the average queue length is

- $\rho =$, $L_Q =$, idle time

Q3 For the M/M/1 queue with service rate $\mu = 10$ customers per hour, consider how L and w increase as the arrival rate λ increases from 5 to 8.64 by increments of 20%, and then to $\lambda = 10$.

λ	5.0	6.0	7.2	8.1	10.0
ρ					
L					
w					

State Parameters of the M/M/1 Queue

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

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

Q4 A service station can hold only three customers: one in service, and two waiting. Additional customers are turned away when the system is full. The offered load is as , namely $a = 2/3.N =$ System Capacity, $a = \lambda/\mu$, $\rho = \frac{\lambda}{c\mu}$

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

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

- P(system is full)

  $P_N = P_3 = \frac{(\frac{2}{3})^3}{1!1^2} P_0 = \qquad P_N \quad \frac{a^N}{c!c^{N-c}} P_0$

- Average of the queue

  $L_Q = \frac{P_0 a^c \rho}{c!(1-\rho)^2}\left[1 - \rho^{N-c} - (N-c)\rho^{N-c}(1-\rho)\right]$

- Effective arrival rate

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

- Queue time

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

- System time, time in shop

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

- Expected number of customers in shop

  $L = \lambda_e w =$

- Probability of busy shop

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

Notes