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

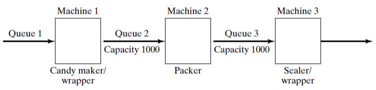
Notes

System	*Customers*	*Server(s)*
Reception desk	People	Receptionist
Repair facility	Machines	Repair person
Garage	Trucks	Mechanic
Airport security	Passengers	Baggage x-ray
Hospital	Patients	Nurses
Warehouse	Pallets	Fork-lift Truck
Airport	Airplanes	Runway
Production line	Cases	Case-packer
Warehouse	Orders	Order-picker
Road network	Cars	Traffic light
Grocery	Shoppers	Checkout station
Laundry	Dirty linen	Washing machines/dryers
Job shop	Jobs	Machines/workers
Lumberyard	Trucks	Overhead crane
Sawmill	Logs	Saws
Computer	Email	CPU, disk
Telephone	Calls	Exchange
Ticket office	Football fans	Clerk
Mass transit	Riders	Buses, trains

Notes

Examples of Queueing Systems

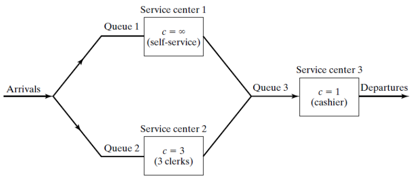
A candy manufacturer has a production line that consists of three machines separated by inventory in-process buffers. The first machine makes and wraps the individual pieces of candy, the second packs 50 pieces in a box, and the third machine seals and wraps the box. The two inventory buffers have capacities 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.
- The nurse attending 5 hospital patients might be represented by M/M/1/5/5.

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 .

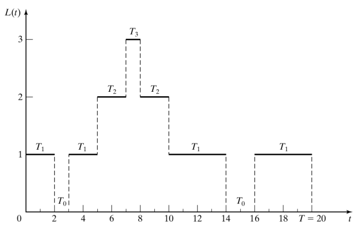


Figure: Simulation of a queueing system over a period of time T

Notes

Long-Run Measures of Performance of Queueing Systems

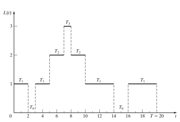


Figure: Simulation of a queueing system over a period of time T

Let T_i denote the total time during $[0, T]$ in which the system contained exactly i customers. In Figure 6, 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

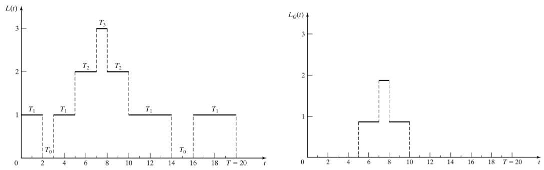
Solution

- 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

Notes