Queueing Theory

February 8, 2010

Modeling a Queueing System

Problem Statement: Consider a single server queueing system which consists of a server fulfilling a specific service to arriving customers who are waiting in a queue. The customer inter-arrival times are *independent and identically distributed* (IID) random variables and the customer service times are also IID random variables. A customer who arrives when the server is busy has to join the queue and follow a specific *queue discipline*. Hence the chief components of a QS are:

- Arrival process/time
- Service mechanism/ service time
- Queue discipline → First In First Out (FIFO) and Last In First Out (LIFO)

Representation

The System State consists of the following:

- Server status (state)
- Number of customers in the queue (state)
- A list of arrival times (state)
- Time of last event (state)
- Time representation (simulation clock)
- Next scheduled event: arrival and departure (event list)
- Statistical counters

Note the following:

- Inter arrival time is denoted by A_i which is the time between arrivals $(i-1)$ and i. Values of A_i are IID from a distribution function F_A
- Service times are denoted by S_i for the $i-th$ customer. Values of S_i are IID from a distribution function F_S.
- e_0 is the very first null event on the time line and is a null event.

- The $i-th$ customer arrives at the time point t_i.
- The length of time that a customer has to wait in line before getting serviced is called a delay and denoted by D_i.
- Hence, the following relationship can be derived:

$$C_i = t_i + D_i + S_i \tag{1}$$

where C_i is the departure time point of the $i-th$ customer. Also, the average customer arrival rate can be expressed as:

$$\lambda = 1/E(A) \tag{2}$$

where $E(A)$ is the *Expected* value of the arrival rate A.

Measuring System Performance

The following parameters can be used to measure performance in the queueing system. (Note: the ˆ hat on a quantity is an estimator of the quantity)

- Expected average delay:

$$\hat{d}(n) = \frac{\sum_{i=1}^{n} D_i}{n} \tag{3}$$

- Expected average number of customers in the queue but not being served:

$$\hat{q}(n) = \frac{\sum_{i=0}^{\infty} iT_i}{T(n)} = \frac{\int_0^{T(n)} Q(t)dt}{T(n)} \tag{4}$$

  where $Q(t)$ is the number of customers in a queue for time point t and $T(n)$ is the time required to observe n delays in the queue. So:

$$0 < t \leq T(n), Q(t) \geq 0 \tag{5}$$

- Expected server utilization:

$$\hat{u}(n) = \frac{\int_0^{T(n)} B(t)dt}{T(n)} \tag{6}$$

  where $\hat{u}(n)$ is the expected proportion of time (in between 0 and $T(n)$) when the server is busy. Hence:

$$0 \geq \hat{u}(n) \leq 1 \tag{7}$$

  and $B(t)$ is a unit step function defined as:

$$B(t) = \begin{cases} 1; \text{ when server is busy} \\ 0; \text{ when server is free} \end{cases} \tag{8}$$

Please note that both $\hat{q}(n)$ and $\hat{u}(n)$ are continuous time averages while $\hat{d}(n)$ is a discrete estimator.

Discussion Points

Please consider the following:

- Time representation
- Discrete and continuous time statistics
- Developing a flowchart for this simulation
- Steady state measures of the statistics
- Termination point of the simulation

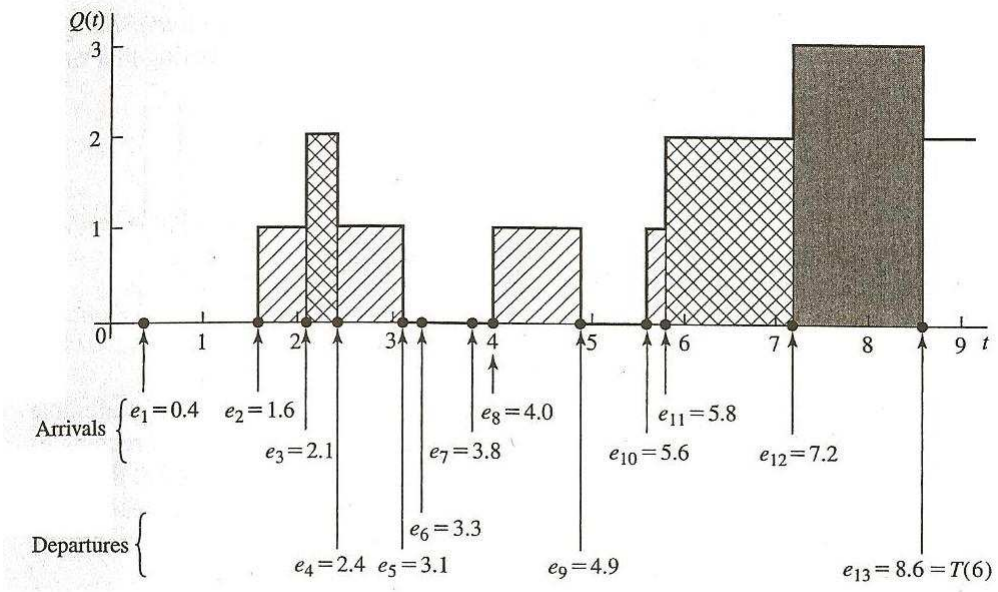

Figure 1: $Q(t)$ arrival and departure events and times in the QS (Law fig1.5)

From the above figure calculate $\hat{d}(1), \hat{q}(1)$ and $\hat{u}(1)$

Some General Results from QT

W_i = Waiting time for ith arrival
S_i = Service time for ith arrival
A_i = Inter-arrival time between *(i-1)*th and ith arrival
λ = Average customer arrival rate

μ = Average customer service rate

Recursive relationship:

$$W_i = W_{i-1} + S_{i-1} - A_i \tag{9}$$

Little's Law

Relates steady state mean system size to steady state average customer waiting times. Given by:

$$L = \lambda W \tag{10}$$

where L is the steady state size of the system given by:

$$L = \lim_{T\to\infty} \frac{\int_0^T N(t)dt}{T} \tag{11}$$

and W is the steady state waiting time in the system given by:

$$W = \lim_{T\to\infty} \frac{\int_0^T N(t)dt}{m} \tag{12}$$

where $N(t)$ is the number of customers in the system at time point t, which lies between 0 and T and m is the total number of customer arrivals between 0 and T. From (11) and (12), (10)can be derived given that the steady state value of λ is:

$$\lambda = \lim_{T\to\infty, m\to\infty} \frac{m}{T} \tag{13}$$

Some general results from Little's Law in a G/G/c QS:

$$\rho = \frac{\lambda}{c\mu} \tag{14}$$

The system is in steady state when $\rho < 1$.

Ergodicity

A stochastic process $N(t)$ is Ergodic if with probability 1 it can be said that all its measures can be determined or well approximated from a single realization $N_0(t)$ of the process. Every process of interest to us is ergodic. That is to say, that the time average of $N(t)$ for a single run of the simulation will be equal to the ensemble average of $N(t)$, the ensemble average being the average taken across an ensemble of simulation runs at steady state. It is in effect the steady state average.

So we can define the probability of $N(t) = n$, i.e. there being n number of customers in the system at the time point t as follows:

$$Pr\{N(t) = n\} = p_n(t) \tag{15}$$

hence, we can say that for a G/G/c QS:

$$L = E[N] = \sum_{n=0}^{\infty} np_n \; and \; L_q = E[N_q] = \sum_{n=c+1}^{\infty} (n-c)p_n \tag{16}$$

where there L_q is the expected number of customers in queue and the number of servers is c.

From Little's Law we can say that $L_q = \lambda W_q$ where W_q is the number of customers in queue. So we can say:

$$L - L_q = \lambda(W - W_q) = \lambda E[S] = \lambda \frac{1}{\mu} = r \tag{17}$$

Here r is the load on the system (expected number of customers in service) at steady state and for c servers we can say that the load on each server is r/c which we denote by ρ - same as (14). Also we can reduce (16) in the following way:

$$L - L_q = E[N] - E[N_q] = \sum_{n=1}^{\infty} np_n - \sum_{n=c+1}^{\infty} (n-c)p_n = c\sum_{n=1}^{\infty} p_n = c(1 - p_0) \tag{18}$$

where p_0 is the probability of there being 0 customers in the system. Combining (17) and (18) we get another general result:

$$\rho = 1 - p_0 \tag{19}$$

In the next class we will discuss the nature of the customer arrival process and the inter-arrival rate.