Notes on Queueing Theory

Queuing theory is the mathematics of waiting lines. It is extremely useful in predicting and evaluating system performance.

Assumptions:

- independent arrivals
- exponential distributions
- customers do not leave or change queues.
- Large queues do not discourage customers.

Queueing theory usually provides reasonable answers even if the above do not exactly hold. There are advanced Queueing theory formulas to handle exceptions to all of the above assumptions.

Basic measurable values of a queueing system

- Arrival rate (λ) — the average rate at which customers arrive. Be careful that you measure lambda in the proper units.
- Service time (s) — the average time required to service one customer. The units for the service time should be the inverse of the units for arrival rate.
- Number in the system (Q) — the average number of customers in the system, both waiting and being serviced.
- Number waiting (W) — the average number of customers waiting. Always less than Q.
- Time in the system (Tq) the average time each customer is in the system, both waiting and being serviced.
- Time waiting (Tw) the average time each customer waits in the queue. Tq = Tw + s

Queueing systems are usually described by three values separated by slashes

Arrival distribution / service distribution / number of servers

where:

M = **M**arkovian or exponentially distributed
D = **D**eterministic or constant.
G = **G**eneral or binomial distribution

The most common queueing system is M/M/1 where the arrival rate is exponentially distributed, the service times are exponentially distributed and there is only one server.

Poisson arrival rate.
If customers are arriving at the exponentially distributed rate λ, then the probability that there will be *k* customers after time *t* is:

$$P_k(t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}$$

Utilization = ρ = λs = faction of time the server is busy.

Little's formula states the queue size equals the arrival rate times the average time in the system.

$$Q = \lambda T_q \quad \text{or} \quad W = \lambda T_W$$

Notes on Queueing Theory

The simplest server is an M/M/1 queue. Both the service time and the arrival rate are variable and exponentially distributed.

$$Tq = \frac{s}{1-\rho} \qquad Q = \frac{\rho}{1-\rho}$$

$$Tw = \frac{s\rho}{1-\rho} \qquad W = \frac{\rho^2}{1-\rho}$$

For some systems the service time is always the same. These systems can be modeled as a M/D/1 queue. In the equations below, note that the wait time (and hence the time in the system) is smaller because there is less variability in the system.

$$Tq = \frac{s(2-\rho)}{2(1-\rho)} \qquad Q = \frac{\rho^2}{2(1-\rho)} + \rho = \frac{\rho}{1-\rho} - \frac{\rho^2}{2(1-\rho)}$$

$$Tw = \frac{s\rho}{2(1-\rho)} \qquad W = \frac{\rho^2}{2(1-\rho)}$$

When there are N servers, we assume that each server is identical. All customers wait in a single queue and use the first available server. The utilization is $\rho = \lambda s / N$. For the calculations, an intermediate value, K is useful.

$$K = \frac{\sum_{i=0}^{N-1} \frac{(\lambda s)^i}{i!}}{\sum_{i=0}^{N} \frac{(\lambda s)^i}{i!}}$$

The probability that all servers are busy is $C = \frac{1-K}{1-\frac{\lambda s K}{N}}$

$$Tq = \frac{Cs}{N(1-\rho)} + s \qquad Q = C\frac{\rho}{1-\rho} + \lambda s$$

$$Tw = \frac{Cs}{N(1-\rho)} \qquad W = C\frac{\rho}{1-\rho}$$

When working on a problem involving queueing theory, it is advisable to following these steps:

- Determine what quantities you need to know. Do you need to know the time in the system or just the waiting time?
- Identify the server. Where are items being queued?
- Identify the queued items. Are the items being queued: processes, bytes, requests, messages or some other object. Once the item is defined, convert all times and rates into these units. For example, if network packets are being queued, then convert the transmission rate into packets/sec.
- Identify the queueing model. How many queues and how many servers are involved. Is the service time constant or random?
- Determine the service time. Calculate the service time in seconds/item.
- Determine the arrival rate. Calculate the arrival rate from all sources in items/sec.
- Calculate ρ, the server utilization.
- Calculate the desired values. Make sure you use the correct equation for the appropriate queueing model.

Notes on Queueing Theory

The probability that there are N customers in the system can be calculated by:

$$\text{Prob}[Q = N] = (1-\rho)\rho^N$$

Summing the probabilities for individual cases gives the probability of N or less customers in the system

$$prob[Q \leq N] = \sum_{i=0}^{N} (1-\rho)\rho^i$$

The probability that there are more than N customers in the system is just 1 minus the above

$$prob[Q > N] = 1 - \sum_{i=0}^{N} (1-\rho)\rho^i$$

When there are multiple exponentially distributed arrivals coming together, you can sum the arrival rate.

When queues are linked serially and the service rate is exponential, the input lambda is the output lambda.