Kinds of queuing systems

Queues are described by a 3- or 4- tuple:
Arrival process/Service Process/Number of servers/Maximum occupancy.
In the arrival process and service process positions:
M – Markovian or memoryless or Poisson (all equivalent)
G – General
D – Deterministic

Probability distributions for M/M/1 queuing systems

Let X be the number of arrivals in an interval of length T and

$$P\{X=n\} = \frac{e^{-\lambda T}(\lambda T)^n}{n!} \text{ -- Poisson distribution}$$

$$E[X] = \lambda T$$

Equivalently, let τ be the time between adjacent arrivals.

$$P\{\tau < x\} = 1 - e^{-\lambda x} \text{ -- and exponential distribution with parameter } \lambda$$

$$E[\tau] = 1/\lambda \text{ -- that is the arrival rate is } \lambda$$

Similarly, the service time, in an M/M/1 queue is given by an exponential distribution with parameter μ. The mean service time is $1/\mu$ and the service rate is μ.

Memoryless property of the exponential distribution:

$$P\{\tau > t + s \mid \tau > t\} = P\{\tau > s\}$$

Analysis of the M/M/1 queuing system

The state of an M/M/1 queuing system is the number of jobs (customers, packets) currently occupying it. We observe that at any time the number of transitions from state n to state $n+1$ is at most one more than the number of transitions from state $n+1$ to state n, from which we conclude that the rates that these two transitions are taken are equal. Let p_n be the probability that the system is in state n. Then the statement that the rates are equal is

$$\lambda p_n = \mu p_{n+1} \text{ , or equivalently } p_{n+1} = \frac{\lambda}{\mu} p_n$$

Since these are probabilities, they must sum to 1 so letting $\rho = \frac{\lambda}{\mu}$ we have

$$1 = \sum_{i=0}^{\infty} p_i = \sum_{i=0}^{\infty} \rho^i p_0 = p_0 \frac{1}{1-\rho} \text{ , therefore } p_0 = 1 - \rho$$

From this we immediately can see that the probability that the system has at least one customer is the probability that it is not in state 0 which is $1 - p_0 = 1 - (1 - \rho) = \rho$. Thus, ρ (that is, λ / μ) is the utilization of the system.

We will now use the above probability analysis together with Little's Theorem to derive several useful characteristics of the M/M/1 system.

The expected number of customers in the system, E[N], is obtained from the definition of the expected value and the probabilities computed above:

$$\text{E[N]} = \sum_{n=0}^{\infty} n p_n = \sum_{n=0}^{\infty} n(1 - \rho)\rho^n = (1 - \rho)\sum_{n=0}^{\infty} n p^n = \frac{\rho}{1 - \rho}$$

Once again we see the importance of the utilization being strictly less than 1.

With E[N] in hand we can proceed using Little's Theorem to compute

$$\lambda \, \text{E[T]} = \text{E[N]}$$

$$\text{E[T]} = \frac{\rho}{\lambda(1 - \rho)} = \frac{1}{\mu(1 - \rho)} = \frac{1}{\mu - \lambda}$$

Now, separating our consideration of the queue and the server and observing that the total time in the system is made of time in the queue plus time being served we have

E[T] = E[time in queue] + E[service time], that is

$$\frac{1}{\mu - \lambda} = \text{E[time in queue]} + \frac{1}{\mu} \text{, so E[time in queue]} = \frac{1}{\mu - \lambda} - \frac{1}{\mu} = \frac{\rho}{\mu - \lambda}$$

And we can again apply Little's Theorem to find the expected number in the queue:

$$\text{E[number in queue]} = \lambda \, \text{E[time in queue]} = \frac{\lambda\rho}{\mu - \lambda} = \frac{\rho^2}{1 - \rho}$$

Personally, I find it much easier to just remember Little's Theorem along with $\text{E[T]} = \frac{1}{\mu - \lambda}$ and derive any of the other results as needed using the numbers given in the problem at hand.

Example

Consider a situation where we have exponentially distributed packet lengths with mean 1250 bytes and a 10Mbit/s link. Assuming Poisson arrivals, for what values of λ, the packet arrival rate, will the average total time in the system be less than or equal to 10ms.

Fact: $\mu = 1000$ packets/second $((10\text{Mbit/s})/(10000\text{bits/packet}))$

Strategy: apply our knowledge of E[T] for M/M/1 systems

$$.010 \geq \frac{1}{\mu - \lambda} = \frac{1}{1000 - \lambda}; \; \lambda \leq 900$$

With the arrival rate known the average number in the system is obtainable directly from Little's Theorem:

$$N = \lambda T = 900(.01) = 9$$

Since the utilization is 0.9, the average number in the queue is N-0.9=8.1.

In a subsequent example we'll look at what happens when the system has a maximum capacity.

Example

Consider what happens when the 10Mbit/s link above is replaced by 100Mbit/s and λ and μ are also scaled up by a factor of 10. – Exercise for the reader.

Example

Compare statistical multiplexing with time or frequency division multiplexing. Assume m separate identically distributed and independent Poisson arrival streams each with parameter λ/m, being statistically multiplexed onto a service with exponential service times with mean $\dfrac{1}{\mu}$. The expected time in the system is $1/(\mu - \lambda)$. Contrast with what happens if the service is time or frequently division multiplexed. This is equivalent to having m servers each with parameter μ/m. So the expected time in the system is

$$\frac{1}{\dfrac{\mu}{m} - \dfrac{\lambda}{m}} = \frac{m}{\mu - \lambda}$$ -- that is, *m* times longer.

M/M/1/B systems

The hypothesis of infinite buffer space in the M/M/1 system is unrealistic. How unrealistic? We need to analyze an M/M/1/B system to find out. The basic analysis is the same as for M/M/1 but now we have only a finite number of states. So where previously we summed probabilities over an infinite number of states, we'll now sum over just the B+1 states of the M/M/1/B system.

$$1 = \sum_{i=0}^{B} p_i = \sum_{i=0}^{B} \rho^i p_0 = p_0$$

$$p_0 = \frac{1-\rho}{1-\rho^{(B+1)}} \text{ and } p_n = p_0 \rho^n$$

Example

Returning to the previous example (1250 byte average packet length, 1000 packets/second average service rate) and remembering that the average number in the system in that case was 9, we now ask: how much capacity do we need in the system (i.e. what value of B) will allow us to drop no more than 1% of the packets due to blocking. Rephrased, this is: for what value of B is $p_b \le 0.01$.

$$p_b = \frac{(1-\rho)\rho^B}{1-\rho^{(B+1)}} \le 0.01$$

$$(1-0.9)0.9^B \le 0.01(1-0.9^{B+1})$$

$$0.9^B \le 0.1-(0.1)0.9^{B+1} \quad (*)$$

approximating

$$1.1(0.9)^B = 0.1$$

$$0.9^B = 0.09$$

$$B \approx 23$$

check in formula (*) which was before the approximation

$$0.9^{23} \approx 0.0886 \underset{?}{\le} 0.1-(0.1)0.9^{24} \approx .092$$

From this example we see that even though the average number in the system is 9 in the case that infinite buffering is available, we need a substantially more buffers than the average in order to avoid losing packets due to blocking. Note also that the average number in the system and average time in the system will be somewhat smaller for the M/M/1/B system than for the M/M/1 system – how much smaller depends on B.