Queueing Theory (2)

- **Distribution of waiting time in *M*/*M*/1**
  - Let T_q be the waiting time in queue of a customer.
  - Then, $P\{T_q = 0\} = P\{\text{server is idle at arrival}\} = 1 - \rho$, and

$$\mathrm{P}\{T_q \le t\} = P\{T_q = 0\} + P\{T_q < t \mid n \ge 1 \text{ customers in system at arrival}\},$$

where for a given $n \ge 1$,

$$P\{T_q < t \mid n \text{ customers in the system}\} = P\{n \text{ service completions within time } t\}.$$

  - To evaluate the last probability we need the probability that the sum of n exponential rvs with is less than t.
  - The following fact evaluates such probability.

Fact. *Let $S_1, S_2, \ldots, S_n$, be n exponential rvs with mean $1/\mu$. Then, $W_n = S_1 + S_2 + \ldots + S_n$ has an Erlang distribution with cdf*

$$P\{W_n < t\} = \int_0^t \mu e^{-\mu u} \frac{(\mu u)^{n-1}}{(n-1)!} du .$$

  - By conditioning on N, the number of people in the system when the customer arrives,

$P\{T_q < t \mid n \geq 1 \text{ customers in the system}\}$

$$= \sum_{n=1}^{\infty} P\{W_n < t \mid N = n\} P\{N = n\}$$

$$= \sum_{n=1}^{\infty} \int_0^t \mu e^{-\mu u} \frac{(\mu u)^{n-1}}{(n-1)!} du \, \rho^n (1-\rho)$$

$$= \sum_{n=1}^{\infty} \int_0^t \mu e^{-\mu u} \frac{(\mu u)^{n-1}}{(n-1)!} \left(\frac{\lambda}{\mu}\right)^n (1-\frac{\lambda}{\mu}) du$$

$$= \int_0^t \left(\frac{\lambda}{\mu}\right)(\mu-\lambda) e^{-\mu u} \sum_{n=1}^{\infty} \frac{(\lambda u)^{n-1}}{(n-1)!} du$$

$$= \int_0^t \rho(\mu-\lambda) e^{-\mu u} e^{\lambda u} du$$

$$= \rho \int_0^t (\mu-\lambda) e^{-(\mu-\lambda)u} du$$

$$= \rho(1 - e^{-(\mu-\lambda)t})$$

- Finally,

$$P\{T_q < t\} = 1 - \rho + \rho(1 - e^{-(\mu-\lambda)t}) = 1 - \rho e^{-(\mu-\lambda)t} .$$

- Therefore,

$$P\{T_q > t\} = \rho e^{-(\mu-\lambda)t} .$$

- Let T be the total time of a customer in the system (in queue plus in service). Then, it can be shown in a similar way that

$$P\{T > t\} = e^{-(\mu-\lambda)t} .$$

- **Example 2**
  - In the bank of Example 1, what is the probability that the waiting time of a customer exceeds 2 hours?
  - $\mathrm{P}\{T_q > t\} = \rho e^{-(\mu-\lambda)t} = 0.9e^{-2} = 0.122.$
  - What is the probability that a customer spends more than 3 hours in the bank?
  - $\mathrm{P}\{T > t\} = e^{-(\mu-\lambda)t} = e^{-3} = 0.05.$

- **Example 3**
  - Consider an Interface Message Processor (IMP). The packet size in bits is exponentially distributed with mean $1/\mu$ bits/packet. The capacity of the communication channel is C bits/sec. Packets arrive at random (i.e. exponential inter arrival times) with arrival rate λ packets/sec. Find the queueing delay for packets at the IMP.
  - Model this as a $M/M/1$ with arrival rate λ packets/sec and service rate $C/(1/\mu) = \mu C$ packets/sec.
  - Then, the mean delay per packet is $W_q = 1/(\mu C - \lambda)$.

- **Example 4 (Message Switching)**
  - Traffic to a message switching center for a corporation arrives in a random pattern (i.e. exponential inter arrival times) at an average rate of 240 messages per minute. The line has a transmission rate of 800 characters per second. The message length distribution (including control characters) is approximately exponential with an average length of 176 characters.
  - What is the average number of messages waiting at the switching center?
  - Model this as a $M/M/1$ with $\lambda = 240$ messages/min = 4 messages /sec, and $\mu = 800/176 = 4.545$ messages/min.
  - Then, $\rho = 4/\ 5.545 = 0.88$ and $L_q = \rho^2/\ (1-\rho\ )$

    $= 6.45$ messages.
  - What is the mean waiting time of a message?
  - From Little's Law, $W_q = L_q/\lambda = 6.45\ /\ 4 = 1.61$ sec.

- **Example 5 (Dedicated vs. shared channels).**
  - Two computers are connected by a 64 kbps line. There are eight parallel sessions using the line. Each session generates Poisson traffic with a mean of 2 packets/sec. The packet lengths are exponentially distributed with mean of 2000 bits. The system designers must choose between giving each session a dedicated 8 kbps piece of bandwidth or having all packets compete for a single 64 kbps shared channel. Which alternative gives better response time (i.e. W)?
  - We need to compare two alternative $M/M/1$ models.
  - The first model has four channels each operating as a $M/M/1$ with $\lambda_1 = 2$ packets/sec and $\mu_1 = 8000/2000 = 4$ packets/sec. Therefore, $W^1 = 1/(\mu_1 - \lambda_1) = 1/2 = 0.5$ sec.
  - The second model has one channel with $\lambda_2 = 8 \times 2 = 16$ packets/sec, and $\mu_2 = 64000/2000 = 32$ packets/sec. Therefore, $W^2 = 1/(\mu_2 - \lambda_2) = 1/16 = 0.063$ sec.
  - The response time in the shared system is 8 times less. This is the better alternative.
  - (Note that both systems have the same $\rho = 0.5$.)
  - Why this result?
  - In the dedicated system, some channels could be idle, while other channels have long queues. This does not happen in the shared system.