M/M/1 Model[©]

This Teaching Note supercedes Section 5.1.1 of the reading by Daniel Mignoli.

The M/M/1 model is characterized by the following assumptions:

- Jobs arrive according to a Poisson process with parameter λt, or equivalently, the time between arrivals, *t*, has an exponential distribution with parameter λ, i.e., for $t \geq 0$, the probability density function is
  $$f(t) = \lambda e^{-\lambda t}. \tag{1}$$
- The service time, *s*, has an exponential distribution with parameter μ, i.e., for $s \geq 0$, the probability density function is
  $$g(s) = \mu e^{-\mu t}. \tag{2}$$
- There is a single server;
- The buffer is of infinite size; and
- The number of potential jobs is infinite.

Definition. The utilization, ρ, is the average arrival rate x average service time.

By (1), the distribution of inter-arrival times is exponential, hence the average inter-arrival time,

$$\bar{t} = \frac{1}{\lambda}. \tag{3}$$

By (2), the distribution of service times is exponential, hence the average service time,

$$\bar{s} = \frac{1}{\mu}, \tag{4}$$

and thus, by (1) and (4),

$$\rho = \frac{\lambda}{\mu}. \tag{5}$$

Assumption. $\rho < 1$.

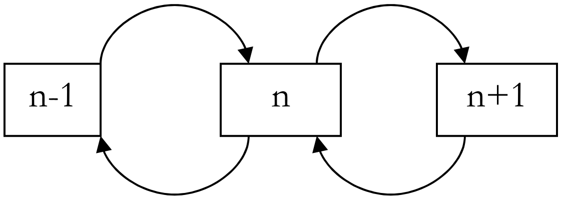
Without this condition, the queue would grow without limit.

Proposition. The probability that there *n* jobs in the system (either queue or process) is

$$P_n = \rho^n [1 - \rho]. \tag{6}$$

[©] 2003, I.P.L. Png

Proof. Consider the state diagram,



In a steady state, the expected number of transitions from *n* up to *n+1* must equal the number of transitions from *n+1* down to *n*, or,[1]

$$\lambda P_n = \mu P_{n+1}. \tag{7}$$

For $n = 0$, we have

$$P_1 = \frac{\lambda}{\mu} P_0 = \rho P_0, \tag{8}$$

using (5). Similarly, by applying (7) repeatedly, we have

$$P_n = \rho^n P_0. \tag{9}$$

To solve for P_0, observe that

$$\sum_{n=0}^{\infty} P_n = 1. \tag{10}$$

Hence, by (9),

$$1 = P_0 \sum_{n=0}^{\infty} \rho^n = P_0 \frac{1}{1-\rho}, \tag{11}$$

since $\rho < 1$. So, (11) implies that

$$P_0 = 1 - \rho. \tag{12}$$

Then, the Proposition follows from (9) and (12). [ ]

The expected number of jobs in the system (either queue or process) is

$$L = \sum_{n=0}^{\infty} nP_n = \sum_{n=0}^{\infty} n\rho^n [1-\rho]. \tag{13}$$

Simplifying, we have

[1] For details, please refer to Frederick S. Hillier and Gerald J. Lieberman, *Introduction to Operations Research*, Section 16.5, "Birth and Death Processes".

$$\begin{aligned} L &= \rho[1-\rho] \sum_{n=0}^{\infty} \frac{d}{d\rho}\rho^n \\ &= \rho[1-\rho]\frac{d}{d\rho}\sum_{n=0}^{\infty}\rho^n \\ &= \rho[1-\rho]\frac{d}{d\rho}\left(\frac{1}{1-\rho}\right) = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu-\lambda} \end{aligned} \tag{14}$$

Since there is a single server, the expected number of jobs in the queue is

$$\begin{aligned} L_q &= \sum_{n=1}^{\infty}[n-1]P_n = \sum_{n=1}^{\infty}[n-1]\rho^n[1-\rho] \\ &= \rho[1-\rho]\sum_{n-1=0}^{\infty}[n-1]\rho^{n-1} = \frac{\rho^2}{1-\rho} = \frac{\lambda^2}{\mu[\mu-\lambda]}. \end{aligned} \tag{15}$$

Little's Formula. In a steady state, the average time spent waiting in the queue,

$$W_q = \frac{L_q}{\lambda}, \tag{16}$$

and the average time spent in the system (in queue or process),

$$W = \frac{L}{\lambda}. \tag{17}$$

(Little's Formula is valid for the steady state of any queueing process.)

Applying Little's Formula,

$$W = \frac{1}{\mu-\lambda}, \tag{18}$$

and

$$W_q = \frac{\lambda}{\mu[\mu-\lambda]}. \tag{19}$$

Example

Consider a switch which has an infinite buffer and an infinite number of users generating messages according to a Poisson process with average inter-arrival time of 800 milliseconds. The switch serves requests with a service time that is exponentially distributed with an average service time of 500 milliseconds. What is the average waiting time? Suppose that the switch is upgraded to reduce average service time to 400 milliseconds. How would that affect the average waiting time?

In this case,

$$\frac{1}{\lambda} = 800,$$

or

$$\lambda = \frac{1}{800}.$$

Further,

$$\mu = \frac{1}{500},$$

and hence, utilization,

$$\rho = \frac{5}{8}.$$

By Little's Formula, the average time,

$$W = \frac{1}{\mu - \lambda} = \frac{1}{\frac{1}{500} - \frac{1}{800}} = \frac{4000}{8-5} = 1333.$$

After the upgrade,

$$\mu = \frac{1}{400},$$

hence utilization falls to

$$\rho = \frac{1}{2}$$

while the average time falls to

$$W = \frac{1}{\mu - \lambda} = \frac{1}{\frac{1}{400} - \frac{1}{800}} = \frac{800}{2-1} = 400.$$

Observe that a 20% reduction in average service time leads to a 70% reduction in average time.

Please note: There is a mistake in the reading by Daniel Mignoli, Section 5.1.5 "A Classic Result When Comparing the M/M/1 to the M/M/c Queue". Please ignore this section.