Queue Theory

1. M/M/1 queue

1.1. Model

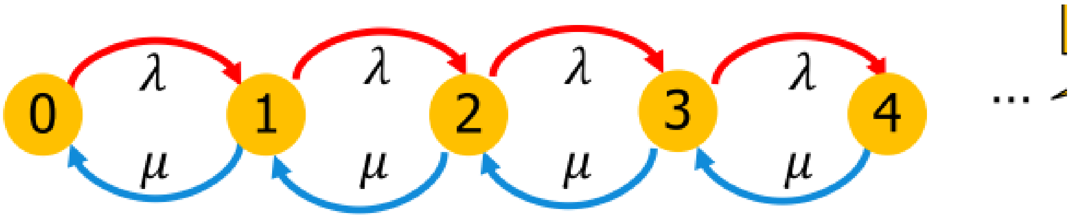
what does M/M/1 stand for?

- The **first letter** is a short hand for the **arrival process**. M stands for **exponential interarrival time**, which is another way of saying the arrival process is a Poisson process.
- The **second letter** is a short hand for the **service time distribution**. The second M therefore means that the service time is **expenentially distributed.**
- The **third number**, 1, is the **number of servers** in the system.
- The default value for the **buffer capacity is infinity**

1.2. Idea and Key questions

- M/M/1 Queue is **Memoryless**

So, we can connect it to the Markov Chain, especially CTMC and try to analyze the steady state.



λ: arriving rate
μ: completing rate

Key Questions:

- How to find the steady state probability of the number of jobs in the queue (Pi) given only λ and μ
- Based on Pi, how to find:
    - the number of requests in the system (N)
    - the (average) response time

1.3. Steady state probility of #jobs

In Queue Theory, we always assume **flows is in equilibrium,** so:

$$\begin{aligned}\lambda P_0 &= \mu P_1 \\ \lambda P_0 &= \mu P_2 \\ &\vdots \\ \lambda P_{n-1} &= \mu P_n\end{aligned}$$

so:

$$\begin{aligned}P_1 &= \rho P_0 \\ P_2 &= \rho^2 P_0 \\ &\vdots \\ P_n &= \rho^n P_0\end{aligned}$$

where $\rho = \frac{\lambda}{\mu}$

and we have: $\sum P_i = 1$

so, we have:

$$P_0 = 1 - \rho$$
$$P_n = \rho^n(1 - \rho)$$

1.4. #requests in the system

$$N = \sum_{n=0}^{\infty} nP_n = \frac{\lambda}{\mu - \lambda}$$

Proof:

$$\sum_{n=0}^{\infty} nP_n = \sum_{n=0}^{\infty} n\rho^n(1-\rho) = (1-\rho)\rho\sum_{n=1}^{\infty} n\rho^{n-1}$$

$$(1-\rho)\rho\boxed{\frac{d}{d\rho}\left(\boxed{\sum_{n=0}^{\infty}\rho^n}\right)} = (1-\rho)\rho\frac{d}{d\rho}\left(\frac{1}{1-\rho}\right)$$

geometric series

$$(1-\rho)\rho\left(\frac{1}{(1-\rho)^2}\right) = \frac{\rho}{(1-\rho)} = \frac{\lambda}{\mu-\lambda}$$

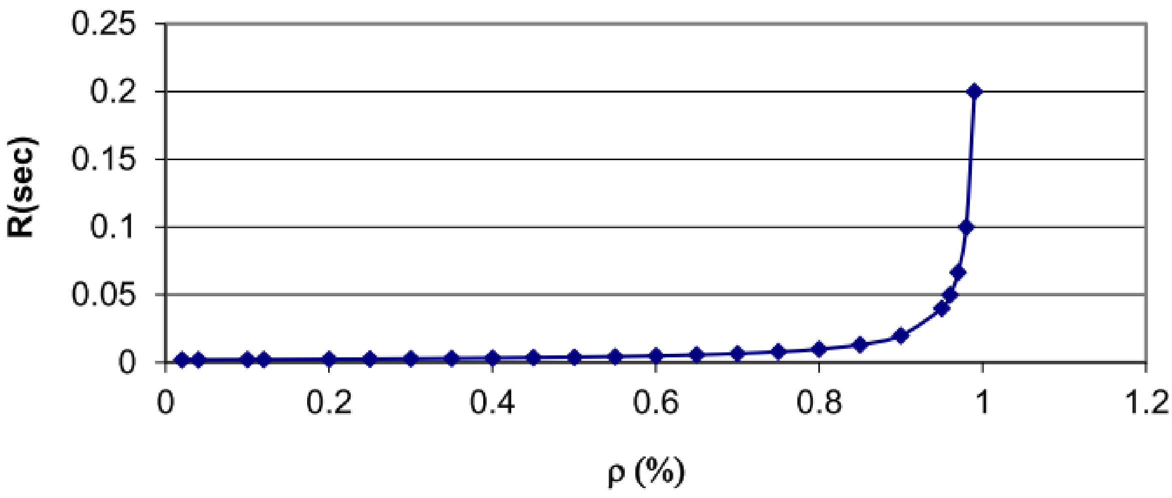
1.5. response time

Use **Littles's Law** $N = XR$

$$R = \frac{N}{X} = \frac{1}{\mu - \lambda}$$

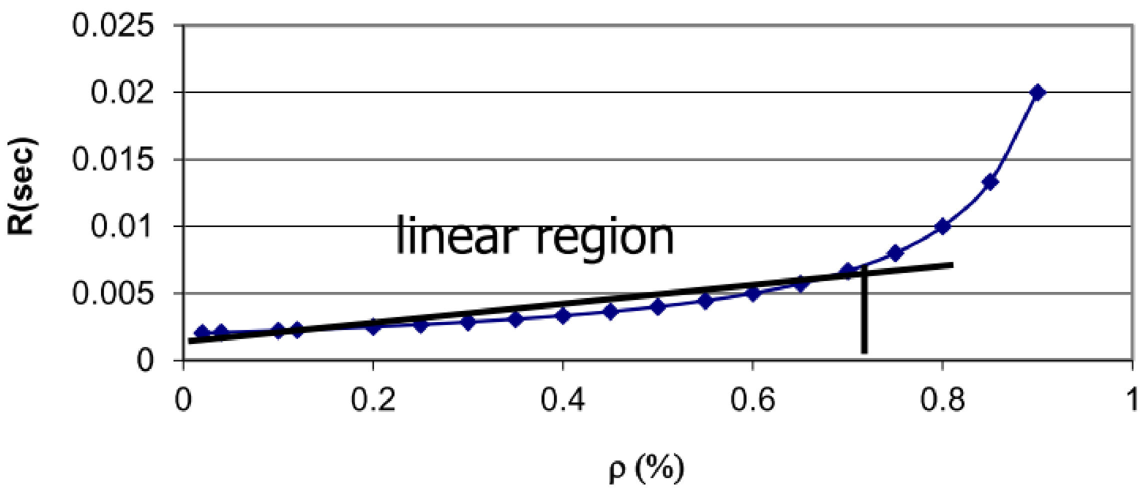
Properties:

- increase rapidly when ρ close to 1, when arrival speed close to service speed



- **Stable Region (linear region)**

can be used to judge how to adjust service speed to influence the response time



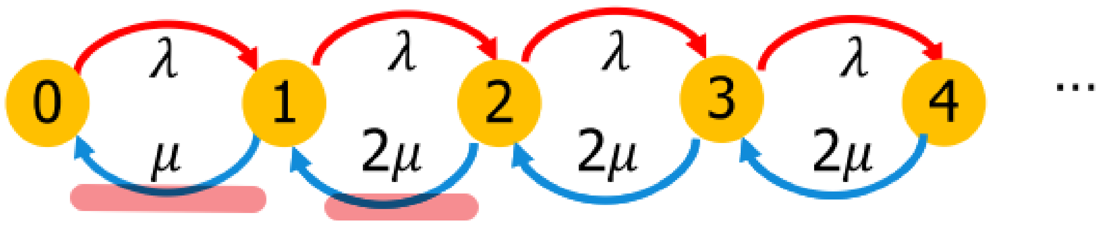
1.6. Main results of M/M/1 queue

Utilization	$U = X\,S = \lambda/\mu = \rho$
Prob. of n clients in the system	$P_n = \rho^n\,(1 - \rho)$
Mean #clients in the system	$N = \rho\,/\,(1\text{-}\rho) = \lambda\,/\,(\mu\text{-}\lambda)$
Mean #clients in the queue	$N_Q = N - (1 - P_0) = N - \rho$
Mean response time	$R = N/\lambda = 1/(\mu\text{-}\lambda) = S/(1\text{-}\rho)$
Mean waiting time	$W = R - S = \rho/(\mu\text{-}\lambda)$

2. First glance at M/M/2 (M/M/c) queue

Considering a M/M/2 queue

2.1. Markov Model



Two keypoints:

- for $N > 1$, the transation from high to low is 2μ
- for $N = 1$, the transition is still μ

The 2μ can be considered as:

- initially, we need average $1/\lambda$ to finish one, but now in $1/\lambda$ we can finish two, the new average is $1/2\lambda$

2.2. Steady state probility of #jobs

$$
\begin{aligned}
\lambda P_0 &= \mu P_1 \\
\lambda P_1 &= 2\mu P_2 \\
&\vdots \\
\lambda P_{n-1} &= 2\mu P_n
\end{aligned}
$$

if we set $\rho = \frac{\lambda}{2\mu}$

$$
\begin{aligned}
P_1 &= 2\rho P_0 \\
P_2 &= 2\rho^2 P_0 \\
&\vdots \\
P_n &= 2\rho^n P_0
\end{aligned}
$$

and we have:

$$
\sum_{n=0}^{\infty} P_n = 1 \Rightarrow 2P_0 \sum_{n=0}^{\infty} \rho^n - P_0 = 1 \Rightarrow \frac{1+\rho}{1-\rho} P_0 = 1
$$

so, we have:

$$
\begin{aligned}
P_0 &= \frac{1-\rho}{1+\rho} \\
P_n &= 2\rho^n \frac{1-\rho}{1+\rho}
\end{aligned}
$$

so

$$
\mathrm{N} = \sum_{n=0}^{\infty} nP_n = \sum_{n=0}^{\infty} 2\mathrm{n}\rho^{\mathrm{n}} \frac{1-\rho}{1+\rho} = \cdots = \frac{2\rho}{(1-\rho^2)}
$$

Main Results of M/M/2 queue

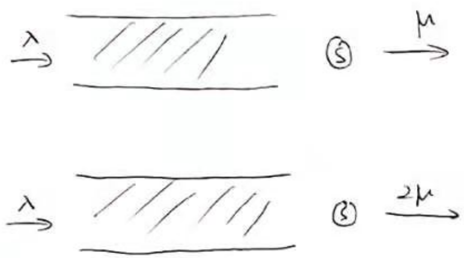
Utilization	$U = 1 - P_0 = 2\rho / (1 + \rho)$
Prob. of n clients in the system	$P_n = 2\rho^n (1 - \rho) / (1 + \rho)$
Mean #clients in the system	$N = 2\rho / (1-\rho^2)$
Mean #clients in the queue	$N_Q = 2\rho^3 / (1-\rho^2)$
Mean response time	$R = N/\lambda = 2 / (\mu (1-\rho^2))$
Mean waiting time	$W = R - 1/\mu = \rho^2 / (\mu (1-\rho^2))$

3. Examples

3.1. Student Waiting Bus example

3.2. Toilet Example

Initial Situation



Analysis:

$$w_1 = \frac{\rho_1}{\mu-\lambda} \qquad \rho_1 = \frac{\lambda}{\mu}$$

$$\therefore \rho_1 = 2\rho_2$$

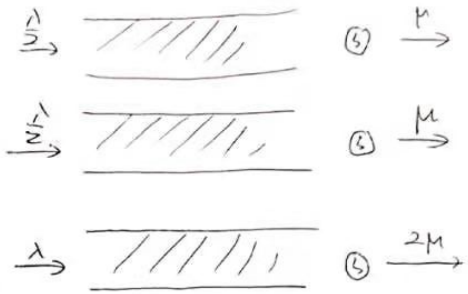
$$w_2 = \frac{\rho_2}{2\mu-\lambda} \qquad \rho_2 = \frac{\lambda}{2\mu}$$

$$\therefore \frac{w_1}{w_2} = \frac{\rho_1}{\mu-\lambda} \cdot \frac{2\mu-\lambda}{\rho_2}$$

$$= \frac{2(2\mu-\lambda)}{\mu-\lambda} \geq \frac{2(2\mu-2\lambda)}{\mu-\lambda} = 4$$

$$\therefore \frac{w_1}{w_2} > 4$$

Situation Variation 1.



Analysis

$$w_1 = \frac{\rho_1}{\mu-\frac{\lambda}{2}} \qquad \rho_1 = \frac{\frac{\lambda}{2}}{\mu}$$

$$\therefore \rho_1 = \rho_2$$

$$w_2 = \frac{\rho_2}{2\mu-\lambda} \qquad \rho_2 = \frac{\lambda}{2\mu}$$

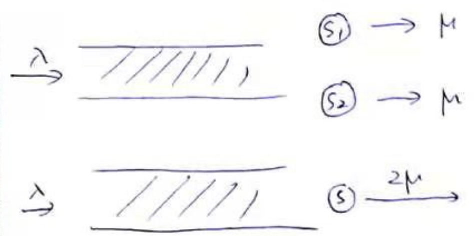
$$\therefore \frac{w_1}{w_2} = \frac{\rho_1}{\mu-\frac{\lambda}{2}} \cdot \frac{2\mu-\lambda}{\rho_2}$$

$$= \frac{2\mu-\lambda}{\mu-\frac{\lambda}{2}} = 2$$

$$\therefore \frac{w_1}{w_2} = 2$$

$\therefore$ not sufficient

Situation Variation 2



Analysis

w_1 now is M/M/2 Model:

$$w_1 = \frac{\rho_1^2}{\mu(1-\rho^2)} \qquad \rho_1 = \frac{\lambda}{2\mu}$$

$$\therefore \rho_1 = \rho_2$$

$$w_2 = \frac{\rho_2}{2\mu-\lambda} \qquad \rho_2 = \frac{\lambda}{2\mu}$$

$$\therefore \frac{w_1}{w_2} = \frac{\rho_1^2}{\mu(1-\rho^2)} \cdot \frac{2\mu-\lambda}{\rho_2} = \frac{\rho_1(2\mu-\lambda)}{\mu(1-\frac{\lambda^2}{4\mu^2})}$$

$$= \frac{\lambda(2\mu-\lambda)}{2\mu^2(1-\frac{\lambda^2}{4\mu^2})} = \frac{2\lambda(2\mu-\lambda)}{4\mu^2-\lambda^2} = \frac{2\lambda(2\mu-\lambda)}{(2\mu+\lambda)(2\mu-\lambda)}$$

$$= \frac{2\lambda}{(2\mu+\lambda)}$$

And $\because$ in steady Queue. $\lambda = \mu$

$\therefore \frac{w_1}{w_2} < 1$

$\therefore$ Less waiting time. Longer dealing time

Win - Win