Introduction to QUEUEING : M/M/c

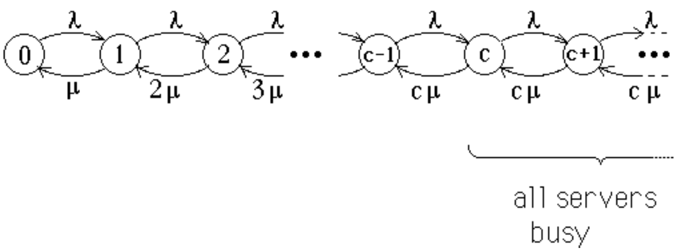
This Hypercard stack was prepared by:
Dennis L. Bricker,
Dept. of Industrial Engineering,
University of Iowa,
Iowa City, Iowa 52242
e-mail: dbricker@icaen.uiowa.edu

author

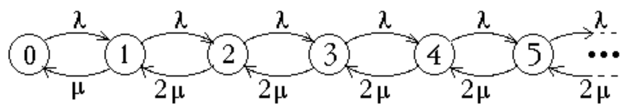
M/M/c

- *Arrival & Service processes are Memoryless, i.e.,*
  interarrival times have Exponential distribution with mean $1/\lambda$
  service times have Exponential distribution with mean $1/\mu$
- *Number of servers is* c
- *Capacity of queueing system is infinite*

M/M/c — Birth/Death Model



Example: M/M/2



$$\frac{1}{\pi_0} = 1 + \frac{\lambda}{\mu} + \left(\frac{\lambda}{\mu}\right)\left(\frac{\lambda}{2\mu}\right) + \left(\frac{\lambda}{\mu}\right)\left(\frac{\lambda}{2\mu}\right)^2 + \left(\frac{\lambda}{\mu}\right)\left(\frac{\lambda}{2\mu}\right)^3 + \cdots$$

$$= 1 + \left(\frac{\lambda}{\mu}\right)\left[1 + \left(\frac{\lambda}{2\mu}\right) + \left(\frac{\lambda}{2\mu}\right)^2 + \left(\frac{\lambda}{2\mu}\right)^3 + \cdots\right]$$

geometric series

$$\frac{1}{\pi_0} = 1 + \left(\frac{\lambda}{\mu}\right)\underbrace{\left[1 + \left(\frac{\lambda}{2\mu}\right) + \left(\frac{\lambda}{2\mu}\right)^2 + \left(\frac{\lambda}{2\mu}\right)^3 + \cdots\right]}$$

geometric series converges to $\frac{1}{1-\lambda/2\mu}$ *if* $\lambda/2\mu < 1$

$$\frac{1}{\pi_0} = 1 + \left(\frac{\lambda}{\mu}\right)\frac{1}{1-\lambda/2\mu}$$

M/M/c

If the arrival rate λ is less than the combined rate $c\mu$ at which the servers can work, then the system will have a ***steadystate*** distribution, given by:

$$\pi_0 = \frac{1}{\sum_{n=0}^{c-1}\frac{(c\rho)^n}{n!} + \frac{(c\rho)^c}{c!}\frac{1}{1-\rho}}$$

$$\pi_j = \frac{(c\rho)^j}{j!}\pi_0 \ , \ j=1,2,\ldots c$$

$$\pi_j = \frac{(c\rho)^j}{c!c^{j-c}}\pi_0 \ , \ j=c,c+1,\ldots.$$

where $\rho = \frac{\lambda}{c\mu} < 1$

Probability that all servers are busy:

$$\sum_{j\geq c}^{\infty} \pi_j = \frac{(c\rho)^c}{c!(1-\rho)}\pi_0 \quad \text{where } \rho = \frac{\lambda}{c\mu} < 1$$

This, then, is the probability that an arriving customer will be required to wait for service!

M/M/c Average Length of Queue

(not including those being served)

$$L_q = \sum_{j\geq c}^{\infty}(j-c)\pi_j \quad \text{where} \quad \pi_j = \frac{(c\rho)^j}{c!c^{j-c}}\pi_0 \ , \ j=c,c+1,\ldots.$$

$$L_q = \sum_{j=0}^{\infty} j\,\pi_{c+j} = \sum_{j=0}^{\infty} j\,\pi_0 \frac{(c\rho)^{c+j}}{c!\,c^j} = \pi_0\frac{(c\rho)^c}{c!}\sum_{j=0}^{\infty} j\,\rho^j$$

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

$$L_q = \pi_0 \frac{(c\rho)^c}{c!} \sum_{j=0}^{\infty} j\,\rho^j = \pi_0 \frac{(c\rho)^c}{c!}\,\rho \underbrace{\sum_{j=0}^{\infty} j\,\rho^{j-1}}_{\text{derivative of a geometric series}}$$

$$\sum_{j=0}^{\infty} j\,\rho^{j-1} = \frac{d}{d\rho}\sum_{j=0}^{\infty} \rho^j = \frac{d}{d\rho}\left(\frac{1}{1-\rho}\right) = \frac{1}{(1-\rho)^2}$$

$$L_q = \pi_0 \frac{(c\rho)^c}{c!}\,\rho \frac{1}{[1-\rho]^2}$$

Average Length of Queue

$$L_q = \frac{\rho\,(c\rho)^c}{c!}\,\pi_0 \left(\frac{1}{1-\rho}\right)^2$$

Once L_q is computed, then we can compute (using Little's formula)

$$W_q = \frac{L_q}{\lambda}, \quad W = W_q + \frac{1}{\mu}, \ \& \ \ L = \lambda W$$

Example: Pooled vs. Separate Servers

Compare two queueing systems:

λ=4/hr ○⟶ ○○▣⟶ μ= 5/hr
λ=4/hr ○⟶ ○○○○▣⟶ μ= 5/hr
separate queue per server

λ=8/hr ○⟶ ○○○○○○ ▣⟶ μ= 5/hr, ▣⟶ μ= 5/hr
pooled servers

two M/M/1 queues

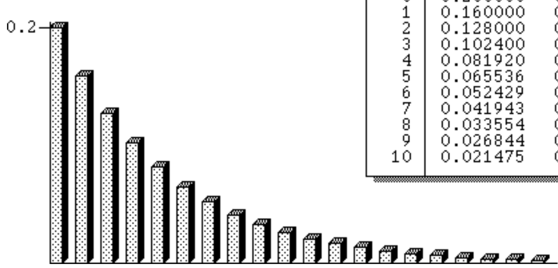
λ=4/hr ○⟶ ○○▣⟶ μ= 5/hr
λ=4/hr ○⟶ ○○○○▣⟶ μ= 5/hr
separate queue per server

Average waiting time: $W_q = \frac{\lambda}{\mu\,(\mu - \lambda)}$

$$W_q = \frac{4/\text{hr}}{(5/\text{hr})(5-4)/\text{hr}} = 0.8 \text{ hr}$$

(48 minutes)

Steady-State Distribution



i	Pi	CDF
0	0.200000	0.200000
1	0.160000	0.360000
2	0.128000	0.488000
3	0.102400	0.590400
4	0.081920	0.672320
5	0.065536	0.737856
6	0.052429	0.790285
7	0.041943	0.832228
8	0.033554	0.865782
9	0.026844	0.892626
10	0.021475	0.914101

single M/M/2 queue

λ=8/hr ○⟶ ○○○○○○ ▣⟶ μ= 5/hr, ▣⟶ μ= 5/hr
pooled servers

Rather than maintaining a separate queue for each server, customers enter a common queue.

$$\rho = \frac{\lambda}{2\mu} = \frac{8/\text{hr}}{2\times 5/\text{hr}} = 0.8 < 1$$ *which implies that a steady state exists!*

single M/M/2 queue

λ=8/hr ○⟶ ○○○○○○ ▣⟶ μ= 5/hr, ▣⟶ μ= 5/hr

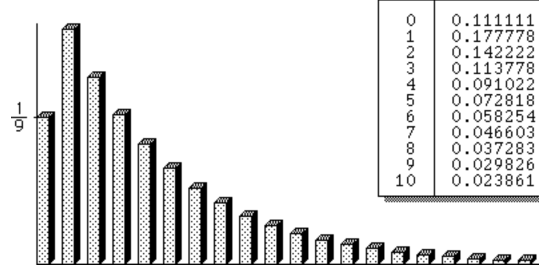
$$\pi_0 = \frac{1}{\frac{(2\times 0.8)^0}{0!} + \frac{(2\times 0.8)^1}{1!} + \frac{(2\times 0.8)^2}{2!\,(1-0.8)}} = \frac{1}{1+1.6+6.4} = \frac{1}{9}$$

$$\pi_0 = 0.111111$$

$$\pi_1 = \frac{(c\rho)^1}{1!}\,\pi_0 = \frac{(2\times 0.8)^1}{1!}\,\frac{1}{9} = 0.1777777$$

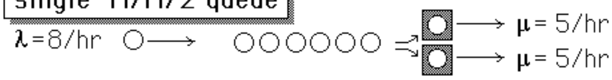
$$P\{\text{both servers busy}\} = 1 - \pi_0 - \pi_1 = 0.7111111$$

Steady-State Distribution



i	Pi	CDF
0	0.111111	0.111111
1	0.177778	0.288889
2	0.142222	0.431111
3	0.113778	0.544889
4	0.091022	0.635911
5	0.072818	0.708729
6	0.058254	0.766983
7	0.046603	0.813586
8	0.037283	0.850869
9	0.029826	0.880695
10	0.023861	0.904556

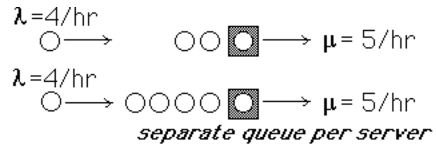
single M/M/2 queue



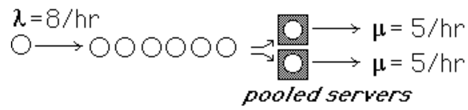
$$L_q = \frac{\rho}{1 - \rho} P\{\text{both servers busy}\}$$

$$= \frac{0.8}{0.2} \ (0.71111111) \quad = 2.844444444$$

$$W_q = \frac{L_q}{\lambda} \quad = 0.35156 \text{ hr.} = 21.1 \text{ minutes}$$



$W_q = 0.8$ hr.
$= 48$ min.



$W_q = 0.352$ hr.
$= 21.1$ min.

By pooling the servers, the average waiting time per customer is reduced by approximately 56%