Stochastic Processes: Lecture 23
Closed queueing networks, QED

Hailun Zhang@SDS of CUHK-Shenzhen

April 21, 2021

Review: $M/M/1$ queue

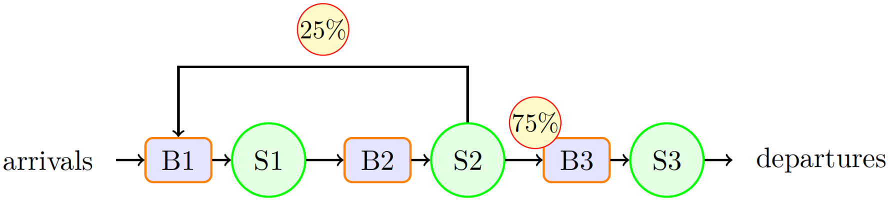
- arrival rate λ, service rate μ
- Define

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

- Assume $\rho < 1$.
- $X = \{X(t), t \geq 0\}$ is a CTMC, where $X(t)$ is the number of jobs in the system.
- Stationary distribution:

$$\pi(n) = (1-\rho)\rho^n \quad n = 0, 1, 2, \ldots$$

Review: a 3-station open network



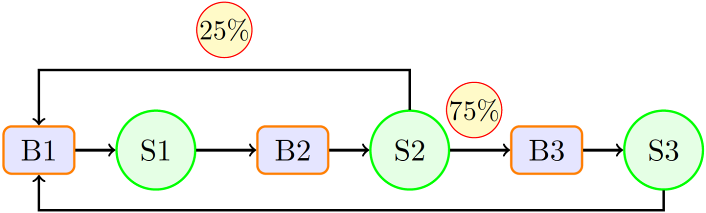
- $\alpha = 1/3$; stationary distribution

$$\pi(4,6,2) = (1-\rho_1)\rho_1^4(1-\rho_2)\rho_2^6(1-\rho_3)\rho_3^2.$$

- $\rho_1 = \lambda_1/\mu_1$, $\lambda_1 = 4/9$, $\lambda_2 = 4/9$, $\lambda_3 = 1/3$

$$\begin{aligned}\lambda_2 &= \lambda_1,\\ \lambda_3 &= .75\lambda_2,\\ \lambda_1 &= \alpha + .25\lambda_2.\end{aligned}$$

Example: a 3-station closed network

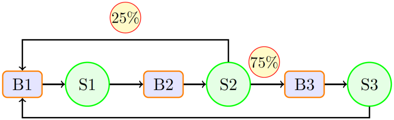


- $N = 10$

$$(10, 0, 0), \quad (9, 1, 0), \quad \ldots$$

- How to find stationary distribution $\pi_{(10,0,0)}$, $\pi_{(9,1,0)}$, $\ldots$?
- Average time in system per job:

A 3-station closed network



- $N = 2$; stationary distribution (Product-form?)

$$\pi(i_1, i_2, i_3) \neq (1-\rho_1)\rho_1^{i_1}(1-\rho_2)\rho_2^{i_2}(1-\rho_3)\rho_3^{i_3}$$

- $\rho_1 = \lambda_1/\mu_1$, $\lambda_1 = 1$, $\lambda_2 = 1$, $\lambda_3 = 3/4$ (infinitely many solutions)

$$\begin{aligned}\lambda_2 &= \lambda_1,\\ \lambda_3 &= .75\lambda_2,\\ \lambda_1 &= \lambda_3 + .25\lambda_2.\end{aligned}$$

Constant C

- $\rho_1 = 1,\ \rho_2 = 1,\ \rho_3 = 3/2.$

$$(2,0,0),(1,1,0),(1,0,1),(0,2,0),(0,0,2),(0,1,1)$$

- Find constant C

$$\pi(2,0,0)+\pi(1,1,0)+\pi(1,0,1) \\ +\pi(0,2,0)+\pi(0,0,2)+\pi(0,1,1)=1.$$

- Find C

$$C+C+C(3/2)+C+C(3/2)^2+C(3/2)=1.$$

$C = \frac{4}{33}$.

- Server 3 utilization:

$$\pi(1,0,1)+\pi(0,0,2)+\pi(0,1,1)=1-3C=\frac{21}{33}.$$

- Stationary distribution

$$\pi_j = \frac{95^j}{j!}\pi_0 \text{ for } j = 0, 1, \ldots, 100,$$

$$\pi_{j+100} = \left(\frac{95}{100}\right)^j \frac{95^{100}}{100!}\pi_0 \text{ for } j = 1, 2, \ldots,$$

- Find π_0

$$1 = \sum_{i=0}^{\infty} \pi_i = \Big[\sum_{i=0}^{100} \frac{95^i}{i!} + \sum_{j=1}^{\infty} \frac{95^{100}}{100!}\rho^j\Big]\pi_0$$

$$= \Big[\sum_{i=0}^{100} \frac{95^i}{i!} + \frac{95^{100}}{100!}\frac{\rho}{1-\rho}\Big]\pi_0$$

The probabilty of an incoming call waits

- The probabilty of an incoming call waits before being answered is

$$\sum_{i=100}^{\infty} \pi_i = \frac{1}{1-\rho}\frac{95^{100}}{100!}\pi_0$$

$$= \frac{\frac{1}{1-\rho}\frac{95^{100}}{100!}}{\sum_{i=0}^{100}\frac{95^i}{i!} + \frac{95^{100}}{100!}\frac{\rho}{1-\rho}}$$

$$= \frac{\frac{1}{1-\rho}}{\frac{\sum_{i=0}^{100}\frac{95^i}{i!}}{\frac{95^{100}}{100!}} + \frac{\rho}{1-\rho}} = \frac{\frac{1}{1-\rho}}{C(100) + \frac{\rho}{1-\rho}},$$

- where

$$C(n) = \frac{\sum_{i=0}^{n}\frac{95^i}{i!}}{\frac{95^n}{n!}} = 1 + (n/95)C(n-1), \quad C(0) = 0.$$

Quality and efficiency-driven (QED) operational regime

$M/M/100$: $\lambda = 95$, $\mu = 1$

- The probability that an incoming call does not wait is 0.4935.
- Average queue size $L_q = \sum_{i=101}^{\infty}(i-100)\pi_i = 9.6227$.
- Average waiting time

$$W_q = L_q/95 = 0.1013 = \sum_{i=1}^{\infty} \frac{i}{100}\pi_{100+i-1} \quad \text{minutes.}$$

- Average utilization per server $\rho = .95$.

For $M/M/1$; $\lambda = 95$, $\mu = 100$

- Average utilization per server $\rho = .95$.
- The probability that an incoming call does not wait is 0.05.
- Average waiting time

$$m\frac{\rho}{1-\rho} = 0.19 \text{ minutes.}$$

Data center design

- Centralized buffer v.s. decentralized buffers
- Routing decisions (load-balancing algorithms) for decentralized buffers
  - random
  - join-shortest-queue
  - “power of two random choices”:

Delay probability

The probability that an incoming customer experiences a delay is

$$\sum_{i=n}^{\infty} \pi_i = \pi_n/(1-\rho) = \frac{\frac{(1/n!)(\lambda/\mu)^n}{1-\rho}}{\sum_{i=0}^{n-1} \frac{1}{i!}(\lambda/\mu)^i + \frac{1/(n!)(\lambda/\mu)^n}{1-\rho}}$$
$$= \frac{\frac{(1/n!)(\lambda/\mu)^n e^{-\lambda/\mu}}{1-\rho}}{\sum_{i=0}^{n-1} \frac{1}{i!}(\lambda/\mu)^i e^{-\lambda/\mu} + \frac{(1/n!)(\lambda/\mu)^n e^{-\lambda/\mu}}{1-\rho}}.$$

Square-root-safety staffing rule: Let $R = \lambda/\mu$ be the offered load.

$$n = R + \beta\sqrt{R}.$$

or

$$R \approx n - \beta\sqrt{n}.$$

Asymptotics

- Stirling formula
$$n! \sim \sqrt{2\pi} n^{n+1/2} e^{-n} \text{ as } n \to \infty,$$
- Taylor expansion
$$\ln(1-x) = -x - \frac{1}{2}x^2 + o(x^2) \quad \text{ as } x \to 0,$$
- Thus
$$\frac{(1/n!)(\lambda/\mu)^n e^{-\lambda/\mu}}{1-\rho} \sim \frac{1}{\sqrt{2\pi}} \frac{1}{\beta} e^{-\beta^2/2} = \frac{1}{\beta}\phi(\beta).$$
- Also
$$\sum_{i=0}^{n-1} \frac{1}{i!}(\lambda/\mu)^i e^{-\lambda/\mu} = \mathbb{P}\{X^{\lambda/\mu} < n\}$$
$$= \mathbb{P}\left\{\frac{X^{\lambda/\mu} - (n-\beta\sqrt{n})}{\sqrt{n-\beta\sqrt{n}}} < \frac{n-(n-\beta\sqrt{n})}{\sqrt{n-\beta\sqrt{n}}}\right\} \to \mathbb{P}\{N(0,1) < \beta\}$$
$$= \Phi(\beta),$$

Delay probability approximation

- the probability of delay is approximated by

$$\frac{\phi(\beta)/\beta}{\Phi(\beta)+\phi(\beta)/\beta} = \frac{1}{1+\beta\Phi(\beta)/\phi(\beta)}$$

  when the number of servers n is large or equivalently the offered load λ/μ is high.

- For $\beta \in [0, .1, .2, .3, .4, .5, .6, .7, .8, .9, 1.0]$, it produces different probabilities of delay:

$$\begin{array}{llllll} 1.0000, & 0.8803, & 0.7714, & 0.6729, & 0.5841, & 0.5045, \\ 0.4335 & 0.3705, & 0.3148, & 0.2660, & 0.2234 & \end{array}$$

Square-root-safety staffing

- For example, if a manager wants to have only 26.6% of her customers experience any delay before being served, she should choose β to be .9.
- With this service level (at 26.6% of delay probability), the staffing rule is
$$n \sim (\lambda/\mu) + \beta\sqrt{\lambda/\mu} = (\lambda/\mu) + (0.9)\sqrt{\lambda/\mu}.$$
- If the offered load is 100, the manager should hire 109 servers.
- If the offered load is 500, the manager should hire 521 servers.
- If the offered load is 1000, the manager should hire 1029 servers.

Utilization with 26.6% delay probability

The following table lists these staffing levels, along with the average utilization per server.

offered load	Number of Servers	Utilization
100	109	91.74%
500	521	96.13%
1000	1029	97.23%