Markov Jump Process Models of Service

Service Engineering - Recitation 12

January 9, 2014

Case 1: Check-in desks at the airport

The arrival rate to the check-in desks at the airport is 282 passengers/hour and the service rate is 60 passengers/hour. There are 5 clerks working in the check-in desks.

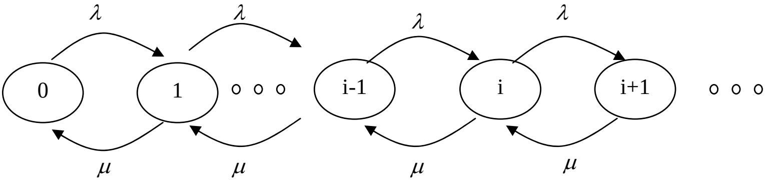
The manager deliberate two designs of queue:

- Five $M/M/1$ queues ($\lambda = 282/5, \mu = 60$, for each).
- One $M/M/5$ queue ($\lambda = 282, \mu = 60$).

Considering a passenger's waiting time, which one of the two designs mentioned above is better?

Case 1: Check-in desks at the airport - $M/M/1$ queue

The transition rate diagram of $M/M/1$ model is:



Using cuts we get:

$$\begin{cases} \lambda \cdot \pi_i = \mu \cdot \pi_{i+1}, \quad i \geq 0 \\ \sum_{i=0}^{\infty} \pi_i = 1 \end{cases}, \qquad \Rightarrow \qquad \begin{cases} \pi_i = \left(\frac{\lambda}{\mu}\right)^i \cdot \pi_0, \quad i \geq 1 \\ \sum_{i=0}^{\infty} \pi_i = 1 \end{cases},$$

Case 1: Check-in desks at the airport - $M/M/1$ queue (2)

Then, number of customers in one desk $L_{System} \sim Geom_0(1-\rho)$, and

$$E(L_{System}) = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu-\lambda} = \frac{0.94}{0.06} = 15\frac{2}{3} \text{ customers.}$$

By Little Law we get that the average time in the system is

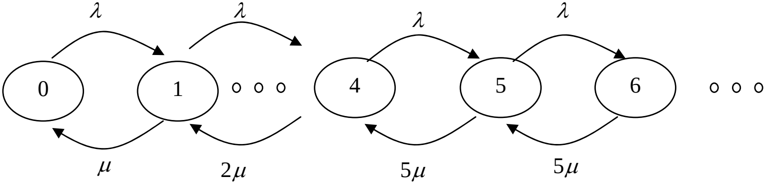
$$E(W_{System}) = \frac{E(L_{System)}}{\lambda} = \frac{15\frac{2}{3}}{\frac{282}{5}} = 0.2778 \text{ hr} = 16.7 \text{ m},$$

and the average waiting time is

$$E(W_q) = E(W_{Sys}) - E(W_{Ser}) = 0.2778 - 1/60 = 0.2611 \text{ hr} = 15.7 \text{ m}.$$

Case 1: Check-in desks at the airport - $M/M/5$ queue

The transition rate diagram of $M/M/5$ model is:



$$
\begin{cases}
\lambda \cdot \pi_i = (i+1) \cdot \mu \cdot \pi_{i+1}, \ i = 0,1,\ldots,4 \ , \\
\lambda \cdot \pi_i = 5\mu \cdot \pi_{i+1}, \ i \geq 5 \ , \\
\sum_{i=0}^{\infty} \pi_i = 1
\end{cases}
\Rightarrow
\begin{cases}
\pi_i = \frac{1}{i!} \left(\frac{\lambda}{\mu}\right)^i \cdot \pi_0, \ i = 0,1,\ldots,4 \ , \\
\pi_i = \frac{1}{5!} \cdot \frac{1}{5^{i-5}} \cdot \left(\frac{\lambda}{\mu}\right)^i \cdot \pi_0, \ i \geq 5 \ , \\
\sum_{i=0}^{\infty} \pi_i = 1
\end{cases}
$$

Case 1: Check-in desks at the airport - $M/M/5$ queue (2)

$$
\begin{aligned}
E(L_q) &= \sum_{i=5}^{\infty}(i-5)\pi_i = \sum_{i=5}^{\infty}(i-5)\frac{1}{5!}\cdot\frac{1}{5^{i-5}}\cdot\left(\frac{\lambda}{\mu}\right)^i\pi_0 \\
&= \frac{1}{5!}\cdot\left(\frac{\lambda}{\mu}\right)^5\cdot\pi_0\cdot\sum_{k=0}^{\infty}k\cdot\left(\frac{\lambda}{5\mu}\right)^k \\
&= \frac{1}{5!}\cdot\left(\frac{\lambda}{\mu}\right)^5\cdot\left(1-\frac{\lambda}{5\mu}\right)^{-1}\cdot\pi_0\cdot\sum_{k=0}^{\infty}k\cdot\left(\frac{\lambda}{5\mu}\right)^k\cdot\left(1-\frac{\lambda}{5\mu}\right) \\
&= \frac{1}{5!}\cdot\left(\frac{\lambda}{\mu}\right)^5\cdot\left(1-\frac{\lambda}{5\mu}\right)^{-1}\cdot\pi_0\cdot E\left[Geom_0\left(1-\frac{\lambda}{5\mu}\right)\right] \\
&= 13.382 \text{ customers.}
\end{aligned}
$$

Then, by Little Law, we get

$$E(W_q) = \frac{E(L_q)}{\lambda} = \frac{13.382}{282} = 0.04745 \text{ hr} = 2.85 \text{ m}.$$

Case 1: Check-in desks at the airport - Comparison

- Considering the average waiting time, the second queue design (one pooled queue) is better.
- This is an example of so called Economies of Scale (EOS)
  - In larger systems (call volume / agents) some of the randomness is eliminated (law of large numbers) and thus better performance is achieved.

Case 2: Bank Teller Conflict

A clerk provides service for two types of customers: face-to-face customers and tele-customers. The service time is exponential with an average duration of 12 minutes for both types. The arrival rate is 3 customers/hour for face-to-face customers, and 4 customers/hour for tele-customers.

- No queueing in the system: a face-to-face customer that arrives when the clerk is serving another customer abandons immediately; a tele-customer that receives a busy signal abandons too.
- The tele-customers enjoy *preemptive priority*: if the teller is serving a face-to-face customer, and the phone rings, he serves the tele-customer and returns to the face-to-face service only after call termination.

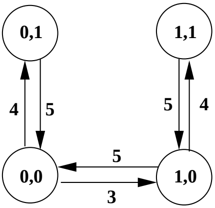
Case 2: Bank Teller (2)

In order to model the system by a Markov process $X = \{X(t),\ t \geq 0\}$, we define

$$X(t) = (X_1(t), X_2(t)),\ t \geq 0,$$

where $X_1(t)$ = number of face-to-face customers in the system (0 or 1); $X_2(t)$ = number of tele-customers in the system (0 or 1).

Transition-rates diagram:



Part 1: Q1

Q1: Solve the steady-state equations for the Markov process X.

Using cuts we get:

$5\pi_{01} = 4\pi_{00}; \quad 3\pi_{00} = 5\pi_{10}; \quad 4\pi_{10} = 5\pi_{11}; \quad \pi_{00} + \pi_{01} + \pi_{10} + \pi_{11} = 1;$

which implies

$$\pi_{00} = 0.347; \qquad \pi_{01} = 0.278; \qquad \pi_{10} = 0.208; \qquad \pi_{11} = 0.167.$$

Part 1: Q2

Q2: Describe the utilization profile of the clerk. Specifically,

- fraction of time devoted to face-to-face service;
- fraction of time devoted to tele-service;
- fraction of time when the server is idle.

$\rightarrow$ The utilization profile:

- face-to-face service: $\pi_{10} = 20.8\%$;
- tele-service: $\pi_{01} + \pi_{11} = 44.4\%$;
- idleness: $\pi_{00} = 34.7\%$;

Part 1: Q3-Q4

Q3: Calculate the average frequency of the following event (per hour): "Face-to-face service is interrupted by a phone call".

$\rightarrow$ The event "Face-to-face service is interrupted by a phone call" corresponds to a jump from (1,0) to (1,1). The frequency of such jumps is equal to: $4 \cdot \pi_{10} = 0.833$ jumps per hour.

Q4: Compute the fraction of customers that arrive to face-to-face service and find that the server is busy.

$\rightarrow$ From PASTA the fraction of customers that arrive to face-to-face service and find the server busy is $1 - \pi_{00} = 0.653$.

Part 1: Q5-Q6

Q5: Compute the overall proportion of customers (both face-to-face and tele-customers) that start service immediately upon arrival.

$\rightarrow$ The overall proportion of customers that start service immediately upon arrival is (PASTA): $\frac{(3+4)\cdot\pi_{00}+0\cdot\pi_{01}+4\cdot\pi_{10}+0\cdot\pi_{11}}{3+4} = 0.466 = 46.6\%$.

Q6: How many customers, on average, are served during a working day (8 hours)? (Assume that the system reaches steady state in a short time.)

$\rightarrow$ The average number of customers, served during a working day (8 hours): $8\cdot\lambda_{eff} = 8\cdot 7\cdot 0.466 = 26.096$.

Part 1: Q7

Q7: Assume that a face-to-face customer is starting service. How many times, on average, will her service be interrupted by phone calls? Calculate the average overall service time of a face-to-face customer that received service, including all interrupts due to phone calls.

$\rightarrow$ Denote the number of interruptions by M. Due to "competition between exponentials" considerations:

$$\mathrm{P}\{M=0\}=\frac{5}{9}; \qquad \mathrm{P}\{M=1\}=\frac{5}{9}\cdot\frac{4}{9}; \qquad \mathrm{P}\{M=k\}=\frac{5}{9}\cdot\left(\frac{4}{9}\right)^k.$$

Hence,

$$M\sim\mathrm{Geom}\left(\frac{5}{9}\right); \qquad \mathrm{E}M=\frac{q}{p}=\frac{4/9}{5/9}=0.8.$$

The overall average service time is equal to the sum of expected service time and expected waiting time:

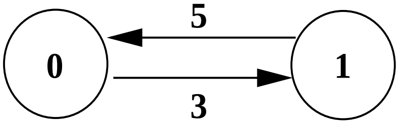
$$12+12\cdot 0.8=21.6 \text{ min.}$$

Part 2

The office manager examines the option of hiring another clerk to separate the two service types: the first server would cater to face-to-face customers and the second would answer phone calls. As before, the customers that encounter a busy server abandon immediately.

Analyze the new system from the servers' point of view (utilization profile) and from the customers' point of view (probability of getting service).

Part 2: Face-to-face service

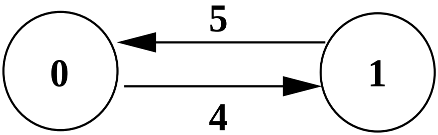


$$5\pi_1 = 3\pi_0; \qquad \pi_1 = \frac{3}{8}\,, \quad \pi_0 = \frac{5}{8}\;.$$

The agent is busy $\pi_1 = 37.5\%$ of time.

The fraction of customers getting service $\pi_0 = 62.5\%$.

Part 2: Tele-service



$$5\pi_1 = 4\pi_0; \qquad \pi_1 = \frac{4}{9}\ , \quad \pi_0 = \frac{5}{9}\ .$$

The agent is busy $\pi_1 = 44.4\%$ of time.

The fraction of customers getting service $\pi_0 = 55.6\%$.

The overall fraction of served customers:

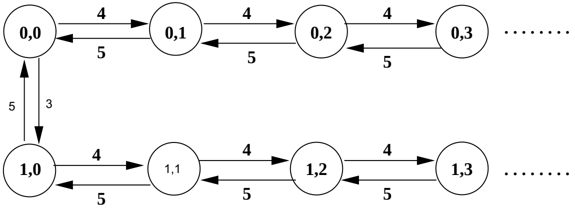
$$\frac{3 \cdot 62.5\% + 4 \cdot 55.6\%}{7} \ = \ 58.5\%\ .$$

Part 3

Another option, considered by the manager, is to add a number of telephone lines to reduce the fraction of customers that get the busy signal to a negligible level. Assume that there are no abandonment. Solve questions from Part 1 in this new setting.

Define the states of the process by (i, j), where i is the number of face-to-face customers and j is the number of tele-customers.

Transition-rates diagram:



Part 3: Q1

Q1: Solve the steady-state equations for the Markov process X.

By the method of cuts:

$$\pi_{0k} = \left(\frac{4}{5}\right)^k \pi_{00} \ , \quad \pi_{10} = \frac{3}{5}\pi_{00} \ , \quad \pi_{1k} = \left(\frac{4}{5}\right)^k \pi_{10} \ .$$

Then

$$\pi_{00} \ = \ \frac{1}{\sum_{k=0}^{\infty}\left(\frac{4}{5}\right)^k + \frac{3}{5}\sum_{k=0}^{\infty}\left(\frac{4}{5}\right)^k} \ = \ \frac{1}{5+3} \ = \ \frac{1}{8} \ ,$$

$$\pi_{10} \ = \ \frac{3}{40} \ .$$

Part 3: Q2

Q2: Describe the *utilization profile* of the clerk. Specifically,

- fraction of time devoted to face-to-face service;
- fraction of time devoted to tele-service;
- fraction of time when the server is idle.

$\rightarrow$ The utilization profile:

- face-to-face service: $\pi_{10} = \dfrac{3}{40} = 7.5\%$;
- tele-service: $1 - \pi_{00} - \pi_{10} = 80\%$;
- idleness: $\pi_{00} = 12.5\%$.

Part 3: Q3-Q4

Q3: Calculate the average frequency of the following event (per hour): "Face-to-face service is interrupted by a phone call".

$\rightarrow$ The frequency of jumps from (1,0) to (1,1):

$$4 \cdot \pi_{10} = 0.3.$$

Q4: Compute the fraction of customers that arrive to face-to-face service and find that the server is busy.

$\rightarrow$ From PASTA the fraction of customers that arrive to face-to-face service and find the server busy is

$$1 - \pi_{00} \;=\; \frac{7}{8} \;=\; 87.5\% \;.$$

Part 3: Q5-Q6

Q5: Compute the overall proportion of customers (both face-to-face and tele-customers) that start service immediately upon arrival.

$\rightarrow$ The overall proportion of customers that start service immediately upon arrival is (PASTA):

$$\frac{(3+4)\cdot\pi_{00}+4\cdot\pi_{10}}{3+4} = 0.168 = 16.8\% - \text{reduced from } 46.6\% \ .$$

Q6: How many customers, on average, are served during a working day (8 hours)? (Assume that the system reaches steady state in a short time.)

$\rightarrow$ The effective arrival rate: $\lambda_{eff} = 4+3\cdot\pi_{00} = 4.375$.

The average number of customers served during a working day: $8\lambda_{eff} = 35$.