SA402 – Dynamic and Stochastic Models Fall 2013
Asst. Prof. Nelson Uhan

Lesson 22. Standard Queueing Models

0 Warm up

Example 1 (Nelson 8.4, modified, cont.)**.** A small ice-cream shop competes with several other ice-cream shops in a busy mall. If there are too many customers already in line at the shop, then potential customers will go elsewhere. Potential customers arrive at a rate of 20 per hour. The probability that a customer will go elsewhere is $j/5$ when there are $j \leq 5$ customers already in the system, and 1 when there are $j > 5$ customers already in the system. The server at the shop can serve customers at a rate of 10 per hour. Approximate the process of potential arrivals as Poisson, and the service times as exponentially distributed.

Last time, we modeled the customer arrival and departure process as a birth-death process with rates:

$$\lambda_i = \begin{cases} 20(1 - i/5) & \text{if } i = 0, 1, \dots, 5 \\ 0 & \text{otherwise} \end{cases} \qquad \mu_i = 10 \quad \text{for } i = 1, 2, \dots$$

Note that the number of servers $s = 1$. The resulting steady-state probabilities are:

i	0	1	2	3	4	5	≥ 6
π_i	0.07	0.14	0.22	0.27	0.21	0.09	0

a. Over the long run, what fraction of the time is the server busy?

b. Over the long run, at what rate are customers lost?

c. Suppose that the shop makes a revenue of \$2 per customer served and pays the server \$4 per hour. What is the shop's long-run expected profit per hour (revenue minus cost)?

1 Standard queueing notation

- Standard notation for queues:

  arrival process / service process / s / n / k / queue discipline

- Arrival process (interarrival times usually assumed to be independent and time stationary)
  - M = Markovian (Poisson process / exponential interarrival times)
  - G = General distribution
  - D = deterministic (fixed interarrival times)
  - E_k = Erlang with k phases
- Service process (service times usually assumed to be independent and time stationary)
  - M = Markovian (exponential service time)
  - D = deterministic (fixed service times)
  - G, E_k
- s = number of servers
- n = system capacity (e.g. number of customers that can be in the shop)
- k = number of potential customers (size of customer population)
- Queueing discipline:
  - FIFO = first in first out (equivalently, FCFS = first come first served)
  - LIFO = last in first out (equivalently, LIFO = last come first served)
  - SIRO = service in random order
- If not specified, default values:
  - $n = \infty$
  - $k = \infty$
  - queueing discipline = FIFO
- Two fundamental standard queues:
  - M/M/∞: birth-death process with
  - M/M/s: birth-death process with

2 The M/M/∞ queue

- Let's apply the formulas for the steady-state probabilities:

$$\pi_j = \frac{d_j}{\sum_{i=0}^{\infty} d_i} \quad \text{for } j = 0, 1, 2, \dots \quad \text{where} \quad d_0 = 1, \quad d_j = \prod_{i=1}^{j} \frac{\lambda_{i-1}}{\mu_i} \quad \text{for } j = 1, 2, \dots$$

- We can simplify d_j:

- Therefore, we can rewrite $\sum_{j=0}^{\infty} d_j$:

- As a result, the steady-state probabilities for a M/M/∞ queue are:

- The number of customers in steady state L is

Example 2. Recall the Massive Mall case: we want to determine the number of parking spaces needed for the new mall by pretending that parking is unlimited, and then investigating how many spaces are sufficient to satisfy demand a large fraction of the time. Assume customers arrive according to a Poisson process with arrival rate $\lambda = 1000$ per hour. In addition, suppose the time that a customer spends at the mall is exponentially distributed with rate $\mu = 1/3$.

a. What is the expected number of cars in the parking lot?
b. What is the expected time a car spends in the parking lot?
c. What is the minimum number of parking spaces needed to hold all cars 99.9% of the time?

3 The M/M/s queue

- Steady-state probabilities: using $\rho = \lambda/(s\mu)$,

$$\pi_0 = \left[\left(\sum_{j=0}^{s} \frac{(s\rho)^j}{j!} \right) + \frac{s^s \rho^{s+1}}{s!(1-\rho)} \right]^{-1} \qquad \pi_j = \begin{cases} \dfrac{(\lambda/\mu)^j}{j!}\pi_0 & \text{for } j = 1, 2, \ldots, s \\ \dfrac{(\lambda/\mu)^j}{s! s^{j-s}}\pi_0 & \text{for } j = s+1, s+2, \ldots \end{cases}$$

- Expected number of customers in queue and expected delay:

$$\ell_q = \frac{\pi_s \rho}{(1-\rho)^2} \qquad w_q = \frac{\ell_q}{\lambda} = \frac{\pi_s \rho}{\lambda(1-\rho)^2}$$

- Expected number of customers in the system and expected waiting time:

$$\ell = \lambda w = \ell_q + \frac{\lambda}{\mu} \qquad w = w_q + \frac{1}{\mu}$$

Example 3. Recall the Darker Image case: we considered adding a second photocopier to the copy shop. Suppose customers arrive according to a Poisson process with rate 4 customers per hour, and that the service time of each photocopier is exponentially distributed with a mean of 12 minutes. Compare the expected delay of customers when there is 1 copier vs. when there are 2 copiers.

4 The G/G/s queue

- When interarrival times and service times are not Markovian (exponentially distributed), things get much harder
- We can still use results from Markovian queues to approximate performance measures, usually with a "correction"
- Setup:
    - G = generic interarrival time random variable with $\lambda = 1/E[G]$
    - X = generic service time random variable with $\mu = 1/E[X]$
    - **Squared coefficients of variation**:

$$\varepsilon_a = \frac{\mathrm{Var}[G]}{E[G]^2} \qquad \varepsilon_s = \frac{\mathrm{Var}[X]}{E[X]^2}$$

- Let $\hat{w}_q$ be the expected delay in this G/G/s queue
- Let w_q be the expected delay in a M/M/s queue with arrival rate λ and service rate μ
- Whitt's (1983) approximation:

$$\hat{w}_q \approx \frac{\varepsilon_a + \varepsilon_s}{2} w_q$$

- We can use this with Little's law (both versions) to find approximations of ℓ_q, ℓ, and w
- This approximation works well when ρ, ε_a, and ε_s are "close" to 1
- When $G \sim \mathrm{Exp}(\lambda)$ and $X \sim \mathrm{Exp}(\mu)$: $\mathrm{Var}(G) =$ □, $\mathrm{Var}(X) =$ □

$\Rightarrow \varepsilon_a =$ □, $\varepsilon_s =$ □ $\Rightarrow$ □

- Note: Whitt's (1983) is one of many approximations that have been proposed for G/G/s queues

Example 4. Consider the Darker Image case again. Suppose now that the service time of each photocopier is uniformly distributed between 2 and 8 minutes. Now compare the expected delay of customers when there is 1 copier vs. when there are 2 copiers.