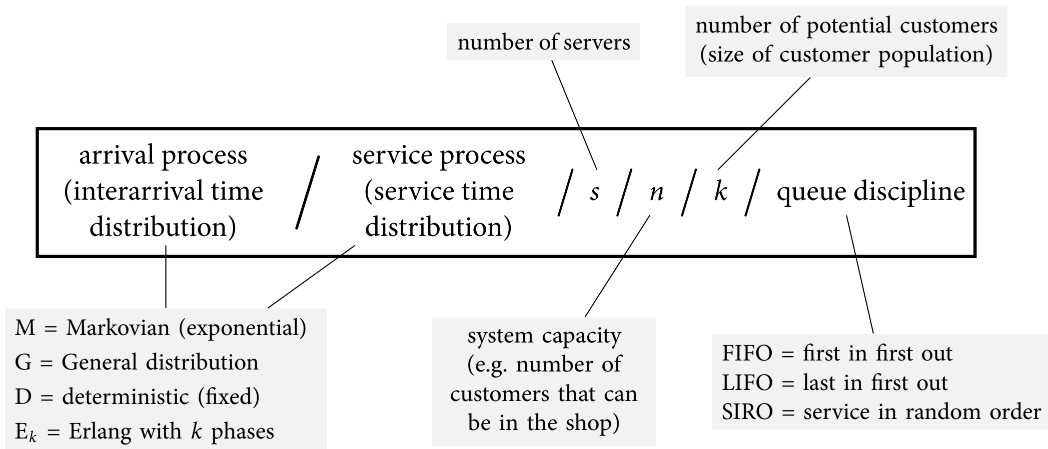
Lesson 14. Standard Queueing Models

1 Standard queueing notation



- If not specified, default values:

$$n = \infty \qquad\qquad k = \infty \qquad\qquad \text{queueing discipline = FIFO}$$

Example 1. What does M/M/∞ mean? Model an M/M/∞ queue as a birth-death process by specifying the arrival and service rates.

Example 2. What does M/M/s mean? Model an M/M/s queue as a birth-death process by specifying the arrival and service rates.

2 The M/M/∞ queue

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

$$\pi_j = \frac{d_j}{D} \quad \text{for } j = 0, 1, 2, \ldots \quad \text{where} \quad d_0 = 1, \quad d_j = \frac{\lambda_0 \lambda_1 \cdots \lambda_{j-1}}{\mu_1 \mu_2 \ldots \mu_j} \quad \text{for } j = 1, 2, \ldots, \quad D = \sum_{i=0}^{\infty} d_i$$

- We can simplify d_j:

- Therefore, we can rewrite D:

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

- The number of customers in steady state L is

Example 3. 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 an arrival rate of 1000 per hour. In addition, suppose the time that a customer spends at the mall is exponentially distributed with a mean of 3 hours.

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 $\rho = \lambda/(s\mu)$,

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

- Expected number of customers in queue and expected delay:

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

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

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

Example 4. The Darker Image copy shop is considering adding a second photocopier. 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
- Usually, we resort to simulation to understand these queues – this is the focus of SA421!
- However, 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**:

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

- Let $\hat{w}_q$ = expected delay in this G/G/s queue
- Let w_q = 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 $G \sim \text{Exp}(\lambda)$ and $X \sim \text{Exp}(\mu)$:

$\text{Var}(G) = \boxed{}$ $\quad \text{Var}(X) = \boxed{}$ $\quad \Rightarrow \quad \varepsilon_a = \boxed{}$ $\quad \varepsilon_s = \boxed{}$

$\Rightarrow \quad \boxed{}$

- Note: This is one of many approximations that have been proposed for G/G/s queues

Example 5. Consider the Darker Image case from Example 4 again. Suppose now that the service time of each photocopier is uniformly distributed between 3 and 21 minutes. Now compare the expected delay of customers when there is 1 copier vs. when there are 2 copiers.

5 Exercises

Problem 1. The Kalman Theater Group (KTG) is building a movie theater mega-complex. They have decided that there will automatic ticket kiosks in front of a single first-come-first-served queue, but they still need to decide how many kiosks to include in the complex design. Based on data that they have collected from their other theaters in similar markets, they have estimated that customers arrive at the kiosks at a rate of 5 per minute, and customers can be served in 2 minutes on average. KTG's standard configuration for mega-complexes in similar markets is 12 kiosks. You have been asked to evaluate this standard configuration.

Assume that the interarrival times and the service times are exponentially distributed.

a. What standard queueing model fits this setting best?
b. What is the traffic intensity in this queueing system?
c. What is the long-run expected fraction of time that all kiosks are unoccupied?
d. Over the long-run, what is the expected time a customer waits in line?