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

Lesson 21. The Birth-Death Process – Performance Measures

1 Last time...

- A **birth-death process** is a Markov process with state space $\mathcal{M} = \{0, 1, 2, \dots\}$ with generator matrix

$$\mathbf{G} = \begin{pmatrix} -\lambda_0 & \lambda_0 & 0 & 0 & \cdots \\ \mu_1 & -(\lambda_1 + \mu_1) & \lambda_1 & 0 & \cdots \\ 0 & \mu_2 & -(\lambda_2 + \mu_2) & \lambda_2 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

- Steady-state probabilities $\pi_0, \pi_1, \pi_2, \dots$:

$$\pi_j = \frac{d_j}{\sum_{i=0}^{\infty} d_i} \quad \text{for } j = 0, 1, 2, \dots$$

  where

$$d_0 = 1, \quad d_j = \prod_{i=1}^{j} \frac{\lambda_{i-1}}{\mu_i} \quad \text{for } j = 1, 2, \dots$$

- Interpretation:

$$\begin{aligned} \pi_j &= \text{probability we find } j \text{ customers in the system in the long run} \\ &= \text{long-run fraction of time that there are } j \text{ customers in the system} \end{aligned}$$

- Today: how to use the steady-state probabilities to compute long-run performance measures

2 System-level performance measures

- How do we measure the workload or congestion in the entire system?
- **Expected number of customers in the system ℓ**

- **Expected number of customers in the queue ℓ_q**

  where s is the number of customers that can be served simultaneously

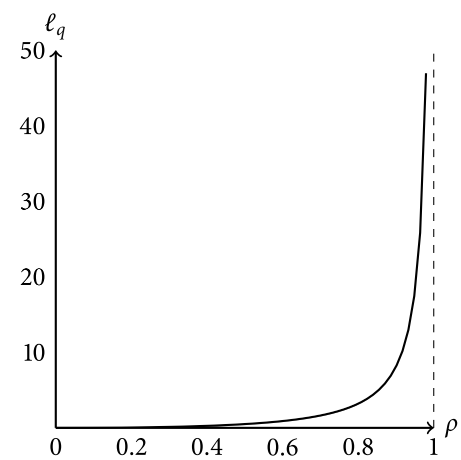
- **Traffic intensity** ρ

where λ is the arrival rate in all states, and μ is the service rate of s identical servers

  - $\rho < 1 \;\Rightarrow\;$ (typically) the system is **stable**: the number of customers does not grow without bound
  - $\rho \geq 1 \;\Rightarrow\;$ customers are arriving faster than they can be served
  - When $\rho < 1$, the traffic intensity ρ is also known as the long-run **utilization** of each server: i.e., the fraction of time each server is busy

- **Offered load** o: the expected number of busy servers

- Relationship between ρ and ℓ_q: ℓ_q explodes as ρ approaches 1



  - This example: Poisson arrivals with rate λ, one server with constant service rate μ: $\ell_q = \dfrac{\rho^2}{1-\rho}$

3 Customer-level performance measures

- **Effective arrival rate** to the system λ_{eff}: "average arrival rate"

- **Little's law** (system-wide)

  where w is the **expected waiting time**: the expected time a customer spends in the system from arrival to departure

  - Deep result, difficult to prove rigorously
  - Intuitively, why does this hold?
    - System is stable $\Rightarrow$ departure rate = arrival rate ("conservation of customers")

    $\Rightarrow$ λ_{eff} should also be the "effective departure rate" from the queueing system

    - Departure rate for an individual customer ≈

    $\Rightarrow$ Departure rate for whole system ≈

    - Therefore,

- **Little's law** (queue only)

  where w_q is the **expected delay**: the expected time a customer spends in the queue

- If the service rate is a constant μ, then waiting time and delay are related like so:

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.

a. Model the process of customer arrivals and departures at this ice-cream shop as a birth-death process (i.e. what are λ_i and μ_i for $i = 0, 1, 2, \ldots$?).
b. Over the long run, how many customers are in the shop? (i.e. what is the probability there are 0 customers in the shop? 1? 2? etc.)
c. On average, how many customers are waiting to be served (not including the customer in service)?
d. Over the long run, what fraction of the time is the server busy?
e. What is the effective arrival rate?
f. What is the expected customer delay?
g. What is the expected customer waiting time?
h. Over the long run, at what rate are customers lost?
i. 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)?