대기행렬 이론 개론

V. M/M/1 Queue

학습에 앞서

- 학습 목표
  - M/M/1, M/M/1/K 대기행렬 모델을 이해한다.
  - 상태전이확률, 정상상태확률을 구하는 능력을 배양한다.

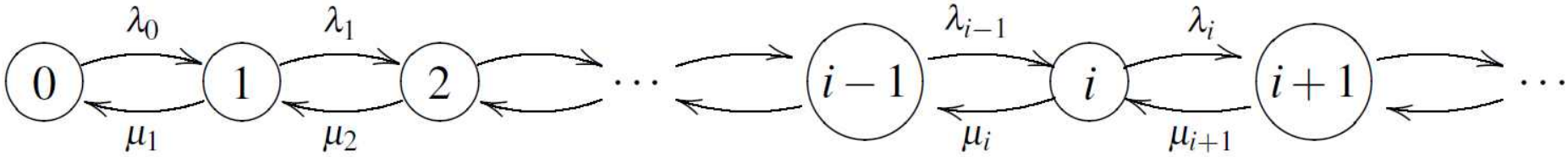
- 목차
  - 1. Continuous-Time Birth-Death Process
  - 2. The M/M/1 Queue
  - 3. The M/M/1/K Queue

1. Continuous-Time Birth-Death Process

Definition A birth and death process $(X(t), t \geq 0)$ is a continuous-time discrete-space (with state-space $\mathbb{N}$) Markov process such that

1. $P(X(t+\Delta t) = i+1 | X(t) = i) = \lambda_i \Delta t + o(\Delta t),\ i \geq 0$
2. $P(X(t+\Delta t) = i-1 | X(t) = i) = \mu_i \Delta t + o(\Delta t),\ i \geq 1$
3. $P(X(t+\Delta t) = i | X(t) = i) = 1 - (\lambda_i + \mu_i)\Delta t + o(\Delta t),\ i \geq 0.$

The r.v. $X(t)$ may be interpreted as the size of the population at time t. In that case, $\lambda_i \geq 0$ gives the birth-rate when the size of the population is i and $\mu_i \geq 0$ gives the death-rate when the size of the population is i with $i \geq 1$. We assume that $\mu_0 = 0$. □



State transition rate diagram for the birth-death process

Steady-state Probability (1/2)

Transition rates

$$q_{i,j} = \begin{cases} \lambda_i & \text{when } j = i+1 \quad \text{probability of birth in interval } \Delta t \text{ is } \lambda_i \Delta t \\ \mu_i & \text{if } j = i-1 \quad \text{probability of death in interval } \Delta t \text{ is } \mu_i \Delta_i t \\ 0 & \text{otherwise} \quad \text{when the system is in state } i \end{cases}$$

Let $\pi_i(t) = P(X(t) = i)$. The time-dependent state probability vector $\pi(t)$ is determined by the equation

$$\frac{d}{dt}\pi(t) = \pi(t) \cdot \mathbf{Q}$$

where

$$\mathbf{Q} = \begin{pmatrix} -\lambda_0 & \lambda_0 & 0 & \cdots & \cdots \\ \mu_1 & -(\lambda_1+\mu_1) & \lambda_1 & 0 & \cdots \\ 0 & \mu_2 & -(\lambda_2+\mu_2) & \lambda_2 & 0 \\ \vdots & 0 & \mu_3 & -(\lambda_3+\mu_3) & \lambda_3 \\ \vdots & \vdots & 0 & \mu_4 & -(\lambda_4+\mu_4) \end{pmatrix}$$

Steady-state Probability (2/2)

Or, from Definition for $i \geq 1$, we have

$$\begin{aligned}
\pi_i(t+\Delta t) &= (\lambda_{i-1}\Delta t + o(\Delta t))\pi_{i-1}(t) + (\mu_{i+1}\Delta t + o(\Delta t))\pi_{i+1}(t) \\
&\quad + (1-(\lambda_i+\mu_i)\Delta t + o(\Delta t))\pi_i(t) + o(\Delta t) \\
\lim_{\Delta t \to 0} \frac{\pi_i(t+\Delta t) - \pi_i(t)}{\Delta t} &= \lambda_{i-1}\pi_{i-1}(t) + \mu_{i+1}\pi_{i+1}(t) - (\lambda_i+\mu_i)\pi_i(t) \\
\frac{d}{dt}\pi_i(t) &= \underbrace{\lambda_{i-1}\pi_{i-1}(t) + \mu_{i+1}\pi_{i+1}(t)}_{\text{flows in}} - \underbrace{(\lambda_i+\mu_i)\pi_i(t)}_{\text{flows out}}.
\end{aligned}$$

An interesting question is the following: what happens when $t \to \infty$. In other words, we are now interested in the *equilibrium* behavior, if it exists. Assume that

$$\pi_i = \lim_{t \to \infty} \pi_i(t)$$

Then, in steady-state,

$$0 = \lambda_{i-1}\pi_{i-1} + \mu_{n+1}\pi_{i+1} - (\lambda_i+\mu_i)\pi_i$$

Balance Equations (1/2)

Result (Balance equations of a birth and death process)

$$\lambda_0 \pi_0 = \mu_1 \pi_1$$
$$(\lambda_i + \mu_i)\pi_i = \lambda_{i-1}\pi_{i-1} + \mu_{i+1}\pi_{i+1} \quad \text{for } i \geq 1$$

the probability flow out of a state = the probability flow into that state

Balance Equations (2/2)

$$\pi_1 = \frac{\lambda_0}{\mu_1}\pi_0$$

$$\pi_2 = \frac{\lambda_0\lambda_1}{\mu_1\mu_2}\pi_0$$

$$\vdots$$

$$\pi_i = \frac{\lambda_0\lambda_1\cdots\lambda_{i-1}}{\mu_1\mu_2\cdots\mu_i}\pi_0 = \frac{\prod_{k=0}^{i-1}\lambda_k}{\prod_{k=1}^{i}\mu_k}\pi_0$$

and from the normalization condition,

$$\sum_{i=0}^{\infty}\pi_i = 1 \qquad \pi_0 + \sum_{i=1}^{\infty}\frac{\prod_{k=0}^{i-1}\lambda_k}{\prod_{k=1}^{i}\mu_k}\pi_0 = 1$$

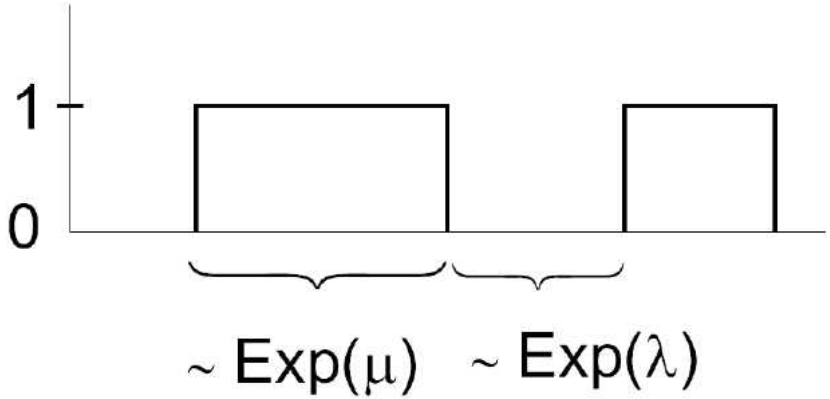
Hence,

$$\pi_0 = \left(1 + \sum_{i=1}^{\infty}\frac{\prod_{k=0}^{i-1}\lambda_k}{\prod_{k=1}^{i}\mu_k}\right)^{-1} \qquad \pi_i = \frac{\prod_{k=0}^{i-1}\lambda_k}{\prod_{k=1}^{i}\mu_k}\pi_0$$

Example: A Single Server System (1/2)

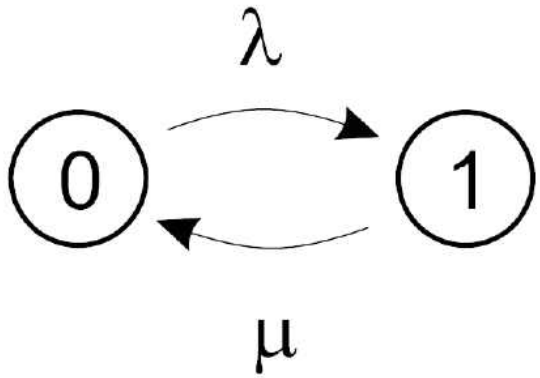
- constant arrival rate λ (Poisson arrivals)
- stopping rate of the service μ (exponential distribution)

The states of the system $\begin{cases} 0 & \text{server free} \\ 1 & \text{server busy} \end{cases}$



Example: A Single Server System (2/2)

$$\begin{cases} \frac{d}{dt}\pi_0(t) &= -\lambda\pi_0(t) + \mu\pi_1(t) \\ \frac{d}{dt}\pi_1(t) &= \lambda\pi_0(t) - \mu\pi_1(t) \end{cases}$$



By adding both sides of the equations

$$\frac{d}{dt}(\pi_0(t) + \pi_1(t)) = 0 \Rightarrow \pi_0(t) + \pi_1(t) = 1 \Rightarrow \pi_1(t) = 1 - \pi_0(t)$$

$$\frac{d}{dt}\pi_0(t) + (\lambda+\mu)\pi_0(t) = \mu \Rightarrow \frac{d}{dt}\left(e^{(\lambda+\mu)t}\pi_0(t)\right) = \mu e^{(\lambda+\mu)t}$$

$$\pi_0(t) = \frac{\mu}{\lambda+\mu} + \left(\pi_0(0) - \frac{\mu}{\lambda+\mu}\right) e^{-(\lambda+\mu)t}$$

$$\pi_1(t) = \underbrace{\frac{\lambda}{\lambda+\mu}}_{\text{equilibrium distribution}} + \underbrace{\left(\pi_1(0) - \frac{\lambda}{\lambda+\mu}\right)}_{\text{deviation from the equilibrium}} \underbrace{e^{-(\lambda+\mu)t}}_{\text{decays exponentially}}$$

2. The *M/M/1* Queue

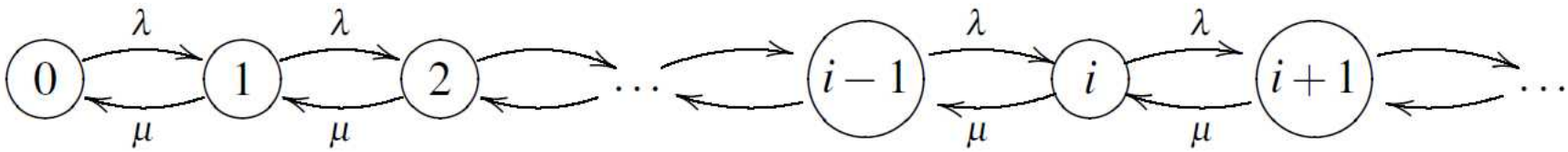
- Introduction

In this queueing system the customers arrive according to a Poisson process with rate λ. The time it takes to serve every customer is exponentially distributed with a mean of $1/\mu$. The service times are supposed to be mutually independent and further independent of the interarrival times.

When a customer enters an empty system his service starts at once; if the system is nonempty the incoming customer joins the queue and waits for service. When a service completion occurs, a customer from the queue, if any, enters the service facility at once to get served.

State Transition Rate Diagram

- System state: $s(i)$, where i is the number of customers in the system.
- Arrival process: $\sim PP(\lambda)$.
- Service process: $\sim Exp(\mu)$.
- Steady-state probability: π_i, $i = 0, 1, 2, \cdots$.



State transition rate diagram for $M/M/1$

The process of $M/M/1$ is a birth-death process with birth rate $\lambda_i = \lambda$ and with death rate $\mu_i = \mu$.

Steady-state Probability

Result (Balance equations for $M/M/1$) In steady-state, we have

$$\lambda \pi_0 = \mu \pi_1$$
$$(\lambda + \mu)\pi_i = \lambda \pi_{i-1} + \mu \pi_{i+1}, \quad i \geq 1$$

$$\pi_1 = \frac{\lambda}{\mu}\pi_0$$
$$\pi_2 = \left(\frac{\lambda}{\mu}\right)^2 \pi_0$$
$$\vdots$$
$$\pi_i = \left(\frac{\lambda}{\mu}\right)^i \pi_0$$

and from the normalization condition,

$$\sum_{i=0}^{\infty} \pi_i = \sum_{i=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^i \pi_0 = 1$$

$$\pi_0 = \left(\sum_{i=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^i\right)^{-1} = 1 - \frac{\lambda}{\mu}$$

Consequently,

$$\pi_i = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^i = (1-\rho)\rho^i,$$

where ρ is the traffic intensity[1], which is given by

$$\rho = \frac{a}{c} = \lambda E(S) = \frac{\lambda}{\mu},$$

because the number of services is $c = 1$ and a is the offered load.

Performance Measure (1/2)

Result (Performance Measure for $M/M/1$)

- Average number of customers in the system:

$$L = E(N) = \sum_{i=0}^{\infty} i\pi_i = \sum_{i=0}^{\infty} i\rho^i(1-\rho) = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu-\lambda}.$$

- Average number of serving customers in the server:

$$\begin{aligned} L_s = E(N_s) &= 0 \cdot P_{\text{idle}} + 1 \cdot P_{\text{busy}} \\ &= 0 \cdot \pi_0 + 1 \cdot (1-\pi_0) \\ &= \rho. \end{aligned}$$

This is equal to the carried load, which is defined as the average processed load, $a' = a(1-P_B) = \lambda E(S) = \rho$.

Performance Measure (2/2)

- Average number of waiting customers in the queue:

$$L_q = E(N_q) = \sum_{i=1}^{\infty} (i-1)\pi_i = \frac{\rho^2}{1-\rho} = \frac{\lambda^2}{\mu(\mu-\lambda)}.$$

  Additionally, $L = L_s + L_q$.

- Average sojourn time in the system: From the Little's formula, $L = \lambda W$, we obtain

$$W = E(T) = \frac{L}{\lambda} = \frac{1}{\mu-\lambda} = \frac{1}{\mu(1-\rho)}$$

- Average waiting time in the queue:

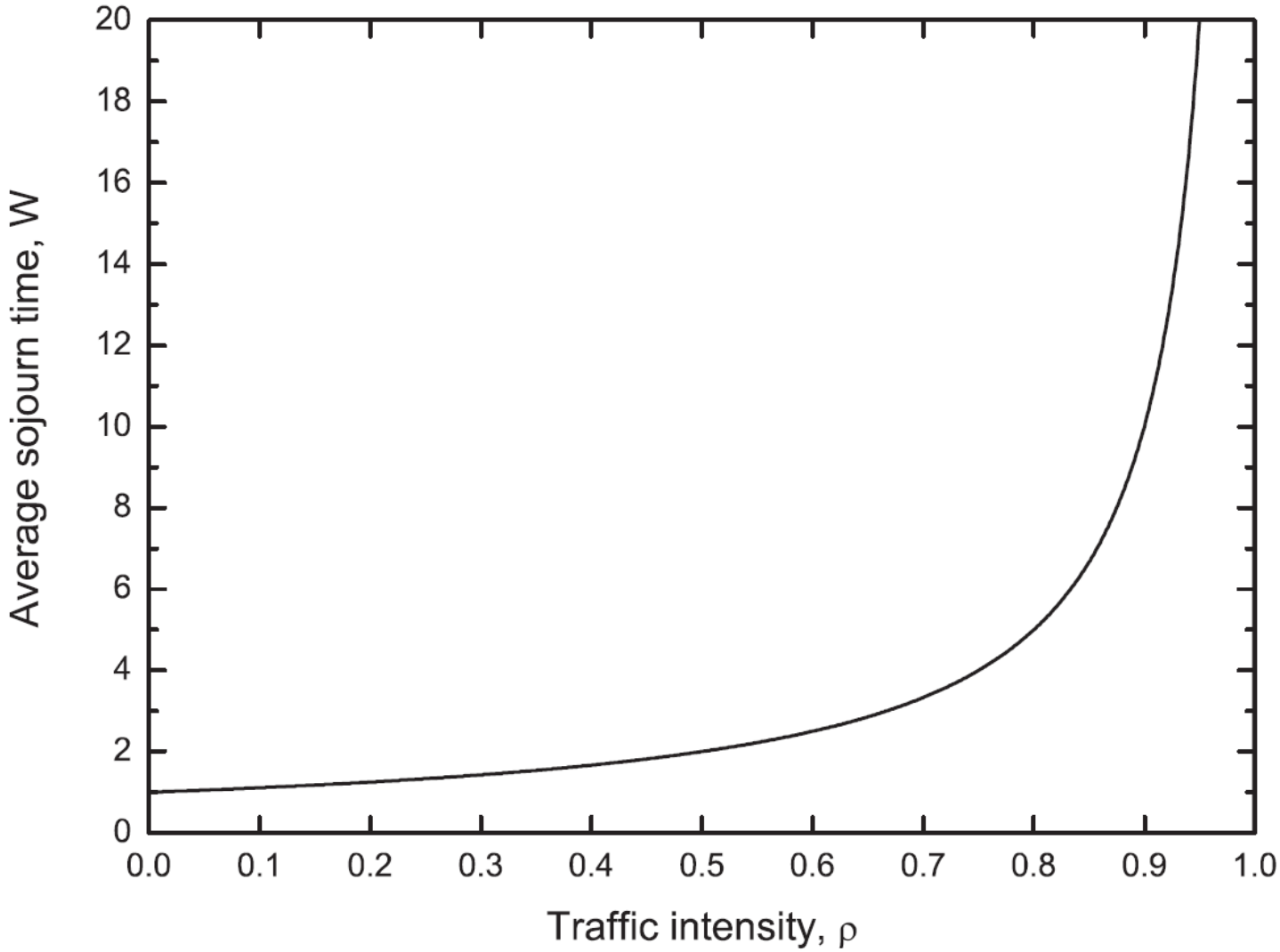
$$W_q = W - E(S) = \frac{1}{\mu(1-\rho)} - \frac{1}{\mu} = \frac{\rho}{\mu(1-\rho)}.$$

  or from the Little's formula, $L_q = \lambda W_q$, we obtain

$$W_q = E(T_q) = \frac{L_q}{\lambda} = \frac{\lambda}{\mu(\mu-\lambda)} = \frac{\rho}{\mu(1-\rho)}.$$

Average Sojourn Time

the average sojourn time, W, with the function of ρ when $\mu = 1.0$.



Application

- We consider an uplink of a wireless communication system with N users and a base station. The N users generate data packets at a Poisson rate λ and the service times of the packets are exponentially distributed with a mean of $1/\mu$. Determine the average delay.

- Solution: The average delay is equal to the average sojourn time which is given by

$$W = E(T) = \frac{L}{\lambda} = \frac{1}{\mu - \lambda} = \frac{1}{\mu(1-\rho)}$$

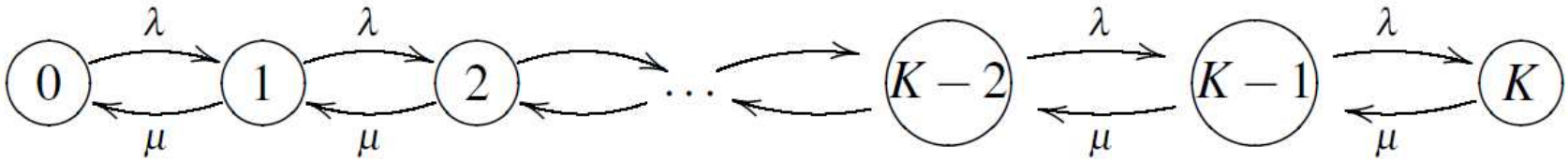
3. The *M/M/1/K* Queue

- Introduction

In practice, queues are always finite. In that case, a new customer is lost when he finds the system full (e.g., telephone calls). The $M/M/1/K$ may accommodate at most K customers, including the customer in the service facility, if any. Let λ and μ be the rate of the Poisson process for the arrivals and the parameter of the exponential distribution for the service times, respectively.

State Transition Rate Diagram

- System state: $s(i)$, where i is the number of customers in the system.
- Arrival process: $\sim PP(\lambda)$.
- Service process: $\sim Exp(\mu)$.
- Steady-state probability: π_i, $i = 0, 1, 2, \cdots, K$.



State transition rate diagram for $M/M/1/K$

Steady-state Probability (1/2)

Result (Balance equations for $M/M/1/K$) In steady-state, we have

$$\begin{aligned}
\lambda \pi_0 &= \mu \pi_1 \\
(\lambda + \mu)\pi_i &= \lambda \pi_{i-1} + \mu \pi_{i+1}, \quad i = 1, 2, \cdots, K-1 \\
\lambda \pi_{K-1} &= \mu \pi_K
\end{aligned}$$

$$\begin{aligned}
\pi_1 &= \frac{\lambda}{\mu}\pi_0 \\
\pi_2 &= \left(\frac{\lambda}{\mu}\right)^2 \pi_0 \\
&\vdots \\
\pi_i &= \left(\frac{\lambda}{\mu}\right)^i \pi_0, \quad i \le K
\end{aligned}$$

and from the normalization condition,

$$\sum_{i=0}^{K} \pi_i = \sum_{i=0}^{K} \left(\frac{\lambda}{\mu}\right)^i \pi_0 = 1$$

$$\pi_0 = \left(\sum_{i=0}^{K} \left(\frac{\lambda}{\mu}\right)^i\right)^{-1} = \frac{1-\rho}{1-\rho^{K+1}}$$

Steady-state Probability (2/2)

Consequently, if $\rho \neq 1 (\lambda \neq \mu)$, $\pi_i = \left(\frac{1-\rho}{1-\rho^{K+1}} \right) \rho^i, \quad 0 \leq i \leq K$

and $\pi_i = 0$ for $i > K$. If $\rho = 1 (\lambda = \mu)$, $\pi_i = \frac{1}{K+1}$.

In particular, an incoming customer will be rejected when he sees the state K. Hence, the probability that an incoming customers is rejected is π_K, which is called *blocking probability*.

Performance Measure (1/2)

Result (Performance Measure for $M/M/1/K$)

- Average number of customers in the system: if $\rho \neq 1$,

$$L = E(N) = \sum_{i=0}^{K} i\pi_i = \frac{\rho}{1-\rho} - \frac{(K+1)\rho^{K+1}}{1-\rho^{K+1}}$$

  and if $\rho = 1$, then $L = K/2$.

- Average number of serving customers in the server:

$$\begin{aligned} L_s = E(N_s) &= 0 \cdot \pi_0 + 1 \cdot (1-\pi_0) \\ &= \frac{\rho(1-\rho^K)}{1-\rho^{K+1}}. \end{aligned}$$

  This is equal to the carried load, $a' = a(1-P_B) = \rho(1-\pi_K)$.

Performance Measure (2/2)

- Average sojourn time in the system: From the Little's formula, $L = \lambda_e W$, we obtain

$$W = E(T) = \frac{L}{\lambda_e} = \frac{1}{\lambda(1-\pi_K)}$$

- Average waiting time in the queue:

$$W_q = W - E(S) = W - \frac{1}{\mu}.$$

or from the Little's formula, we obtain

$$W_q = E(T_q) = \frac{L_q}{\lambda_e} = \frac{L_q}{\lambda(1-P_B)}.$$

Application

- We consider an uplink of a wireless communication system with N users and a base station. The N users generate data packets at a Poisson rate λ and the service times of the packets are exponentially distributed with a mean of $1/\mu$. The base station can store K packets in the queue. Determine the average delay.

- Solution: The average delay is equal to the average sojourn time which is given by

$$W = E(T) = \frac{L}{\lambda_e} = \frac{1}{\lambda(1-\pi_K)}$$

Summary

- Queueing Models
  - Continuous-Time Birth-Death Process
  - The M/M/1 Queue
  - The M/M/1/K Queue

- Steady-state Probability

  Transition rates

$$q_{i,j} = \begin{cases} \lambda_i & \text{when } j = i+1 \text{ probability of birth in interval } \Delta t \text{ is } \lambda_i \Delta t \\ \mu_i & \text{if } j = i-1 \quad \text{probability of death in interval } \Delta t \text{ is } \mu_i \Delta_i t \\ 0 & \text{otherwise} \quad \text{when the system is in state } i \end{cases}$$

$$\frac{d}{dt}\pi(t) = \pi(t) \cdot \mathbf{Q}$$

- Performance Measure
  - Average number of customers, Average sojourn time, Average waiting time