Scheduling and Control of Queueing Networks - 1-1

생성자	㊊ 재환 김
태그	엔지니어링

단일 대기열 시스템(Single Queue)

- **단일 대기열의 정의 및 분석 가능성**
  이 책의 첫 번째 부분에서는 단일 대기열 시스템을 다룹니다. 고객이 서비스 시스템에 도착하여 하나의 서비스를 요구하며, 이 서비스는 하나 이상의 서버에서 제공됩니다. 특히 정확한 분석이 가능한 몇 가지 특별한 대기열 시스템의 특성을 연구합니다.
- **1장: 단일 대기열의 정의와 다루기 쉬운 예시**
  1장에서는 단일 대기열을 정의하고 다양한 기호와 특성을 소개합니다. 이 장에서 다루는 예시로는 수학적으로 분석하기 쉬운
  생사 대기열(birth and death queue) 같은 시스템이 있습니다. 또한, 대기열 시뮬레이션에 대해서도 논의합니다.
- **2장: M/G/1 시스템**
  2장에서는 기억이 없는 푸아송 도착(Memoryless Poisson Arrivals)과 일반 분포 처리 시간을 가진 시스템, 즉
  M/G/1 시스템을 다룹니다. 이 시스템의 성능 측정값은 **작업 보존 원리(work conservation)**와 **PASTA 속성(Poisson Arrivals See Time Averages)**을 통해 정확히 도출할 수 있습니다.
- **3장: 우선순위 및 배치 처리 대기열**
  3장에서는 작업의 배치 처리 스케줄링과 정적 작업 스트림의 대기열을 다룹니다. 또한, **우선순위 대기열(priority queues)**과 다양한 서비스 정책에 대해 논의합니다.

1장: 대기열 및 생사 대기열의 시뮬레이션

대기열 시스템의 개요

- **단일 대기열**은 고객들의 도착 흐름으로 구성됩니다. 각 고객은 단일 서비스 작업을 요구하며, 이 서비스는 하나 이상의 서버가 제공하는 시스템에서 이루어집니다.

- 이러한 대기열을 설명하기 위해 몇 가지 기호가 도입됩니다. 두 가지 기본 관계식이 유도됩니다:
    1. 첫 번째는 **대기열 길이(queue length)**로, 이는 도착 수에서 출발 수를 뺀 값입니다.
    2. 두 번째는 **Lindley의 방정식**으로, 이는 연속적인 고객들의 대기 시간을 재귀적으로 계산하는 방법을 제공합니다.

대기열 시뮬레이션

- Lindley 방정식을 활용하여 대기열 시뮬레이션을 논의합니다. 푸아송 도착(Poisson Arrivals)과 지수 분포를 따르는 서비스 시간을 가진 단일 대기열 시스템의 예를 연구합니다. 이를 통해 **마르코프 생사 과정**으로 모델링된 시스템에서 대기열 길이의 정상 상태 분포를 유도합니다. 여기서는 **상세 균형 방정식(detailed balance equations)**을 사용하여 해를 구합니다.

1.1 단일 대기열의 정의

- 대기열 시스템은 두 부분으로 나눌 수 있습니다:
    1. **수요 측**에서는 고객의 서비스 요청 흐름이 존재합니다.
    2. **공급 측**에서는 하나 이상의 서비스 스테이션이 있으며, 각 스테이션에는 하나 이상의 서버가 운영됩니다.
- 여기서는 단일 고객 흐름을 고려한 단일 서비스 스테이션을 모델링합니다. 이 시스템에서 각 고객은 단일 서비스를 요구하고, 서비스 스테이션은 그 서비스를 제공합니다.
- 고객의 도착은 확률적 점 과정(stochastic point process)인 A(t)로 모델링됩니다. 이는 t > 0에서, 0부터 t까지 시간 동안 발생한 도착 수를 나타냅니다.
- **도착 시간**을 나타내는 값들은 다음과 같습니다:
    - A_n: n번째 고객의 도착 시간
    - $T_n = A_n - A_{n-1}$: 연속 도착 간의 간격
- 이때 T_n은 독립적으로 분포된 확률 변수로 가정되며, 기댓값이 1/λ인 유한한 분포를 따릅니다. 따라서 도착 과정 A(t)는 도착률이 λ인 **갱신 과정(renewal process)**입니다.

도착 및 서비스 시간 분포

- **도착 시간 분포**: 도착은 포아송 프로세스로 모델링되며, 도착률은 λ입니다.
- **서비스 시간 분포**: n번째 고객의 서비스 시간 X_n은 도착 시간과 독립적인 분포를 따릅니다.

서비스 시간은 분포 G를 따르며, 평균 $m = 1/\mu$를 가집니다.

대기열 시스템의 모델링

- **단일 서비스 스테이션**: 고객들은 단일 서비스 스테이션에서 서비스를 받으며, 이 스테이션은 하나 이상의 서버로 구성될 수 있습니다. 일반적으로 도착한 고객은 대기열에 합류한 뒤, 가용한 서버로 이동하여 서비스를 받습니다.

켄달 표기법

- **D.G. Kendall**이 도입한 표기법은 대기열 시스템을 세 가지 요소로 표현합니다:
  1. 도착 분포
  2. 서비스 분포
  3. 서비스 스테이션 수
- 예시:
  - **M/M/1**: 포아송 도착, 지수 분포 서비스 시간, 단일 서버 시스템. 여기서 M은 "memoryless(무기억성)"를 의미합니다.

    MM
  - **M/D/s**: 포아송 도착, 결정적 서비스 시간, s개의 서버.

    ss
  - **G/G/∞**: 일반 도착, 일반 서비스 시간, 무한한 수의 서버. 이 시스템에서는 고객들이 도착 즉시 서비스를 받으며 대기열이 형성되지 않습니다.
  - **G/·/·**: 일반 도착, 간격이 일정하지 않은 일반적인 도착 분포를 따르는 시스템. 이 시스템은 특정 시점에 최대 고객 수를 수용할 수 있습니다.
  - **M/M/K/K**: 포아송 도착, 지수 분포 서비스 시간, K개의 서버와 K명의 총 고객을 수용하는 대기열. **Erlang 손실 시스템**이라고도 불리며, 전화 교환기 등에서 사용됩니다. 이 시스템에서는 모든 서버가 사용 중일 때 추가로 도착한 고객은 손실됩니다.

서비스 정책

- **FCFS (First Come First Serve)**: 선입선출 방식. 고객이 도착한 순서대로 서비스가 이루어집니다. FIFO(First In First Out)라고도 합니다.
- **LCFS (Last Come First Serve)**: 후입선출 방식. 마지막에 도착한 고객이 먼저 서비스를 받습니다. LIFO(Last In First Out)라고도 합니다.
- **PS (Processor Sharing)**: 서버가 시스템 내의 모든 고객에게 동일한 처리 능력을 분배하여 서비스합니다.

대기열 시스템의 기본 관계식

- 대기열 이론의 기초를 이루는 두 가지 주요 관계식:
    1. **고객의 동태**: 시스템 내 고객 수를 나타내는 **대기열 길이 Q(t)**는 특정 시간 t에 대기 중이거나 서비스를 받고 있는 모든 고객의 수입니다.

1.1 단일 대기열

1. 대기열 길이의 관계식

대기열 길이 $Q(t)$는 시간 t에 시스템에 남아있는 고객 수를 나타냅니다. 이를 다음과 같은 공식으로 표현할 수 있습니다:

$$Q(t) = Q(0) + A(t) - D(t)$$

- $Q(0)$: 초기 시간 0에서의 대기열 길이, 즉 처음에 시스템에 있던 고객 수
- $A(t)$: 시간 0부터 t까지 도착한 고객 수
- $D(t)$: 시간 0부터t까지 시스템을 떠난(서비스를 완료한) 고객 수

이 관계식은 시스템에 남아있는 고객 수가 초기 고객 수에 도착한 고객 수를 더하고 떠난 고객 수를 뺀 값임을 보여줍니다. 그림 1.1에서 이를 시각적으로 설명하고 있습니다.

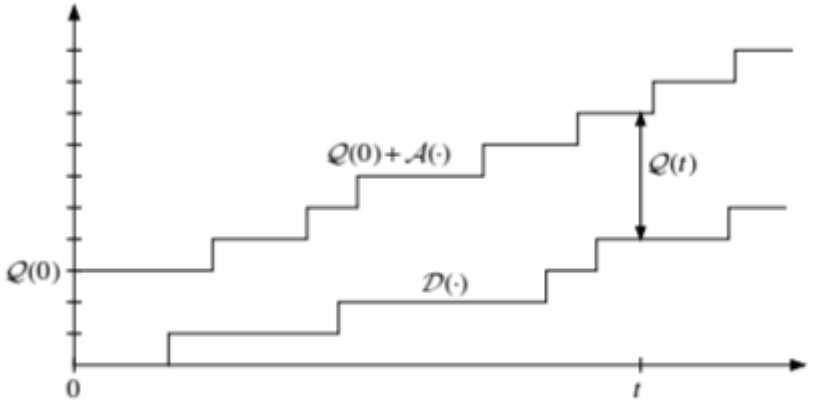

Figure 1.1 Queue length is arrivals minus departures.

그림 1.1은 대기열 길이가 고객의 도착과 출발에 따라 어떻게 변화하는지를 단계적으로 보여줍니다.

2. Lindley 방정식 (대기 시간 계산)

Lindley 방정식은 선입선출(FCFS, First Come First Serve) 방식에서 고객의 대기 시간을 계산하는 데 사용됩니다.

고객 n+1의 대기 시간을 V_{n+1}이라고 정의하면, 다음과 같은 관계식을 따릅니다:

$V_{n+1} = (V_n + X_n - T_{n+1})^+$

여기서,

- V_n: n번째 고객의 대기 시간
- X_n: n번째 고객의 서비스 시간
- T_{n+1}: n+1번째 고객이 도착할 때까지의 간격
- $(x)^+ = max(0, x)$: 양의 부분 함수로, 결과값이 음수일 경우 0으로 설정

Lindley 방정식의 설명

- 고객 n은 도착 후 $V_n + X_n$ 시간 단위 후에 시스템을 떠납니다.
- 고객 n+1은 T_{n+1} 시간 단위 후에 도착합니다. T_{n+1}이 $V_n + X_n$보다 크다면, 즉 고객 n이 서비스를 완료한 후에 도착한다면, 고객 n+1은 즉시 서비스를 받게 됩니다.
- 반대로, T_{n+1}이 $V_n + X_n$보다 작다면, 고객 n+1은 고객 n의 서비스가 끝날 때까지 기다리게 됩니다.

이 과정은 대기 시간을 재귀적으로 계산할 수 있는 방법을 제공합니다. Lindley 방정식으로 고객의 대기 시간을 계산하는 방식은 그림 1.2에서 설명되고 있습니다.

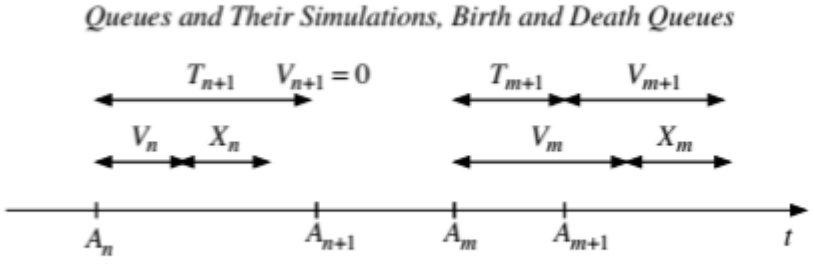

Figure 1.2 Waiting time calculation using Lindley's equation.

1.2 대기열의 시뮬레이션

시뮬레이션의 중요성

- **대기열 시스템을 연구하는 도구**로서 시뮬레이션은 매우 강력합니다. 복잡한 대기열 시스템을 수학적으로 분석하기 어려운 경우가 많아, 다양한 성능 지표를 시뮬레이션을 통해 추정할 수 있습니다.

- 시뮬레이션은 **정리나 방정식을 통한 분석과는 다른 관점을 제공**하여, 대기열 시스템의 동작을 새로운 방식으로 이해할 수 있게 합니다.

재귀 관계를 이용한 시뮬레이션

- **Lindley 방정식**과 같은 재귀 관계식을 이용하여 대기열을 시뮬레이션할 수 있습니다. 단일 서버에서 선입선출(FCFS) 방식으로 운영되는 시스템의 경우, 시뮬레이션은 다음과 같이 진행됩니다:
  1. **시작 상태 설정**: 시스템을 초기화하고, 모든 고객에 대한 서비스 완료 후 서버의 이용 가능 시간을 결정합니다.
  2. **도착 간격 및 서비스 시간 생성**: 각 고객의 도착 간격과 서비스 시간을 순차적으로 생성합니다.
  3. **대기 시간, 서비스 시작 및 종료 시간 계산**: 각 고객의 대기 시간, 서비스 시작 및 종료 시간을 계산하여 **대기 시간**과 **체류 시간**(대기 시간 + 서비스 시간)을 산출합니다.
  4. **대기열 길이 계산**: 도착한 고객 수와 떠난 고객 수를 비교하여 실시간으로 대기열 길이를 계산합니다.

예시 1.1: 대기열 시뮬레이션

- **상황 설정**: 17번째 고객이 시간 39.1에 도착하고, 서버는 시간 42.8에 해당 고객의 서비스를 시작할 수 있습니다.
- **변수 정의**:
  - S_n: n번째 고객의 서비스 시작 시간
  - D_n: n번째 고객의 서비스 종료(출발) 시간
  - W_n: n번째 고객의 체류 시간 (도착부터 출발까지 걸린 시간)
  - V_n: n번째 고객의 대기 시간 (도착부터 서비스 시작까지 걸린 시간)
- **시뮬레이션 절차**:
  - **도착 간격**과 **서비스 시간**은 지정된 분포에 따라 **의사난수(pseudorandom)** 방식으로 생성됩니다.
  - 이 값들을 바탕으로 재귀적 계산을 수행하여 고객의 도착 시간, 대기 시간, 서비스 시작 및 종료 시간을 결정합니다.

계산 방법

- 각 고객의 도착 시간 A_n은 이전 고객의 도착 시간에 도착 간격 T_n을 더해 계산합니다:

$$A_n = A_n - 1 + T_n$$

- 서비스 시작 시간 S_n은 이전 고객의 종료 시간과 도착 시간 중 더 큰 값으로 계산됩니다:

  $$S_n = max(A_n, D_{n-1})$$

- 대기 시간 V_n은 서비스 시작 시간과 도착 시간의 차이로 계산됩니다:

  $$V_n = S_n - A_n$$

- 종료 시간 D_n은 서비스 시작 시간에 서비스 시간 X_n을 더한 값입니다:

  $$D_n = S_n + X_n$$

- 체류 시간 W_n은 대기 시간과 서비스 시간을 더한 값입니다:

  $$W_n = V_n + X_n$$

1.3 생사 대기열 (Birth and Death Queues)

고객의 도착 및 출발 시간

Customer	T_n	A_n	X_n	Start	Depart	Sojourn	Wait
17		39.1	2.2	42.8	45.1	6.0	3.8
18	2.8	41.9	2.7	45.1	47.7	5.9	3.2
19	4.3	46.1	1.0	47.7	48.7	2.6	1.6
20	2.5	48.6	0.3	48.7	49.0	0.4	0.1
21	4.1	52.8	1.5	52.8	54.2	1.5	0.0
22	4.8	57.6	2.0	57.6	59.5	2.0	0.0
23	1.3	58.9	3.9	59.5	63.5	4.6	0.6

이미지의 상단 표에서는 17번부터 23번까지의 고객에 대해 다음과 같은 변수를 나타냅니다:

- T_n: 두 고객 사이의 도착 간격 (도착 시간 간격)
- A_n: 각 고객의 실제 도착 시간
- X_n: 각 고객의 서비스 시간
- **Start**: 각 고객의 서비스가 시작된 시간
- **Depart**: 각 고객이 시스템을 떠난 시간
- **Sojourn**: 각 고객이 시스템 내에 머무른 총 시간 (대기 시간 + 서비스 시간)
- **Wait**: 고객이 서비스 시작을 기다린 시간

이 표에서 고객의 대기 시간이 각 고객마다 다르게 나타나는 것을 확인할 수 있습니다. 예를 들어, 17번째 고객은 3.8 단위 시간 동안 대기해야 했지만, 21번째 고객은 대기 시간이 0이었습니다.

대기열 길이 계산

표 아래에서 설명된 방식은 대기열 길이를 계산하는 방법을 설명합니다.

- 각 고객의 도착 시간에는 **+1**을, 각 고객의 출발 시간에는 **1**을 부여하여 대기열의 길이를 계산합니다.
- 이 방식으로 대기열의 길이가 시간이 지나면서 어떻게 변하는지 추적할 수 있습니다.

표의 두 번째 부분은 시간에 따른 대기열 길이를 보여줍니다.

각 고객의 도착 및 출발 이벤트가 발생할 때마다 대기열의 변화가 기록됩니다.

대기열의 길이는 도착 이벤트에서 증가하고, 출발 이벤트에서 감소합니다. 예를 들어, 17번 고객이 도착했을 때 대기열 길이는 3이었고, 그 후 고객이 출발하면서 2로 감소합니다.

그림 1.3: 대기열 길이 시뮬레이션

Time	39.1	41.9	42.1	42.8	45.1	46.1	47.7	48.6	48.7
Customer	17	18	-15	-16	-17	19	-18	20	-19
Queue	3	4	3	2	1	2	1	2	1

Time	49.0	52.8	54.2	57.6	58.9	59.5	59.8	61.6	63.5
Customerr	-20	21	-21	22	23	-22	24	25	-23
Queue	0	1	0	1	2	1	2	3	2

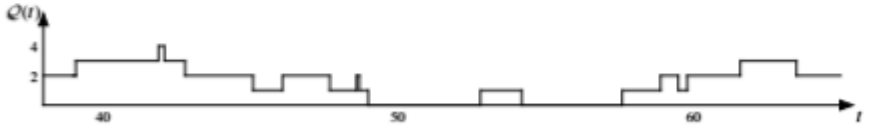

Figure 1.3 Simulation of queue length.

이 그림에서는 시간에 따른 대기열 길이 변화를 그래프로 나타냅니다.

대기열 길이는 시간이 지남에 따라 도착과 출발에 의해 변화하며, 고객들이 도착할 때마다 증가하고, 서비스를 마치고 떠날 때마다 감소하는 패턴을 보여줍니다.

생사 대기열의 정의

- **푸아송 도착 및 지수 분포 서비스 시간**을 가진 대기열 시스템을 설명합니다.

- 이 시스템에서는 임의의 시간 t에서 다음 도착까지 남은 시간과 서비스 중인 고객들의 서비스 시간이 모두 지수 분포를 따릅니다.

  tt

- 중요한 점은 **현재 시스템에 있는 모든 고객의 남은 처리 시간**과 **다음 도착까지의 남은 시간**이 서로 독립적이라는 것입니다.
- 즉, 시스템에 있는 고객의 상태는 과거의 어떤 사건과도 독립적으로 작동합니다. 이는 푸아송 과정의 중요한 속성입니다.

생사 대기열 모델: 연속 시간 마코프 체인

- 대기열 길이 과정 $Q(t)$는 **연속 시간 마코프 체인(Continuous Time Markov Chain)**으로 모델링됩니다. 이 시스템의 상태는 m = 0, 1, 2, ...로 나타나며, 이는 시스템 내 고객 수를 의미합니다.
- 상태가 $Q(t) = m$일 때, 다음 이벤트는 고객의 도착(생성)이나 서비스 완료(소멸)입니다. 도착이나 완료가 발생하면 상태가 +1 또는 -1로 변합니다. 이와 같은 대기열은 **생사 대기열(Birth and Death Queue)**로 불립니다.
- 상태 전이는 다음과 같은 **전이율(transition rate)**에 의해 설명됩니다:

  $q(m, m+1) = \lambda_m$

  $q(m, m-1) = \mu_m$

  - 여기서 λm은 고객이 시스템에 m명이 있을 때 발생하는 도착률, 즉 **생성률**입니다.
  - μm은 시스템에 m명의 고객이 있을 때 발생하는 서비스 완료율, 즉 **소멸률**입니다.
  - 만약 도착이 푸아송 분포를 따른다면,$\lambda_m = \lambda$, $\mu_m = \mu$로 일정하게 가정할 수 있습니다. 그러나 더 일반적인 모델에서는 이러한 비율이 상태 m에 따라 달라질 수 있습니다.

대기열 시스템의 주요 특성: 정상 상태 분포

- 대기열 시스템의 주요 기술 요소 중 하나는 **정상 상태 분포(Stationary Distribution)** $\pi(m)$입니다.
  - 이는 시간이 무한대에 가까워질 때 대기열이 특정 상태에 머무를 확률로 정의됩니다:

    $\pi(m) = lim_{t \to \infty} P(Q(t) = m), m \geq 0$

  - 이 분포는 종종 **한계 분포(Limiting Distribution)** 또는 **장기 분포(Long-Run Distribution)**라고 불리며, 시스템이 더 이상 초기 상태에 의존하지 않게 되었을 때의 상태를 나타냅니다.

- 정상 상태 분포가 존재한다면, 이를 통해 대기열 시스템의 안정성을 정의할 수 있습니다.

상세 균형 방정식 (Detailed Balance Equation)

- **생사 과정(Birth and Death Process)**은 시간이 지남에 따라 가역적(reversible)이며, 그 확률은 다음과 같은 **상세 균형 방정식**을 만족합니다:

  $\pi(m)q(m,m+1) = \pi(m+1)q(m+1,m)$

  이는 상태 m에서 상태 m+1로의 전이율과 그 반대 전이율이 균형을 이루는 것을 의미합니다.

  이를 식으로 표현하면:

  $\pi(m+1) = \pi(m)\frac{\lambda m}{\mu_{m+1}}$

  - 이 방정식은 두 상태 사이에서 전이되는 속도가 동일하게 유지되어야 함을 나타냅니다. 이는 전기 네트워크에서 **플럭스(flux)**로 차용된 개념입니다.

정상 상태 분포 유도

- 상세 균형 방정식을 통해 **생사 대기열의 정상 상태 분포**를 다음과 같이 유도할 수 있습니다:

  $\pi(m) = \pi(0)\frac{\lambda_0\lambda_1\ldots\lambda_{m-1}}{\mu_1\mu_2\ldots\mu_m}$

  - 여기서 $\pi(0)$은 **정규화 상수(normalizing constant)**로, 전체 확률의 합이 1이 되도록 만들어주는 상수입니다.

수렴성과 정상 상태의 조건

- 정상 상태 분포가 존재하기 위한 충분 조건은 마코프 체인이 **재발성(ergodicity)**을 만족해야 한다는 것입니다. 이는 체인이 양의 확률로 모든 상태를 반복 방문하고, 시스템이 양의 재발률을 가진다면 정상 상태 분포가 존재할 수 있음을 의미합니다.

M/M/1 대기열 모델

1. 기본 개념

- **M/M/1 대기열**은 도착이 푸아송 분포를, 서비스 시간이 지수 분포를 따르는 시스템입니다. 여기서 λ는 도착률(단위 시간당 고객 도착 속도), μ는 서비스율(단위 시간당 고객 처리 속도)입니다.
- 이 시스템의 **로딩(부하)율** ρ는 다음과 같이 정의됩니다:

$$\rho = \frac{\lambda}{\mu}$$

- ρ는 단위 시간당 시스템에 도착하는 평균 작업량을 나타냅니다. 예를 들어, ρ = 0.5는 시스템이 50% 용량으로 사용되고 있음을 의미합니다.

2. 상태 전이 및 전이율

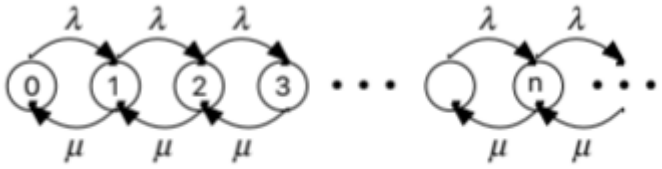

Figure 1.4 The M/M/1 queue, states and transition rates.

그림 1.4는 M/M/1 대기열의 상태 전이와 전이율을 시각적으로 보여줍니다.

각 상태는 시스템 내 고객 수를 나타내며, 도착 이벤트에 따라 상태가 증가하고 서비스 완료에 따라 감소합니다.

- **고객 도착 시**: 상태가 n에서 n+1로 이동하며, 전이율은 λ입니다.
- **고객 출발 시**: 상태가 n에서 n−1로 이동하며, 전이율은 μ입니다.

3. 정상 상태 분포 (Stationary Distribution)

M/M/1 대기열의 **정상 상태 분포**는 다음과 같습니다:

$$\pi(n) = (1-\rho)\rho^n, n = 0, 1, 2, \ldots, \rho < 1$$

- 여기서 $\pi(n)$은 시스템에 n명의 고객이 있을 확률을 나타냅니다.
- 정상 상태 분포는 기하 분포를 따르며, 매개변수는 1 - ρ입니다. 시스템이 안정적이려면 ρ가 1보다 작아야 합니다. 이는 대기열 길이가 시간에 따라 무한정 증가하지 않고 일정 범위 내에서 유지됨을 의미합니다.
- 대기열의 평균 길이는 다음과 같습니다:

  $$\frac{\rho}{1-\rho}$$

4. 고객의 체류 시간 (Sojourn Time)

정리 1.3에 따르면, M/M/1 대기열에서 선입선출(FCFS) 방식을 따를 때, 고객의 체류 시간 W_n은 지수 분포를 따릅니다:

$$W_n \sim Exp(\mu - \lambda)$$

체류 시간은 고객이 시스템에 머무는 총 시간으로, 대기 시간과 서비스 시간을 포함합니다.

체류 시간의 평균은 다음과 같습니다:

$$\frac{1}{\mu-\lambda}$$

5. 체류 시간의 확률 밀도 함수 (PDF)

체류 시간의 확률 밀도 함수는 다음과 같습니다:

$$f_W(t) = \Sigma_{j=0}^{\infty}(1-\rho)\rho^j \frac{\mu^{j+1}t^j}{j!}e^{-\mu t} = \mu(1-\rho)e^{-(1-\rho)\mu t}$$

- 이 식은 시스템 내 j명의 고객이 있을 때 고객의 대기 시간과 전체 체류 시간을 설명합니다.
- 체류 시간은 서비스율 μ와 로딩율 ρ에 따라 결정됩니다.

M/M/∞ 대기열 모델 (무한 서버 대기열)

1. M/M/∞ 대기열의 정의

- **M/M/∞ 대기열**은 도착률 λ를 따르는 푸아송 도착과 서비스율 μ를 따르는 지수 분포 서비스 시간이 적용되는 시스템입니다. 이 시스템의 주요 특징은 **무한한 수의 서버**가 존재하여 고객이 도착하면 즉시 서비스를 받을 수 있으며, 대기열에서 기다릴 필요가 없다는 점입니다.
- **도착한 고객 수**는 푸아송 도착률 λ에 따라 결정되며, 각 고객은 지수 분포를 따르는 서비스 시간 동안 서비스를 받습니다.

2. 대기열 길이

- 이 시스템에서는 대기열 길이 $Q(t)$가 시스템 내에 존재하는 고객 수, 즉 **현재 서비스를 받고 있는 고객 수**와 동일합니다.

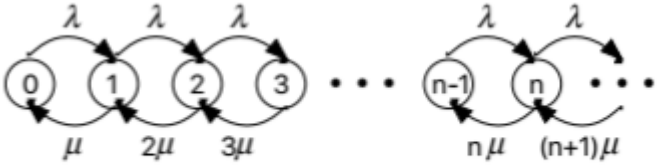

Figure 1.5 The M/M/∞ queue, states and transition rates.

- 그림 1.5는 시스템 상태 전이 및 전이율을 시각적으로 설명합니다:
  - 0 고객일 때는 도착 시 λ로 상태가 1로 변합니다.
  - 상태가 n일 때 고객이 도착하면 상태가 n+1로 증가하고, 서비스가 완료되면 상태가 n-1로 감소합니다.

3. M/M/∞ 대기열의 안정성

- M/M/∞ 대기열은 항상 안정적입니다. 시스템에 무한한 수의 서버가 존재하기 때문에 도착하는 모든 고객이 즉시 서비스를 받을 수 있습니다.

4. M/M/∞ 대기열의 정상 상태 분포 (Stationary Distribution)

- M/M/∞ 대기열의 정상 상태 분포는 **푸아송 분포**를 따릅니다. 이때 파라미터는 $\rho = \frac{\lambda}{\mu}$ 입니다.

  정상 상태 분포는 다음과 같이 주어집니다:

  $$\pi(n) = \frac{\rho^n e^{-\rho}}{n!}, n = 0, 1, 2, ...$$

  여기서 $\pi(n)$은 시스템 내에 n명의 고객이 있을 확률을 나타냅니다.

- 평균과 분산은 모두 ρ로 주어집니다. 즉, 고객 수의 평균과 분산이 모두 ρ입니다.

5. 자원 풀링(Resource Pooling)의 개념

- **자원 풀링**은 여러 대기열을 하나로 통합하여 자원을 효율적으로 사용하는 개념입니다.

  예를 들어, s개의 **M/M/1 대기열**을 각각 s배 더 빠른 서비스율을 가진 시스템으로 통합하면, 고객의 체류 시간은 동일하게 유지됩니다. 하지만 전체 시스템에 머무는 고객의 수는 1/s로 감소합니다.

- 즉, 여러 개의 대기열 시스템을 통합함으로써 자원 사용을 최적화할 수 있으며, 이를 통해 시스템의 성능을 개선할 수 있습니다. 이러한 자원 풀링 개념은 대기열 이론에서 중요한 역할을 하며, 다양한 실제 응용에서 사용됩니다.

M/M/K/K 대기열 (얼랑 손실 시스템)

1. M/M/K/K 대기열의 정의

- **M/M/K/K 대기열**은 도착률 λ를 따르는 푸아송 도착과 서비스율 μ를 따르는 지수 분포 서비스 시간을 가진 시스템입니다.
- 이 시스템은 최대 K명의 고객을 수용할 수 있습니다. 모든 서버가 사용 중일 때 고객이 도착하면, 해당 고객은 시스템에 진입하지 못하고 서비스를 받지 못한 채 **손실**됩니다.
- 이 모델은 **얼랑의 손실 모델**로도 알려져 있으며, 주로 **전화 교환기 시스템**에 적용되었습니다. K개의 전화 회선이 있고, 모든 회선이 사용 중일 때 추가로 도착하는 통화 요청은 연결되지 못하고 손실됩니다.

2. 상태 전이 및 전이율

Figure 1.6 The M/M/K/K queue, states and transition rates.

그림 1.6은 M/M/K/K 대기열의 상태 전이와 전이율을 보여줍니다.

- n명의 고객이 있을 때 새 고객이 도착하면 상태가 n+1로 증가하며, 이때의 전이율은 λ입니다.
- 서비스가 완료되면 상태가 n−1로 감소하며, 그 전이율은 nμ입니다. (여기서 n은 현재 서비스 중인 고객 수입니다.)

3. 정상 상태 분포 (Stationary Distribution)

이 시스템의 **정상 상태 분포**는 다음과 같습니다:

$$\pi(n) = \frac{\rho^n}{n!} \Big/ \sum_{m=0}^{K} \frac{\rho^m}{m!}, \quad n = 0, 1, 2, \ldots, K$$

여기서 $\rho = \frac{\lambda}{\mu}$는 로딩율(부하율)을 나타내며, $\pi(n)$은 시스템 내에 n명의 고객이 있을 확률을 의미합니다.

4. 성능 측정 지표

두 가지 주요 성능 지표가 있습니다:

1. **고객 손실 확률**: 모든 서버가 사용 중일 때 새로 도착한 고객이 시스템에 진입하지 못할 확률입니다. 이는 K개의 서버가 모두 사용 중일 때 도착한 고객이 거부되는 확률로, 다음과 같이 계산됩니다:

   $$\pi(K) = \frac{\rho^K}{K!} \Big/ \sum_{m=0}^{K} \frac{\rho^m}{m!}$$

2. **평균 사용 중인 서버 수**: 시스템에서 평균적으로 사용 중인 서버 수는 다음과 같이 계산됩니다:

   $$\rho(1 - \pi(K))$$

M/M/s 대기열

1. M/M/s 대기열의 정의

- **M/M/s 대기열**은 s개의 서버가 있는 대기열 시스템으로, 도착은 푸아송 분포를 따르고, 서비스 시간은 지수 분포를 따릅니다.

- 이 시스템은 s개의 서버를 넘는 고객이 대기열에 줄을 서게 됩니다. 모든 서버가 사용 중일 때 도착한 고객은 대기열에서 기다립니다.

2. 상태 전이 및 전이율

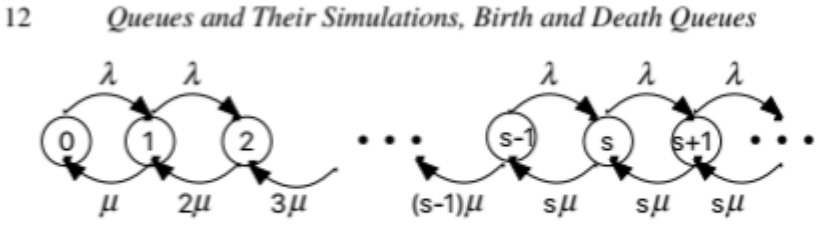

Figure 1.7 The M/M/*s* queue, states and transition rates.

그림 1.7에서는 M/M/s 대기열의 상태 전이와 전이율을 보여줍니다.

- 고객 도착 시 상태가 n+1로 증가하며, 그 전이율은 λ입니다.
- 서비스가 완료되면 상태가 n−1로 감소하며, 그 전이율은 현재 서비스 중인 고객 수에 따라 nμ입니다.

3. 안정성 조건 및 정상 상태 분포

M/M/s 대기열에서 안정성이 유지되려면, 로딩율 $\rho = \frac{\lambda}{s\mu}$가 1보다 작아야 합니다. 즉, 시스템에 도착하는 고객 수가 서버들이 처리할 수 있는 능력을 초과하지 않아야 합니다.

정상 상태 분포는 다음과 같이 주어집니다:

$$\pi(n) = \begin{cases} \frac{\rho^n}{n!}\pi(0), & n \leq s \\ \frac{\rho^n}{s^s(n-s)!}\pi(0), & n > s \end{cases}$$

여기서 $\pi(0)$는 정상 상태에서 대기열에 고객이 없는 상태일 확률로, 이를 통해 다른 모든 확률을 구할 수 있습니다.

1.3.1 마코비안 대기열의 시뮬레이션

마코프 체인 기반 대기열 시뮬레이션에서는 도착과 서비스 완료 이벤트를 시뮬레이션합니다. 이를 위해 **통합화 접근법(uniformization approach)**을 사용할 수 있습니다.

통합화 접근법의 주요 단계:

1. **포아송 과정**을 이용해 도착률 θ를 시뮬레이션합니다. θ는 도착 및 서비스 이벤트의 총합입니다.θ\thetaθ\theta
2. **각 이벤트**가 도착인지 서비스 완료인지 결정합니다.

3. 이벤트 발생 시 해당하는 **상태 변이**를 수행합니다. 도착 시 고객이 추가되고, 서비스 완료 시 고객이 시스템에서 제거됩니다.
4. 서비스 완료 이벤트가 발생했으나 대기열이 비어 있을 경우, 이를 **더미 이벤트**로 간주하고 상태 변화를 일으키지 않습니다.

이 시뮬레이션 방법은 복잡한 대기열 시스템, 특히 **연속 시간 마코프 체인(CTMC)**으로 모델링된 시스템에 적합하며, 다양한 대기열 네트워크에서도 활용됩니다.

예시 1.7: 직렬 연결된 두 서버 시스템의 시뮬레이션

시스템 설명:

- 이 시스템에서는 도착률 λ를 따르는 푸아송 도착이 발생하며, 두 개의 서버가 직렬로 연결되어 있습니다.
- 각 서버는 지수 분포를 따르는 서비스율을 가집니다. 첫 번째 서버의 서비스율은 μ_1, 두 번째 서버의 서비스율은 μ_2입니다.
- 각 고객은 첫 번째 서버에서 서비스를 받은 후, 두 번째 서버로 이동하여 추가 서비스를 받습니다. 모든 고객은 두 번째 서버에서 서비스를 완료한 후 시스템을 떠납니다.

시뮬레이션 과정:

- 시뮬레이션에서 도착 및 서비스 이벤트는 포아송 과정으로 모델링되며, 각 이벤트는 다음과 같은 확률로 처리됩니다:
  - 도착 확률은 $\frac{\lambda}{\theta}$입니다. 여기서 $\theta = \lambda + \mu_1 + \mu_2$는 시스템 내 모든 이벤트의 총 발생률입니다.
  - 첫 번째 서버에서 두 번째 서버로의 이동은 $\frac{\mu_1}{\theta}$의 확률로 발생합니다. 첫 번째 서버가 비어 있을 경우, 이는 더미 이벤트로 처리됩니다.
  - 두 번째 서버에서 시스템을 떠나는 확률은 $\frac{\mu_2}{\theta}$입니다. 두 번째 서버가 비어 있을 경우, 이 역시 더미 이벤트로 간주됩니다.

1.4 역사적 배경, 출처 및 확장

대기열 이론의 기원

- **대기열 이론**은 **Agner Krarup Erlang**이 창안했습니다. 그는 **코펜하겐 전화 회사**에서 근무하며 전화 교환 시스템을 모델링하기 위해 이 이론을 개발했습니다.
  - Erlang의 주요 논문 두 편은 1909년과 1917년에 발표되었습니다.

Kendall의 표기법과 Lindley 방정식

- **Kendall 표기법**은 1953년 Kendall이 도입했습니다.
- **Lindley 방정식**은 1952년 Lindley가 제안했으며, 생사 과정(birth and death processes) 이론의 기초를 마련했습니다.

추가 참고 문헌

- **Sheldon Ross**의 저서 **Stochastic Models**는 생사 대기열 이론을 소개하고 대기열 이론의 기본을 다룹니다.

정상 상태 분포와 과도 상태 분포

- **정상 상태 분포(Stationary Distribution)**는 시간이 무한히 지나 대기열이 안정된 상태에서의 고객 수 분포를 나타냅니다. 이는 시스템 성능 지표 계산에 사용됩니다.
- 특정 시간 **t**에서의 대기열 길이 분포를 알고자 할 때는 **Kolmogorov의 전방 방정식**이나 **후방 방정식**을 이용할 수 있지만, 이는 복잡한 과정입니다.
- Baccelli와 Massey(1989)의 연구는 이러한 과도 상태 분포(transient distribution)를 구하는 방법을 다룹니다.

실세계 대기열의 복잡성

- 현실 세계의 대기열 시스템은 이론적 모델보다 복잡합니다. 예를 들어, 도착률 λ(t)와 서비스율 μ(t)은 시간에 따라 변할 수 있습니다.
- 이러한 복잡한 시스템을 다루기 위해 **시뮬레이션**이 필요합니다. Asmussen과 Glynn(2007)의 저서 **Stochastic Simulation: Algorithms and Analysis**는 이러한 복잡한 시스템의 시뮬레이션 방법을 상세히 설명합니다.

연습 문제 (Exercises)

1.1 Excel을 사용한 시뮬레이션

- **(i)** 두 서버가 있는 단일 대기열을 시뮬레이션합니다.도착률은 푸아송 분포를 따르며 λ=0.2이고, 서비스 시간은 지수 분포를 따르며 평균이 8입니다.
- **(ii)** 직렬로 연결된 두 서버가 있는 시스템을 시뮬레이션합니다.도착률은 푸아송 분포를 따르며 λ=0.25이고, 첫 번째와 두 번째 서버의 서비스 시간은 각각 평균이 3인 지수 분포를 따릅니다.

1.2 다중 서버에 대한 Lindley 방정식 유도

- s개의 서버를 가진 시스템에서, 선입선출(FCFS) 방식을 따르는 경우에 대해 **Lindley 방정식**의 유사한 공식을 유도하세요.

1.3 M/M/s 대기열에서 대기 시간과 체류 시간 유도

- M/M/s 시스템에서 고객의 대기 시간과 체류 시간을 유도하세요. 정상 상태에서 시스템 내 고객 분포를 통해 성능 지표를 계산하는 방법을 사용하세요.

1.4 M/M/K/K 대기열에서 사용 중인 서버의 평균 수 계산

- Erlang 손실 시스템인 M/M/K/K 대기열에서 평균적으로 사용 중인 서버의 수를 계산하세요.

1.5 택시 대기줄 시뮬레이션

- 다음 조건의 택시 대기열 시스템을 시뮬레이션하세요:
  - 택시는 최대 2대까지 대기할 수 있고, 승객은 최대 3명까지 대기할 수 있습니다.
  - 택시는 6분마다 한 대씩, 승객은 8분마다 한 명씩 도착합니다.

연습 문제 1.5: 택시 대기 시스템

이 문제는 택시 대기 시스템을 생사 대기열로 모델링하는 것을 요구합니다.

시나리오 설명:

- 승객이 도착하면 대기 중인 택시가 있을 경우 즉시 탑승합니다. 택시가 없다면 승객은 대기열에 합류하거나, 대기 공간이 없으면 시스템을 떠납니다.
- 택시가 도착하면 대기 중인 승객이 있을 경우 즉시 승객을 태우고 출발합니다. 승객이 없다면 택시는 대기열에 합류하거나, 대기 공간이 없으면 떠납니다.

요구사항:

- **(i)** 시스템의 상태와 전이율을 나타내는 다이어그램을 작성하세요.
- **(ii)** 정상 상태 분포를 계산하세요. 특히, 대기열이 비어 있는 경우, 승객이 1, 2, 또는 3명인 경우, 택시가 1 또는 2대 있는 경우에 대한 확률을 구하세요.
- **(iii)** 승객이 도착 즉시 택시를 이용할 확률을 계산하세요.
- **(iv)** 서비스를 받지 못하고 떠나는 승객의 비율을 계산하세요.
- **(v)** 서비스를 받는 승객의 대기 시간 분포를 구하세요.

연습 문제 1.6: K개의 기계와 M명의 수리공이 있는 대기열 시스템

이 문제는 **K개의 기계와 M명의 수리공**을 가진 시스템을 생사 대기열로 모델링하는 문제입니다.

시나리오 설명:

- K개의 기계는 일정 시간이 지나면 고장 나며, 수리공이 수리해야만 다시 운영이 가능합니다.
- 각 기계의 고장률은 λ이고, 수리공의 수리 속도는 μ입니다. 시스템에서 동시에 수리할 수 있는 기계는 최대 M대입니다.λ\lambdaμ\mu

요구사항:

- 이 시스템을 생사 대기열로 모델링하고, 상태 전이와 전이율을 나타내는 다이어그램을 작성하세요.
- 시스템의 정상 상태 행동을 유도하고, 성능 지표로 사용 가능한 수리공의 가동률과 기계의 운영 확률을 계산하세요.
- 특히, μ/λ=4인 경우에 대해 K와 M의 값이 각각 3, 7일 때 성능 지표를 표로 작성하세요.μ/λ=4\mu / \lambda = 4

연습 문제 1.7: 주유소 대기열 시스템

이 문제는 **주유소 대기 시스템**을 모델링하는 문제입니다.

시나리오 설명:

- 두 개의 주유 펌프가 있으며, 총 4대의 차량이 주차할 수 있습니다.
- 차량은 시간당 20대의 비율로 푸아송 분포에 따라 도착합니다.
- 주유 시간은 평균 10분인 지수 분포를 따릅니다.
- 주유소가 가득 차 있으면 도착한 차량은 서비스를 받지 못하고 떠납니다.

요구사항:

- **(i)** 상태와 전이율을 나타내는 다이어그램을 작성하세요.
- **(ii)** 시스템의 정상 상태 분포를 계산하고, 차량이 주유소에 도착했을 때 대기 공간이 없는 경우를 포함하여 각 상태의 확률을 구하세요.
- **(iii)** 서비스를 받지 못하고 떠나는 차량의 비율을 계산하세요.
- **(iv)** 주유 서비스를 받은 차량의 체류 시간 분포와 그 평균을 계산하세요.

연습 문제 1.8: 주유소 시뮬레이션

이 문제는 연습 문제 1.7에서 주어진 주유소 시스템을 시뮬레이션하는 문제입니다.

요구사항:

- 다음 두 가지 조건에 따라 주유소 시스템을 시뮬레이션하세요:
  - **(i)** 주유 시간이 평균 10분인 지수 분포를 따르는 경우
  - **(ii)** 주유 시간이 8분에서 12분 사이의 균등 분포를 따르는 경우
- 각 시나리오에 대해 1100대의 차량을 시뮬레이션한 후, 처음 100대를 제외하고 나머지 1000대의 결과를 분석하세요.
- **(i)** 서비스를 받지 못하고 떠난 차량의 비율을 계산하세요.

연습 문제 1.9: 포기하는 대기열 시스템 (Reneging)

시나리오 설명:

- 단일 서버 대기열에서 고객은 도착률 λ에 따라 푸아송 분포로 도착합니다.
- 서비스 시간은 서비스율 μ를 따르는 지수 분포를 가집니다.
- 고객의 인내심은 한계가 있어, 대기 시간이 평균 1/θ를 초과하면 서비스를 받지 않고 대기열을 떠납니다(포기 또는 이탈).

요구사항:

- 이 시스템을 **마코프 생사 과정**으로 모델링하세요.
- 정상 상태 분포를 구하고, 서비스를 받는 고객들의 평균 대기 시간을 계산하세요.

연습 문제 1.10: 대기열을 피하는 시스템 (Balking)

시나리오 설명:

- 단일 서버 대기열에서 고객은 도착률 λ를 따르는 푸아송 분포로 도착합니다.
- 서비스 시간은 서비스율 μ를 따르는 지수 분포입니다.
- 고객은 대기열에 n명의 다른 고객이 있을 경우 확률 $p(n) = p^n$으로 대기열에 합류합니다. 그렇지 않으면, 시스템에 합류하지 않고 떠납니다(대기열 회피).

요구사항:

- 이 시스템을 **마코프 생사 과정**으로 모델링하세요.
- 정상 상태 분포를 구하고, 대기열에 합류하지 않는 고객 비율과 대기열에 합류한 고객들의 평균 대기 시간을 계산하세요.

연습 문제 1.11: 이스라엘식 대기 시스템 (The Israeli Queue)

시나리오 설명:

- 이 문제는 1950년대 이스라엘의 영화 티켓 대기열 시스템을 모델링합니다.
- p의 확률로 대기열에 있는 고객들은 서로를 알고 있을 수 있습니다.
- 새로 도착한 고객은 대기열을 순차적으로 살펴보며, 아는 사람이 있으면 그 사람과 함께 줄을 서고, 그렇지 않으면 대기열의 끝에 섭니다.
- 고객 도착률은 λ이며, 서비스율은 μ입니다.

요구사항:

- 이 시스템을 **마코프 생사 과정**으로 모델링하세요.
- 정상 상태 분포를 구하고, 대기열에서 아는 사람을 만나 함께 줄을 서는 고객 비율을 계산하세요.
- 고객 대기 시간의 분포를 계산하세요.

연습 문제 1.12: 휴가를 가는 서버가 있는 대기열 시스템 (Queue with Vacations)

시나리오 설명:

- M/M/1 대기열 시스템에서 서버는 고객이 없을 때 일정한 확률로 **휴가**를 갑니다.
- 휴가 시간은 지수 분포 $Exp(\theta)$를 따릅니다.
- 서버가 휴가 중일 때 고객이 도착하면 서비스가 지연됩니다.

요구사항:

- 이 시스템을 **마코프 체인**으로 모델링하고, 정상 상태 분포를 유도하세요.