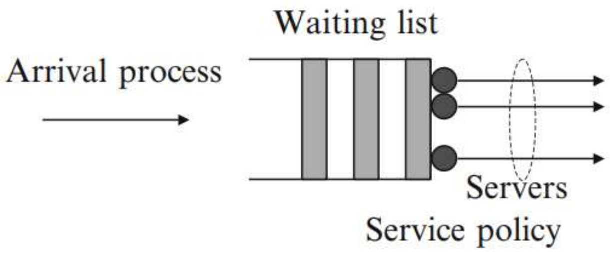
4.4) Notations for Queuing Systems

<Queue>



- 요청의 도착 과정

[은행에 고객들이 도착하는 과정]

- 서비스를 기다리는 요청 목록

[은행에 도착한 고객들은 창구가 비어 있을 때까지 대기]

- 목록에 있는 다양한 요청에 대한 서비스 정책

[예를 들어, 은행은 FIFO와 같은 서비스 정책을 채택해 가장 먼저 온 고객이 가장 먼저 서비스를 받음]

- 동시에 처리할 수 있는 요청의 최대 수를 나타내는 서버의 수

[서버는 은행의 각 창구로 여러 창구가 있으면 여러 고객을 동시에 처리]

- 각 요청의 서비스 지속 시간에 대한 통계

[각 고객의 서비스 처리 시간은 다를 수 있음]

<켄달 표기법> - 큐잉 시스템의 구조를 나타내는 표기법

$$A/B/S/\Delta/E$$

1) A : 도착 과정의 유형

- A = M : Poisson Process
- A = GI : Renewal Arrival

[은행에 도착하는 각 고객의 도착 사이의 시간은 동일한 확률 분포를 따르고, 각각의 도착은 이전 도착에 의존하지 않음]

- A = D : Deterministic Process

[은행에 고객이 도착하는 시간의 간격은 변하지 않고 일정함]

2) B : 요청의 서비스 시간

- B = M : Exponentially distributed service duration
- B = G : Generally distributed service process (일반적인 확률 분포를 사용. 지수 분포, 정규 분포, 균등 분포 등 다양한 분포 가능)

- B = D : Deterministic service time

3) S : 서버의 수

4) Δ : 현재 서비스되고 있는 요청을 포함하여 요청에 대한 방의 수 (Δ ≥ S) (대기열의 크기)

5) E : 서비스 요청을 생성할 수 있는 출처의 수

< Compound Arrival Process (복합 도착 과정) >
: 여러 개의 요청이 그룹으로 도착하는 상황
: $A^{[\text{comp}]}/B/S...$ 형태의 모델로 표현됨.
: 첨자 "[comp]"는 도착 그룹의 분포
: 이때 B는 도착 그룹의 분포에 따라 달라짐.

< Batched Services >
: 여러 요청이 함께 처리됨.
: $A/B^{[b]}/S...$ 형태의 모델로 표현됨.
: 첨자 "[b]"는 함께 서비스되는 요청의 수

< Work-Conserving >
: 서버가 대기열에 요청이 있는 한 비어 있지 않는 경우
: 서버가 때때로 '휴가 중'일 수 있어서 시스템에 작업이 있음에도 불구하고 서버가 잠시 동안 그것을 처리하지 않음.

<트래픽>
: 통신 시스템에서 트래픽은 데이터, 음성, 이미지, 비디오 등의 정보가 네트워크를 통해 전송되는 과정을 의미

$$\rho = \lambda E[X]$$

- **트래픽 강도 (ρ)**: 네트워크의 부하를 나타내는 지표. 이 값이 1에 가까워질수록 네트워크는 포화 상태에 이르며, 이는 지연이 증가하고 성능이 저하됨을 의미
  : '**Erlangs**' 라는 단위로 측정됨. ('Erlang'은 통신 시스템에서 도착율과 서비스 시간을 기반으로 한 트래픽의 양을 나타내는

단위)
- **평균 도착률 (λ)**: 단위 시간당 서비스를 요청하는 이벤트의 평균 수
- **평균 서비스 지속 시간 (E[X])**: 각 요청을 처리하는 데 걸리는 평균 시간

4.5) Little Theorem and Insensitivity Property

<리틀의 법칙>

1) 정의

대기열은 다음 두 가지로 특징지어질 수 있음 :

① N : 서비스 중인 것을 포함하여 대기열에 있는 요청의 평균 수

② T : 요청이 대기열에 들어간 순간부터 서비스가 끝날 때까지의 평균 시스템 지연

2) 가정

① **경계 조건** : 대기열은 일정 시간에 비어 있어야 한다.

② **고객의 보존** : 모든 도착하는 고객(즉, 시스템에 들어오는 요청)은 결국 서비스를 완료하고 시스템을 떠날 것이다.

3) 식

$$N = \lambda T$$

- 평균 도착률 (λ)

<무감도 특성>

1) 정의

① **평균 대기 시간 (T)** : 평균 지연 시간과 같은 의미로 대기열 규율에 관계없이 평균 대기 시간(T)은 동일하다.

: 예를 들어 FIFO, LIFO, 랜덤 등의 규율에서 평균 대기 시간은 같다.

② **지연의 다른 측면** : 그러나 지연의 다른 측면(예: 지연 분산)은 대기열 규율에 따라 달라질 수 있다. (지연 분산은 지연 시간이 평균 값으로부터 얼마나 크게 변동하는지를 나타냄.)

2) 가정

① 서비스 정책(예: FIFO, LIFO)이 서비스 시간과 독립적이어야 한다.

② 서비스 정책이 작업 보존적이어야 한다. 즉, 대기열에 작업이 있는 한 서버가 비어 있지 않아야 한다. (은행에 손님이 대기 중인데 창구 직원이 놀고 있으면 안됨.)

4.6) M/M/1 Queue Analysis

<M/M/1>

1) 정의

① **포아송 도착 과정**

: 주어진 시간 간격 내에 일정 수의 이벤트가 발생할 확률이 일정함.

② **지수 분포된 서비스 시간**

: 서비스가 얼마나 진행되었는지에 상관없이 남은 서비스 시간의 분포는 항상 동일

: 예를 들어 서비스에 평균적으로 5분이 걸린다면, 서비스 시간은 지수 분포를 따라 무작위로 결정된다. 서비스가 이미 3분 진행된 후에도 남은 시간은 여전히 평균 5분의 지수 분포를 따른다. 즉, 서비스가 이미 어느 정도 진행되었다고 해서 남은 시간이 줄어들거나 바뀌지 않는다.

③ **단일 서버**

2) 그림으로 표현

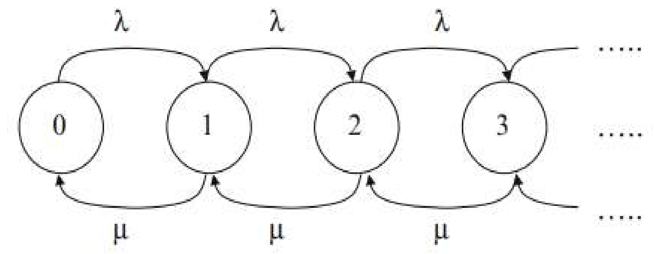


Fig. 4.11: Continuous-time Markov chain modeling an M/M/1 queue

- **λ** : 시스템에 새로운 요청이 도착하는 비율 (포아송 분포의 매개변수)
- **μ** : 시스템이 요청을 처리하고 완료하는 비율 (서비스 시간이 따르는 지수 분포의 평균 비율)
- **입력 트래픽 강도 (ρ)** : $\frac{\lambda}{\mu}$ (ρ<1 일 때 대기열이 안정하므로 1미만)

3) 각 상태의 확률 계산

$$P_i = P_0\left(\frac{\lambda}{\mu}\right)^i = P_0\rho^i$$
$$P_0 = \frac{1}{1+\sum_{i=1}^{\infty}\rho^i} = \frac{1}{\sum_{i=0}^{\infty}\rho^i} = 1-\rho \quad \text{(normalization)}$$

- P_i : 시스템이 i개의 요청을 처리하고 있는 상태의 확률
- P_0 : 시스템에 아무 요청도 없을 확률
  : 시스템이 비어 있을 확률이 1에서 트래픽 강도를 뺀 값과 같음
- 안정성 조건(ρ<1)을 따르면 P_0>0 이 된다. 따라서, 대기열이 안정적이라면, 때때로 비어 있어야 한다.

(∴) $\boxed{P_i = (1-\rho)\rho^i}$

4) 상태 확률 분포의 확률 생성 함수(PGF)

$$P(z) = \sum_{i=0}^{\infty}(1-\rho)\rho^i z^i = \frac{1-\rho}{1-z\rho}$$

- **P(z)** : 확률 생성 함수(PGF)이다. (PGF는 각각의 상태 확률을 합쳐서 하나의 함수로 나타내는 데 사용됨. PGF의 첫 번째 미분은 확률 변수의 평균)

$$N = \sum_{i=0}^{\infty} i\,(1-\rho)\rho^i = \left.\frac{\mathrm{d}P(z)}{\mathrm{d}z}\right|_{z=1} = \frac{\rho}{1-\rho}$$

- **N** : 시스템 내 요청의 평균 수. (PGF의 도함수로 구할 수 있음)

5) 평균 지연 시간(T)

$$T = \frac{N}{\lambda} = \frac{1}{\mu-\lambda}$$

6) 대기열에 의해 처리되는 트래픽

$$\gamma = \sum_{i=1}^{\infty} \mu (1-\rho)\rho^i = \mu \sum_{i=1}^{\infty} (1-\rho)\rho^i = \mu (1-P_0)$$

7) M/M/1 대기열에 i개 이상의 요청이 포함될 확률

$$\begin{aligned} \mathrm{Prob}\{n > i\} &= 1 - \sum_{j=0}^{i} (1-\rho)\,\rho^j \\ &= 1 - (1-\rho)\,\frac{1-\rho^{i+1}}{1-\rho} = \rho^{i+1} \end{aligned}$$