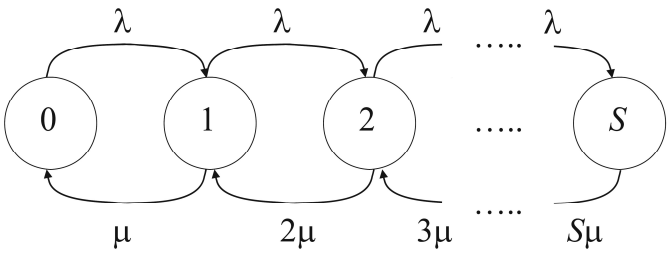
4.9 M/M/S/S Queue Analysis

4.9.1 M/M/S/S 대기열이란?

M/M/S/S : 패킷 등 요청이 도착하는 확률 모델이 포아송 프로세스를 따른다.
M/**M**/S/S : 단일 요청을 처리하는 데 걸리는 시간이 지수 분포를 따른다.
M/M/**S**/S : 요청을 동시에 S개까지 처리할 수 있다. (= 서비스가 S개다.)
M/M/S/**S** : 요청이 S개까지 대기할 수 있다.



M/M/S/S 대기열의 연속 시간 마르코프 체인 모델
Continuous-time Markov chain modeling and M/M/S/S queue

- 요청이 대기열에 0개부터 S개까지 대기할 수 있으므로 S+1개의 상태가 존재함.

- 출생률(Birth-rate) : $\lambda_i = \lambda \quad (i = 0, 1, \cdots, S-1)$

- 사망률(Death-rate) : $\mu_i = i\mu \quad (i = 1, \cdots, S)$

- 임의의 상태에 대한 확률(P_i) : $P_i = \frac{\rho^i}{i!} P_0 \qquad \left(\rho = \frac{\lambda}{\mu},\ i = 1, 2, \cdots, S\right)$
  (여기서 ρ는 트래픽 강도(Traffic inensity)라고 한다.)

- 대기 중인 요청이 0일 확률(P_0) : $P_0 = \frac{1}{\sum_{i=0}^{S} \frac{\rho^i}{i!}}$

- S개의 대기 공간이 꽉 차서 새로 온 요청이 거절될 확률(Blocking Probability, P_B) :

$$P_B(S, \rho) \equiv P_S = \frac{\frac{\rho^S}{S!}}{\sum_{i=0}^{S} \frac{\rho^i}{i!}}$$

  (이 공식은 Erlang-B 모델을 따른다.)

- 거절된 요청들은 일정 시간이 지나면 다시 요청하게 되는데, 해당 요청까지의 시간 간격은 이전의 거절당한 요청과는 무관하게 λ라는 rate을 유지한다.

4.9.2 S의 최솟값을 구하는 문제와 Erlang-B 테이블

- ρ가 주어졌을 때, P_B가 일정 값 이하가 되기 위한 S의 최솟값을 구하는 문제에서, $P_B(0,\rho)=1$임을 이용해서 다음과 같이 재귀적으로 비교적 간단히 계산할 수 있다.

$$\frac{1}{P_B(i,\rho)} = 1 + \frac{i}{\rho P_B(i-1,\rho)}$$

- 이 계산 방식을 더 효율적으로 이용하기 위해서 Erlang-B 테이블을 사용할 수 있다.

Erlang-B 테이블

S	1%	2%	3%	5%	7%
1	0.0101	0.0204	0.0309	0.0526	0.0753
2	0.153	0.223	0.282	0.381	0.470
3	0.455	0.602	0.715	0.899	1.06
4	0.869	1.09	1.26	1.52	1.75
5	1.36	1.66	1.88	2.22	2.50
6	1.91	2.28	2.54	2.96	3.30
7	2.50	2.94	3.25	3.74	4.14
8	3.13	3.63	3.99	4.54	5.00
9	3.78	4.34	4.75	5.37	5.88
10	4.46	5.08	5.53	6.22	6.78
11	5.16	5.84	6.33	7.08	7.69
12	5.88	6.61	7.14	7.95	8.61
13	6.61	7.40	7.97	8.83	9.54
14	7.35	8.20	8.80	9.73	10.5
15	8.11	9.01	9.65	10.6	11.4
16	8.88	9.83	10.5	11.5	12.4
17	9.65	10.7	11.4	12.5	13.4
18	10.4	11.5	12.2	13.4	14.3
19	11.2	12.3	13.1	14.3	15.3
20	12.0	13.2	14.0	15.2	16.3
21	12.8	14.0	14.9	16.2	17.3
22	13.7	14.9	15.8	17.1	18.2
23	14.5	15.8	16.7	18.1	19.2
24	15.3	16.6	17.6	19.0	20.2
25	16.1	17.5	18.5	20.0	21.2
26	17.0	18.4	19.4	20.9	22.2
27	17.8	19.3	20.3	21.9	23.2
28	18.6	20.2	21.2	22.9	24.2
29	19.5	21.0	22.1	23.8	25.2
30	20.3	21.9	23.1	24.8	26.2

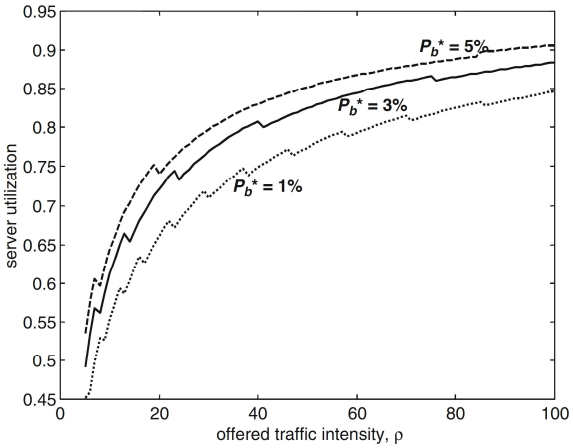
- 예) 전송 밀도가 $\rho=6$이면서 요청이 거절될 확률이 3% 이하가 되려면, 서비스 수가 최소 11이 되어야 한다.

Q) 전송 밀도가 $\rho=12$인 환경에서 요청이 거절될 확률이 5% 이하가 되려면, 서비스 수가 얼마나 되어야 하는가?

4.9.3 서비스 이용률

- 서비스 이용률(Utilization factor, ϕ) : 서버 하나의 평균 이용률(서버가 얼마나 바쁜가?)

$$\phi = \frac{\rho}{S}(1 - P_B)$$



M/M/S/S 대기열에서의 전송 밀도 대비 서버 이용률

- 이 그래프가 연속적이지 못 한 구간이 발생하는 이유는, ϕ의 값을 결정하는 S가 ρ와 P_B값에 따라 달라지기 때문이다. P_B의 허용치와 ρ의 값에 따라 S의 최솟값이 달라지기 때문이다.

- ρ값이 증가하면, 서비스 처리 속도(사망률)가 신규 요청 발생 속도(출생률)보다 훨씬 더 빨라지므로, 서버가 그만큼 더 바빠지기에 ϕ 값이 커진다.

- P_B의 허용치가 증가하면, 서버가 그만큼 더 바빠져도 되기에 ϕ 값이 커진다.

4.9.4 M/M/S/S queue에서도 적용되는 리틀 법칙

- 대기 중인 요청 수의 기댓값(N) : $N = \sum_{i=0}^{S} iP_i = \rho(1-P_S)$
- 이 시스템에서의 평균 arrival rate(λ_s) : $\lambda_s = \overline{\lambda} = \sum_{i=0}^{S-1} \lambda_i P_i = \lambda(1-P_S)$
- 시스템의 stability를 만족하기 위해서 mean arrival rate(λ_s)와 mean traffic carried rate(γ)가 같아야 한다.

- 주의) 평균적으로 요청이 거절되는 빈도 또한 단순하게 λP_B로 대기열에 있는 요청이 S개일 때의 arrival rate가 되지만, 실제로 요청이 거절되는 것과 요청을 수락하는 것의 확률 모델은 대기열에 있는 요청 개수에 의존하기 때문에 푸아송 분포를 따르지 않는다.

- 리틀 법칙 : $N = \lambda_s T = \lambda(1-P_S)T \rightarrow N = \frac{1}{\mu}T$
  (M/M/S/S에서의 모든 요청은 대기열에 들어가자마자 바로 서비스를 받게 되거나 대기열에 못 들어가면 거절되어서 대기열 내 요청을 처리율은 μ로 동일)

- M/G/S/S 대기열의 경우, 무감도 특성을 따른다고 가정하면, 새로운 요청이 거절될 확률의 모델이 되는 Erlang-B 모델을 똑같이 적용할 수 있다.

● **부록 : 공식 유도**

$$\frac{1}{P_B(S,\rho)} = \frac{\sum_{i=0}^{S} \frac{\rho^i}{i!}}{\frac{\rho^S}{S!}} = \frac{\frac{\rho^S}{S!} + \sum_{i=0}^{S-1} \frac{\rho^i}{i!}}{\frac{\rho^S}{S!}} = 1 + \frac{\sum_{i=0}^{S-1} \frac{\rho^i}{i!}}{\frac{\rho^S}{S!}} = 1 + \frac{S}{\rho} \times \frac{\sum_{i=0}^{S-1} \frac{\rho^i}{i!}}{\frac{\rho^{S-1}}{(S-1)!}} = 1 + \frac{S}{\rho P_B(S-1,\rho)}$$

$$\therefore \frac{1}{P_B(i,\rho)} = 1 + \frac{i}{\rho P_B(i-1,\rho)}$$

$$N = \sum_{i=0}^{S} iP_i = \sum_{i=0}^{S} i\frac{\rho^i}{i!}P_0 = \rho\sum_{i=1}^{S} \frac{\rho^{i-1}}{(i-1)!}P_0 = \rho\sum_{i=1}^{S} P_{i-1} = \rho(1-P_S)$$

$$\therefore N = \rho(1-P_S)$$

$$\lambda_s = \overline{\lambda} = \sum_{i=0}^{S-1} \lambda_i P_i = \sum_{i=0}^{S-1} \lambda P_i = \lambda \sum_{i=0}^{S-1} P_i = \lambda(1-P_S)$$

$$\therefore \lambda_s = \lambda(1-P_S)$$

4.10 M/M/S/S/P Queue Analysis

4.10.1 M/M/S/S/P 대기열이란?

- 기존의 M/M/S/S 대기열 시스템에서 요청을 할 수 있는 주체가 무한대가 아니라 P 만큼 유한한 조건이 추가된 경우다.

- P가 요청을 처리하는 대기열 이론의 서버 수(S)보다 더 적으면, 요청을 거절할 수 있는 확률이 없으니 $P > S$의 경우만을 고려하겠다.

예) 전화

- 전화를 하는 사람의 수는 무한대가 아니다. 이때, 전화 요청을 할 수 있는 그 사람의 인원이 P에 해당한다.

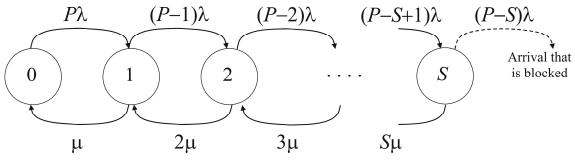
- 가정

1) P 명의 사람들의 똑같이 λ의 빈도로 하루에 통화를 하는 시간을 가짐
2) P 명의 사람들이 통화하는 여부가 서로에게 독립이라서 다른 사람들의 통화 상태 여부가 자신의 통화 상태 여부에 영향을 전혀 주지 않음

→ 서로 독립인 푸아송 프로세스의 합으로 인해, 전체 인원(P)에 의한 통화 빈도는 λP 가 된다.

- 하지만, 여기서 i명의 사람이 이미 통화 중이라면, i명만큼 현재 전체 인원에 대한 통화 빈도에서 제외된다. **다시 말해서 Poisson arrival process를 따르지 않는다.**

$$\lambda_i = (P - i)\lambda$$



M/M/S/S/P 대기열의 연속 시간 마르코프 채인 모델
Markov chain model for the M/M/S/S/P queue

- **PASTA 특성을 가지지 않음** : 전체 요청 중 전체 서버의 과부하로 거절된 요청의 비율 (Call Congestion, P_B)와 전체 시간 중 서버 전체가 과부하 상태인 시간의 비율 (Time Congestion, E)가 다르다. P_B는 P에 영향을 받고, E는 S에 영향을 받기 때문이다.

정리 : M/M/S/S/P에서는 Poisson arrival process가 적용되지 않아 PASTA 특성이 없음

4.10.2 M/M/S/S/P 대기열에서 각 상태의 확률과 요청을 거절할 확률

- 그럼에도 불구하고 균형 조건(Equilibrium condition)을 이용해서 각 상태의 확률을 구할 수 있다. ($i = 1, 2, \cdots, S$)

$$P_i = \binom{P}{i}\rho^i P_0 = \frac{\binom{P}{i}\rho^i}{\sum_{i=0}^{S}\binom{P}{i}\rho^i}$$

$$P_0 = \frac{1}{\sum_{i=0}^{S}\binom{P}{i}\rho^i}$$

- 통화를 하는 시간과 통화를 하지 않는 시간이 지수 분포를 따르고, 각각의 mean rate를 λ와 μ로 볼 수 있다. 통화를 하는 상태로 있을 확률을 P_{ON}이라고 하면, 다음과 같이 계산된다.

$$P_{ON} = \frac{\lambda}{\lambda+\mu} = \frac{\rho}{\rho+1}$$

$$P_i = \frac{\binom{P}{i}P_{ON}^i(1-P_{ON})^{P-i}}{\sum_{i=0}^{S}\binom{P}{i}P_{ON}^i(1-P_{ON})^{P-i}}$$

- P_{ON}을 이용해서 구한 식 P_i을 보면, P_{ON}의 확률로 통화 상태를 가진 P명의 독립적인 사용자 중 i명이 통화를 현재 진행 중인 확률로 P_i를 정의할 수 있다.
  그리고 P_i가 이항분포(binomial distribution)를 가졌음을 알 수 있다.

- 앞에서 설명했듯이 PASTA 특성이 없어서 모든 서비스가 현재 전화 요청을 처리하는 확률인 $P_{full} = P_S$는 요청을 거절할 확률(P_B)과 다르다.

1. Arrival rate의 기댓값 : $\lambda_{avg} = (P - N)\lambda$

- 여기서 N은 대기열에 있는 요청의 기댓값이다. ($N = \sum_{n=0}^{S} nP_n$)
- 앞서 설명했듯이 대기열에 있는 요청의 수에 따라서 arrival rate이 달라져서 λ_{avg}는 P의 값과 N의 값에 의해서 결정된다.

2. 그리고 $n = S$인 상태에서 거절되는 요청의 arrival rate인 block rate :

$$\lambda_{block} = (P - S)\lambda P_S$$

3. Mean arrival rate accepted :

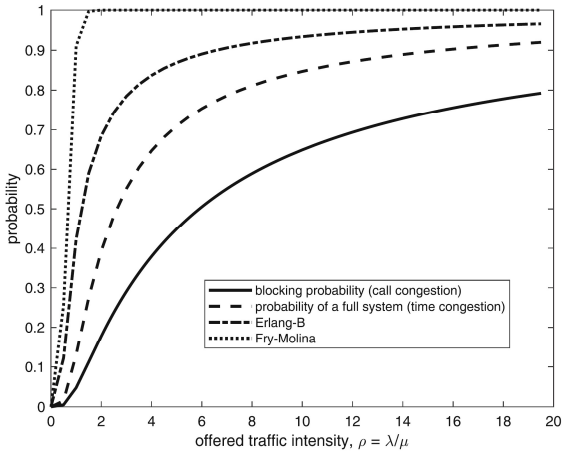
$$\lambda_s = \overline{\lambda} = \sum_{i=0}^{S-1} (P - i)\lambda P_i = \sum_{i=0}^{S} (P - i)\lambda P_i - (P - S)\lambda P_s = \lambda_{avg} - \lambda_{block}$$

4. Blocking Probability : $P_B = \frac{\lambda_{block}}{\lambda_{avg}} = P_S \frac{P - S}{P - N} = \frac{\binom{P-1}{S}\rho^S}{\sum_{i=0}^{S}\binom{P-1}{i}\rho^i}$

- 더 간단하게 계산하고 싶다면, 최대 트래픽 강도(Traffic inensity)가 $\rho_{\max} = P\rho$인 M/M/S/S 대기열로 근사화해서 Erlang-B의 공식을 이용해 P_B를 계산할 수 있다.

$$P_B = \frac{\rho_{\max}^S}{S! \sum_{i=0}^{S} \frac{\rho_{\max}^i}{i}}$$

- 중요한 점은, Blocking Probability를 계산하는 방식이 여러 가지가 존재한다는 것이다.



다양한 방식으로 계산한 M/M/S/S/P 대기열의 Blocking Probability 값의 비교

● **부록 : 공식 유도**

$$P\lambda P_0 = \mu P_1 \;\rightarrow\; P_1 = P\rho P_0$$
$$(P-1)\lambda P_1 = 2\mu P_2 \;\rightarrow\; P_2 = \frac{P-1}{2}\rho P_1$$
$$\vdots$$
$$\lambda P_{S-1} = S\mu P_S \;\rightarrow\; P_S = \frac{1}{S}\rho P_{S-1}$$
$$\therefore P_i = \frac{P!}{(P-i)!i!}\rho^i P_0 = \binom{P}{i} P_0$$

$$P_{ON} = \frac{\rho}{\rho+1} = 1 - \frac{1}{\rho+1}$$
$$\frac{1}{\rho+1} = 1 - P_{ON}$$
$$\rho+1 = \frac{1}{1-P_{ON}}$$
$$\rho = \frac{P_{ON}}{1-P_{ON}}$$

$$P_i = \frac{\binom{P}{i}\rho^i}{\sum_{i=0}^{S}\binom{P}{i}\rho^i} = \frac{\binom{P}{i}P_{ON}^i(1-P_{ON})^{-i}}{\sum_{i=0}^{S}\binom{P}{i}P_{ON}^i(1-P_{ON})^{-i}} = \frac{\binom{P}{i}P_{ON}^i(1-P_{ON})^{P-i}}{\sum_{i=0}^{S}\binom{P}{i}P_{ON}^i(1-P_{ON})^{P-i}}$$

$$\lambda_{avg} = \sum_{n=0}^{S}(P-n)\lambda P_n \qquad (N = \sum_{n=0}^{S} nP_n)$$
$$= P\lambda\sum_{n=0}^{S}P_n - \lambda\sum_{n=0}^{S}nP_n$$
$$= (P-N)\lambda$$