각 고객 class 별 서버의 수에 제한이 있는 M/M/2 대기행렬모형 분석
- An analysis of M/M/2 system with restriction to the number of servers for each customer class -

정 재 호*
Jung, Jae Ho
허 선**
Hur, Sun

Abstract

In this paper, we model a two-server queueing system with priority, to which we put a restriction of the number of servers for each customer class. A group of customers is divided into two different classes. The class 1 customers has non-preemptive priority over class 2 customers. We use the method of PGF depending on the state of server. We find the PGF of the number of customers in queue, server utilization, mean queue length and mean waiting time for each class of customers.

1. 서론

도착과정이 포아송 과정(Poisson process)을 따르고, 서비스 시간이 지수분포를 따르며 복수의 서버를 갖는 모형을 M/M/c 대기행렬모형이라 한다. M/M/c 대기행렬모형은 현실 모형을 분석하는 데 부족한 점이 많다. 현실 시스템에서는 도착하는 고객의 유형이 다른 경우가 많고, 이러한 경우 유형에 따라 고객의 도착률이 다를 수 있으며, 각 서버의 서비스율도 다를 수 있다. 또한, 고객의 유형에 따라 서비스를 받을 수 있는 서버의 수가 제한될 수도 있다. 이와 같이 보다 현실적인 모형으로 접근하기 위해서는 M/M/c 대기행렬시스템에 대한 변형된 모형의 연구가 필요하다.

기존 연구를 살펴보면, 이호우[1]과 Cooper[2]에서는 일반적인 M/M/c 대기행렬모형에 관하여 분석하였다. 또한, Cooper[2]에서는 여러 형태의 순서적 서버 사냥모형에 관한 분석 기법이 제시되어 있다. Kella[3]에서는 비축출형 우선순위를 가지는 M/M/c 대기행렬모형에 대한 분석을 하였다. Wagner[4]에서는 비축출형 우선순위 서비스를 적용하는 시스템 용량이 유한한 복수 서버 모형을 분석하였다.

* 한양대학교 정보경영공학과 박사과정
** 한양대학교 정보경영공학과 교수

본 논문에서는 순서적 서버 사냥 모형을 적용하되, 고객 집단을 두 가지 class로 나누어, class 1의 고객이 class 2의 고객에 대하여 우선순위를 가질 때, 이 고객의 유형에 따라 서비스를 받는 서버의 수에 제한이 있는 M/M/2 대기행렬모형에 대하여 분석한다. 이 시스템에서 안정 상태에서의 고객 class 별 대기고객수의 PGF(probability generating function)를 유도하고, 각 서버의 서버 이용률, 고객 class 별 평균대기고객수와 평균대기시간 등을 유도하는 것을 본 연구의 목적으로 한다.

2. 대기고객수 분석

2.1 시스템 가정

본 연구는 두 가지 유형의 class의 고객들을 서비스하는 서버의 수에 제한이 있는 M/M/2 대기행렬시스템을 연구 대상으로 한다. 본 연구에 사용될 가정은 다음과 같다.

i) 시스템에 도착하는 고객의 class 는 두 가지이다.

ii) 각 class 고객은 도착률이 λ_i $(i=1,2)$이고 서로 독립인 포아송 과정으로 시스템에 도착한다.

iii) 각 서버는 서로 독립이고 서비스율이 μ_i $(i=1,2)$인 지수분포를 따른다.

iv) <그림 2-1>에서와 같이 각 class의 고객이 서비스를 받을 수 있는 서버의 수에 제한을 둔다. 즉, class 1의 고객은 두 서버 중에서 하나의 서버를 선택하여 서비스를 받을 수 있는 반면, class 2의 고객은 서버 2에서만 서비스를 받을 수 있다.

v) 고객이 기다리는 대기열의 수는 하나이다.

vi) 두 서버가 모두 유휴할 때 도착한 class 1의 고객은 서버 1을 찾아 먼저 서비스를 받는다.

vii) 두 class의 고객이 대기열에 함께 있을 때에는 class 1의 고객이 먼저 서비스를 받으며, class 1의 고객이 대기열에 없고 서버 2가 유휴하면 비로소 class 2의 고객이 서비스를 받는다. class 2의 고객이 먼저 서비스를 받고 있을 때에는 class 1의 고객이 도착하더라도 class 2의 고객을 축출하지 않고 서비스가 끝날 때까지 기다린다.

viii) 시스템의 용량과 고객의 모집단의 크기는 무한하다.

ix) 기타 사항들은 일반적인 대기행렬모형의 가정에 준한다.

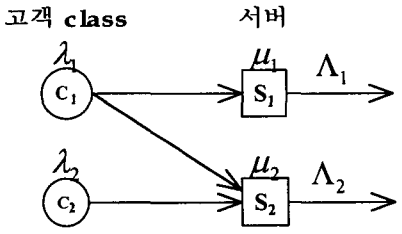

그림 2-1. class 별 서버의 수에 제한이 있는 M/M/2 대기행렬모형

2.2 기호 설명

본 논문에서 사용하게 될 변수 및 기호를 아래와 같이 정의한다.

· λ_i : class i 고객의 도착률, $i=1,2$

· μ_i : 서버 i의 서비스율, $i=1,2$

· ρ_i : 서버 i의 서버 이용률, $i=1,2$

· Λ_i : 서버 i로부터의 고객 이탈률, $i=1,2$

· s : 서버의 상태, 즉, $s=\begin{cases} 0, & \text{두 서버가 모두 유휴한 경우} \\ 1, & \text{서버 1만 바쁜 경우} \\ 2, & \text{서버 2만 바쁜 경우} \\ 3, & \text{두 서버가 모두 바쁜 경우} \end{cases}$

· $P_{i,j,s}$: 안정 상태에서 서버의 상태가 s일 때, class 1의 대기고객수가 i명, class 2의 대기고객수가 j명일 확률

· $P(z,w,\theta)$: $P_{i,j,s}$의 결합확률생성함수(joint probability generating function : PGF)

· L_i : class i 고객의 평균대기고객수, $i=1,2$

· W_i : class i 고객의 평균대기시간, $i=1,2$

3. 대기고객수 PGF

3.1 전이율 다이어그램(transition diagram)

본 모형에 대한 시스템 상태를 아래와 같이 서버의 상태와 대기고객수로 정의한다.

상태 : (i, j, s)

여기서, i = 대기 공간에 있는 class 1의 고객수, $i=0,1,2,\ldots$

j = 대기 공간에 있는 class 2의 고객수, $j=0,1,2,\ldots$

s = 서버의 상태 = 0,1,2,3

상태 정의를 바탕으로 전이율 다이어그램을 <그림 3-1>과 같이 나타낼 수 있다.

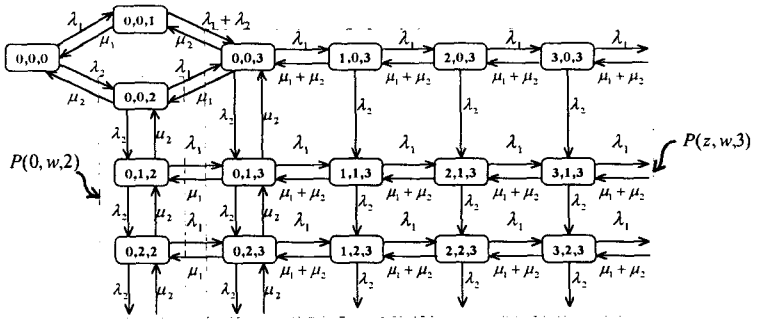

그림 3-1. 전이율 다이어그램(transition diagram)

3.2 안정상태 방정식

<그림 3-1>로부터 안정상태 방정식들을 유도하면 식 (3-1)~(3-8)과 같다.

$$(\lambda_1+\lambda_2)P_{0,0,0} = \mu_1 P_{0,0,1}+\mu_2 P_{0,0,2} \tag{3-1}$$

$$(\mu_1+\lambda_1+\lambda_2)P_{0,0,1}=\lambda_1 P_{0,0,0}+\mu_2 P_{0,0,3} \tag{3-2}$$

$$(\mu_2+\lambda_1+\lambda_2)P_{0,0,2}=\lambda_2 P_{0,0,0}+\mu_1 P_{0,0,3}+\mu_2 P_{0,1,2} \tag{3-3}$$

$$(\mu_1+\mu_2+\lambda_1+\lambda_2)P_{0,0,3}=(\lambda_1+\lambda_2)P_{0,0,1}+\lambda_1 P_{0,0,2}+(\mu_1+\mu_2)P_{1,0,3}+\mu_2 P_{0,1,3} \tag{3-4}$$

$$(\lambda_1+\lambda_2+\mu_2)P_{0,j,2}=\lambda_2 P_{0,j-1,2}+\mu_1 P_{0,j,3}+\mu_2 P_{0,j+1,2},\quad j\geq 1 \tag{3-5}$$

$$(\lambda_1+\lambda_2+\mu_1+\mu_2)P_{0,j,3}=\lambda_1 P_{0,j,2}+\lambda_2 P_{0,j-1,3}+(\mu_1+\mu_2)P_{1,j,3}+\mu_2 P_{0,j+1,3},\quad j\geq 1 \tag{3-6}$$

$$(\lambda_1+\lambda_2+\mu_1+\mu_2)P_{i,0,3}=\lambda_1 P_{i-1,0,3}+(\mu_1+\mu_2)P_{i+1,0,3},\quad i\geq 1 \tag{3-7}$$

$$(\lambda_1+\lambda_2+\mu_1+\mu_2)P_{i,j,3}=\lambda_1 P_{i-1,j,3}+\lambda_2 P_{i,j-1,3}+(\mu_1+\mu_2)P_{i+1,j,3},\quad i\geq 1,\ j\geq 1 \tag{3-8}$$

여기서 구하고자 하는 값들은 각 대기고객수 확률인 $P_{i,j,s}$ 값들이지만, 위 식들로부터 구하기 어려우므로 PGF를 이용하여 대기고객수 PGF를 유도한다.

3.3 대기고객수 PGF

3.2 절에서 구한 안정상태 방정식들 식 (3-1)~(3-8)을 이용하여 대기고객수 PGF를 유도한다. 필요한 PGF를 정의하면 식 (3-9), (3-10)과 같다.

$$P(z,w,3)=\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}P_{i,j,3}\,z^i w^j \tag{3-9}$$

$$P(0,w,s)=\sum_{j=0}^{\infty}P_{0,j,s}\,w^j,\quad s=2,3 \tag{3-10}$$

여기서, $P(z,w,3)$은 두 서버가 바쁠 때 각 class의 대기고객수 PGF를 나타내며, $P(0,w,2)$는 서버 2만 바쁠 때 class 2의 대기고객수 PGF를, $P(0,w,3)$은 두 서버가 모두 바쁠 때 class 2의 대기고객수 PGF를 나타낸다.

식 (3-1)~(3-8)을 이용하여 필요한 PGF를 유도하면 식 (3-11), (3-12)와 같다.

$$P(0,w,2)=\frac{P_{0,0,1}\left\{\lambda_1+\lambda_2-\frac{\lambda_1+\lambda_2+\mu_2}{w}-\frac{\mu_2}{w^2}+\frac{\mu_1+\mu_2}{w f(w)}\right\}+P_{0,0,0}\left\{\frac{\lambda_1}{w}+\left(\frac{\mu_2}{w}-\frac{\mu_1+\mu_2}{f(w)}\right)\frac{\lambda_1+\lambda_2(1-w)}{w\mu_1}\right\}}{\lambda_1+\frac{1}{\mu_1}\left(\frac{\mu_2}{w}-\frac{\mu_1+\mu_2}{f(w)}\right)\left(\lambda_1+\lambda_2-\lambda_2 w+\mu_2-\frac{\mu_2}{f(w)}\right)} \tag{3-11}$$

$$P(z,w,3)=\frac{(1/z-1/f(w))(\mu_1+\mu_2)}{(\lambda_1+\lambda_2-\lambda_1 z-\lambda_2 w+\mu_1+\mu_2+(\mu_1+\mu_2)/z)K(w)}\cdot\frac{w-1}{w}$$
$$\cdot\left[P_{0,0,1}\{(\lambda_1+\lambda_2)\lambda_1/\mu_1+(1-((\lambda_1+\lambda_2)/\mu_1)(w-1))(\lambda_2-\mu_2/w)\}+P_{0,0,0}\cdot\lambda_1\mu_2/(\mu_1 w)\right]$$

여기서, $K(w)=\lambda_1+(1/\mu_1)(\mu_2/w-(\mu_1+\mu_2)/f(w))\cdot(\lambda_1+\lambda_2+\mu_2-\lambda_2 w-\mu_2/w)$

$$f(w)=[\lambda_1+\lambda_2(1-w)+\mu_1+\mu_2-\sqrt{(\lambda_1+\lambda_2-\lambda_2 w+\mu_1+\mu_2)^2-4\lambda_1(\mu_1+\mu_2)}]\ /\ (2\lambda_1). \tag{3-12}$$

식 (3-11)와 식 (3-12)과 같이 본 모형에 필요한 PGF $P(0,w,2)$와 $P(z,w,3)$을 구하였다. 여기서, 미지수인 $P_{0,0,1}$와 $P_{0,0,0}$은 4장에서 구하게 될 시스템 이용률로부터 얻을 수 있다.

4. 시스템 이용률

4.1 시스템 이용률 관계식

먼저 <그림 2-1>에서와 같이 고객 class와 서버와의 관계를 이용하여 각 서버에서의 고객 이탈률 Λ_1, Λ_2에 관한 관계식을 유도한다.

$$\Lambda_1=\lambda_1\cdot(\mu_1/(\mu_1+\mu_2))\Pr(\text{서버 1,2 모두 } busy)+\lambda_1\Pr(\text{서버 1 } idle) \tag{4-1}$$

$$\Lambda_2=\lambda_2+\lambda_1\cdot(\mu_2/(\mu_1+\mu_2))\Pr(\text{서버 1,2 모두 } busy)+\lambda_1\Pr(\text{서버 1 만 } busy) \tag{4-2}$$

각 서버의 이용률을 ρ_1, ρ_2라 하면 각각 식 (4-3)와 식 (4-4)와 같이 표현된다.

$$\rho_1=\Lambda_1/\mu_1=(\lambda_1/(\mu_1+\mu_2))\Pr(\text{서버 1,2 모두 } busy)+(\lambda_1/\mu_1)(1-\rho_1) \tag{4-3}$$

$$\rho_2=\Lambda_2/\mu_2=\lambda_2/\mu_2+(\lambda_1/(\mu_1+\mu_2))\Pr(\text{서버 1,2 모두 } busy)+(\lambda_1/\mu_2)\Pr(\text{서버 1만 } busy) \tag{4-4}$$

식 (4-3), (4-4)를 이용하여 다음 식들을 유도할 수 있다.

$$\Pr(\text{서버 1,2 모두 } busy) = (\mu_1+\mu_2)/\lambda_1)(1+\lambda_1/\mu_1)\rho_1-(\mu_1+\mu_2)/\mu_1 \quad (4\text{-}5)$$

$$\Pr(\text{서버 1만 } busy) = \rho_2-(1+\lambda_1/\mu_1)\rho_1-\lambda_2/\mu_2+\lambda_1/\mu_1/(\lambda_1/\mu_2) \quad (4\text{-}6)$$

$$\Pr(\text{서버 2만 } busy) = \rho_2-((\mu_1+\mu_2)/\lambda_1)(1+\lambda_1/\mu_1)\rho_1-(\mu_1+\mu_2)/\mu_1 \quad (4\text{-}7)$$

$$\Pr(\text{서버 1,2 모두 } idle) = ((\mu_1+\mu_2)/\lambda_1+\mu_2/\mu_1+\mu_1/\mu_2)\rho_1-(\lambda_1+\lambda_2)/\mu_2-\mu_2/\mu_1 \quad (4\text{-}8)$$

시스템에 입력되는 고객들은 모두 시스템을 나가게 되므로, 다음 관계식이 성립한다.

$$\Lambda_1+\Lambda_2=\lambda_1+\lambda_2 \quad (4\text{-}9)$$

여기서, $\Lambda_1=\mu_1\rho_1$, $\Lambda_2=\mu_2\rho_2$이므로 식 (4-9)으로부터 ρ_1과 ρ_2의 관계식을 유도하여 ρ_2에 관하여 정리하면 식 (4-10)과 같다.

$$\rho_2=-(\mu_1/\mu_2)\rho_1+(\lambda_1+\lambda_2)/\mu_2 \quad (4\text{-}10)$$

식 (4-10)을 식 (4-7), (4-8)에 대입하면 $P_{0,0,1}$, $P_{0,0,0}$를 ρ_1에 관한 관계식으로 표현할 수 있다.

$$P_{0,0,1}=-(1/\lambda_1+\mu_2/\lambda_2+\mu_2/\mu_1)\rho_1+1+\mu_2/\mu_1 \quad (4\text{-}11)$$

$$P_{0,0,0}=((\mu_1+\mu_2)/\lambda_1+\mu_2/\mu_1+\mu_1/\mu_2)\rho_1-(\lambda_1+\lambda_2)/\mu_2-\mu_2/\mu_1 \quad (4\text{-}12)$$

여기서, $\Pr$(서버 1만 *busy*), $\Pr$(서버 1,2 모두 *idle*)은 각각 $P_{0,0,1}$, $P_{0,0,0}$와 같고 이는 3장에서 구한 PGF 식들에서의 미지수들이므로, $P(0,w,2)$와 $P(z,w,3)$을 구할 수 있게 된다.

4.2 시스템 이용률

이 절에서는 3장에서 구한 대기고객수 PGF와 4.1 절에서 얻은 식들로부터 시스템 이용률 ρ_1과 ρ_2를 유도한다. 먼저 3장에서 구한 PGF $P(z,w,3)$의 식 (3-12)에 w 대신 1을 대입하고 z 대신 1을 대입하면 0/0형태가 되므로 로피탈의 정리를 이용하여 푼 후, 식 (4-11), (4-12)을 대입하여 ρ_1에 관하여 풀면 식 (4-13)을 얻을 수 있다.

$$\rho_1=(A_3/\mu_1-A_2)/(A_1+A_3\cdot(1+\lambda_1/\mu_1)/\lambda_1)$$

여기서, $A_1=-(1/\lambda_1-\mu_2/\lambda_1-\mu_2/\mu_1)\{(\lambda_1+\lambda_2)\lambda_1/\mu_1+\lambda_2-\mu_2\}+(\mu_1+\mu_2)\mu_2+\mu_2^2\lambda_1/\mu_1+\lambda_1\mu_1$,

$A_2=(1+\mu_2/\mu_1)\{(\lambda_1+\lambda_2)\lambda_1/\mu_1+\lambda_2-\mu_2\}-(\lambda_1+\lambda_2+\mu_2^2\lambda_1/\mu_1)$

$$A_3=\{\lambda_1\mu_2+(\lambda_2-\mu_2)(\mu_1+\mu_2)\}\cdot\lambda_1/\mu_1+(\mu_1+\mu_2-\lambda_1)(\lambda_2-\mu_2) \quad (4\text{-}13)$$

그리고, 식 (4-12)에 식 (4-13)을 대입하면 ρ_2를 구할 수 있다.

$$\rho_2=-(\mu_1/\mu_2)\cdot(A_3/\mu_1-A_2)/(A_1+A_3\cdot(1+\lambda_1/\mu_1)/\lambda_1)+(\lambda_1+\lambda_2)/\mu_2 \quad (4\text{-}14)$$

5. 평균대기고객수와 평균대기시간

5.1 class 1의 평균대기고객수와 평균대기시간

식 (3-12)를 이용하여 class 1의 평균대기고객수를 구하면 식 (5-1)과 같다.

$$L_1=\frac{d}{dz}P(z,1,3)|_{z=1}=-\frac{\lambda_1(\mu_1+\mu_2)\left[P_{0,0,1}\left\{\frac{(\lambda_1+\lambda_2)\lambda_1}{\mu_1}+\lambda_2-\mu_2\right\}+P_{0,0,0}\frac{\lambda_1\mu_2}{\mu_1}\right]}{\left[\frac{\{\lambda_1\mu_2+(\lambda_2-\mu_2)(\mu_1+\mu_2)\}\lambda_1}{(\mu_1+\mu_2-\lambda_1)\mu_1}+(\lambda_2-\mu_2)\right](\mu_1+\mu_2-\lambda_1)^2} \quad (5\text{-}1)$$

식 (5-1)에서 $P_{0,0,1}$, $P_{0,0,0}$ 대신 각각 식 (4-11), (4-12)을 대입하면 값을 얻을 수 있다.

class 1 평균대기시간은 Little의 법칙을 이용하여 식 (5-2)에 대입하면 구할 수 있다.

$$W_1 = L_1/\lambda_1 \tag{5-2}$$

5.2 class 2의 평균대기고객수와 평균대기시간

3장에서 유도한 대기고객수 PGF $P(0,w,2)$와 $P(z,w,3)$을 이용하면 식 (5-3)을 얻을 수 있다.

$$\frac{d}{dw}P(1,w,3)|_{w=1} = [(-4f^{(2)}(1)K^{(1)}(1) + f^{(1)}(1)\{3K^{(2)}(1) + 4K^{(1)}(1)(f^{(1)}(1)+1)\})C(1)$$
$$+ 10f^{(1)}(1)K^{(1)}(1)C^{(1)}(1)] \cdot (\mu_1+\mu_2)/(12\lambda_2[K^{(1)}(1)]^2)$$

여기서, $f^{(1)}(1) = \lambda_2/(\mu_1+\mu_2-\lambda_1)$, $f^{(2)}(1) = (2\lambda_2^2(\mu_1+\mu_2))/((\mu_1+\mu_2-\lambda_1)^3)$,

$C^{(1)}(1) = [(\lambda_1+\lambda_2)(\mu_2+\lambda_1-\lambda_2)/\mu_1+\lambda_2]P_{0,0,1}$, $K^{(1)}(1) = [-\mu_2+(\mu_1+\mu_2)f^{(1)}(1)]\frac{\lambda_1}{\mu_1}+\lambda_2-\mu_2$,

$K^{(2)}(1) = [2\mu_2+(\mu_1+\mu_2)\{f^{(2)}(1)-2(f^{(1)}(1))^2\}]\lambda_1/\mu_1 + 2[-\mu_1+f^{(1)}(1)(\mu_1+\mu_2)](\mu_2-\lambda_1)/\mu_1 + 2\mu_2$ (5-3)

식 (4-7), (4-8), (5-1), (5-3)을 식 (5-4)에 대입하면 class 2의 평균대기고객수를 얻을 수 있다.

$$L_2 = \Pr(\text{서버 2만 } busy) \cdot \frac{d}{dw}P(0,w,2)|_{w=1} + \Pr(\text{서버 1,2 모두 } busy) \cdot \frac{d}{dw}P(1,w,3)|_{w=1} \tag{5-4}$$

class 2의 평균대기시간은 식 (5-4)을 식 (5-5)에 대입하면 구할 수 있다.

$$W_2 = L_2/\lambda_2 \tag{5-5}$$

6. 결론

본 연구에서는 고객의 집단을 두 가지 class로 나누어 각 class 별로 사용할 수 있는 서버의 개수를 다르게 두는 M/M/2 대기행렬모형의 변형된 형태를 분석하였다. 일반적인 M/M/2 모형에서는 모든 고객이 균등하게 두 서버를 임의로 선택하여 서비스를 받을 수 있는 반면, 본 모형에서 한 고객 class는 두 서버 중 하나를 선택하여 서비스를 받을 수 있지만, 다른 한 고객 class는 하나의 서버에서만 서비스를 받도록 제한을 두었으며, 서비스 규칙으로 비축출형 우선순위를 적용하였다.

본 연구에서는 시스템의 전체 고객수를 고려하기보다는 대기열에 기다리는 고객수를 중심으로 안정상태에서의 각 성능치들을 분석하였다. 서버의 상태에 따른 대기고객수 PGF를 유도하였고, 대기고객수와 서버 이용률과의 관계를 이용하여 서버 이용률을 유도하였으며, 대기고객수 PGF로부터 각 class 별 평균대기고객수와 평균대기시간을 유도하였다.

7. 참고 문헌

[1] 이호우, 대기행렬이론, 개정판, 시그마프레스, 1998.

[2] Cooper, R. B., Introduction to Queueing Theory, 3rd ed., CEEPress Books, 1990.

[3] Kella, O. and Yechiali, U., "Waiting times in the non-preemptive priority M/M/c queue," Commun. Statist.-Stochastic Model 1(2), pp. 252-262, 1985.

[4] Wagner, D., "Waiting times of a finite-capacity multi-server model with non-preemptive priorities," European Journal of Operational Research 102, pp. 227-241, 1997.