

# 고객이 다수의 서버를 원하는 대기행렬의 수행척도에 관한 연구

## (A Study on the Measures of Performance of M/M/s Queues Where Customers Demanding Multiple Channel Use)

김 성 식<sup>\*\*\*</sup>

### Abstract

Applying the matrix method for the two dimensional Markovian queues, this paper offers a procedure to calculate the exact values of measures of performance of M/M/s queues where customers demand multiple number of servers. A method of obtaining steady state probabilities is illustrated as well. An example problem is also presented.

### 서 론

대기행렬 시스템의 한 특수한 형태로 고객들이 서로 다른수의 서버를 요구하는 시스템을 들수있다. 이 경우 시스템은 한명의 고객을 처리하기 위하여서는 그가 요구하는 수만큼의 서버를 동시에 제공 하여야만 한다. 따라서 원하는 수만큼의 서버가 동시에 놓고있지 않는한 비록 놓고있는 서버가 있다고 하더라도 그고객은 대기열에서 기다리게 된다. Green [4]은 이러한 종류의 시스템을 서버들의 독립된 정도에 따라 세가지로 분류 하였는데 독립서버 모델 (models with independent servers), 동시서버 서비스 모델 (joint service model) 그리고 불변서버서비스 모델 (constant service rate model) 그것이다. 본 연구는 이 세가지 모델중 동시서버 서비스 모델을 택하여 이의 수행척도를 구하는 계산절차를 제시하는데 목적을 두고 있다. 동시서버 서비스 모델이란 한 특정 고객에게 할당된 서버들은 부여된 작업을 동시에 같이 수행 하여 끝내는 모델을 말한다. 이러한 타입의 대기행렬은 여러분야에서 관찰 될수있다. 우선 서로다른 수의 기억용량을 요구하는 작업들이 도착하고 이들은 각각 자기가 필요한 기억용량이 할당될 때까지 buffer에서 기다리게 되는 컴퓨터가 그 예가 될수 있으며 병원에서 의 수술시 여러명의 의사가 동시에 필요한 경우 또는 통신센터에서 정보가 여러가지 형태 (통화는 1개 회선 데이터 또는 영상은 여러회선)로 들어오는 경우도

그예로 볼 수 있다. 본 대기행렬 현상이 이렇게 실생활에서 자주 일어남에도 불구하고 이현상에 대한 대기행렬이론의 관점에서의 연구는 매우 드물다.

대기행렬이론이 자주 응용되는 분야는 컴퓨터시스템에 관련된 대기현상 분야이다. 이러한 이유로 multi-resource 라는 관점에서의 대기현상연구 특히 대기행렬 네트워크를 해석한 연구는 상당히 많다 (Kleinrock [7]). 그러나 대부분의 연구들에서는 고객타입의 분류는 이들의 처리시간의 분포에 의하여 이루어질뿐 그들이 요구하는 resource 의 양에 의하여 분류되지 않는다. 이들중 예외는 Kenneth [5]의 연구에서 볼 수 있는데 이 논문에서는 고객이 각기 다른양의 resource 를 요구할 경우에 발생하는 대기열의 안정조건 (equilibrium condition)을 선형계획법에 의하여 구하고 있다. 컴퓨터응용을 떠나서 본 연구와 비슷한 주제를 다루고있는 논문은 Green [3]의 독립서버 모델에 관한 연구이다. 이 연구에서는 포아손도착, 지수분포 서비스 모델을 다루었고 본 논문의 모델과 다른점은 한 고객에 할당된 서버가 각자 자기자신의 서버를 끝내면 다른 서버의 상태에 관계없이 즉시 놓고있는 상태로 환원되는 점이다. Green은 이모델을 M/G, 1 모델로 변환시켜 결과를 얻고있다. 이외에 동시서버 서비스모델의 서버서비스속도를 조사하여 (김성식 [1]) 이 모델의 상략확률의 계산방법 (approximation) [2] 을 연구한것도 같은 부류에 속한다.

이와같이 본 시스템을 조사한 연구의 수가 적은 이

\* 고려대학교

\*\* 본 연구는 문교부 지원에 의하여 이루어졌음.

유는 모델의 복잡성에 있다고 볼수 있다. 대기행렬을 연구할 때는 먼저 상태공간(state space)의 정의에서부터 시작한다. 고객이 여러명의 서버를 요구할 경우 상태를 정의하기는 상당히 복잡하다. 상태를 정의할때 시스템내의 이들 타입의 조합은 물론 누가 서비스를 받고 있는가 하는 점이 반영 되어야 한다. 더구나 상태공간이 정의 되었다 하더라도 이들간의 관계를 나타내기 위하여는 많은수의 식과 변수가 필요하게 되고 이것들이 수리적인 해법을 발견하는데 많은 복잡함과 어려움을 주고있다. 우리는 본 연구의 대상 시스템이 이러한 문제를 갖고 있기 때문에 아직 까지 이에 대한 연구가 그리 많지 않았다고 믿고있다.

이론적으로 한시스템의 도착과 서버서비스 형태가 마코비안 일경우 일상적인 안정상태 방정식을 쓸수있고 이를 풀 수가 있다. 그러나 일반적으로 다변수로 표시되는 상태를 갖인 복잡한 시스템에서는 실제 문제 해결시 이들식을 풀기는 매우 어렵고 불가능한 경우가 많다. 고객이 다수의 서버를 요구하는 시스템도 그 중 한가지 경우이다. 따라서 상태방정식을 형성하고 이를 푸는 어려움을 극복하기 위하여 다른 접근방법이 필요하게된다. 우리는 이 논문에서 2차원 마코비안 대기행렬의 행렬에 의한 해법(Kim [6])을 문제의 해결방안으로 사용하고자 한다. 이 방법에서는 연관된 상태들이 벡터로 정리되어 고려해야할 상태의 수가 줄어들게 된다. 또한 상태간의 관계식을 덩어리로 분리된(block partitioned) 확률행렬로 표시하게 되며 이때 이행률의 원소들은 대기행렬과정의 transition 확률을 표시하게 된다. 이러한 행렬식의 일반적인 여러성질은 Neuts [9]와 이에따른 연구 [8], [10]들에서 자세히 관찰 될수 있다.

## 모델형성 및 상태의 정의

본연구에 다룬 대기행렬 시스템은  $s$  ( $s \geq 1$ ) 명의 서버가 있고 도착은 평균치가  $\lambda$  인 포아손분포에 의하여 일어나며 도착하는 고객의 종류는  $K$  ( $K \geq 1$ ) 가지 타입이며 타입은 각각자기가 원하는 서버의 수  $d_k$  ( $k=1, \dots, K$ ),  $d_1 < d_2 < \dots < d_K$  의하여 정하여 진다. 서버는 대기행렬에 제일앞에 있는 고객이 원하는수 이상의 서버가 놓기되면 시작되고 서버서비스 시간의 분포는 모든타입의 고객에 대하여 평균치가  $1/\mu$  인 지수분포를 하며 FIFO로 서버서비스가 이루어진다. 대기열의 길이에 제한이 없다고 보았다. 또한 타입 $k$  고객의 도착율은  $\lambda_k$  ( $k=1, \dots, K$ ),  $\sum \lambda_k = \lambda$ , 고객타

입이  $k$ 일확률을  $P_k$ 라 하고  $g$ 를 서버서비스를 받고있는 고객수라고 하자. 즉  $g = \max_i i, i \leq n(d_1 + \dots + d_{c_i} \leq s$  여기서  $n$ 은 시스템에 있는 고객수,  $c_i$ 는  $n$  고객중  $i$ 번째로 시스템에 도착한 고객의 타입을 표시한다.

이제 시스템의 상태를 완전히 정의하기 위해서는 시스템이 비어있는 상태(0)으로 표시하고  $n$  ( $n \geq 1$ ) 명이 있을 경우  $(n, c_1, c_2, \dots, c_n)$ 으로 표시할 수 있고 여기서 각 벡터는 시스템내의 고객수  $n$ 과 그들의 타입을 표시한다. 그러나 서버의 관점에서 볼 때 각 고객의 타입이 시스템에 도착할때 정해지던가 대기열의 제일앞에 도달하는 순간 정하여 지던가 상관없다. 따라서 대기열의 선두고객을 제외한 나머지 고객의 타입이  $k$ 일 확률은 단순히  $P_k$ 로 쓸수있고 따라서  $n \geq g+2$  일 경우

$P_r[\text{시스템상태}(n, c_1, c_2, \dots, c_n)] = P_r[\text{상태}(n, c_1, c_2, \dots, c_{g+1}) | P_{c_{g+1}}, \dots, P_{c_n}]$  이 된다. 따라서 확률이라는 관점에서 시스템의 상태는 시스템내의 고객의 수와 먼저 도착한  $g+1$  명의 고객의 타입을 나타내는 벡터들로 완전히 정해진다.  $g+1$  이 갖일수 있는 제일큰수는  $\lfloor s/d_1 \rfloor + 1$  이되며 여기서  $\lfloor A \rfloor$ 는  $A$  보다 작은 정수(integer) 중 제일 큰 수를 말한다. 그러나 서버들이  $\lfloor s/d_1 \rfloor$  명의 고객을 서버서비스 하고있는 경우 타입에 관계없이 대기열 선두의 고객은 서버서비스 받을 수 없게된다 따라서 대기열 선두의 고객이 타입  $k$ 일 확률은  $P_k$  ( $k=1, \dots, K$ )로 주어지게 되고 이제  $U = \lfloor s/d_1 \rfloor$ ,  $f = \min(n, U)$ 라고 정의하면 안전상태의 확률(steady-state probability)이 존재할 경우 이들은 상태들(0),  $(n, c_1, \dots, c_f)$ 의 확률이 정해지면 모두 정해질 수 있게된다. 이제 이러한 확률들을 구하기 위하여 우리는 2차원 대기행렬의 행렬에 의한 해법이 사용될 수 있도록 먼저 상태들을 그 체제에 맞도록 정리하여야 한다. 시스템에  $n$ 명의 고객이 있을경우 먼저 도착한  $f$ 명의 고객이 갖일 수 있는 방법의수는  $K^n$ 이다.

이제  $J_m$  ( $m=1, \dots, K^f$ )이 그러한 타입을 나타낸다고 하고 다음과 같은 형태로 정리 되어있다. 하자.

$$J_1 = (1, 1, \dots, 1)$$

$$J_2 = (1, 1, \dots, 2)$$

⋮

$$J_{K^f} = (K, K, \dots, K)$$

여기서 벡터  $J_m$ 들은  $m$ 이 1씩 증가 할때마다 제일 뒤의 요소는 1씩 증가하고 뒤에서 부터  $i$ 번째 요소는  $K^{i-1}$  번 씩 같은 값을 갖은후 1씩증가한다. 또 모든 요소들에서  $K$ 다음 갖은 값은 1이다. 이제  $J_m$ 을 시스템에  $n$ 명이 있을때의  $m$ 번째 고객타입의 조합이라하자



$$V^{n+1} = V_0 \cdot V_1 \cdot \dots \cdot V_{n+1}$$

이라 하자.

또  $Y_0 = 1/\lambda$ 이고  $Y_i$  ( $i=1, \dots, U$ )을  $(j, j)$  번째 요소가 상태  $J_m^n$ 에서의 평균 머무르는 시간(mean sojourn time)인  $K^m \times K^m$  대각선 행렬이라 하면

$$Q(i) = \pi(i) \cdot Y_i \cdot \sum_{n=0}^{\infty} \pi(n) \cdot Y_n e$$

로 주어지게 되고

$$Q(n) = \lambda Q(0) V(n) Y_n \quad 1 \leq n \leq U$$

$$Q(n) = \lambda Q(0) V(U) R^{n-U} \cdot Y_U \quad n > U$$

또

$$Q(0) = [1 + \lambda \sum_{n=0}^{\infty} V(n) Y_n e + \lambda V(U) \times (I - R)^{-1} \cdot Y_U e]^{-1}$$

으로 임의시간에서의 상태 확률을 구할 수 있다.

### 수행척도 계산 및 예

앞에서 구한  $q(J_m^n)$ 로 이제 우리는 안정상태 하에서의 시스템의 수행척도를 구할 수 있다. 먼저  $J_m^n(k)$ 를 시스템이 상태  $(n, J_m^n)$ 에 있을 때 앞의  $f$ 명의 고객에 포함된 타입  $k$ 의 고객수라 하고  $J_m^n$ 를 그 상태에서 실제로 서어비스 받고 있는 고객수라고 하자. 그러면

$$\sum_k J_m^n(k) = I_m^n, \quad \sum_k J_m^n(k) = f$$

가 됨을 쉽게 알 수 있고, 또

$$D_m^n = [J_m^n(1), J_m^n(2), \dots, J_m^n(K)]$$

$$\hat{D}_m^n = [J_m^n(1), J_m^n(2), \dots, J_m^n(K)]$$

$$m=1, 2, \dots, K^f, \quad n \leq 1$$

이라 하면  $D_m^n$ 과  $\hat{D}_m^n$ 은 각각 그들의 원소가 첫번  $f$  고객 중에 포함된 타입  $k$ 의 고객 수, 서어비스 받고 있는 고객 중에 포함된 타입  $k$ 의 고객수를 나타내는 벡터이다.

이제  $L_k$ 를 시스템에 있는  $k$ 타입 고객 수의 기대치라 하면,

$$L_k = \sum_{n=0}^{\infty} \sum_{m=1}^{K^f} q(J_m^n) \cdot J_m^n(k) + \sum_{n=U+1}^{\infty} (n-U) \cdot Q(n) P_k$$

가 되며 여기서 우변의 두번째 항은 타입이 정의 안된 뒤의 고객중 타입  $k$ 의 고객 수의 기대치이다.

이제  $L = [L_1, L_2, \dots, L_K]$ 라 하고  $P = [P_1, \dots, P_K]$ 라 하면,

$L = \sum_{n=0}^{\infty} Q(n) \cdot D(n) + \sum_{n=U+1}^{\infty} (n-U) Q(n) e \cdot p$ 로 주어지며, 여기서  $D(n)$ 은  $D(n) = [D_1^n, D_2^n, \dots, D_{K^f}^n]$ 인  $K^f \times K$  행렬이다.

이제  $L$ 을 다시  $n \leq U$ 인 경우와  $n > U$ 인 경우로 분리하여 정리하면,  $D_m^n = D_m^U, \quad n > U$ 이므로

$$L = \sum_{n=0}^U Q(n) D(n) + \lambda Q(0) V(U) \times \sum_{n=U}^{\infty} R^{n-U} Y_U D(U) + [\lambda Q(0) \times \sum_{n=U+1}^{\infty} (n-U) V(U) R^{n-U} \cdot Y_U e] \cdot P$$

$$= [\lambda Q(0) V^U] R(1-R)^{-1} Y_U e] P$$

로 주어지며 각타입의 고객수의 기대치가 벡터형태로 구해진다. 따라서 시스템에 있는 전체고객수의 기대치는  $TL = Le$ 로 주어진다. 이제 대기열에서 기다리는 고객수를 구하기 위하여  $z_k$ 를 서어비스 받고 있는 타입  $k$ 고객수의 기대치라 하면

$$z_k = \sum_{n=0}^{\infty} q(J_m^n) J_m^n(k) \text{ 이 되고}$$

$Z = (z_1, z_2, \dots, z_K), \quad \hat{D}(n) = [\hat{D}_1^n, \dots, \hat{D}_{K^f}^n]$ 인  $K^f \times K$  행렬이라 하면

$$Z = \sum_{n=0}^{\infty} Q(n) \hat{D}(n)$$

$$= \sum_{n=0}^{\infty} Q(n)$$

$$= \sum_{n=0}^{U-1} Q(n) \hat{D}(n) + \lambda Q(0) V^U (I-R)^{-1}$$

$Y_U \hat{D}(U)$ 로 계산 되어 대기열에서 기다리는 타입  $k$ 인 고객수의 기대치는

$$\hat{L} = (\hat{L}_1, \hat{L}_2, \dots, \hat{L}_K)$$

$$= L - Z \text{ 로 구하여진다.}$$

또한  $z_1 + z_2 + \dots + z_K = Ze$ 는 서어비스를 받고 있는 고객수의 기대치  $Le$ 는 대기열에서 기다리는 전체 고객수가 된다.

또한 Little's의 법칙에 의하여 고객이 시스템에서의 대기시간은

$$W = \lambda \cdot TL$$

대기열에서의 대기시간은

$$Wq = \lambda \hat{L} e$$

로 주어지나 각각 타입의 고객의 개별적인 대기시간 현재의 상태에서 구해지지는 않으면 앞으로 더욱 연구하여야할 문제이다.

이제 계산예를 보이기 위하여  $S=7, K=3,$

$$(d_1, d_2, d_3) = (2, 3, 4), \quad (P_1, P_2, P_3) = (0.3,$$

$0.4, 0.3] \quad \lambda=1$  그리고  $\mu=1$ 인 시스템을 생각해 보자. 이모델에서  $U=[7/2]=3, K^U=3^3=27$ 이 되고 상태들과 불변확률벡터  $d$ 의 값은

$$J_m^1: (1), (2), (3)$$

$$J_m^2: (1, 1), (1, 2), (1, 3), (2, 1), (2, 2), (2, 3), (3, 1), (3, 2), (3, 3)$$

$$J_m^3: (1, 1, 1), (1, 1, 2), (1, 1, 3), (1, 2, 1), (1, 2, 2), (1, 2, 3), (1, 3, 1), (1, 3, 2), (1, 3, 3), (2, 1, 1), (2, 1, 2), (2, 1, 3), (2, 2, 1), (2, 2, 2), (2, 2, 3), (2, 3, 1), (2, 3, 2), (2, 3, 3), (3, 1, 1), (3, 1, 2), (3, 1, 3), (3, 2, 1), (3, 2, 2), (3, 2, 3), (3, 3, 1), (3, 3, 2), (3, 3, 3)$$

$$d = (.0236, .0315, .086, .0315, .0473, .0355, .020266, .0355, .0266, .0315, .0473, .035, .0473, .063, .0631, .0473, .0355, .0473, .0355, .0266, .0355, .0266, .0355, .0473, .0473, .0355, .0355, .0473, .0355]$$

이된다. 이제 필요한 계산을 계속하면

$$\rho = .5225 \text{ (indicates stationarity)}$$

$$V^{(1)} = [.6110, .8097, .5793]$$

$$V^{(2)} = [.1429, .1874, .1388, .1868, .248, .1844, .1342, .1790, .1342]$$

$$V^{(3)} = [.0211, .0282, .0211, .0282, .0376, .0282, .0211, .0281, .0211, .0283, .0377, .0283, .0376, .0502, .0376, .0282, .0376, .0282, .0216, .288, .0216, .0278, .0283, .0288, .0283, .0377, .0283]$$

$$Q(1) = .317$$

$$Q(1) = [.9693, .1284, .0918]$$

$$Q(2) = [.01511, .01982, .01468, .01975, .02600, .01950, .01420, .01890, .02130]$$

$$Q(3) = [168, 224, 224, 224, 397, 298, 223, 297, 223, 225, 399, 299, 398, 531, 398, 299, 398, 299, 398, 299, 225, 304, 298, 304, 406, 304, 449, 592, 449]10^{-5}$$

그리고  $Q(n)$ 's ( $= Q(n)e$ ) 들은

$$Q(0) = .317$$

$$Q(1) = .317$$

$$Q(2) = .169$$

그리고

$$Q(3) = .088$$

$$Q(4) = .048$$

$$Q(5) = .027$$

$$Q(6) = .015$$

$$Q(7) = .0087$$

$$Q(8) = .0051$$

또한

$$L = [.627, .858, .728]$$

$$TL = 2.213$$

$$Z = [.415, .576, .516]$$

$$\hat{L} = [.212, .282, .212]$$

$$\hat{L}e = .706$$

이구해진다.

결론

우리는 이 논문에서  $M/M/s$  시스템에서 고객이 다수의 불특정수의 서버를 원하는 경우 수행책도의 계산절차를 소개 하였다. 여기서 문제점은 상태를 정의

하는것이 실제의 경우 복잡하고 그 수가 매우 많다는 사실이다. 상태의 수를 줄이는 방법에 대한 연구가 더욱 요망되며 아울러 우선순위가 부여된 경우의 연구도 뒤따라야 할것이다. 특히 이 모델이 컴퓨터 시스템을 연구하는데 도움을 준다는 관점에서 더욱 그러하다.

### (참고 문헌)

1. 김성식, 장진익., "Multiserver Demanding M/M/s 대기행렬의 Service Rate 변화곡선에 관한 연구" 대한 산업공학회지, Vol. 6, No. 1, 1980.
2. 김성식., "고객이 다수의 서버를 원하는 M/M/s 시스템의 계산방법" 대한산업공학회지, Vol. 7, No. 2, 1971.
3. Green, L., "A Queuing System in Which Customers Require Random Number of Servers.," Open. Research, Vol. 28, No. 6, 1980.
4. ———, "Queues in Which Customers Require a Random Number of Servers.," Manag. Science, Vol. 27, No. 1, 1981.
5. Kenneth, J. O., "Capacity Bound for Multiresource Queues.," J. of A. C. M., Vol. 24, No. 4, 1877.
6. Kim, S. S., "A Matrix Method for the Analysis of Two Dimensional Markovian Queues.," J. K. I. I. E., Vol. 8, No. 2, 1982.
7. Kleinrock, L., Queuing System Vol. 2: Computer Applications, John Wiley and Sons, 1976.
8. Neuts, M. F., "The Single Server Queue with Poisson Input and Semi-Markov Service Times.," J. Appl. Prob., Vol. 3, 1966
9. ———, "The Markov Renewal Branching Processes," Proc. Conf. on Math. Methods in the Theory of Queues, Kalamazoo, M. I., Springer Verlag, N. Y., 1974.
10. ———, "Markov Chain with Application in Queuing Theory, Which Have A Matrix-Geometric Invariant Probability Vector.," Adv. Appl. Prob., Vol. 10, 1978.