대기행렬이론과 그 응용(1)

1. 서론

대기행렬이론(queueing theory)은 줄 또는 열에 서서 기다리는 것에 대한 연구로서 서비스를 받으려는 고객들의 불규칙한 도착으로 인해 발생하는 대기행렬(waiting line or queue)로 특징 지워진다. 고객의 불규칙한 도착과 서비스 시간의 불균일로 인하여 기다리는 상태를 초래하게 되고, 이러한 대기상태를 개선하기 위한 연구가 대기행렬 이론이다. 대기행렬 이론은 A. K. Erlang 의 전화시설에 대한 연구에서 시작하였으며, 그 후 1953년에 D. G. Kendall 이 다수 서비스시설에 대한 연구와 공통기호(notation)를 소개하였다. 이 기호는 1966년 A. M. Lee의 보충으로 현재 사용되고 있다. 지금도 대기행렬이론에 대한 연구는 계속되고 있다.

서비스할 수 있는 능력보다 서비스 수요가 많으면 열을 지어서 기다리는 것이 일반적인 현상이다. 흔히 산업분야에서 서비스시설의 수자에 대하여 의사 결정을 내려야 하는 경우가 종종 있으나 언제 고객이 도착하며, 얼마만한 시간을 요할 것인지를 정확하게 예측할 수 있는 경우가 흔치 않으므로 이러한 의사결정은 매우 어려운 경우가 많다. 너무 많은 서비스를 제공한다면 비용이 많이 들것이고 또 반대로 충분한 서비스 능력을 마련해 두지 않으면 때때로 기다리는 행렬이 지나치게 길어질 것이다. 서비스시설이 많아서 고용인을 놀리거나, 서비스시설이 적어서 고객을 지나치게 기다리게 하는 것은 사회적인 낭비이다. 그러므로 궁극적인 목표는 서비스에 대한 고용주의 비용과 서비스를 받기 위해 줄을 서서 기다리는 고객의 낭비를 경제적인 면에서 균형을 취하는데 있다. 대기행렬이론 그 자체가 직접 이러한 문제를 해결해 주지는 않으나 대기하는 행렬의 여러 가지 성질, 예를 들면 평균대기시간을 예측해 줌으로써 의사 결정에 커다란 공헌을 하고 있다.

본고에서는 대기행렬이론에 대하여 일반적인 언급을 한 후, 대기하는 행렬의 상태를 기술하는 몇 가지 기초적인 수학적 모델과 대기행렬의 성질을 예측해 주는 기본적 결과를 다룬 다음, 그 적용실례를 들고자 한다.

2. 대기행렬의 분석법

대기행렬현상은 적어도 3 가지의 각각 다른 방법으로 분석할 수 있다. 우선 대기행렬 실제 시스템에서 물리적으로 실험하여 그 결과를 관찰할 수 있다. 예를 들자면, 한 관리자가 제직공정에서 틀잡이 1 인당 적정담당틀수를 결정하고자 한다면 여러 대체안(alternatives)을 가지고 직접 현장에서 실험하여, 직기의 평균정대시간과 생산량 증감으로 인한 이익(또는 손해) 및 틀잡이 고용비용을 비교, 분석하여 최선의 대안을 결정하는 것이다. 이러한 의사결정방법은 물론 비용이 많이 들지만 아직 대기행렬모델을 분석하는데 가장 보편적으로 이용되는 접근법이다.

시뮬레이션(Simulation)은 대개행렬문제를 분석하는 두 번째 방법이다. 이 기법은 도착과 서비스 양상, 비용, 고객행동 등에 대한 정보를 수집할 것을 필요로 한다. 실제로 시스템은 구성하고 있는 기본구성 요소들로 하나의 모델을 설계하고, 이 모델을 Monte Carlo technique 등을 사용하여 분석 한다.

분석수단으로는 컴퓨터나 종이, 연필이 사용된다. 시뮬레이션은 현실적 시스템에 관해 실험에 의한 예측을 해 줌으로써 문제 해결에 중요한 진전을 가져올 수 있어서 특히 문제의 구조가 복잡해서 수식으로 나타내기 곤란한 경우에 빈번히 이용되고 있다.

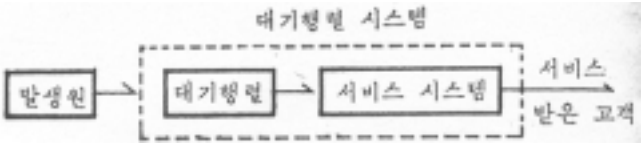
끝으로 대기행렬문제는 확률이론의 활용을 통해서 직접 수학적 해결을 찾을 수 있다. 이 수학적 분석법이 바로 대기행렬이론으로서 본고에서는 이 수학적 분석법에 대하여만 논하고자 한다.

3. 대기행렬 모형의 구조

3-1. 기본적 대기행렬 과정

대부분의 대기행렬 모형에서 가정되는 기본적인 과정은 다음과 같다. 서비스를 필요로 하는 고객들이 발생원(Calling Source)에 의해서 항상 발생된다. 이러한 고객들은 대기행렬 체계(시스템)에 들어와서는 줄을 서서 기다리게 된다. 적당한 시간에 서비스규칙(service discipline)이라는 어떤 규칙에 의하여 한 사람씩 서비스를 받게 된다.

그러면 서비스 시스템에 의해 필요한 서비스를 받은 고객은 대기행렬 시스템을 떠나게 된다. 이러한 진행과정이 그림 1에 표시되어 있다.



<그림 1> 기본적인 대기행렬 시스템

대기행렬 과정의 여러 요소에 대해서는 일반적으로 여러 가지의 가정이 있다. 이들을 자세히 설명하면 다음과 같다.

1) 발생원

발생원(calling source) 또는 모집단(calling population)의 한가지 특성은 그 크기이다. 이 크기란 때때로 찾아올 수 있는 고객의 수자, 즉 지금은 오지 않지만 미래의 언제가에는 올 수 있는 고객의 총 수자로 나타내지며, 무한일 수도 있고 유한일 수도 있다. 크기를 무한대로 가정하면 계산이 훨씬 쉽기 때문에 실제 크기가 비교적 큰 유한수자일 때는 무한대인 경우로 가정하는 경우가 많다. 유한 크기의 경우는 항상 대기행렬 시스템 안에 있는 고객의 수가 잠재 고객의 수에 영향을 미치기 때문에 분석하기가 어려워진다. 만약 발생원이 새 고객을 창출하는 속도가 대기행렬체계 안에 있는 고객의 수에 의해 영향을 받는다면 유한 크기를 가정해야만 한다.

시간적으로 고객이 창출되는 통계적 유형도 구체화해야 한다. 고객이 도착하는 것은 대개 포아송(Poisson) 분포를 따른다고 가정된다. 이 가정은 고객이 대개행렬 체계에 도달하는 것이 확률적이기는 하지만 어떤 평균적인 비율로 도착한다는 것이다. 이 가정은 도착사이의 시간간격이 지수(exponential) 분포를 따른다는 가정과 같다. 도착사이의 시간간격을 도착간격시간(interarrival time)이라고 한다.

고객들의 행동에 대한 일반적인 아닌 가정 또한 모두 규정해야 한다. 이러한 한 가지 예가 도착하고도 대기행렬체계에 들어오지 않는 경우인데 이것은 대기행렬이 너무 길기 때문에 고객이 들어가기 싫어하기 때문이다. 이 때는 고객이 도착하지 않은 것으로 간주한다.

2) 대기행렬(Queue)

대기행렬은 시스템에 허용 가능한 최대의 고객수로 특징지울 수 있다. 대기행렬은 이 수자가 무한이냐 유한이냐에 따라서 대기행렬이 무한 또는 유한이라고 한다.

3) 서비스 규칙(Service discipline)

서비스규칙은 대기행렬 중에서 서비스 해주기 위해 고객을 선택하는 순서를 말한다.

예를 들면 선착순(FCFS)으로 할 수도 있고, 임의로(random) 할 수도 있고, 어떤 우선순위(priority)에 의해 할 수도 있다. 별도로 명시하지 않는 한 대기행렬모형에서는 대개 선착순을 가정한다.

4) 서비스 기구(service mechanism)

서비스 기구에는 하나 또는 여러개의 서비스 설비(service facilities)가 연속(series)적으로 배치되어 있고, 각각의 서비스 설비에는 한 개 또는 여러

개의 병렬 서비스 기구가 있는데 이를 창구라고 한다. 만일 두 개 이상의 서비스 설비가 있으면 고객은 연속적으로 이 서비스 설비들의 서비스를 받을 것이다. (직렬 서비스 창구). 하나의 설비에서 고객은 병렬 서비스 창구 중의 하나에 들어서며, 이 창구에서 서비스를 완전히 받는다. 대기행렬 모형에서는 설비들의 배열을 명시해야 하고 각각의 설비에 포함된 창구의 수도 반드시 명시해야 한다. 가장 기초적인 모형은 한 개 혹은 한정된 몇 개의 창구를 가지는 하나의 설비를 가정하는 경우이다.

서비스가 시작되고 나서부터 끝날 때까지 경과한 시간을 서비스 시간 또는 점유시간(holding time)이라고 한다. 대기행렬 모형의 각각의 창구에 대하여 서비스 시간의 확률 분포를 명시하여야 하며(아마도 고객 의형에 따라서도 각각 명시해야 할 것이다), 개개의 창구가 모두 똑같은 분포를 갖는다고 가정하는 것이 통례이다. 흔히 Erlang분포나 그 특수 경우인 지수(exponential) 분포를 서비스 시간으로 택한다.

3-2. 대기행렬의 기본요소

재기행렬모형을 결정하는 기본요소는 다음의 6가지로 요약할 수 있다.

(1) 투입요소, 즉 도착시간의 확률분포

(2) 산출요소, 즉 서비스시간의 확률분포

(3) 서비스 창구의 수

(4) 시스템에 허용하는 최대고객 수

(5) 투입요소의 발생원

Kendall-Lee의 기호를 사용하여 표시할 경우 앞의 6가지 각 요소를 일반적으로 표시하면,

(1 /2 /3) : (4 /5 /6)

이며, 위의 (1)과 (2)에 해당하는 것으로서 자주 사용되는 확률분포 기호는 다음과 같다.

M=포아송 (Poisson)분포

D=도착이나 서비스 사이의 시간간격이 일정한 경우

Ek=얼랑(Erlang) 분포나 감마(Gamma)분포

G=일반적인 분포, 즉 모든 분포에 적용되는 경우.

서비스 규칙인 (4)에 대한 기호는 다음과 같다.

FCFS=first come, first served ; 먼저 온 순으로 서비스를 하는 경우

LCFS=last come, first served ; 나중 온 순으로 서비스를 하는 경우

SIRO=service in random order ; 임의대로 서비스를 하는 경우

GD=general service discipline ; 일반작인 규칙, 즉 모든 규칙을 적용할 수 있는 경우

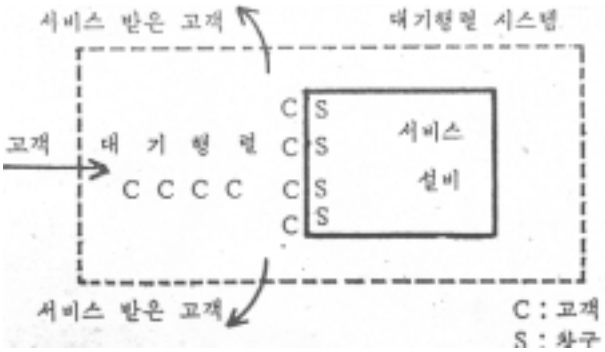
NPRP=nonpreemptive ; 서비스에 우선순위가 있는 경우 중의 한 경우

이상의 6가지 요소들을 이용하여 대기행렬모형의 특수한 상황을 규정지울 수 있다.

3-3. 일반적인 대기행렬 과정

앞에서 언급한 바와 같이 대기행렬이론은 여러 종류의 대기행렬 상태에 대하여 연구한다. 그러나 주로 다음과 같은 상황, 즉 단일 서비스 설비 앞에 단일 대개행렬(때로는 아무도 대기하지 않을 수도 있음)이 있고, 여기에 하나 또는 그 이상의 창구가 설치되어 있는 그러한 생활에 연구가 집중되어 왔다.

발생원에 의해 창출된 각각의 고객은 대기행렬에 서서 약간 기다리다가 창구 중의 하나에서 서비스를 받고 대기행렬 시스템을 떠난다. 이러한 대기행렬 시스템이 그림 2에 나와 있다.



<그림 2> 기본 대기행렬 시스템

창구가 단일한 사람으로 되어 있어야 할 필요는 없으며 한 집단의 사람일 수도 있다. 예를 들면 고객이 요구하는 일을 동시에 하기 위하여 힘을 합해야 하는 수리반이 있다. 창구는 꼭 사람이 아니라도 좋다. 여러 경우에 있어 창구는 필요할 때 불러 쓰는 지게차와 같이 기계나 장비일 수도 있다. 마찬가지로 대기행렬 중의 고객 또한 사람이어야 할 필요는 없다. 예를 들어 정방, 권사, 제직공정에서의 사절일 수도 있고, 조립하기 위해 기다리는 부품일 수도 있고, 유료 고속도로에서 요금지불을 위해 기다리는 자동차일 수도 있다. 서비스 설비 앞에서 실제로 꼭 대기행렬을 지어 서 있을 필요도 없다. 말하자면 사절로 인하여 정대하고 있는 직기와 같이 서비스를 기다리는 고객(직기)은 일정한 장소에 위치해 있고, 창구(틀잡이)가 찾아가서 서비스를 하는 경우도 있다.

주어진 지역 배당된 한 사람의 창구 또는 한 그룹의 창구는 그 지역의 서비스 설비를 구성하게 된다. 그리고 대기행렬이론에 의하면 그룹으로 기다리거나 아니거나 간에 차이가 없으므로 평균 대기자의 수, 평균 대기시간 등의 자료를 얻을 수 있다. 대기행렬이론이 응용되기 위한 필수요구조건은 주어진 서비스를 기다리는 고객의 수자 변화가 그림 2(또는 적절한 다른 패턴)에서와 같이 되어야 한다는 것이다.

3-4. 용어와 기호 (Terminology and Notation)

달리 명시하지 않는 한 다음의 표준 용어와 기호를 사용한다.

시스템의 상태 = 대기행렬시스템 내의 고객의 수

대기행렬의 길이 = 서비스를 기다리는 고객의 수

= 시스템의 상태-서비스 받는 고객의 수

N(t) = 시간 t 의 대기시스템의 고객 수 (t 0)

Pn(t) = 시간 0 의 고객수가 주어졌을 때 시간 t에 대기시스템 내에 있는 고객의 수가 n일 확률

s=창구의 수

λn=시스템 내에 n사람의 고객이 있을 때 새로운 고객의 평균 도착속도

(단위시간당 도착하는 고객의 기대 값)

μn=시스템 내에 n사람의 고객이 있을 때 전체 시스템의 평균 서비스속도

(단위 시간당 서비스 받는 고객수의 기대 값)

λn이 모든 n에 대해 상수일 때 이 상수는 λ로 표시한다. 서비스하고 있는 창구의 평균 서비스속도가 n 1에 대하여 상수이면 이 상수는 μ로 표시한다.

(이 경우 n s이고 모든 창구가 서비스 중이면 μn=sμ이다). 이러한 상태에서는 1 / λ는 기대 도착시간 간격이라 하고, 1 / μ를 기대 서비스시간이라고 한다. 또 λn/μn 은 서비스 설비의 이용도(utilization factor), 즉 도착하는 고객(λ)과 시스템 서비스 능력(sμ)과의 비율, 말하자면 창구 활동시간의 기대비율이다.

또한 안정상태의 결과를 기술하기 위하여 몇 개의 기초가 필요하다. 대기시스템이 일을 시작한지 얼마되지 않았을 때 그 시스템의 상태(시스템 내의 고객 수)는 처음 시작했을 때의 상태나 경과한 시간에 크게 영향을 받는다. 이런 경우의 시스템은 일시상태(transient state)라고 부른다. 그러나 충분한 시간이 경과하면 시스템의 상태는 필연적으로 초기 상태나 경과한 시간과는 무관하게 된다(비정상 상황인 λn/μn 이 1보다 큰 경우는 예외).

따라서 상태는 이제 안정상태(steady state)에 도달하게 된다. 일시 상태의 경우 수학적인 계산에서 약간 어렵다는 것이 부분적인 이유이기는 하지만 대기행렬이론은 주로 안정된 상태에 역점을 둔다. 다음의 기호는 시스템이 안정상태에 있을 경우를 가정한 것이다.

Pn = n명의 고객만이 대기시스템에 있을 확률

Ls = 대기시스템 내에 있는 고객수의 기대값

Ws = 시스템 내에서 기다려야 하는 시간의 기대값(서비스 받는 시간 포함)

Wg = 대기열에서 기다려야 하는 시간의 기대값(서비스 받는 시간은 제외)

3-5. L과 W의 관계

λn이 모든 n에 대해 상수λ라고 가정하자.

안정상태의 대기행렬 과정

$$Ls = \lambda Ws \quad (1)$$

임이 1961년 J. D. C. Little에 의해 증명되었다. 또 이 증명을 통해

$$LQ = \lambda WQ \quad (2)$$

임을 알수 있다.

만일 λn이 모두 같지 않다면 이 공식에λ대신 장시간 동안의 도착속도를 평균한 값을 대입할 수 있다.

모든 n 1에 대해 평균 서비스 시간이 상수 1/μ이라고 가정하자, 그러면

$$Ws = WQ + (1/\mu) \quad (3)$$

이다.

Ls, Ws, LQ, WQ의 4값은 어느 한 가지를 해석적인 방법으로 구하면 다른 값도 식(1) (3)에 의하여 곧바로 구할 수 있으므로 이 관계식은 아주 중요하다.