한국시뮬레이션학회 논문지
Vol. 24, No. 3, pp. 27-35 (2015. 9)
http://dx.doi.org/10.9709/JKSS.2015.24.3.027
ISSN 1225-5904

시스템 내 고객 수에 따라 서비스율과 도착율을 조절하는 M/G/1/K 대기행렬의 분석

최두일 · 임대은 *

Analysis of an M/G/1/K Queueing System with Queue-Length Dependent Service and Arrival Rates

Doo-Il Choi · Dae-Eun Lim*

ABSTRACT

We analyze an $M/G/1/K$ queueing system with queue-length dependent service and arrival rates. There are a single server and a buffer with finite capacity K including a customer in service. The customers are served by a first-come-first-service basis. We put two thresholds L_1 and $L_2(\geq L_1)$ on the buffer. If the queue length at the service initiation epoch is less than the threshold L_1, the service time of customers follows S_1 with a mean of μ_1 and the arrival of customers follows a Poisson process with a rate of λ_1. When the queue length at the service initiation epoch is equal to or greater than L_1 and less than L_2, the service time is changed to S_2 with a mean of $\mu_2(\geq \mu_1)$. The arrival rate is still λ_1. Finally, if the queue length at the service initiation epoch is greater than L_2, the arrival rate of customers are also changed to a value of $\lambda_2(\leq \lambda_1)$ and the mean of the service times is μ_2. By using the embedded Markov chain method, we derive queue length distribution at departure epochs. We also obtain the queue length distribution at an arbitrary time by the supplementary variable method. Finally, performance measures such as loss probability and mean waiting time are presented.

Key words : M/G/1/K, Controllable Service Rate, Controllable Arrival Rate, State-dependent Service Rate, State-dependent Arrival Rate

요 약

대기행렬 시스템에는 고객들의 대기시간이 지나치게 길어지는 것을 막기 위해 다양한 정책들이 적용되는데, 본 연구에서는 고객숫자에 따른 제어 정책을 갖는 유한용량 $M/G/1/K$ 대기행렬을 분석한다. 고객의 숫자에 따라 서버의 서비스율과 고객의 도착율을 조절하는 정책이다. 두 개의 한계점(thresholds) L_1과 $L_2(\geq L1)$를 설정하고 시스템 내 고객의 숫자가 L_1보다 작을 때는 시스템은 보통(또는 상대적으로 느린)의 서비스율(service rate)과 보통의 도착율(arrival rate)을 갖는다. 고객의 숫자가 증가하여 L_1이상이고 L_2보다 작으면 도착율은 그대로 이지만 서비스율을 증가시켜 빠르게 서비스한다. 이후 고객의 숫자가 더욱 증가하여 L_2이상이면 고객의 도착율도 작은 값으로 바꾸어 고객을 덜 입장시킨다. 위 정책을 갖는 $M/G/1/K$ 대기행렬을 내재점 마코프 체인과 준-마코프 과정을 이용하여 분석하고 수치예제를 제시한다.

주요어 : M/G/1/K, 서비스율 조절 대기행렬, 도착률 조절 대기행렬

1. 서 론

대기행렬 시스템에 입장한 고객들의 대기시간이 적절한 수준을 넘지 않도록 관리하는 것은 서비스의 품질보장 측면에서 높은 수준의 관심이 요구된다. 대기시간을 관리하기 위해 다양한 종류의 제어 정책들이 대기행렬에 적용

Received: 6 April 2015, **Revised:** 14 September 2015, **Accepted:** 15 September 2015

***Corresponding Author:** Dae-Eun Lim
E-mail: del@kangwon.ac.kr
Kangwon National University, Department of System and Management Engineering

되고 있는데 고객숫자(또는 대기열의 길이, queue length)에 따라 서비스의 속도(또는 서비스율, service rate)를 조절하는 것도 대표적인 정책 중 하나이다. 고객수에 따라 서비스 속도(서비스율)를 조절하는 정책을 Banik[1]이 소개한 반도체 제조라인의 예를 통해 살펴보자. 성능이 좋지만 운용 비용이 비싼, 그리고 성능은 떨어지지만 비용은 저렴한 제조 설비가 각 한 대 씩 있다. 여기서 좋은 성능은 높은 서비스율을 의미한다. 처리를 기다리는 lot의 숫자가 일정 수준 이하일 때는 낮은 성능의 설비를 저렴하게 가동한다. 대기하는 lot의 숫자가 일정 수준을 넘어서면 저성능의 설비는 가동을 멈추고 고성능의 설비를 가동하는 정책이다. 고성능 설비는 대기시간을 줄이는데 기여할 수 있으나 운영비용 측면에서는 부담이 될 수 있어 저성능 설비도 적절히 이용하는 것이다. 제어 정책 중에는 고객의 도착율(arrival rate)을 조절하는 방법도 있는데 대기 중인 고객수가 일정 수준을 넘어서면 고객을 덜 오게 하여 대기 고객숫자를 줄이는 것이다. 대기행렬망(queueing network)에서 특정 서버에 고객이 몰리는 경우 다른 서버로 분산시키는 형태로 설명할 수 있으며 통신 시스템이나 제조 시스템에서 높은 수준의 쓰루풋(throughput) 요구를 만족시키기 위해 사용되는 정책이다[2,3]. 도착률을 조절하는 정책의 예로, 통신 시스템에서 라우터 등에 의해 여러 서버로 분산시키는 것에 해당한다. 물류 운반이 자동화돼 있는 반도체 제조라인에서 운반수단인 OHT(overhead hoist transfer)들이 특정 구간에 모여들어 혼잡이 발생할 수 있다. 혼잡 구간의 통과가 요구될 경우 OHT를 우회시키는 방식으로 혼잡 구간 진입을 제어할 수도 있다. 이와 같이 고객수에 따라 서비스율과 도착율을 조절하는 정책은 오래전부터 통신과 제조 시스템에서 응용되어 왔다. 최근에는 데이터 센터(data center)의 운영에 적용되는 사례들을 찾아 볼 수 있다[4,5].

본 연구는 서버의 서비스율과 고객의 도착율을 모두 조절하는 대기행렬을 분석한다. 대기하는 고객수가 증가할 때 먼저 서비스율을 조절하고, 그래도 고객수가 증가하는 경우에는 도착율을 조절하는 정책으로 이를 해석적(analytic)으로 분석한다. 앞선 반도체 제조라인의 예에서 대기시간 증가에 대한 조치로, 현장 수준(floor level)에서 먼저 저성능 설비에서 고성능 설비로 전환했는데도 불구하고 대기하는 lot의 숫자가 일정수준 이상이라면 상위 수준의 디스패칭 모듈에서 lot들을 다른 설비로 보내는 상황을 생각해볼 수 있다. 기존 연구들 중에도 본 연구와 유사하게 하나의 시스템에서 서비스율과 도착율이 모두 조절되는 것들이 있으나, 기존 논문들은 주로 고객의 입장 제어(admission control) 및 가격(pricing) 정책에 관심이 있어 본 연구와는 관심사항이 다르다[6,7]. 본 연구의 중요성은 다음과 같이 설명할 수 있다. 제시되는 모형은 기존 연구들에서 소개되지 않았던 것으로 서비스율과 도착율을 모두 조절하는 혼잡도 조절정책을 갖는다. 분석되지 않았던 모형을 해석적으로 분석하여 고객수 분포 및 대기시간 등의 성능 지표를 제시하며, 여러 혼잡도 제어정책이 혼합 적용되는 현실을 볼 때 기존 모형 대비 보다 높은 응용력을 기대할 수 있다.

본 논문은 다음과 같이 구성된다. 2장에서는 기존 연구들을 간략히 살펴본다. 3장에서는 분석 대상 모형을 자세히 설명하고 고객 이탈시점과 임의시점에서의 고객수 분포를 도출한다. 4장에서는 이를 이용하여 다양한 수치예제를 제시하고 5장에서는 내용을 정리하며 본 연구를 끝맺는다.

2. 기존연구

서비스율과 도착율을 조절하는 연구는 크게 두 가지로 구분할 수 있다. 먼저 서비스율과 도착율 가운데 하나만을 조절하는 모형들에 대해서는 주로 $M/G/1$ 형태의 대기행렬들이 분석되었으며 서비스율을 조절하는 연구의 숫자가 상대적으로 많다. 먼저 서비스율 조절 모형들 중, Nishimura and Jiang[8]은 한계점 (threshold)이 한 개인 $M/G/1$ 대기행렬을 분석했다. 고객숫자가 한계점 이상이면 빠른 속도로 서비스하고, 한계점 보다 작은 값을 가지면 느린 속도로 서비스하는 형태이다. 이후 집단(batch) 도착 $M^X/G/1$ 모형도 분석[9]되었으며 도착 프로세스도 MAP (Markovian Arrival Process)[10], BMAP (Batch MAP)[1] 및 DMAP(Discrete-time MAP)[11]으로 확장되었고, 특히 Choi[10]의 연구에서는 한계점의 개수도 복수개로 확장되었다.

고객숫자에 따라 도착율을 조절하는 모형에 대해서도 주로 $M/G/1$ 형태의 대기행렬 모형이 분석되었다. 비교적 최근 Choi et al.[12]은 도착과정을 MMPP (Markovian Modulated Poisson Process)로 확장한 연구결과를 선보였다.

본 연구는 앞서 소개된 모형들과 달리 서비스율과 도착율 두 가지 모두를 조절하는 새로운 정책에 관한 것이다. 두 가지 모두를 조절하는 정책 자체가 새로운 것은 아니지만, 기존 연구들과는 접근방법과 관심사가 다르다. 두 가지 모두를 조절하는 정책에 관한 기존 연구들은 시스템 상태 (state)에 따른 고객들의 입장 제어와 가격책정을 주

로 마코프 의사결정과정(Markov decision process)를 이용하여 모델링했다. 따라서 고객 대기시간 및 평균 고객수 등을 구하는 본 연구와는 구하려는 성능지표에서 차이가 있다. 최근 결과인 Lee and Kulkarni[5], Yoon and Lewis[7] 그리고 Adusumilli and Hasenbein[13] 등을 참고해볼 수 있다. 마코프 의사결정과정이 아닌 본 연구와 유사한 모델링 방식의 최근 연구는 Choi et al.[14]에서 찾아볼 수 있다. Choi et al.[14]은 고객수에 따라 도착율과 서비스율이 모두 변하는 모형을 분석하여 평균 고객수 등을 구했는데 이 모형은 대기공간 없이 재시도 공간(retrial orbit)만을 고려하고 있다. 따라서 평균 고객수가 순서대로 차례를 기다리는 사람들의 숫자가 아닌 재시도 공간에서 재시도하고 있는 고객의 숫자이다. 본 연구의 모형에는 재시도 공간이 없으며 대기공간을 가정하여 분석한다.

정리하면, 본 연구는 기존 연구들과 달리 도착율과 서비스율을 모두 조절하는 $M/G/1/K$ 모형에 관한 새로운 분석이다. 이 정책은 보다 현실적인 것으로, 서론의 반도체 제조라인의 예시와 같이 현실 문제에서는 혼잡도 해소를 위해 한 가지의 정책만을 적용하는 것이 아니라 여러 정책도 함께 또는 차례대로 적용하는 것을 반영한 것이다.

3. 모형 분석

본 장에서는 먼저 분석 대상 모형을 소개하고 기호들을 정의한다. 그리고 서비스 종료시점(이탈시점)을 내재점으로 갖는 내재점 마코프 체인(embedded Markov chain)을 이용하여 이탈 시점의 고객수 분포를 도출한다. 이에 부가변수법(supplementary variable technique)을 적용하여 임의시점의 고객수 분포를 유도한다. 그리고 평균 대기시간, 시스템 내 평균 고객수와 시스템이 꽉 차서 도착하는 고객의 입장이 거부될 확률 등의 성능지표들을 제시한다.

3.1 분석 대상 모형

본 연구에서는 시스템 내 고객 숫자에 따라 고객의 도착율(arrival rate)과 서버의 서비스율(service rate)이 조절되는 유한용량 $M/G/1/K$ 대기행렬을 다룬다. 먼저 시스템에는 서비스 받는 고객을 포함하여 최대 K명까지 입장이 가능하며 도착 순서에 따라 서비스를 제공(First-Come, First-Served)한다. 시스템에 고객이 총 K명일 때 도착하는 고객은 대기 공간에 입장할 수 없다. 서비스 시간의 분포는 두 종류의 임의의 일반 분포(general distribution)를 가정한다. 두 종류의 서비스 시간에 대한 확률변수는 각각 S_1과 S_2로 표시하며, 이들의 라플라스 변환(Laplace transform)은 $G_1^*(s)$와 $G_2^*(s)$으로 정의한다. 또한 G_1과 G_2는 각각 이들의 분포 함수 (distribution function)이며 $E[S_1]=\mu_1$과 $E[S_2]=\mu_2$이다. 고객의 도착도 두 종류의 포아송 과정 (Poisson process)를 가정하는데 이 둘은 동일한 과정이지만 발생률만 다르며 각각 λ_1과 λ_2로 발생률을 표시한다. 시스템에서 고객수에 대해서는 두 개의 한계점(threshold)을 설정했는데 각각 L_1과 $L_2(0<L_1\leqq L_2<K)$로 표시한다. 대상 모델은 구체적으로 다음과 같이 작동한다. 시스템 내 서비스 받는 고객을 포함하여 고객수가 L_1보다 적으면 서비스 시간은 S_1이며 고객의 도착률은 λ_1이다. 고객수가 증가하여 L_1이상이 되면 먼저 서비스 시간을 S_1에서 S_2로 전환한다. 고객 숫자가 증가했으니 신속히 처리하려는 것을 가정하여 $\mu_2\leqq\mu_1$으로 설정한다. 그래도 고객수가 증가하여 L_2이상이 되면 이제는 고객의 도착률을 λ_2로 전환하는데 상대적으로 적은 숫자의 고객이 도착하게 하여 대기고객수를 줄여야 하므로 $\lambda_2\leqq\lambda_1$을 가정한다. 고객 수는 모두 서비스 종료시점의 숫자이다.

3.2 고객 이탈시점의 고객수 분포(queue length distribution at departure epochs)

제안된 모형의 분석은 Choi et al.[15]등에서 사용된 기존 연구의 틀을 적용하여 비슷한 순서로 진행할 수 있다. 그러나 산출되는 식은 본 연구에서 처음 제시되는 결과이다.

$\tau_n(n\geqq 1)$을 n번째 고객의 이탈시점이라고 하자 (단, $\tau_0=0$). N_n은 τ_n^+ 시점, 즉 n번째 고객의 이탈 직후 시점에서의 고객수를 나타낸다고 정의하면, $\{N_n, n\geqq 0\}$은 유한한 상태공간 $\{0,1,\cdots,K-1\}$을 갖는 마코프 체인이 된다. 마코프 체인 $\{N_n, n\geqq 0\}$의 정상확률(stationary probability)을 유도하기 위해 x_k와 $\mathbf{x}$를 다음과 같이 정의한다.

$$x_k=\lim_{n\to\infty}\Pr\{N_n=k\},\ 0\leq k\leq K-1,$$
$$\mathbf{x}=(x_0,x_1,\cdots,x_{K-1}).$$

이제 x_k를 구하기 위해 확률 $a_n^r(r=1,2)$과 b_n을 다음과 같이 정의한다.

$$a_n^r=\Pr\left\{\begin{array}{l}\text{서비스시간 } S_r\text{동안 도착률이 } \lambda_r\text{인}\\ \text{포아송과정에 의해 } n\text{명 도착}\end{array}\right\}$$
$$=\int_0^\infty \frac{e^{-\lambda_r x}(\lambda_r x)^n}{n!}dG_r(x)\ ,r=1,2,$$

$$b_n = \Pr\left\{\begin{array}{l}\text{서비스시간 } S_2\text{동안 도착률이 } \lambda_1\text{인}\\ \text{포아송과정에 의해 } n\text{명 도착}\end{array}\right\}$$
$$= \int_0^\infty \frac{e^{-\lambda_1 x}(\lambda_1 x)^n}{n!} dG_2(x).$$

그리고 $\overline{a_n^r} = \sum_{k=n}^{\infty} a_k^r$, $\overline{b_n} = \sum_{k=n}^{\infty} b_k$이다. 이들 정의를 이용하여 마코프 체인 $\{N_n, n \geq 0\}$의 1단계 전이확률행렬(one-step transition probability matrix)을 다음의 행렬 P와 같이 표현할 수 있다. 정상확률 벡터 $\mathbf{x}$의 값은 아래 두 종류의 방정식을 푸는 것으로 얻을 수 있는데, 이 값들은 이탈시점에 고객이 $k(0 \leq k \leq K-1)$명 있을 확률이다.

$$\mathbf{x}P = \mathbf{x},\ \mathbf{xe} = 1$$

여기서 $\mathbf{e} = (1,1,\cdots,1)^T$인 벡터이다.

3.3 임의시점의 고객수 분포(queue length distribution at an arbitrary epoch)

이제 임의시점 t에서의 고객수 분포를 유도한다. $N(t)$는 t시점에서 서비스 받는 고객수를 포함한 시스템 내 전체 고객수를 의미하며 $\xi(t)$는 t시점에서 진행되고 있는 서비스 시간 분포를 의미한다. 즉,

$$\xi(t) = \begin{cases} 1, t\text{시점에서 서비스시간이 } S_1\text{인 경우} \\ 2, t\text{시점에서 서비스시간이 } S_2\text{인 경우} \end{cases}$$

임의시점의 정상확률 y_n은 다음과 같이 정의한다.

$$y_n = \lim_{t\to\infty} \Pr\{N(t) = n\},\ 0 \leq n \leq K.$$

먼저, 핵심갱신정리 (key renewal theorem)에 의해 y_0를 쉽게 구할 수 있다.

$$y_0 = \frac{1}{\lambda_1 E} x_0$$

위 식에서 E는 고객들의 이탈간격의 평균(mean inter departure time)을 의미하며 $E = x_0\left[\frac{1}{\lambda_1} + \mu_1\right] + \sum_{n=1}^{L_1-1} x_n\mu_1 + \sum_{n=L_1}^{K-1} x_n\mu_2$으로 구할 수 있다. 다음으로 $\{y_n, n \geq 1\}$을 구하기 위해 $\hat{T}$과 $\tilde{T}$라는 두 개의 부가변수를 정의하여 부가변수법을 적용한다. $\hat{T}$과 $\tilde{T}$은 각각 고객의 남은 서비스 시간 (remaining service time)과 진행된 (elapsed) 서비스 시간을 의미한다. $\alpha_{n,r}(x)$은 고객의 숫자, 잔여 서비스 시간과 서비스 형태에 관한 결합확률분포이고 $\alpha_{n,r}^*(s)$는 $\alpha_{n,r}(x)$의 라플라스 변환으로 다음과 같이 정의된다$(r = 1,2)$.

$$\alpha_{n,r}(x)dx = \lim_{t\to\infty} \Pr\left\{\begin{array}{l} N(t) = n, x < \hat{T} \leq x + dx, \\ \xi(t) = r \end{array}\right\},$$
$$\alpha_{n,r}^*(s) = \int_0^\infty e^{-sx}\alpha_{n,r}(x)dx.$$

$\alpha_{n,r}^*(s)$을 구하려면 $\tilde{T}$동안 도착한 고객수가 필요한데 이에 관한 결합확률분포 $\beta_r(n,x)$와 이들의 라플라스 변환 $\beta_r^*(n,s)(r = 1,2)$는 다음과 같이 정의할 수 있다.

$$P = \begin{pmatrix}
a_0^1 & a_1^1 & a_2^1 & \cdots & a_{L_1-1}^1 & a_{L_1}^1 & a_{L_1+1}^1 & \cdots & a_{L_2-1}^1 & a_{L_2}^1 & a_{L_2+1}^1 & \cdots & a_{K-3}^1 & a_{K-2}^1 & \overline{a}_{K-1}^1 \\
a_0^1 & a_1^1 & a_2^1 & \cdots & a_{L_1-1}^1 & a_{L_1}^1 & a_{L_1+1}^1 & \cdots & a_{L_2-1}^1 & a_{L_2}^1 & a_{L_2+1}^1 & \cdots & a_{K-3}^1 & a_{K-2}^1 & \overline{a}_{K-1}^1 \\
0 & a_0^1 & a_1^1 & \cdots & a_{L_1-2}^1 & a_{L_1-1}^1 & a_{L_1}^1 & \cdots & a_{L_2-2}^1 & a_{L_2-1}^1 & a_{L_2}^1 & \cdots & a_{K-4}^1 & a_{K-3}^1 & \overline{a}_{K-2}^1 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & a_1^1 & a_2^1 & a_3^1 & \cdots & a_{L_2-L_1+1}^1 & a_{L_2-L_1+2}^1 & a_{L_2-L_1+3}^1 & \cdots & a_{K-L_1-1}^1 & a_{K-L_1}^1 & \overline{a}_{K-L_1+1}^1 \\
0 & 0 & 0 & \cdots & b_0 & b_1 & b_2 & \cdots & b_{L_2-L_1} & b_{L_2-L_1+1} & b_{L_2-L_1+2} & \cdots & b_{K-L_1-2} & b_{K-L_1-1} & \overline{b}_{K-L_1} \\
0 & 0 & 0 & \cdots & 0 & b_0 & b_1 & \cdots & b_{L_2-L_1-1} & b_{L_2-L_1} & b_{L_2-L_1+1} & \cdots & b_{K-L_1-3} & b_{K-L_1-2} & \overline{b}_{K-L_1-1} \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & b_1 & b_2 & b_3 & \cdots & b_{K-L_2-1} & b_{K-L_2} & \overline{b}_{K-L_2+1} \\
0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & a_0^2 & a_1^2 & a_2^2 & \cdots & a_{K-L_2-2}^2 & a_{K-L_2-1}^2 & \overline{a}_{K-L_2}^2 \\
0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & a_0^2 & a_1^2 & \cdots & a_{K-L_2-3}^2 & a_{K-L_2-2}^2 & \overline{a}_{K-L_2-1}^2 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & a_1^2 & a_2^2 & \overline{a}_3^2 \\
0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & a_0^2 & a_1^2 & \overline{a}_2^2 \\
0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & a_0^2 & \overline{a}_1^2
\end{pmatrix}$$

$$\beta_r(n,x)dx = \lim_{t\to\infty} \Pr\left\{ \begin{array}{l} \widetilde{T}\text{동안 도착률 } \lambda_r\text{로 } n\text{명 도착}, \\ x < \widetilde{T} \leqq x+dx, \xi(t) = r \end{array} \right\},$$

$$\beta_b(n,x)dx = \lim_{t\to\infty} \Pr\left\{ \begin{array}{l} \widetilde{T}\text{동안 도착률 } \lambda_1\text{로 } n\text{명 도착}, \\ x < \widetilde{T} \leqq x+dx, \xi(t) = 2 \end{array} \right\}.$$

$$\beta_r^*(n,s) = \int_0^\infty e^{-sx}\beta_r(n,x)dx,\ r=1,2,$$

$$\beta_b^*(n,s) = \int_0^\infty e^{-sx}\beta_b(n,x)dx.$$

$\alpha_{n,r}^*(s)$에 관한 식은 준-마코프 과정 (semi-Markov process)의 기본적 특성에 의해 다음의 방정식들을 얻을 수 있다. 먼저, $1 \leqq n < L_1$에 대해서는 다음이 성립한다.

$$a_{n,1}^*(s) = \frac{\mu_1}{E}\left\{ x_0\,\beta_1^*(n-1,s) + \sum_{k=1}^{\min(n,L_1-1)} x_k\,\beta_1^*(n-k,s) \right\},$$

$L_1 \leqq n < K$에 대해서는 다음과 같이 쓸 수 있다.

$$a_{n,2}^*(s) = \frac{\mu_2}{E}\left[\sum_{k=L_1}^{\min(n,L_2-1)} x_k\beta_b^*(n-k,s) + \sum_{k=L_2}^{n} x_k\,\beta_2^*(n-k,s)1_{\{n\geq L_2\}} \right].$$

$\beta_r^*(n,s)$는 계수를 비교하는 것으로 아래와 같이 얻을 수 있는데 이는 기존 연구[15]의 방법을 그대로 적용한 결과이다.

$$\beta_r^*(n,s) = \frac{1}{\mu_r}\left[\sum_{l=0}^{n} a_l^r R_{n-l}^r(s) - G_r^*(s)R_n^r(s) \right],\ r=1,2$$

$$\beta_b^*(n,s) = \frac{1}{\mu_2}\left[\sum_{l=0}^{n} b_l R_{n-l}^1(s) - G_2^*(s)R_n^1(s) \right].$$

단, $R_n^r(s) = (s-\lambda_r)^{-1}\left\{\lambda_r(\lambda_r-s)^{-1}\right\}^n$으로 정의된다.

$\beta_r^*(n,s)\,(r=1,2)$를 대입하고 $s=0$을 대입하면 최종적으로 $y_n(=\alpha_{n,1}^*(0)+\alpha_{n,2}^*(0))$을 얻을 수 있다. 이는 임의시점에서 $n(1\leqq n\leqq K)$명이 있을 확률이다.

$$y_n = \frac{1}{\lambda_1 E}\left\{ x_0\left(1-\sum_{l=0}^{n-1}a_l^1\right) + \sum_{k=1}^{\min(n,L_1-1)} x_k\left[1-\sum_{l=0}^{n-k}a_l^1\right] \right\} + \frac{1}{\lambda_1 E}\sum_{k=L_1}^{\min(n,L_2-1)} x_k\left[1-\sum_{l=0}^{n-k}b_l\right]1_{\{n\geq L_1\}} + \frac{1}{\lambda_2 E}\sum_{k=L_2}^{n} x_k\left[1-\sum_{l=0}^{n-k}a_l^2\right]1_{\{n\geq L_2\}},\quad 1\leqq n<K,$$

$$y_K = 1-\sum_{k=0}^{K-1} y_k.$$

3.4 시스템의 성능지표

앞 절에서 구한 y_n을 이용하여 여러 가지 시스템의 성능지표들을 구할 수 있다. 먼저, 시스템이 먼저 도착한 고객들로 꽉 차있어서 새로 도착하는 고객의 입장이 거부되는 blocking 확률은 다음과 같이 구할 수 있다.

$$P_{Loss} = y_K.$$

또한, 시스템 내 평균 고객수와 리틀의 법칙(Little's law)에 의해 평균 시스템 체류시간(sojourn time)은 다음과 같다.

$$L = \sum_{i=0}^{K} i\,y_i,$$

$$W = \frac{L}{\lambda^*(1-P_{Loss})}.$$

이 때 $\lambda^* = \sum_{k=0}^{L_2-1} x_k\lambda_1 + \sum_{k=L_2}^{K-1} x_k\lambda_2$로 유효 도착률(effective arrival rate)를 의미한다.

4. 수치 예제(numerical examples)

이번 장에서는 실험 조건을 설명하고 조건을 변화시키며 실험하고 발견되는 시스템의 특성을 설명한다. 크게 세 가지의 조건을 변화시키며 실험했다: i) 시스템 내 최대 고객수 K의 크기, ii) 두 한계점 L_1과 L_2의 값이다. iii) 또한 서비스 시간 분포의 변동 계수(Coefficient of Variation, CV)의 값에 따른 시스템의 특성도 고려한다. 주요 성능지표로는 시스템 내 평균 고객수(또는 시스템 체류시간)과 blocking 확률이다.

4.1 상세 실험 조건

서비스 시간의 분포는 CV의 값이 1인 지수분포와, 모수가 5인 얼랑 분포(Erlang distribution)를 고려한다. 모수가 5인 얼랑 분포는 CV값이 $\sqrt{5}^{-1}$로 지수분포의 1보다 작아 서비스 시간들의 변동성이 지수분포보다 상대적으로 적다. 서비스율 $1/\mu_1$과 $1/\mu_2$는 각각 6과 8로, 고객 도착률 λ_1과 λ_2는 각각 5와 3으로 설정한다. 이는 시스템 내 고객수가 $L_1(<L_2\leqq K)$ 이하이면 도착률과 서비스율은 λ_1과 $1/\mu_1$이다. 고객수가 L_1보다 크고 L_2이하이면 도착률과 서비스율은 λ_1과 $1/\mu_2$이며, 고객수가 L_2보다 많고 K보다 적으면 λ_2과 $1/\mu_2$로 작동한다. 시스템 내 최

Table 1. K and L_1 values ($L_1 \leq L_2 < K$)

K	L_1
5	1, 3
10	2, 4, 6, 8
15	2, 4, 6, 8, 10, 12, 14
20	2, 4, 6, 8, 10, 12, 14, 16, 18

대 입장 가능 고객수는 5, 10, 15 및 20으로 설정하며 L_1과 L_2는 주어진 K값의 범위 내에서 값을 갖도록 했는데 이는 Table 1에 정리했다.

4.2 실험 결과

4.2.1 분포 (CV)에 따른 지표들의 변화

분포를 달리하면 변동계수의 값에 따른 시스템 내 평균 고객수 L과 blocking 확률의 경향을 볼 수 있다. Fig. 1는 변동계수가 서로 다른 두 분포의 평균 고객수에 대한 산포도이다. 그림에서 Erlang은 모수가 5인 얼랑 분포의, Exp는 지수분포의 결과값을 의미한다. 두 가지 확률분포에 대해 Table 1에서 주어진 각각의 K,L_1과 L_2에 조합에 대해 평균 대기고객수와 제공 로드(offered load)를 계산한 뒤 이를 산포도로 그린 것이다. 각각의 점들은 서로 다른 K, L_1과 L_2 값들의 조합을 갖고 있다. 시스템 내 평균 고객수는 변동계수의 값이 큰 지수분포의 경우가 비슷한 수준의 제공로드에서 얼랑 분포보다 큰 것을 알 수 있다. Fig. 2를 보면 시스템 내 입장 가능한 고객숫자가 커질수록 blocking 확률은 확연히 줄어드는 당연한 결과를 확인할 수 있다. 즉, $K=20$의 경우에는 모든 K, L_1과 L_2에 조합에 대해 blocking 확률의 값이 0.01이하이며 분포에 따른 차이도 크지 않다. 그러나 시스템에 적은 숫자의 고객이 입장할 수 있는 경우 (즉, K의 값이 작은 경우)에는 지수분포의 blocking 확률이 얼랑분포의 것보다 대체로 큰 편이다. 비슷한 조건에서 변동계수가 크면 blocking확률도 증가하는 것을 볼 수 있다.

4.2.2 두 한계점 L_1과 L_2에 따른 지표의 변화

다음의 Table 2는 $K=20$일 때 L_1과 L_2의 값에 따른 시스템 내 평균 고객수와 평균 대기시간의 변화를 나타내며 지수분포(Exp로 표시)와 모수가 5인 얼랑분포(Erlang-5로 표시)에 관한 결과이다. 먼저 L_1의 값이 작을수록 L값과 W 값도 작은데 이는 고객숫자가 적을 때부터 빠른 속도로 서비스한 결과로 볼 수 있다. 또한 같은 L_1에서 L_2의 값이 작을수록 L값과 W값이 작은데 이도 역시 고객을 덜 오게 한(도착율을 낮춘) 당연한 결과로 볼 수 있다. 그런데 같은 L_1값에 대해 L_2의 값이 커질수록 L값 등은 커지는데 그 차이가 적은 편이다. 구체적으로, $(L_1,L_2)=(2,6)$일 때 L값은 1.5963이고 $(L_1,L_2)=(4,6)$의 경우 L값은 2.0172이다. L_1의 수준이 2만큼 커질 때 고객수는 약 0.4명 증가했다. 반면, $(L_1,L_2)=(2,12)$인 경우, 즉 L_2의 값을 6만큼 크게 한 경우에는 L값이 1.6400으로, (2,6)인 경우 대비 0.04명 정도 증가로 큰 차이가 없다. 즉, 본

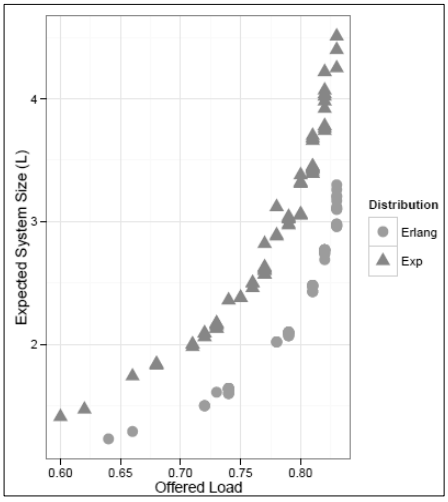

Fig. 1. Average system size over offered load

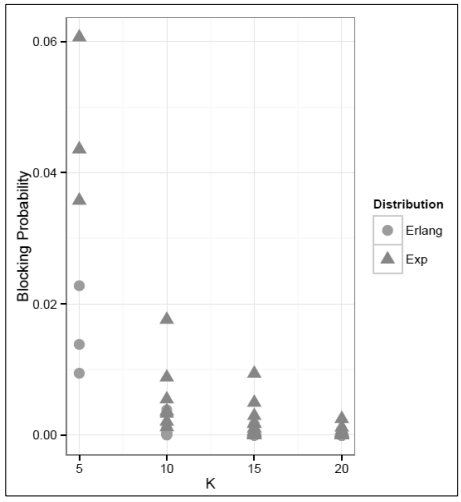

Fig. 2. Blocking probabilities over buffer size K

Table 2. System size (L) and sojourn time (W) over L_1 and L_2 values ($K=20$ case)

		Average System Size (L)		Average System Time (W)				Average System Size (L)		Average System Time (W)	
L_1	L_2	Erlang-5	Exp	Erlang-5	Exp	L_1	L_2	Erlang-5	Exp	Erlang-5	Exp
2	4	1.5029	1.8388	0.3096	0.3904	8	10	2.7408	3.3190	0.5499	0.6724
2	6	1.5963	1.9969	0.3214	0.4086	8	12	2.7635	3.3908	0.5531	0.6817
2	8	1.6279	2.0875	0.3261	0.4212	8	14	2.7700	3.4270	0.5541	0.6868
2	10	1.6373	2.1346	0.3276	0.4284	8	16	2.7717	3.4437	0.5544	0.6893
2	12	1.6400	2.1575	0.3280	0.4321	8	18	2.7721	3.4506	0.5544	0.6904
2	14	1.6407	2.1682	0.3281	0.4339	10	12	2.9568	3.6819	0.5924	0.7427
2	16	1.6409	2.1730	0.3282	0.4347	10	14	2.9715	3.7378	0.5945	0.7502
2	18	1.6409	2.1749	0.3282	0.4350	10	16	2.9755	3.7648	0.5952	0.7541
4	6	2.0172	2.3844	0.4085	0.4929	10	18	2.9765	3.7761	0.5953	0.7557
4	8	2.0723	2.5036	0.4156	0.5073	12	14	3.1039	3.9791	0.6214	0.8005
4	10	2.0894	2.5682	0.4182	0.5163	12	16	3.1133	4.0218	0.6228	0.8063
4	12	2.0943	2.6005	0.4189	0.5212	12	18	3.1157	4.0403	0.6232	0.8090
4	14	2.0956	2.6157	0.4191	0.5236	14	16	3.2014	4.2165	0.6406	0.8468
4	16	2.0960	2.6226	0.4192	0.5247	14	18	3.2071	4.2463	0.6415	0.8509
4	18	2.0961	2.6254	0.4192	0.5251	16	18	3.2641	4.3970	0.6530	0.8822
6	8	2.4339	2.8858	0.4898	0.5889	18	19	3.2994	4.5100	0.6601	0.9055
6	10	2.4691	2.9779	0.4945	0.6004						
6	12	2.4794	3.0258	0.4960	0.6071						
6	14	2.4822	3.0490	0.4965	0.6106						
6	16	2.4830	3.0596	0.4966	0.6122						
6	18	2.4832	3.0639	0.4966	0.6129						

실험조건 하에서는 서비스율을 변화시킨 것이 시스템 내 고객수를 줄이는데 상당히 효과적인 것으로 생각해볼 수 있다.

4.2.3 속도를 조절하지 않는 $M/G/1/K$ 모형과의 비교

이번에는 시스템 내 고객 숫자에 따른 제어 정책이 없는 $M/G/1/K$ 모형과 본 연구에서 소개한 제어정책을 적용한 $M/G/1/K$ 모형의 성능을 비교한다. 서비스 시간의 분포는 모수가 5인 얼랑 분포와 지수분포를 가정하여 제어정책이 있는 모형과 없는 모형 모두에 적용하였다. 그리고 $K=15$도 공통적으로 적용되었다. 제어정책이 없는 모형은 도착률과 서비스율이 변하지 않으므로 각각 5와 8을 가정하였다. 제어정책이 있는 모형에 대해서는 도착률은 $\lambda_1=5$과 $\lambda_2=3$로, 서비스율은 $1/\mu_1=6$과 $1/\mu_2=8$로 설정했다. 마지막으로 제어정책이 있는 $M/G/1/K$ 모형에 대해 L_1과 L_2는 Table 1에 있는 값들을 갖도록 했다.

Fig. 3에서 'STD-Erl5'와 'STD-Exp'는 각각 제어정책이 없고 서비스 시간의 분포가 얼랑 또는 지수분포인 기본 (standard) $M/G/1/K$ 모형을 의미한다. 'VS-Erl5'와 'VS-Exp'는 본 연구에서 소개한 정책을 갖는 $M/G/1/K$ 모형으로 서비스시간의 분포가 각각 모수가 5인 얼랑분포와 지수분포인 경우이다. Fig. 3에 표시된 점들은 L_1과 L_2의 조합에 따른 제공로드와 평균 시스템 사이즈를 의미한다. Table 3은 Fig. 3을 숫자로 정리한 것으로 L_1과 L_2의 조합에 따른 L값의 변화를 보여준다. 도착률과 서비스율의 전환이 늦을수록 즉, L_1과 L_2의 값이 클수록 평

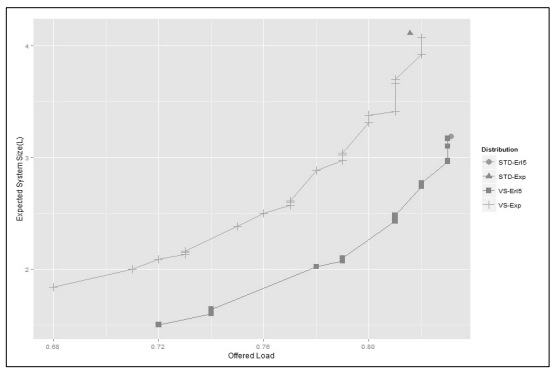

Fig. 3. Average system size over offered load (with or without the overload control mechanism)

Table 3. System size (L) with the overload control method over L_1 and L_2 values ($K=15$ case)

L_1	L_2	Erl-5	Exp	L_1	L_2	Erl-5	Exp
2	4	1.5	1.84	6	8	2.43	2.88
2	6	1.6	2	6	10	2.47	2.97
2	8	1.63	2.09	6	12	2.48	3.02
2	10	1.64	2.13	6	14	2.48	3.04
2	12	1.64	2.15	8	10	2.74	3.31
2	14	1.64	2.16	8	12	2.76	3.38
4	6	2.02	2.38	8	14	2.77	3.41
4	8	2.07	2.5	10	12	2.96	3.66
4	10	2.09	2.57	10	14	2.97	3.7
4	12	2.09	2.6	12	14	3.1	3.92
4	14	2.1	2.61	14	15	3.17	4.07

균 고객수가 증가한다는 앞 절의 결과를 다시 확인할 수 있다.

Fig 3과 Table 3에서 확인할 수 있는 것은 어떠한 정책이라도 적용된 경우에는 적용하지 않을 때보다 L의 값이 작다는 점이다. 이는 다소 당연한 결과이나 추후 고객 유지비용(holding cost)과 서버 운영 비용(operating cost)를 포함하는 비용분석을 통해 최적의 L_1과 L_2 조합을 찾는 것이 필요하다. 평균 시스템 크기의 값을 작게 하여 유지비용을 줄이기 위해서는 L_1과 L_2의 값이 둘 다 작은 조합을 택할 수 있으나 서비스 속도를 빠르게 하는 것은 과다한 운영 비용을 발생시킬 수 있다. 또한 고객 도착률을 조절하는 것도 불만족을 유발할 수 있기 때문이다.

5. 결 론

본 연구에서는 시스템에 있는 고객의 숫자에 따라 서비스율과 도착률을 조절하는 $M/G/1/K$ 대기행렬을 해석적으로 분석하고 수치예제를 통해 시스템의 특성을 살펴보았다. 이러한 정책은 실제 통신 시스템이나 제조 시스템에서 높은 수준의 쓰루풋을 달성하기 위해 적용되고 있다. 분석 방법은 다음과 같다. 먼저 이탈시점의 고객수 확률분포는 평형방정식을 통해 쉽게 얻는다. 그리고 준마코프 과정의 특성들을 이용하여 임의시점 고객수 분포를 얻는다. 임의시점 고객수 분포를 이용하여 시스템에 있는 고객들 숫자의 평균값과 평균 대기시간을 구할 수 있다. 이후 수치예제를 통해 확률분포의 변동 계수 및 한계점에 대한 정책을 변화시키면서 시스템의 특성을 연구하였다. 본 연구는 기존 연구들에서 분석되지 않은 새로운 모형을 분석하고 수치예제를 통해 정책의 특성을 연구했다는 점에서 기여도를 찾을 수 있을 것이며, 추후 연구로 적절한 비용 구조(cost structure)를 설계한 뒤 보다 경제적인 운용방법을 고안해보는 것도 기대할 수 있을 것이다.

References

1. A.D. Banik (2014), “Some Aspects of Stationary Characteristics and Optimal Control of the BMAP/G−G/1/N (∞) Oscillating Queueing System”, *Applied Stochastic Models in Business and Industry*, Vol. 31, No. 2, pp. 204-230.
2. H. Li, Y. Zhu and P. Yang (1995), “Computational Analysis of M(n)/G/1/N Queues with Setup Time”, *Computers & Operations Research*, Vol. 22, pp. 829-840.
3. H. Li and Y. Zhu (1997), “M(n)/G/1/N Queues with Generalized Vacations”, *Computers & Operations Research*, Vol. 24, pp. 301-316.
4. C. Lim and A. Tang (2011), “Dynamic Speed Scaling and Load Balancing of Interconnected Queues”, in Information Theory and Applications Workshop (ITA), IEEE, pp. 1-10.
5. N. Lee and V.G Kulkarni (2014), “Optimal Arrival Rate and Service Rate Control of Multi-Server Queues”, *Queueing Systems*, Vol. 76, pp. 37-50.
6. M.Y. Kitaev and R.F. Serfozo (1999), “M/M/1 Queues with Switching Costs and Hysteretic Optimal Control”, *Operations Research*, Vol. 47, pp. 310-312.
7. S. Yoon and M.E. Lewis (2004), “Optimal Pricing and Admission Control in a Queueing System with Periodically Varying Parameters”, *Queueing Systems*, Vol. 47, pp. 177-199.
8. S. Nishimura and Y.Jiangm (1995), “An M/G/1 Vacation Model with Two Service Modes”, *Probability in the Engineering and Informational Sciences*, Vol. 9, pp. 355-374.
9. A.N. Dudin, “Optimal Control for an M^X/G/1 Queue with Two Operation Modes”, *Probability in the Engineering and Informational Sciences*, Vol. 11, pp. 255-265.
10. D.I. Choi (1999), “MAP/G/1/K Queue with Multiple Thresholds on Buffer”, *Communications-Korean Mathematical Society*, Vol. 14, pp. 611-625.
11. U.C. Gupta, S.K. Samanta and V. Goswami (2014), “Analysis of a Discrete-time Queue with Load Dependent Service under Discrete-time Markovian Arrival Process”, *Journal of the Korean Statistical Society*, Vol. 43, No. 4, pp. 547-557.

12. D.I. Choi, T.-S. Kim and S. Lee (2008), "Analysis of an MMPP/G/1/K Queue with Queue Length Dependent Arrival Rates, and its Application to Preventive Congestion Control in Telecommunication Networks", *European Journal of Operational Research*, Vol. 187, pp. 652-659.
13. K.M. Adusumilli and J.J. Hasenbein (2010), "Dynamic Admission and Service Rate Control of a Queue", *Queueing Systems*, Vol. 66, pp. 131-154.
14. B.D. Choi, Y.H. Chung and A.N. Dudin (2001), "The BMAP/SM/1 Retrial Queue with Controllable Operation Modes", *European Journal of Operational Research*, Vol. 131, pp. 16-30.
15. D.I. Choi, T.S. Kim and S. Lee (2007), "Analysis of a Queueing System with a General Service Scheduling Function, with Applications to Telecommunication Network Traffic Control", *European Journal of Operational Research*, Vol. 178, pp. 463-471.

최 두 일 (dichoi@halla.ac.kr)

1996 KAIST 응용수학과 이학박사
1997 미국 일리노이 주립대학교 Post- Doc.
1998～현재 한라대학교 교수

관심분야 : 대기행렬 이론 및 응용, 정보통신시스템

임 대 은 (del@kangwon.ac.kr)

2004 고려대학교 산업공학과 학사
2006 KAIST 산업공학과 석사
2009 KAIST 산업및시스템공학과 박사
2015～현재 강원대학교 시스템경영공학과 조교수

관심분야 : 대기행렬이론, 제조 시스템 시뮬레이션