The Korean Communications in Statistics **Vol. 14** No. 2, 2007, pp. 287–299

Fast Simulation of Overflow Probabilities in Multiclass Queues*

Jiyeon Lee[1] and Kyungsoon Bae[2]

Abstract

We consider a multiclass queue where queued customers are served in their order of arrival at a rate which depends on the customer type. By using the asymptotic results obtained by Dabrowski *et al.* (2006) we calculate the sharp asymptotics of the stationary distribution of the number of customers of each class in the system and the distribution of the number of customers of each class when the total number of customers reaches a high level before emptying. We also obtain a fast simulation algorithm to estimate the overflow probability and compare it with the general simulation and asymptotic results.

Keywords: Multiclass queues; fast simulation; change of measures; stationary distributions; overflow probabilities.

1. 서론

생산 시스템, 컴퓨터 시스템, 통신 네트워크 등에 응용되는 대기행렬 (Queues)에 대한 연구는 컴퓨터, 통신 산업의 발달과 더불어 더욱 활발히 연구되고 있는 분야이다. 특히 시스템의 기획 단계에서 서비스 품질 (Quality of Service, QoS)은 중요한 요건으로 간주되어 평균 서비스 지연 시간과 과부하 (overflow) 발생 확률 등은 시스템의 중요한 성능 척도로 사용된다. 하지만 시스템이 복잡해질수록 시스템의 상태에 대한 확률 분포를 정확히 분석하는 것은 거의 불가능하고, 그 대신 확률의 점근적인 결과 (asymptotic result)를 유도하거나 또는 시뮬레이션을 통해 성능을 분석하게 된다.

점근적 결과는 시스템의 성능을 나타내는 함수 $f(\ell)$에 대해 $\lim_{\ell\to\infty} f(\ell)/g(\ell) = 1$을 만족하여 $f(\ell) \sim g(\ell)$의 관계가 되는 함수 $g(\ell)$을 찾는 것이다. 대기행렬 시스템의 점

* This work was supported by the Korea Research Foundation Grant funded by Korea Government (MOEHRD, Basic Research Promotion Fund)(KRF-2005-204-C00016).

1) Professor, Department of Statistics, Yeungnam University, 214-1 Dae-Dong, Kyeongsan, Kyeongbuk 712-749, Korea.
Correspondence : leejy@yu.ac.kr

2) Graduate Student, Department of Statistics, Yeungnam University, 214-1 Dae-Dong, Kyeongsan, Kyeongbuk 712-749, Korea.
E-mail : nate831026@nate.com

근적 결과는 주로 단일계층 (single class)의 고객들만 존재하는 시스템에 한정되었고 (McDonald, 1999; Lee, 2004), 다계층 (multiclass) 대기행렬 시스템에 대한 연구는 거의 전무한 실정이다. 다계층 대기행렬이란, 한 서버에 다양한 형태 (계층)의 고객들이 도착하는 대기행렬 시스템으로, 예를 들어 은행의 창구에 비교적 서비스 시간이 짧은 예금 입출금 고객, 조금 시간이 필요한 카드 관련 고객 그리고 시간이 많이 소요되는 대출관련 고객들이 섞여서 도착하는 대기행렬을 생각할 수 있다. 일반적으로 이런 다계층 대기행렬에서는 계층별로 부여되는 우선권 (priority)의 설정여부에 관심이 많은데, 본 논문에서는 각 계층의 고객 수의 분포에 대한 점근적 분석에 주안점을 둔다.

한편, 안정된 대기행렬 시스템에서 고객의 수가 일정한 수를 초과하는 과부하는 극히 희박하게 발생하는 희귀 사건 (rare event)에 해당된다. 이러한 희귀 사건의 확률을 일반적인 시뮬레이션 방법으로 추정하게 되면 하나의 희귀 사건을 발생시키기 위해 시뮬레이션 시행 횟수를 아주 많이 해야 하기 때문에, 그 비용도 많이 들고 오랜 시행동안 효과적으로 작동하는 난수 발생기의 적용도 어렵다. 이런 경우에 적은 시뮬레이션 횟수로 희귀 사건의 확률을 효과적으로 추정할 수 있는 중요 표본 추출방법 (importance sampling method)을 사용하게 된다. 이것은 희귀 사건을 발생하기 쉬운 사건으로 확률측도를 변화 (change of probability measures) 시킨 후, 변화된 확률측도 하에서 확률 값을 추정한 다음, 이를 원래의 희귀 사건의 확률로 재변환하는 하는 방법이다. 이것을 빠른 시뮬레이션 (fast simulation, Heidelberger, 1995; Lee와 Kweon, 2001)이라고 한다. 빠른 시뮬레이션을 적용하기 위해서는 과부하가 일어나는 발생 경로를 고려하여 최적의 확률측도 변화를 찾는 것이 중요하다. 이 변화된 확률측도의 대기행렬 시스템을 시뮬레이션하면 과부하가 쉽게 발생하게 되고 따라서 적은 수의 시뮬레이션으로 결과를 얻을 수 있게 된다.

서비스율이 고객의 계층에 따라 다르고 FIFO (First In First Out)로 서비스가 진행되는 대기행렬은 준-역행가능 (quasi-reversible, McDonald, 2004) 하지 않기 때문에 각 계층의 고객 수의 정상확률분포를 곱의 형태 (product form)로 얻을 수는 없다. 다만, Boxma와 Takine (2003)에 의해 정상상태에서의 각 계층의 고객 수에 대한 적률생성함수 (moment generating function)가 얻어졌고, Choi 등 (2000)는 재입장 (feedback)이 있는 다계층 대기행렬에서의 정상확률에 대한 적률생성함수를 구하는 복잡한 반복식 (iterative formula)을 유도하였다. 그리고 최근에 Dabrowski 등 (2006)은 기존 시스템의 확률측도를 변화시켜 쉽게 과부하가 발생되도록 한 후, 다계층 대기행렬의 총 고객 수와 현재 서비스 받고 있는 고객의 계층에 대한 결합 정상확률의 점근적 결과를 유도하였다. 본 논문에서는 Dabrowski 등 (2006)의 결과를 이용하여 각 계층의 고객 수에 대한 정상확률분포의 점근적 결과를 얻고, 총 고객의 수가 0을 출발해서 처음으로 적정수준을 초과하여 과부하가 발생할 때의 각 계층의 고객 수에 대한 확률분포의 점근적 결과도 찾고자 한다. 또한, 변화된 확률측도를 적용하여 과부하 확률 값을 추정하는 빠른 시뮬레이션 알고리즘도 제안하고자 한다.

2장에서는 다계층 대기행렬모형을 소개하고 Dabrowski 등 (2006)이 얻은 전체 고객 수에 대한 정상확률과 과부하 확률의 점근적 결과를 정리하고 이를 이용하여 각 계층별

고객 수에 대한 정상확률과 과부하가 발생할 때의 각 계층의 고객 수에 대한 분포의 점근적 결과를 유도한다. 3장에서는 변화된 확률측도를 이용하여 과부하 확률값을 빠른 시뮬레이션을 통해 추정하고, 일반 시뮬레이션 및 2장에서 얻어진 점근적 결과와 서로 비교한다.

2. 다계층 대기행렬의 점근적 결과

서비스 제공자가 한 명이 있는 대기행렬 시스템에 C개의 서로 다른 계층의 고객들이 서로 독립적으로 도착한다. 이 때, 계층 c의 고객들은 도착률 λ_c의 포아송 과정 (Poisson process)을 따르며 도착한다. 서버의 서비스는 FIFO로 진행되며 계층 c에 속하는 고객의 서비스 차례가 되면 평균 $1/\mu_c$의 지수분포 (exponential distribution)의 시간만큼 서비스를 제공하게 된다. 그러면 전체 고객의 총 도착률은 $\lambda := \sum \lambda_c$이고 이 시스템의 안정성 (stability)을 보장하기 위해 총 로드 ρ에 대해

$$\rho = \sum_c \frac{\lambda_c}{\mu_c} < 1$$

을 가정한다. 그리고 일반성을 잃지 않고 $\sum_c(\lambda_c + \mu_c) = 1$로 가정하여 균일화 기법 (uniformization method, Walrand, 1988)을 이용하여 본 시스템과 동일한 정상확률을 갖는 이산 시간형 마코프 체인 (Markov chain)을 고려한다. 마코프 체인의 상태는 고객들의 계층을 이용하여 벡터 $X(t) = (X_0(t), X_1(t), \ldots, X_{n-1}(t))$와 $|X(t)| = n$으로 나타낸다. 이 때, $X_0(t)$는 시점 t에서 대기열의 맨 앞에서 현재 서비스를 받고 있는 고객의 계층을 나타내고, $X_1(t)$는 그 뒤에 대기하는 고객의 계층을, ..., $X_{n-1}(t)$은 마지막에 도착하여 기다리고 있는 고객의 계층을 표시한 것으로써 총 고객의 수는 n명임을 나타낸 것이다.

Dabrowski 등 (2006)은 마코프 체인의 전이확률에 대한 조화함수 (harmonic function)를 이용하여 원래의 안정된 마코프 체인을 불안정한 마코프 체인으로 확률측도를 변화시키고, 변화된 확률측도를 이용하여 정상확률의 점근적 결과를 유도하였다. 그 내용을 정리하면 다음과 같다.

상태벡터 $X(t)$의 값 $x = (x_0, x_1, \ldots, x_{n-1})$에 대하여 $N_c(x) := \#\{x_k = c, k \geq 0\}$를 정의한다. 이것은 상태 x에 대하여 현재 서비스 중이거나 대기 중인 고객 중 계층이 c인 고객의 수를 나타낸다. 이를 이용하여 다음과 같은 조화함수 h를 정의한다. 조화함수는 마코프 체인의 전이확률행렬 (transition probability matrix) $K(x, y)$에 대해

$$h(x) = \sum_y K(x, y)h(y), \quad \text{모든 } x\text{에 대해}$$

를 만족하는 함수를 말한다 (McDonald, 1999). 만약 고객이 한명도 없으면 이 함수값은 1로, 그 외의 상태 x에 대해서

$$h(x) := \prod_{c \in C} \exp(\gamma_c N_c(x)) \tag{2.1}$$

로 두면, 이 함수가 $\{x : |x| > 0\}$의 영역에서 조화함수가 되기 위해서는 모든 계층 a에 대해

$$\sum_c \lambda_c \exp(\gamma_c) + \mu_a \exp(-\gamma_a) = \sum_c \lambda_c + \mu_a$$

를 만족해야 한다. 특별히 $a = 1$일 때, 위의 관계식으로부터

$$1 = \sum_c \frac{\lambda_c}{(\exp(-\gamma_1) - 1)\mu_1 + \mu_c} \tag{2.2}$$

을 얻고, 이 식을 풀면 양의 γ_1의 값을 구할 수 있다. γ_1의 값을 이용하여 $\exp(-\gamma_c) = (\exp(-\gamma_1)\mu_1 + \mu_c - \mu_1)/\mu_c$의 관계로부터 나머지 γ_c의 값들을 구한다. 이때 모든 γ_c는 양의 값을 갖는다.

식 (2.1)의 조화함수 h를 이용하여 다음과 같이 변화된 확률측도를 갖는 이산 시간형 마코프 체인을 생성한다. 계층 c의 고객이 서비스를 받고 있는 중일 때, 계층 a의 고객이 도착할 확률은 $\tilde{\lambda}_a := \lambda_a h((c, x_1, \ldots, x_{n-1}, a))/h((c, x_1, \ldots, x_{n-1})) = \lambda_a \exp(\gamma_a)$이고, 서비스를 받고 있는 계층 c의 고객이 서비스를 마치고 시스템을 이탈할 확률은 $\tilde{\mu}_c := \mu_c h((x_1, \ldots, x_{n-1}))/h((c, x_1, \ldots, x_{n-1})) = \mu_c \exp(-\gamma_c)$로 둔다. 그러면 총 도착율은 $\tilde{\lambda} = \sum \tilde{\lambda}_c$이고, 계층 c 고객의 로드는 $\tilde{\rho}_c = \tilde{\lambda}_c/\tilde{\mu}_c$가 되며 따라서 새 시스템의 총로드는 $\tilde{\rho} := \sum \tilde{\rho}_c$가 된다. 그러면 식 (2.2)은

$$1 = \sum_c \frac{\lambda_c}{\tilde{\mu}_c}$$

으로 바꾸어 나타낼 수 있다. 따라서 모든 계층 c에 대해 $\gamma_c > 0$이므로 변화된 확률측도를 갖는 마코프 체인의 총로드 $\tilde{\rho}$에 대해

$$\tilde{\rho} = \sum_c \frac{\tilde{\lambda}_c}{\tilde{\mu}_c} = \sum_c \frac{\lambda_c \exp(\gamma_c)}{\tilde{\mu}_c} > \sum_c \frac{\lambda_c}{\tilde{\mu}_c} = 1$$

임을 알 수 있다. 즉, 변화된 확률측도를 갖는 시스템은 더 이상 안정 (stable)적이지 않고 쉽게 과부하가 발생하는 시스템으로 변형된 것이다. 이러한 변화된 확률측도를 이용하여 Dabrowski 등 (2006)은 다음과 같은 연구결과를 얻었다.

첫째, 시스템의 정상확률분포 π에 대해

$$\pi(\{x : |x| = \ell\}) \sim C_0 \exp(-\Gamma\ell).$$

단, $C_0 = (1 - \rho)(\exp(\Gamma) - 1)/(\tilde{\rho} - 1)$이고 $\exp(\Gamma) = \tilde{\lambda}/\lambda$이다. 그리고 현재 서비스 받고 있는 고객의 계층과의 결합분포로는 $\ell \to \infty$일 때,

$$\pi(\{x : |x| = \ell, x_0 = a\}) \sim \left(\frac{\lambda_a}{\tilde{\mu}_a}\right) C_0 \exp(-\Gamma\ell) \tag{2.3}$$

을 얻었다.

둘째, 과부하 사건 H_ℓ을 총 고객의 수가 0을 벗어난다고 했을 때 다시 0이 되기 전에 처음으로 충분히 큰 수 ℓ에 도착하는 사건이라고 하고, τ_ℓ을 고객의 수가 처음으로 ℓ이 될 때까지의 최초 도달 시간이라고 하면, 과부하 확률 $P_0(H_\ell)$에 대해

$$P_0(H_\ell) \sim C_0 \exp(-\Gamma(\ell - 1)) \tag{2.4}$$

을 구했고 과부하가 발생할 때 서비스 중에 있는 고객의 계층을 함께 고려한 과부하 확률 $P_0(H_\ell \cap \{X_0(\tau_\ell) = a\})$의 점근적 결과로써

$$P_0(H_\ell \cap \{X_0(\tau_\ell) = a\}) \sim C_0 \exp(-\Gamma(\ell - 1)) \frac{\lambda_a/\tilde{\mu}_a - \rho_a}{1 - \rho} \tag{2.5}$$

를 얻었다.

위의 결과를 이용하면, 각 계층의 고객 수에 대한 정상확률과 총 고객의 수가 적정수준을 초과할 때 각 계층의 고객 수에 대한 분포의 점근성을 다음의 정리로 구할 수 있다.

정리 2.1 다계층 대기행렬의 각 계층의 고객 수에 대한 정상확률 π는 $\ell = \sum_c \ell_c$일 때,

$$\begin{aligned}&\pi(\{x : N_1(x) = \ell_1, \ldots, N_C(x) = \ell_C\}) \\ &\sim C_0 \exp(-\Gamma\ell) \left(\sum_c \frac{\lambda}{\tilde{\mu}_c} \frac{\ell_c}{\ell}\right) \binom{\ell}{\ell_1 \cdots \ell_C} \prod_c \left(\frac{\lambda_c}{\lambda}\right)^{\ell_c}.\end{aligned}$$

여기서 $\binom{\ell}{\ell_1 \cdots \ell_C}$는 $\ell!/\ell_1! \cdots \ell_C!$로 정의되는 다항계수 (multinomial coefficient)이다.

증명: 먼저 현재 서비스를 받고 있는 고객의 계층 c에 대해 총확률법칙 (law of total probability)을 적용하면,

$$\begin{aligned}&\pi(\{x : N_1(x) = \ell_1, \ldots, N_C(x) = \ell_C\}) \\ &= \sum_c \pi(\{x : N_1(x) = \ell_1, \ldots, N_C(x) = \ell_C, x_0 = c\})\end{aligned}$$

이고, 서비스를 받고 있는 고객을 제외한 대기열에서 대기 중인 총 $\ell - 1$명의 고객들에 대해서는 고객이 소속된 계층에 대한 분포는 확률이 $(\lambda_1/\lambda, \ldots, \lambda_C/\lambda)$인 다항분포 (multinomial distribution)을 따르게 된다 (Boxma와 Takine, 2003). 따라서 식 (2.3)을 이용하면

$$\begin{aligned}&\sum_c \pi(\{x : N_1(x) = \ell_1, \ldots, N_C(x) = \ell_C, x_0 = c\}) \\ &= \sum_c \pi(\{x : |x| = \ell, x_0 = c\}) \binom{\ell - 1}{\ell_1 \cdots \ell_c - 1 \cdots \ell_C} \left(\frac{\lambda_1}{\lambda}\right)^{\ell_1} \cdots \left(\frac{\lambda_c}{\lambda}\right)^{\ell_c - 1} \cdots \left(\frac{\lambda_C}{\lambda}\right)^{\ell_C} \\ &\sim \sum_c C_0 \exp(-\Gamma\ell) \left(\frac{\lambda_c}{\tilde{\mu}_c}\right) \binom{\ell - 1}{\ell_1 \cdots \ell_c - 1 \cdots \ell_C} \left(\frac{\lambda_1}{\lambda}\right)^{\ell_1} \cdots \left(\frac{\lambda_c}{\lambda}\right)^{\ell_c - 1} \cdots \left(\frac{\lambda_C}{\lambda}\right)^{\ell_C}\end{aligned}$$

$$= C_0 \exp(-\Gamma\ell) \left(\sum_c \frac{\lambda}{\tilde{\mu}_c} \frac{\ell_c}{\ell} \right) \binom{\ell}{\ell_1 \cdots \ell_C} \prod_c \left(\frac{\lambda_c}{\lambda} \right)^{\ell_c}.$$

□

정리 2.2 다계층 대기행렬의 총 고객의 수가 0을 벗어난다는 조건 하에서 총 고객의 수가 다시 0이 되기 전에 적정수준 ℓ을 초과하여 과부하가 발생할 때, 각 계층의 고객 수에 대한 분포 $P_0(\{N_1(X(\tau_\ell)) = \ell_1, \ldots, N_C(X(\tau_\ell)) = \ell_C\}|H_\ell)$은 $\ell = \sum_c \ell_c$일 때,

$$\begin{aligned} & P_0(\{N_1(X(\tau_\ell)) = \ell_1, \ldots, N_C(X(\tau_\ell)) = \ell_C\}|H_\ell) \\ & \sim \frac{\lambda}{1-\rho} \left(\sum_c (\frac{1}{\tilde{\mu}_c} - \frac{1}{\mu_c}) \frac{\ell_c}{\ell} \right) \binom{\ell}{\ell_1 \cdots \ell_C} \prod_c \left(\frac{\lambda_c}{\lambda} \right)^{\ell_c}. \end{aligned} \tag{2.6}$$

증명: 먼저 시점 τ_ℓ에서 현재 서비스를 받고 있는 고객의 계층 c와 식 (2.5)를 이용하면,

$$\begin{aligned} & P_0(H_\ell \cap \{N_1(X(\tau_\ell)) = \ell_1, \ldots, N_C(X(\tau_\ell)) = \ell_C\}) \\ & = \sum_c P_0(H_\ell \cap \{N_1(X(\tau_\ell)) = \ell_1, \ldots, N_c(X(\tau_\ell)) = \ell_c - 1, \ldots, \\ & \quad N_C(X(\tau_\ell)) = \ell_C, X_0(\tau_\ell) = c\}) \\ & = \sum_c P_0(H_\ell \cap \{|X(\tau_\ell)| = \ell, X_0(\tau_\ell) = c\}) \\ & \quad \times \binom{\ell - 1}{\ell_1 \cdots \ell_c - 1 \cdots \ell_C} \left(\frac{\lambda_1}{\lambda} \right)^{\ell_1} \cdots \left(\frac{\lambda_c}{\lambda} \right)^{\ell_c - 1} \cdots \left(\frac{\lambda_C}{\lambda} \right)^{\ell_C} \\ & \sim \sum_c C_0 \exp(-\Gamma(\ell - 1)) \frac{\frac{\lambda_c}{\tilde{\mu}_c} - \rho_c}{1 - \rho} \\ & \quad \times \binom{\ell - 1}{\ell_1 \cdots \ell_c - 1 \cdots \ell_C} \left(\frac{\lambda_1}{\lambda} \right)^{\ell_1} \cdots \left(\frac{\lambda_c}{\lambda} \right)^{\ell_c - 1} \cdots \left(\frac{\lambda_C}{\lambda} \right)^{\ell_C} \\ & = C_0 \exp(-\Gamma(\ell - 1)) \frac{\lambda}{1 - \rho} \left(\sum_c \left(\frac{1}{\tilde{\mu}_c} - \frac{1}{\mu_c} \right) \frac{\ell_c}{\ell} \right) \binom{\ell}{\ell_1 \cdots \ell_C} \prod_c \left(\frac{\lambda_c}{\lambda} \right)^{\ell_c} \end{aligned}$$

을 얻을 수 있다. 단, 위의 두번째 등식은 서비스 중인 고객을 제외한 나머지 $\ell - 1$명의 고객의 계층분포가 다항분포임을 이용하여 유도한 것이다. 위의 식을 (2.4)의 $P_0(H_\ell)$로 나누어주면 식 (2.6)를 얻을 수 있다. □

예제 2.1 만약 모든 계층 c에 대해 $\mu_c = \mu$로 동일하다면, 모든 c에 대해 $\gamma_c = \Gamma = \log(\mu/\lambda)$가 된다. 따라서 모든 c에 대해 $\tilde{\mu}_c = \lambda$가 되고 $C_0 = 1 - \lambda/\mu$이므로, 정리 2.1에 적용하면

$$\begin{aligned} & \pi(\{x : N_1(x) = \ell_1, \ldots, N_C(x) = \ell_C\}) \\ & \sim \left(\frac{\lambda}{\mu} \right)^{\ell} \left(1 - \frac{\lambda}{\mu} \right) \binom{\ell}{\ell_1 \cdots \ell_C} \prod_c \left(\frac{\lambda_c}{\lambda} \right)^{\ell_c} \end{aligned}$$

을 얻을 수 있다. 이 식은 서비스 분포가 모두 모수 μ인 지수분포로 동일할 때, Walrand (1988)의 식 (3.34)에서 구한 각 계층의 고객 수에 대한 정상확률과 일치한다. 한편, 총 고객의 수가 처음으로 ℓ이 될 때, 각 계층의 고객 수에 대한 분포는 정리 2.2에 의해

$$P_0(\{N_1(X(\tau_\ell)) = \ell_1, \ldots, N_C(X(\tau_\ell)) = \ell_C\}|H_\ell) \sim \binom{\ell}{\ell_1 \cdots \ell_C} \prod_c \left(\frac{\lambda_c}{\lambda}\right)^{\ell_c}$$

로 얻어져서 모수가 $(\lambda_1/\lambda, \ldots, \lambda_C/\lambda)$인 다항분포를 따른다.

3. 과부하 확률 추정을 위한 빠른 시뮬레이션

균일화된 (uniformized) 이산 시간형 마코프 체인이 t번의 상태 변이 (transition)를 하는 동안, $A_c(t)$을 계층 c에 속하는 고객의 총 도착 수, $D_c(t)$을 서비스를 마치고 시스템을 떠나는 계층 c에 속하는 총 고객의 수를 나타내고, $M_c(t)$는 계층 c의 고객이 대기 열의 맨 앞에서 서비스를 받고 있을 때 발생하는 귀무 전이 (null transition)의 횟수라고 정의하자. 그러면 $t = \sum_{c=1}^{C}[A_c(t) + D_c(t) + M_c(t)]$가 만족하고, 최초 도달 시간 τ_ℓ은 사건 H_ℓ이 발생할 때까지의 총 전이 횟수가 된다. 사건 H_ℓ에 속하는 모든 샘플 경로 (sample path) ω에 대해 제일 처음 도착하는 고객의 도착확률은 각 계층 c에 대해 λ_c/λ이므로 과부하 확률 $P_0(H_\ell)$은 다음과 같이 나타낼 수 있다.

$$P_0(H_\ell) = \frac{1}{\lambda} \sum_{\omega \in H_\ell} \left(\prod_{d=1}^{C} \prod_{c=1}^{C} \prod_{b=1}^{C} \lambda_d^{A_d(\tau_\ell)(\omega)} \mu_c^{D_c(\tau_\ell)(\omega)} \left(\sum_{a \neq b} \mu_a\right)^{M_b(\tau_\ell)(\omega)} \right).$$

2장에서 정의된 변화된 확률측도 $\tilde{\lambda}_c, \tilde{\mu}_c$, $c = 1, 2, \ldots, C$를 이용하면, $P_0(H_\ell)$은

$$\begin{aligned}
P_0(H_\ell) &= \frac{1}{\lambda} \sum_{\omega \in H_\ell} \left(\prod_{d=1}^{C} \prod_{c=1}^{C} \prod_{b=1}^{C} (\lambda_d \exp(\gamma_d))^{A_d(\tau_\ell)(\omega)} (\mu_c \exp(-\gamma_c))^{D_c(\tau_\ell)(\omega)} \right. \\
&\quad \left. \times \left(\sum_{a \neq b} \mu_a\right)^{M_b(\tau_\ell)(\omega)} \exp(-\gamma_d)^{A_d(\tau_\ell)(\omega)} \exp(\gamma_c)^{D_c(\tau_\ell)(\omega)} \right) \\
&= \frac{1}{\lambda} \sum_{\omega \in H_\ell} \left(\prod_{d=1}^{C} \prod_{c=1}^{C} \prod_{b=1}^{C} \tilde{\lambda}_d^{A_d(\tau_\ell)(\omega)} \tilde{\mu}_c^{D_c(\tau_\ell)(\omega)} \left(\sum_{a \neq b} \mu_a\right)^{M_b(\tau_\ell)(\omega)} \right) \\
&\quad \times \exp\left(-\sum_d \gamma_d A_d(\tau_\ell)(\omega) + \sum_c \gamma_c D_c(\tau_\ell)(\omega) \right) \\
&= \frac{\tilde{\lambda}}{\lambda} \tilde{E}_0 \left[1_{H_\ell} \cdot \exp\left(-\sum_c \gamma_c (A_c(\tau_\ell) - D_c(\tau_\ell)) \right) \right]
\end{aligned}$$

가 된다. 여기서 1_A는 사건 A의 지시확률변수 (indicator random variable)이고 $\tilde{E}_0$는 변화된 확률측도를 전이확률로 갖는 새 마코프 체인에서 계산되는 기대값을 나타낸다. 새

마코프 체인에서도 제일 처음 도착하는 c 계층 고객의 도착확률이 $\tilde{\lambda}_c/\tilde{\lambda}$이기 대문에 마지막 식에 $\tilde{\lambda}$가 곱해졌다. 모든 c에 대해 $A_c(\tau_\ell) - D_c(\tau_\ell) = N_c(X(\tau_\ell))$이므로 과부하 확률은 식 (2.1)에 정의된 조화함수 h에 의해

$$P_0(H_\ell) = \frac{\tilde{\lambda}}{\lambda}\tilde{E}_0[1_{H_\ell}/h(X(\tau_\ell))] \tag{3.1}$$

이 된다.

식 (3.1)의 우변의 기대값을 계산하는 새 마코프 체인의 시스템은 2장에서 살펴보았듯이 총로드 $\tilde{\rho}$가 1보다 크다. 따라서 시스템이 불안정하여 전체 고객의 수가 빨리 증가하므로 작은 횟수의 시뮬레이션을 통해서도 그 확률값을 추정할 수 있게 한다.

다음은 식 (3.1)을 이용한 과부하 확률 추정 알고리즘을 정리한 것이다.

과부하 확률 $P_0(H_\ell)$ 추정을 위한 빠른 시뮬레이션 알고리즘

step 1 식 (2.2)의 해인 γ_c을 구하여 변화된 확률측도 $\tilde{\lambda}_c$와 $\tilde{\mu}_c$의 값을 구한다. $(c = 1, 2, \ldots, C)$

step 2 초기 고객 수 0에서 시작하여 변화된 확률측도를 갖는 균일화된 (uniformized) 마코프 체인을 생성한다.

step 3 첫 고객이 도착하면 바쁜 기간 (busy period)의 수 n을 1 증가시킨다.

step 4 바쁜 기간 중 전체 고객의 수가 ℓ이 되면, 그 때의 상태 $x = (x_0, x_1, \ldots, x_{\ell-1})$에 대하여

$$h_n(x) = \prod_{c \in C} \exp(\gamma_c N_c(x))$$

을 계산하고 고객 수를 초기 상태 0으로 되돌린다.

step 5 step 2 $\sim$ step 4를 계속 반복한다.

step 6

$$\hat{p}_\ell = \left(\frac{\tilde{\lambda}}{\lambda}\right)\frac{\sum_{i=1}^{n} 1/h_i(x)}{n}$$

은 바쁜 기간 중 전체 고객의 수가 ℓ을 초과하여 과부하가 발생하는 확률 $P_0(H_\ell)$의 추정값이 된다.

위와 유사하게 변화된 확률측도를 이용하면

$$P_0(H_\ell \cap \{N_1(X(\tau_\ell)) = \ell_1, \ldots, N_C(X(\tau_\ell)) = \ell_C\})$$
$$= \exp\left(-\sum_c \gamma_c \ell_c\right)\left(\frac{\tilde{\lambda}}{\lambda}\right)\tilde{P}_0(H_\ell \cap \{N_1(X(\tau_\ell)) = \ell_1, \ldots, N_C(X(\tau_\ell)) = \ell_C\})$$

을 얻을 수 있다. 여기서 $\tilde{P}_0$는 변화된 확률측도를 갖는 새 마코프 체인에서 계산되는 확률이다. 따라서 정리 2.2의 총 고객의 수가 처음으로 ℓ이 되었을 때 각 계층의 고객 수에 대한 분포

$$p(\ell_1, \ell_2, \ldots, \ell_C) := P_0(\{N_1(X(\tau_\ell)) = \ell_1, \ldots, N_C(X(\tau_\ell)) = \ell_C\}|H_\ell)$$

을 추정하기 위한 빠른 시뮬레이션은 다음과 같다.

과부하가 발생했을 때 각 계층의 고객 수의 분포 $p(\ell_1, \ell_2, \ldots, \ell_C)$ 추정을 위한 빠른 시뮬레이션 알고리즘

step 1 식 (2.2)의 해인 γ_c을 구하여 변화된 확률측도 $\tilde{\lambda}_c$와 $\tilde{\mu}_c$의 값을 구한다. $(c = 1, 2, \ldots, C)$

step 2 초기 고객 수 0에서 시작하여 변화된 확률측도를 갖는 균일화된 (uniformized) 마코프 체인을 생성한다.

step 3 첫 고객이 도착하면 바쁜 기간 (busy period)의 수 n을 1 증가시킨다.

step 4 바쁜 기간 중 전체 고객의 수가 ℓ이 되면, 그 때의 상태 $x = (x_0, x_1, \ldots, x_{\ell-1})$에 대하여

$$h_n(x) = \prod_{c \in C} \exp(\gamma_c N_c(x))$$

을 계산한다. 그리고 만약 그 때의 상태 $x = (x_0, x_1, \ldots, x_{\ell-1})$에 대하여 $N_1(x) = \ell_1, N_2(x) = \ell_2, \ldots, N_C(x) = \ell_C$이면

$$p_n = 1$$

로 두고 그 외의 경우에는

$$p_n = 0$$

으로 계산한 후 고객 수를 초기 상태 0으로 되돌린다.

step 5 step 2 ~ step 4를 계속 반복한다.

step 6 추정량

$$\hat{p}(\ell_1, \ell_2, \ldots, \ell_C) = \frac{\exp\left(-\sum_c \gamma_c \ell_c\right) \sum_{i=1}^{n} p_i}{\sum_{i=1}^{n} 1/h_i(x)}$$

은 n개의 바쁜 기간 중 전체 고객의 수가 ℓ이 되었을 때 각 계층의 고객 수에 대한 분포 확률의 추정값이 된다.

예제 3.1

고객의 계층이 2개이고 각 모수가 $\lambda_1 = 0.05, \lambda_2 = 0.2, \mu_1 = 0.2, \mu_2 = 0.55$로 총로드가

표 3.1: 과부하 확률

ℓ	일반 시뮬레이션	빠른 시뮬레이션	점근적 결과
1	1.000000000	1.000000000	.220501853
2	.361138616	.360531445	.153733258
3	.179577673	.180108735	.107182386
4	.102905479	.102784795	.074727251
5	.063723055	.062466666	.052099625
6	.041308807	.040406700	.036323709
7	.027526189	.027101879	.025324785
8	.018648510	.018390560	.017656368
9	.012761712	.012579981	.012309969
10	.008790000	.008816822	.008582475
11	.006080654	.006082443	.005983677
12	.004218208	.004193264	.004171802
13	.002930854	.002929010	.002908569
14	.002039221	.002037969	.002027846
15	.001418691	.001426282	.001413808
16	.000988086	.001000210	.000985703
17	.000688697	.000699796	.000687229
18	.000480306	.000484533	.000479134
19	.000334309	.000338722	.000334051
20	.000233090	.000231585	.000232899
21	.000162258	.000159471	.000162377
22	.000113156	.000111054	.000113209
23	.000078765	.000077490	.000078929
24	.000054610	.000053614	.000055029
25	.000038066	.000037825	.000038366
26	.000026619	.000026818	.000026749
27	.000018610	.000019144	.000018649
28	.000012948	.000013159	.000013002
29	.000009051	.000009188	.000009065
30	.000006356	.000006457	.000006320

표 3.2: 시뮬레이션 비교

방법	일반 시뮬레이션			빠른 시뮬레이션		
바쁜 기간의 수	250	50000	10^9	250	700	15000
과부하 확률 $P_0(H_{25})$	0.00000	0.0000800	0.000038066	0.000041944	0.000036909	0.000037825

$\rho = 0.6136$인 안정된 다계층 대기행렬을 고려한다. 이 경우 $\gamma_1 = 0.7947, \gamma_2 = 0.2199$로 얻어져서 변화된 확률측도는 $\tilde{\lambda}_1 = 0.1094, \tilde{\lambda}_2 = 0.2492, \tilde{\mu}_1 = 0.0914, \tilde{\mu}_2 = 0.4414$이고 총 로드는 $\tilde{\rho} = 1.7610$가 된다.

표 3.1은 각 ℓ에 대한 과부하 확률 $P_0(H_\ell)$을 일반 시뮬레이션, 빠른 시뮬레이션 그리고 2장에서 얻어진 점근적 결과인 식 (2.4)를 이용해서 추정한 값들이다. 일반 시뮬레이션은 총 10억개의 바쁜 기간을 생성해서 추정한 값이고, 빠른 시뮬레이션은 바쁜 기간을 15,000개만 생성해서 추정한 값으로 두 값이 무척 비슷함을 알 수 있고, 점근적 결과는 적당히 큰 ℓ에 대해 높은 근사력을 보이고 있다.

표 3.2는 $\ell = 25$일 때의 과부하 확률 $P_0(H_{25})$에 대한 일반 시뮬레이션과 빠른 시뮬레이션의 추정값을 생성한 바쁜 기간의 수에 따라 비교한 것이다. 빠른 시뮬레이션의 경우 훨씬 작은 생성 횟수로 과부하 확률을 추정할 수 있음을 볼 수 있다.

총 고객의 수가 처음으로 $\ell = 15$가 되었을 때 계층 1에 해당되는 고객 수에 대한 확률분포를 일반 시뮬레이션, 빠른 시뮬레이션 그리고 정리 2.2의 점근적 결과를 이용하여 추정한 것을 표 3.3와 그림 3.1에 나타내었다. 여기서 일반 시뮬레이션은 총 10억개의 바쁜 기간을 생성해서 추정하였고, 빠른 시뮬레이션은 바쁜 기간을 1000만개 생성해서 추정하였다. 고객의 계층을 구별하여 추정하기 때문에 총 고객수에 대한 과부하 확

표 3.3: 총 고객 수가 $\ell = 15$일 때, 계층 1의 고객 수에 대한 분포

ℓ_1	일반 시뮬레이션	빠른 시뮬레이션	점근적 결과
0	0.0099852611	0.0104816495	0.0101817388
1	0.0696367285	0.0695025973	0.0694348121
2	0.1762096186	0.1757799824	0.1762041815
3	0.2501644121	0.2503204125	0.2501388953
4	0.2323867565	0.2323545219	0.2320424455
5	0.1517800564	0.1515311392	0.1520643958
6	0.0733655179	0.0735849007	0.0735439360
7	0.0269769809	0.0269670571	0.0269124773
8	0.0074596935	0.0075446655	0.0075464581
9	0.0017276489	0.0016328069	0.0016264883
10	0.0002643282	0.0002642344	0.0002678415
11	0.0000394730	0.0000327939	0.0000331488
12	0.0000035244	0.0000030484	0.0000029884
13	0.0000000000	0.0000001833	0.0000001854
14	0.0000000000	0.0000000071	0.0000000071
15	0.0000000000	0.0000000000	0.0000000001

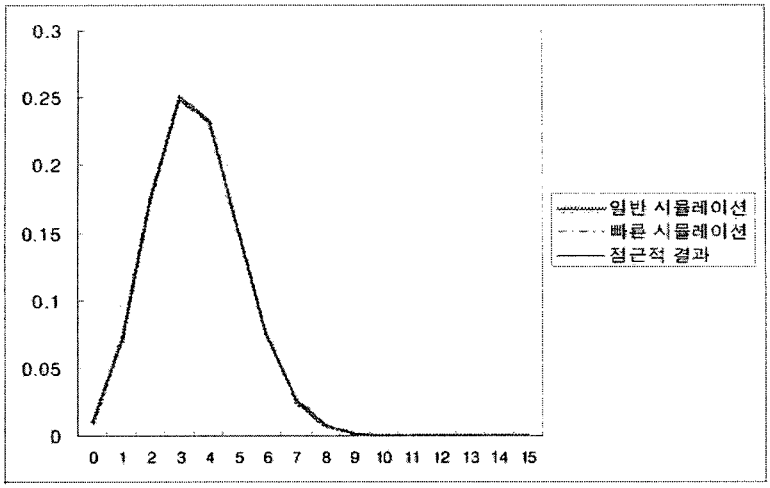

그림 3.1: 총 고객 수가 $\ell = 15$일 때, 계층 1의 고객 수에 대한 분포

률 추정보다 더 많은 시뮬레이션 횟수가 이용되었고, 추정된 세 개의 값은 매우 가까움을 알 수 있다.

4. 결론

본 논문은 고객의 계층에 따라 서비스율이 서로 다른 다계층 대기행렬에서 Dabrowski 등 (2006)이 얻은 전체 고객 수에 대한 정상확률과 과부하 확률의 점근적 결과를 이용하여 각 계층별 고객 수에 대한 정상확률분포의 점근적 결과를 얻었다. 이 점근적 결과는 서비스율이 모두 동일한 경우에는, 이미 알려진 정확한 정상확률분포와 일치함을 확인하였다. 또한 전체 고객의 수가 적정수준 ℓ을 초과하여 과부하가 발생할 때의 각 계층의 고객 수에 대한 확률분포의 점근적 결과도 얻었다. 이러한 점근적 결과는 적당히 큰 ℓ의 경우에는 높은 근사력으로 일반 시뮬레이션을 통한 추정보다 훨씬 유용함을 알 수 있었다.

안정된 대기행렬에서는 과부하가 극히 희박하게 발생하는 희귀사건이기 때문에 일반 시뮬레이션을 통한 확률 추정에는 엄청난 시뮬레이션 횟수가 요구된다. 본 논문에서는 과부하가 발생하는 확률값과 과부하 발생 시 각 계층의 고객 수에 대한 확률분포를 추정하는 빠른 시뮬레이션 알고리즘을 소개하고 일반 시뮬레이션보다 훨씬 적은 수의 시뮬레이션 횟수로 추정가능함을 보였다.

참고문헌

Boxma, O. J. and Takine, T. (2003). The $M/G/1$ FIFO queue with several customer classes. *Queueing Systems*, **45**, 185–189.

Choi, B. D., Kim, B. and Choi, S. H. (2000). On the M/G/1 Bernoulli feedback queue with multi-class customers. *Computers & Operations Research*, **27**, 269–286.

Dabrowski, A., Lee, J. and McDonald, D. (2006). Large deviations of multitype queues. preprint.

Heidelberger, P. (1995). Fast simulation of rare events in queueing and reliability models. *ACM Transactions on Modeling and Computer Simulation*, **5**, 43–85.

Lee, J. and Kweon, M. H. (2001). Estimation of overflow probabilities in parallel networks with coupled inputs. *The Korean Communications in Statistics*, **8**, 257–269.

Lee, J. (2004). Asymptotics of Overflow Probabilities in Jackson Networks. *Operations Research Letters*, **32**, 265–272.

McDonald, D. R. (1999). Asymptotics of first passage times for random walk in an orthant. *The Annals of Applied Probability*, **9**, 110–145.

McDonald, D. (2004). *Elements of Applied Probability for Engineering, Mathematics and Systems Science*. World Scientific, River Edge, NJ.

Walrand, J. (1988). *An Introduction to Queuing Networks*, (GL Jordan, ed.), Prentice Hall, England Cliffs, NJ.

[Received December 2006, Accepted March 2007]