Journal of Korea Multimedia Society Vol. 12, No. 1, January 2009(pp. 31-40)

분할방식에 의한 N-설계 콜센터의 근사 성능분석

박철근[†], 성수학[††], 정해[†††]

요 약

콜센터는 회사와 고객들을 연결하는 주요 접속점이 되고 있다. 최근 계속해서 진보하는 통신 기술로 인해 콜센터의 수나 규모도 극적으로 성장하고 있다. 총 운영비의 많은 부분을 차지하는 인건비를 미루어 볼 때 효율적인 상담원 배치 계획은 콜센터의 경제적이고 성공적인 경영을 좌우한다. 그러므로 상담원의 수를 효과적으로 결정하는 것이 무엇보다 중요하다. 이러한 의미에서 콜센터 운영 및 관리는 큐잉 이론을 이용하는 수리적 최적화 문제로 모델링 할 수 있다. 본 논문에서는 대기 중 중도포기를 하는 두 종류의 고객을 가지며 두 대기 큐들의 용량이 유한인 N-설계 콜센터의 근사적 분석을 상태분할 방법을 이용해 다루기로 한다. 콜센터의 성능 측도에 대한 시스템 파라미터들의 영향을 알아보기 위해 수치계산 예를 보여준다.

Approximate Performance Analysis of an N-design Call Center by the Decomposition Method

Chul Geun Park[†], Soo-Hak Sung[††], Hae Chung[†††]

ABSTRACT

Call centers have become the prevalent contact points between companies and their customers. By virtue of recent advances in communication technology, the number and size of call centers have grown dramatically. As a large portion of the operating costs are related to the labor costs, efficient design and workforce staffing are crucial for the economic success of call centers. Therefore it is very important to determine the adequate number of agents. In this context, the workforce staffing level can be modeled as mathematical optimization problem using queueing theory. In this paper, we deal with an approximate analysis of an N-design call center with two finite queues and two types of reneging customers by using the state decomposition method. We also represent some numerical examples and show the impact of the system parameters on the performance measures of the call center.

Key words: Call center(콜센터), N-design(N-설계), State decomposition(상태분할), Queueing(큐잉), Performance(성능분석)

1. 서 론

고객지원센터(contact center)는 전화, 팩스, 전자우편, 채팅 및 다른 통신채널을 통하여 서비스를 제공하는 고객들을 위한 정보지원 서비스 구조이다. 특히 중요한 고객지원센터 중 하나가 콜센터인데, 주로 고객의 전화 호를 서비스한다. 정보통신 기술의 진보로 말미암아 정보 지원센터의 수, 크기 및 범위뿐만 아니라 그곳에 고용되거나 고객으로 그것을 사용하는 사람의 수 또한 폭발적으로 성장하고 있다. 예를

※ 교신저자(Corresponding Author) : 박철근, 주소 : 충남 아산시 탕정면 선문대학교(336-708), 전화 : 041)530-2358, FAX : 041)530-2910, E-mail : cgpark@sunmoon.ac.kr
접수일 : 2008년 4월 14일, 완료일 : 2008년 10월 7일
[†] 종신회원, 선문대학교 정보통신공학부 교수
[††] 배재대학교 전산수학콘텐츠학과 교수 (E-mail : sungsh@pcu.ac.kr)
[†††] 금오공과대학교 전자공학부 부교수 (E-mail : hchung@kumoh.ac.kr)
※ 본 논문은 2007년도 정부(과학기술부)의 재원으로 한국과학재단의 지원을 받아 수행된 연구임(No. R01-2007-000-20053-0).

들어 2000년 전후로 영국의 콜센터 상담원수는 약 60만, 네덜란드는 약 20만에 달하며 독일의 콜센터 고용원수는 약 28만에 이르렀다[1]. 미국의 2000년 콜센터 통계에 따르면 고객과 사업상 상담을 위한 상호 작용은 약 70% 정도 콜센터에서 일어나며 미국 전체 고용자의 약 3%가 콜센터 고용자이며, 콜센터 상담원수는 150만을 능가하는 것으로 평가하고 있다[2,3].

콜센터의 운용비용의 60~70%가 상담원들의 인건비에 해당한다[3]. 그러므로 상담원을 효율적으로 배치하고 고객의 서비스 수준을 고려하면서 적절한 상담원의 수를 결정하는 문제는 콜센터의 성공적인 운영과 밀접한 관계가 있다. 이런 의미에서 콜센터 운영 및 관리는 큐잉 이론을 이용하는 수리적 최적화 문제로 모델링 할 수 있고 각국에서 이에 대한 많은 연구가 진행 중에 있다. 특히 간단한 수리적 큐잉 모델들은 이미 연구가 많이 진행되어있다.

모든 회선이 점유되었을 때 도착하는 고객은 통화중 신호를 만나게 된다. 이 고객은 블록 되거나(블록 손실호) 조금 후에 재시도 한다. 모든 상담원이 서비스 중일 때 한동안 회선연결을 유지하는 고객들은 큐에 위치한다. 그러나 큐에서 대기해야 한다는 사실을 시스템으로부터 아는 고객은 대기 없이 바로 포기하기도 하는데 이를 즉시포기(balking)라 한다. 이후에 고객들은 조금 후에 재시도하거나 손실(즉시포기 손실호)된다. 큐에 대기 중인 고객이 서비스 개시 전에 인내를 다한다면, 그들은 연결을 끊는다. 이와 같은 중도포기(reneging) 이후에 고객들은 조금 후에 재시도하거나 손실(중도포기 손실호)될 것이다[4].

기본적인 모델에서는 한 종류의 고객만이 한 상담그룹에 의해 서비스된다. 이런 콜센터의 성능은 Erlang-C 큐잉모델(M/M/N)을 사용하여 종종 분석된다. 이 모델은 성능 측도들은 쉽게 계산하지만 무한 대기 인내고객과 무한 회선수를 비현실적으로 가정한다. 이 기본적인 큐잉 모델은 고객의 지수분포 인내시간을 갖는 M/M/N+M 큐잉모델로 확장되는데, 이 모델 역시 분석 가능하다[1]. M/M/N+M 큐잉 모델의 확장으로 일반분포 대기시간을 갖는 콜센터 모델인 M/M/N+G 모델도 분석가능하다[5].

본 논문에서는 이미 연구가 이루어진 대부분의 모델과 달리 중도포기는 고려하기로 하며, 숙련도 기반 라우팅을 고려하여 분석하기로 한다. 상담원들은 그들의 숙련도에 따라 구분되는데 숙련도는 상담원이 그 서비스를 얼마나 잘 그리고 빨리 제공하느냐를 기술한다[3].

숙련도 기반 라우팅을 갖는 모델들의 예로, 소위 N-설계 모델, X-설계 모델, W-설계 모델과 M-설계 모델 콜센터가 있다[6]. N-설계 모델은 두 유형의 고객이 두 그룹의 상담원에게 서비스를 받는데 이중 한 상담원 그룹은 두 유형의 고객을 모두 서비스 할 수 있는 모델이다. 두 개의 무한 대기공간을 갖는 N-설계 모델의 근사적 성능 분석의 결과는 이미 나와 있다[7]. 본 논문에서는 이와는 달리 두 고객 모두 큐에서 대기 중 중도포기가 가능하며 두 대기 큐가 모두 유한인 N-설계 모델을 다루기로 한다. 각 큐에서 대기고객은 각기 다른 지수분포 인내시간을 가진 후 중도포기 한다. 숙련도 기반 라우팅을 갖는 모델들의 성능 분석을 위해 엄밀한 확률과정론적 방법을 사용하면 성능 측도들을 수치적으로 구하는데 계산과정의 부담이 상담원의 수와 서버의 이용도와 함께 급속하게 증가하게 된다[8].

본 논문에서는 N-설계 모델에 대한 시스템의 상태공간을 여러 개의 하위 공간으로 나누는 소위 분할방법을 이용한 근사적 분석 기법을 이용한다. 근사기법은 필수 계산시간을 충분히 줄이면서 상당한 정확성을 제공한다는 것은 잘 아려져 있다[7,8]. 근사 시스템의 상태 공간의 크기는 서버의 수가 증가함에 따라서 천천히 증가하며 상태 공간의 크기는 시스템의 이용도에 따라서는 증가하지 않는다. 본 논문의 나머지 부분은 다음과 같이 구성되어있다. 제2절에서는 고려하는 N-설계 콜센터의 큐잉 모델을 기술한다. 제3절에서는 이 큐잉 모델의 상태와 상태공간을 묘사하고 분할방법을 이용하여 시스템을 분석하는 방법을 기술한다. 제4절에서는 이것을 바탕으로 수치실험으로 시스템의 동작을 연구하고 상담원의 서로 다른 할당을 논의한다. 끝으로 제5절에서는 결론과 추후연구 주제를 언급한다.

2. 시스템 모델

이 절에서는 단순화된 모델을 갖는 N-설계 콜센터의 시스템 모델과 라우팅 과정을 설명한다. 그림 1에 나타난 것처럼 A와 B 두 종류의 고객과 두 종류의 상담원(A전담원, 일반원) 그룹을 갖는 콜센터를

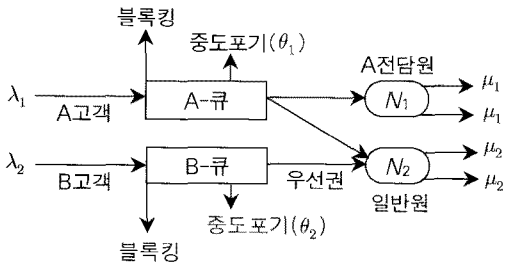

그림 1. 두 상담원 그룹의 N-설계 콜센터 모델

모델링한다. 이 큐잉 모델은 고객의 도착과정, 상담원의 서비스 특성, 시스템의 라우팅 정책 및 대기 큐의 크기제한 등으로 잘 기술된다.

A와 B고객은 각각 도착률 λ_1과 λ_2를 갖는 독립 포아송 과정에 따라 시스템에 도착한다. 두 종류 모두 비인내 고객이나 즉시포기는 고려하지 않고 중도포기만 고려하기로 한다. A고객과 B고객은 서비스가 시작되지 않으면 각각 평균 θ_1^{-1}과 θ_2^{-1}을 갖는 지수분포 인내시간 후에 중도포기 한다. 포기하는 고객은 손실되며 재시도는 없는 것으로 가정한다. 두 상담원 그룹은 각기 서로 다른 숙련도를 갖는다고 가정한다. A상담원 그룹은 A고객을 전담으로 서비스하는 N_1명의 상담원(A전담원)을 갖는다. A전담원의 서비스 시간은 평균 μ_1^{-1}의 지수분포를 갖는다. 일반원의 서비스 시간은 평균 μ_2^{-1}의 지수분포를 갖고 B고객에게 서비스 우선권을 준다.

두 종류의 고객은 대응하는 자신의 큐를 갖고 있다. 대기하거나 서비스 받는 고객을 포함하여 시스템에 있는 A고객의 최대 수는 K_1명이다. 시스템에 있는 B고객의 최대 수는 K_2명이고 역시 유한이다. 시스템에 있는 고객의 최대 수는 콜센터에 두 종류 고객들을 위해 서로 독립적으로 시설되어 있는 전화 회선수를 반영한다. 말하자면 A고객 K_1명이 시스템에 있다면 이때 도착하는 고객은 통화중 신호를 받고 손실되어 시스템을 떠난다. 마찬가지로 시스템에 있는 B고객도 최대 K_2명을 넘을 수 없다.

도착하는 A고객은 먼저 해당 A전담원에게 서비스를 받는다. 그렇지 않고 모든 A전담원이 서비스 중이고 일반원만 가용이면 일반원에게 서비스를 받는다. 모든 A전담원과 모든 일반원이 서비스 중일 때 도착하는 A고객은 자신의 해당 큐에서 대기한다. 대기 중인 A고객이 인내를 다하면 위에 언급된 포기율로 중도포기 한다. 일반원의 고객호 선택 규칙은 고객 유형에 따라 다르다. A전담원은 자신의 고객 유형에 대해 선입선출 규칙을 따른다. 서비스 가능한 일반원은 B큐를 먼저 보고 선입선출 규칙에 따라 대기 중인 B고객을 서비스하고, B큐가 비게 되면 A큐를 보고 대기 중인 A고객을 서비스한다. 두 큐 모두에 고객이 없으면 서비스 가능한 일반원은 유휴가 된다. 결국, 일반원이 B고객을 위해 우선순위 서비스 정책을 갖는 N-설계 라우팅 규칙을 갖는다. 더구나 도착하는 B고객은 서비스 중인 A고객의 서비스를 가로채지 않는다.

이제 시스템의 근사적 분석과정을 설명하기 위해 시스템의 상태를 2차원 마르코프 과정으로 나타내기로 한다. 고려하는 모델에서 K_1과 K_2가 유한이므로 시스템 상태공간은 유한이 되고 이차원 마르코프 연쇄는 안정 상태 확률분포를 가진다. 안정 상태에서 X_1을 A전담원에게 서비스 중이거나 A큐에 대기 중인 A고객의 전체수라 하고, X_2를 일반원에게 서비스 중이거나 B큐에 대기 중인 B객을 포함하고 일반원에게 서비스 중인 A고객도 포함하는 고객의 수라 두자. 근사 성능분석을 위한 분할방법을 설명한다[7,8]. 상태공간을 4개의 분할 공간 $S_1=\{X_1 \le N_1\} \bigcap \{X_2 < N_2\}$, $S_2=\{N_1 < X_1 \le K_1^*\} \bigcap \{X_2 < N_2\}$, $S_3=\{X_1 \le N_1\} \bigcap \{N_2 \le X_2 \le K_2^*\}$ 및 $S_4=\{N_1 < X_1 \le K_1^*\} \bigcap \{N_2 \le X_2 \le K_2^*\}$로 나눈다. 분명하게도 S_2는 금지영역이다. 여기서 K_1^*와 K_2^*는 확률변수로 다음 절에서 상세히 설명한다. 수치계산 예에서는 평균 $K_A = E[K_1^*]$와 $K_B = E[K_2^*]$를 사용한다. 그러므로 이들을 상수로 생각해도 무방하다. 분할방법의 핵심은 다음 근사 확률 값들을 구하는 것이다.

$$P(X_1=i|X_2=j) \approx P(X_1=i|X_2<N_2) \equiv p_{1,i}, \quad (1)$$

$$P(X_1=i|X_2=j) \approx P(X_1=i|N_2 \le X_2 \le K_2^*) \equiv p_{2,i}, \quad (2)$$

$$P(X_2=j|X_1=i) \approx P(X_2=j|X_1 \le N_1) \equiv q_{1,j}, \quad (3)$$

$$P(X_2=j|X_1=i) \approx P(X_2=j|N_1 < X_1 \le K_1^*) \equiv q_{2,j}. \quad (4)$$

3. 분할 모델 분석

3.1 분할 영역 S_1에서 $p_{1,i}$의 계산

조건 $\{X_2 < N_2\}$가 주어지면 $\{N_1 < X_1 \le K_1^*\}$인 경

우는 발생하지 않는다. $\{X_2 < N_2\}$에서 A큐는 비고 모든 도착하는 A객은 조건 $\{X_1 \le N_1\}$일 동안 A전담원에게 서비스를 받는다. 이 경우 대기 공간이 없는 다중서버 모델인 $M/M/N_1/N_1$(Erlang-B) 큐잉 시스템으로 근사가능하다. 구하려는 $p_{1,i}$는 이 생성-소멸 과정의 안정 상태 확률분포로 주어진다[9].

3.2 분할 영역 S_4에서 $q_{2,j}$의 계산

조건 $\{N_1 < X_1 \le K_1^*\}$가 주어지면 모든 A전담원과 일반원은 서비스 중이다. 여기서 K_1^*는 $K_1 - N_2$(모든 일반원이 A고객을 서비스)에서 K_1(모든 일반원이 B고객을 서비스)까지 변할 수 있다. $\{X_2 < N_2\}$인 경우는 발생하지 않고, $\{N_2 < j \le K_2^*\}$일 경우 $X_2 = j$에서 $X_2 = j-1$로의 천이는 $N_2\mu_2$로 발생하는 반면, $X_2 = j$에서 $X_2 = j+1$로의 천이는 비율 λ_2로 발생한다. 대기중인 B고객은 서비스가 시작되지 않으면 평균 θ_2^{-1}을 갖는 지수분포 인내시간 후에 중도포기 한다. 이 분할 영역에서 B큐를 도착율 λ_2, 서비스율 $N_2\mu_2$, 고객의 인내시간 분포 $\tau \sim \exp(\theta_2)$ 및 유한 대기 큐에 의해 잘 표시되는 $M/M/1/K_2^* + M$ 큐로 모델링할 수 있다. 여기서 K_2^*는 최소 $K_2 - N_2 + 1$에서 최대 $K_2 + 1$까지 변할 수 있다. 그러나 중요한 계산 요소인 평균 큐 길이는 다음 절에서 구할, 일반원에게 서비스 중인 B형 고객의 수에 대한 분포를 통하여 알 수 있다.

3.3 분할 영역 S_3에서 $q_{1,j}$의 계산

조건 $\{X_1 \le N_1\}$이 성립할 때 B고객은 우선순위에 따라 N_2명의 일반원에게 A고객보다 우선권을 갖고 처리율 μ_2로 서비스 받게 된다. 이 조건에서 A큐는 비었으며 A고객은 A전담원이 모두 서비스 중($\{X_1 = N_1\}$)이고 서비스 가능한 일반원이 있을 때($\{X_2 < N_2\}$) 일반원에게 서비스 받을 수 있다. 이 영역에서 A고객은 위에서 설명된 $M/M/N_1/N_1$ 시스템에서 일반원에게로 넘친다. 이와 같은 오버플로우 트래픽을 하나의 차단 포아송 과정(IPP: Interrupted Poisson Process)으로 모델링할 수 있다[10].

차단 포아송 과정은 하나의 지수분포를 갖는 시간동안 ON(ON주기)으로, 또 다른 지수분포를 갖는 시간동안 OFF(OFF주기)로 교대로 바뀌는 일종의 포아송 과정이다. ON주기 동안 고객의 도착간격 시간은 지수 분포를 따른다. 반면에 OFF주기 동안 고객의 도착은 없다. 언급된 세 가지 지수분포는 상호 독립이다[11,12]. IPP를 표현하기 위해 필요한 매개변수를 γ_A^{-1}와 γ_S^{-1} 및 λ로 각각 ON주기와 OFF 주기의 평균 길이 및 ON주기 동안 고객의 포아송 도착률이라 정의한다. 그리고 Q_I를 IPP 도착과정의 천이를 표시하는 지배 마르코프 연쇄(underlying Markov chain)의 무한소 생성자(infinitesimal generator)라 두고, Λ_I을 ON주기 및 OFF 주기 동안 도착률을 나타내는 대각행렬이라 두면 IPP는 (Q_I, Λ_I)로 완전히 표현되고 Q_I와 Λ_I는 다음과 같이 표시된다.

$$Q_I = \begin{pmatrix} -\gamma_A & \gamma_A \\ \gamma_S & -\gamma_S \end{pmatrix}, \quad \Lambda_I = \begin{pmatrix} \lambda & 0 \\ 0 & 0 \end{pmatrix}. \tag{5}$$

주목하는 $M/M/N_1/N_1$ 시스템에 부과되는 트래픽 강도(traffic intensity)는 $\rho_1 \equiv \lambda_1/\mu_1$이므로 오버플로우 트래픽은 IPP로 쉽게 모델링된다[10].

오버플로우 트래픽이 IPP 과정으로 모델링되었다고 하면 분할 시스템에서 일반원에게 서비스를 받고자 하는 고객의 입력은 두 가지로 하나는 IPP 과정으로 표시된 A고객들의 입력과 나머지 하나는 상담원에게 서비스 우선권을 갖는 B고객들의 포아송 입력이다. IPP 과정과 포아송 과정의 중첩은 마르코프 변조 포아송 과정(MMPP: Markov Modulated Poisson Process)이 된다는 것은 이미 잘 알려져 있다[13]. 중첩과정을 모델링하는 MMPP를 표시하기 위해 Q를 MMPP의 도착과정의 천이를 나타내는 마르코프 과정의 무한소 생성자라하고, Λ를 도착률을 나타내는 대각행렬이라고 하면 MMPP는 (Q,Λ)로 완전히 표현되고 Q와 Λ는 다음과 같이 표시된다.

$$Q = Q_I, \quad \Lambda = \Lambda_I + \Lambda_2, \tag{6}$$

여기서 Λ_I는 IPP의 도착률 행렬이고 $\Lambda_2 = diag(\lambda_2, \lambda_2)$는 B형 고객의 포아송 도착률 λ_2을 나타내는 도착률 행렬이다.

이제 조건 $\{X_1 \le N_1\}$이 성립할 때, $q_{1,j} \equiv P(X_2 = j | X_1 \le N_1)$를 계산한다. 분명하게도 1)유휴 일반원이 있을 때는 두 종류의 고객이 선입선출 형태로 유휴 일반원에게 서비스를 받고, 2)모든 일반원이 서비스 중일 때는 우선순위 규칙에 따라 B큐에 대기 중인

B고객이 우선적으로 일반원에게 서비스 받게 된다.

먼저 1)의 경우, 안정상태 확률 $q_{1,j}$는 $MMPP/M/N_2/K_2^*$큐잉 시스템에서 도착하는 순간에 서비스 중인 일반원의 수가 j명일 확률로 근사할 수 있다. 여기서 K_2^*는 최소 K_2에서 최대 K_2+N_2까지 변할 수 있다. 중요한 계산요소인 평균 큐길이는 상담원에게 서비스 중인 B고객 수의 분포를 통하여 알 수 있다. 큐길이의 안정 상태 확률을 구하기 위해 $\{(X_2,Z)\}=\{(j,k)|1\le j\le K_2^*, k=1,2\}$를 $MMPP/M/N_2/K_2^*$ 큐잉 시스템의 상태를 나타내는 마르코프 과정이라 둔다. 여기서 $Z=k$는 무한소 생성자 Q를 갖는 MMPP입력과정의 상태를 나타내고, $X_2=j$는 시스템에 있는 A고객과 B고객의 수를 나타낸다. 확률과정 $\{(X_2,Z)\}$의 무한소 생성자를 Q^*라고 하면, 다음과 같다.

$$Q^*=\begin{pmatrix} Q_1 & Q_2 \\ O & Q_3 \end{pmatrix}, \tag{7}$$

여기서 O는 $(K-N_2-1)\times(N_2+1)$ 영행렬을 나타내고 Q_1, Q_2 및 Q_3는 다음 식들로 주어진다.

$$Q_1=\begin{bmatrix} Q_1(1) & \Lambda & \cdots & 0 & 0 \\ \mu_2 I & Q_1(2) & \cdots & 0 & 0 \\ 0 & 2\mu_2 I & \cdots & 0 & 0 \\ \cdots & \cdots & \ddots & \cdots & \cdots \\ 0 & 0 & \cdots & Q_1(N_2) & \Lambda \\ 0 & 0 & \cdots & N_2\mu_2 I & Q_1(N_2+1) \\ 0 & 0 & \cdots & 0 & c_1 I \end{bmatrix},$$

$$Q_2=\begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \cdots & \cdots & \ddots & \cdots & \cdots \\ 0 & 0 & \cdots & 0 & 0 \\ \Lambda_2 & 0 & \cdots & 0 & 0 \\ Q_3(1) & \Lambda_2 & \cdots & 0 & 0 \end{bmatrix},$$

$$Q_3=\begin{bmatrix} Q_3(2) & \Lambda_2 & \cdots & 0 & 0 \\ c_2 I & Q_3(3) & \cdots & 0 & 0 \\ 0 & c_4 I & \cdots & 0 & 0 \\ \cdots & \cdots & \ddots & \cdots & \cdots \\ 0 & 0 & \cdots & Q_3(K_2^*-N_2-1) & \Lambda_2 \\ 0 & 0 & \cdots & c_{K_2^*-N_2} I & Q-c_{K_2^*-N_2} I \end{bmatrix}. \tag{8}$$

위 행렬에서 $Q_1(i)=Q-\Lambda-(i-1)\mu_2 I$, $c_k=N_2\mu_2+k\theta_2$, $k=1,2,\ldots,K_2^*-N_2$, $Q_3(i)=Q-\Lambda_2-c_i I$이고 Q와 Λ및 Λ_2는 식 (5)와 (6)에 주어져 있다. 확률 $q_{1,j}$를 구하기 위해 $\pi Q^*=0$, $\pi e=1$을 만족하는 Q^*의 정상확률 벡터 π를 $j=0,1,2,\ldots,K_2^*$에 대해 다음과 같이 정의한다.

$$\pi=(\pi_0,\pi_1,\pi_2,\ldots,\pi_{K_2^*}),\ \pi_j=(\pi_{j1},\pi_{j2}). \tag{9}$$

구하려는 확률인 $q_{1,j}$는 임의 도착 순간에 서비스 중인 일반원의 수에 대한 안정상태 확률이므로 다음과 같은 식으로 주어진다[9].

$$q_{1,j}=\pi_{j2}/\sum_{k=0}^{K_2^*}\pi_{k2},\ j=0,1,\ldots,K_2^*. \tag{10}$$

Little의 정리에 의해 A큐로부터 넘쳐 일반원에게 서비스를 받는 A고객의 평균수는 다음과 같다.

$$N_A=\sum_{j=0}^{N_2-1}\pi_{j1}\lambda/\mu_2. \tag{11}$$

3.4 분할 영역 S_4에서 $p_{2,i}$의 계산

조건 $\{N_2\le X_2\le K_2^*\}$가 성립할 때 모든 일반원은 서비스 중이다. 일반원은 B고객을 서비스하거나 B큐가 비었다면 A고객을 서비스한다. A고객은 N_1명의 전담원에게 개별 처리율 μ_1을 갖고 안정적으로 서비스를 받는다. A전담원이 모두 서비스 중이면, B큐가 빌 경우($X_2=N_2$)에만 일반원에게 서비스를 받는다. 이 경우 A고객은 일반원에게 서비스 받을 수 있고 A큐를 지나 일반원 그룹으로 라우팅 된다.

한편, 도착하는 B고객은 먼저 N_2명의 일반원에게 개별 처리율 μ_2을 갖고 안정적으로 서비스를 받는다. 일반원이 모두 서비스 중이면 B큐에 대기한다. A고객이 일반원에게 서비스 가능할 때는 처리율 $N_2\mu_2$를 갖는 단일 서버 시스템으로 이용가능하고, 서비스 불가능한 시간은 $M/M/1/K_2^*+M$ 큐의 서비스 기간인 바쁜 주기(busy period)에 해당한다. 여기서 큐의 크기 K_2^*는 일반원에게 서비스 받는 A고객의 수에 따라 달라지는 확률변수이다. 바쁜 주기의 정확한 해석을 위해서는 확률변수 K_2^*의 정확한 분포를 알아야 한다. $M/M/1/K_2^*+M$ 큐의 바쁜 주기는 2개의 지수 확률변수의 조합으로 이루어지는 초지수(hyper- exponental) 함수로 근사된다[7].

바쁜 주기를 계산하기 위해 $L(t)$를 시각 0에서 시작하여 시각 t에 시스템에 있는 고객의 수라하고 τ를 첫 바쁜 주기의 길이라고 하면 먼저 다음의 Laplace 변환을 구해야 한다.

$$\phi_n(s)=E[e^{-s\tau}|L(0)=n],\ n=1,2,\cdots,K,\ |x|\le 1,\ s>0, \tag{12}$$

여기서 경계치 조건은 $\phi_{K+1}(s)=\phi_K(s)$, $\phi_0(s)\equiv 1$ 이다. 결국 $\phi_1(s)$가 구하려는 Laplace 변환이며 [14]을 이용하여 다음과 같이 구할 수 있다.

$$\phi_1(s)=\frac{c_2N_2\mu_2f_z(K)-N_2\mu_2\lambda_2f_z(K-1)}{c_2^2f_z(K+1)-c_2(\lambda_2+\theta_2)f_z(K)+\lambda_2\theta_2f_z(K-1)}, \quad (13)$$

여기서 $f_z(K)=z_1^K-z_2^K$이고 $c_2=N_2\mu_2+\theta_2$이며 z_1과 z_2는 복호동순으로 다음과 같다.

$$z_1,z_2=\frac{(\lambda_2+c_2+s)\pm\sqrt{(\lambda_2+c_2+s)^2-4\lambda_2c_2}}{c_2}. \quad (14)$$

바쁜 주기의 분포를 초지수 분포로 근사하기 위해 $\phi_1(s)$의 첫 세 모멘트를 m_1, m_2와 m_3라 하고, τ를 바쁜 주기의 길이라 하고 다음과 같이 초지수 함수를 정의한다.

$$h(\tau)=\alpha_2\gamma_1e^{-\gamma_1\tau}+(1-\alpha_2)\gamma_2e^{-\gamma_2\tau}, \quad (15)$$

여기서 τ, α, γ_1과 γ_2는 음수가 아니며 바쁜 주기의 첫 세 개의 모멘트와 초지수함수의 첫 세 모멘트는 다음 식들로 잘 대응된다[9].

$$\gamma_1,\gamma_2=\frac{v_1\pm\sqrt{v_1^2-4v_2}}{2},\ \alpha_2=\frac{\gamma_1(1-\gamma_2m_1)}{\gamma_1-\gamma_2}, \quad (16)$$

여기서 v_1과 v_2는 다음과 같이 주어진다.

$$v_2=\frac{6m_1^2-3m_2}{(3/2)m_2^2-m_1m_3},\ v_1=m_1^{-1}+\frac{m_2v_2}{2m_1}. \quad (17)$$

다시 $p_{2,i}=P\{X_1=i|N_2\le X_2\le K_2\}$의 계산을 위해 두 조건을 고려한다. 1) 첫째, A큐에 대기가 있을 때는 B큐가 비었을 때($X_2=N_2$)에만 포아송 입력율 λ_1과 식 (15)로 주어지는 초지수함수 서비스 시간과 서비스율 $\gamma=\alpha_2\gamma_1+\alpha_2^c\gamma_2$를 갖는 $M/G_1/1/(K_2^*-N_1+1)+M$ 큐와 $M/M/1/(K_2^*-N_1+1)+M$큐 둘 다에 지배를 받는다. 2) 둘째, A고객은 포아송 입력율 λ_1과 지수함수 인내 시간 후 개별 중도 포기율 θ_1을 갖고, 지수함수 서비스 시간과 개별 서비스율 μ_1을 갖는 N_1명의 A전담원에게 안정적으로 서비스를 받는 $M/M/N_1/N_1$ 큐에 지배를 받는다.

1)의 경우, $X_1=i>N_1$이면 A전담원은 지수분포[$\exp(N_1\mu_1)$]서비스 시간을 갖고 A고객을 서비스한다. 중도포기까지 고려하면 $B_1\sim\exp(N_1\mu_1+\theta_1)$의 서비스 시간을 갖고 시스템을 떠난다고 할 수 있다. 일반원은 초지수 분포[$H\sim h(x)$;식 (15)]로 A고객을 서비스한다. 그러므로 A고객은 지수 분포(B_1)와 초지수 분포(H) 중 더 짧은 서비스 시간[$Min(B_1,H)$] 동안 서비스를 받고 시스템을 떠나므로 $M/G_2/1/(K_2^*-N_1+1)+M$에 지배를 받는다.

임의의 순간에 시스템에 있는 고객 수에 대한 확률은 PASTA(Poisson Arrivals See Time Average) 정리에 의해 고객 도착 순간 시스템 상태 확률과 같으므로 결국 확률 $p_{2,i}$, $i=1,\cdots,K_2^*-N_1-1$에 대해 다음 결과를 얻는다[15].

$$p_{2,i+N_1}=\frac{\pi_{i+1}}{\pi_0+\lambda_1/(\gamma+\theta_1+N_1\mu_1)},$$

$$p_{2,K}=1-\frac{1}{\pi_0+\lambda_1/(\gamma+\theta_1+N_1\mu_1)}, \quad (18)$$

여기서 $\gamma=\alpha_2\gamma_1+(1-\alpha_2)\gamma_2$이고 γ_1, γ_2 및 α_2는 식 (16)에 주어져 있다.

2)의 경우, $\{X_1\le N_1\}$이면 대응되는 큐잉 모델은 $M/M/N_1/N_1$이다. $L=\{L(t),t\ge 0\}$는 마르코프 생성-소멸 과정이다. 그러므로 구하려는 확률분포 $p_{2,i}$는 이 생성-소멸 과정의 안정상태 확률분포이다. 위 식 (18)과 정규화 조건으로부터 $p_{2,i}$는 다음 식으로 주어진다.

$$p_{2,i}=\frac{1}{i!}\left(\frac{\lambda_1}{\mu_1}\right)^i\left(1-\sum_{i=N_1+1}^{K_2^*}p_{2,i}\right)\Big/\sum_{i=0}^{N_1}\frac{1}{j!}\left(\frac{\lambda_1}{\mu_1}\right)^j \quad (19)$$

3.5 성 측도들의 계산

지금까지 구한 $\{p_{1,i}\},\{p_{2,i}\},\{q_{1,j}\}$ 및 $\{q_{2,j}\}$와 조건부 확률을 이용하여 $P(X_1=i)$, $i=0,1,\cdots,K_1^*$와 $P(X_2=j)$, $j=0,1,\cdots,K_2^*$를 얻는다. 이들을 이용하여 각종 시스템 성능 측도들을 구할 수 있다. A큐와 B큐의 평균 큐 크기는 식 (11)에 의해서 다음 식으로 주어진다.

$$K_A=K_1-N_1-N_A,K_B=K_2-N_2+N_A. \quad (20)$$

A고객과 B고객의 블록킹 확률들은 각기 다음과 같이 주어진다.

$$P_A=(X_1=K_A+N_1),P_B=(X_2=K_A+N_2). \quad (21)$$

A큐와 B큐에서 각각의 평균 대기시간은 Little의

공식에 의해 다음과 같다.

$$W_{qA} = \frac{1}{\lambda_1(1-P_A)} \sum_{i=N_1+1}^{K_A+N_1} (i-N_1)P(X_1=i),$$

$$W_{qB} = \frac{1}{\lambda_2(1-P_B)} \sum_{j=N_2+1}^{K_A+N_2} (j-N_2)P(X_2=j), \qquad (22)$$

4. 수치계산 결과

이 절에서는 N-설계 콜센터의 시스템 매개변수들의 변화에 따른 지연 및 블록킹 확률등의 변화에 대하여 알아본다. A와 B 두 종류 고객을 위해 독립적으로 시설된 전화 회선수를 각각 K_1 =70과 K_2 =50으로 고정한다. 또한, A전담원의 수는 N_1 =30명, 일반원의 수는 N_2 =40명으로 고정한다. A고객과 B고객의 도착률 λ_1과 λ_2도 적절한 트래픽 이용도를 얻기 위해 변화한다. A전담원은 일반원보다 숙련된 상담기술을 가지고 있으며 두 유형 고객의 서비스 시간 μ_1과 μ_2는 분당 1에서 1/3까지 변한다. 일반적으로 지수분포 인내시간의 평균 θ_1^{-1}와 θ_2^{-1}는 120~240초 사이에서 선택한다[1,3,5]. 아래 그림 2부터 그림 5까지에서 A전담원과 일반원에 대한 지수분포 평균 서비스 시간은 각각 μ_1^{-1} =2분과 μ_2^{-1} =3분이다. 지수분포 인내시간의 평균은 θ_1^{-1} =2분과 θ_2^{-1} =4분으로 택하기로 한다. 블록킹 확률 P_A와 P_B 및 평균대기 시간 W_{qA}와 W_{qB}는 식 (21)과 (22)에 나타나 있다.

그림 2는 A고객의 도착률이 분당 3명에서 26명까지 변할 때 일반원을 점유하는 A고객의 평균수의 변화를 나타낸다. 그림의 범례에서 'lam2'는 B고객의 B큐에 도착률을 의미하며 4, 8, 12, 16명의 4가지 고정된 각각의 값에 대하여 하나의 곡선이 대응된다.

예를 들어 'lam2=8'일 때 A고객의 도착률이 분당 20명(가로축)이면 일반원을 점유하는 A고객의 평균수는 15개(세로축) 이므로 40명의 일반상담원(N_2 = 40) 중에서 평균적으로 15명은 A고객을 서비스하고 나머지 25명의 일반원이 B고객을 서비스 하게 되는 것이다. 이 경우, 식 (20)에서 보듯이, A고객의 연결을 유지하는 회선수는 K_1 =70이고 A전담원은 N_1 = 30명이므로 A큐의 평균 크기는 K_A =25(=70-30-15)가 되고, B고객의 연결을 유지하는 회선수는 K_2 =50이고 일반원은 N_2 =40명이므로 B큐의 평균 크기는 K_B =25(=50-40+15)이 된다.

그림 3은 A고객의 도착률이 변할 때 A고객의 블록킹 확률의 변화를 나타내는 그림이다. 4가지 고정된 B고객의 도착률 각각에 대해 A고객의 도착률이 커지면서 A고객의 블록킹 확률은 지수적으로 커진다는 것을 알 수 있다. 특이한 점은 B고객의 도착률이 큰 경우(lam2=12, 16)에 A고객의 블록킹 확률은 비슷하다. 이것에 대한 이유는 B고객의 도착률이 클 경우 일반원을 점유하는 A고객의 평균수가 적어짐에 따라 A큐의 평균 크기가 커지면서 A고객의 블록킹 확률이 작아지게 되기 때문이다.

그림 4는 A고객의 도착률이 변할 때 B고객의 B큐에서 블록킹 확률의 변화를 나타내는 그림이다. A고객의 도착률이 B고객의 블록킹 확률에 거의 영향을

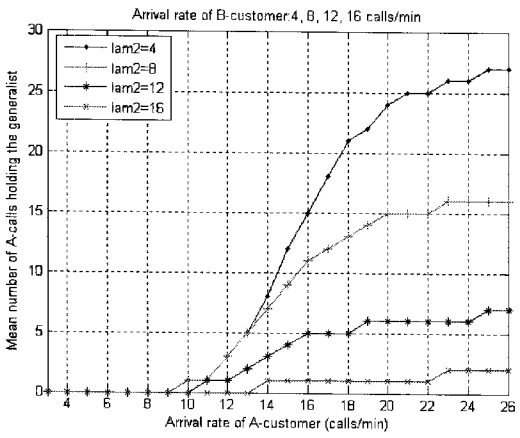

그림 2. A고객 도착율 대 일반원 점유 A고객 수

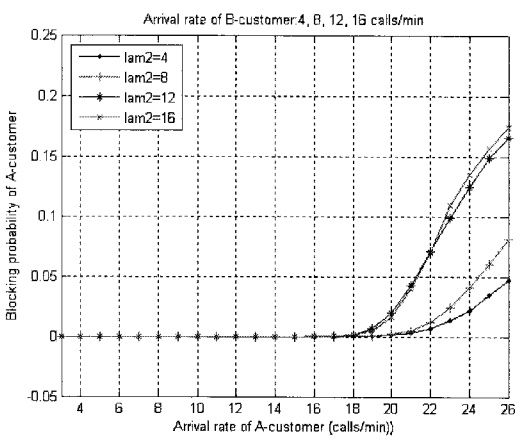

그림 3. A고객 도착율 대 블록킹율 P_A

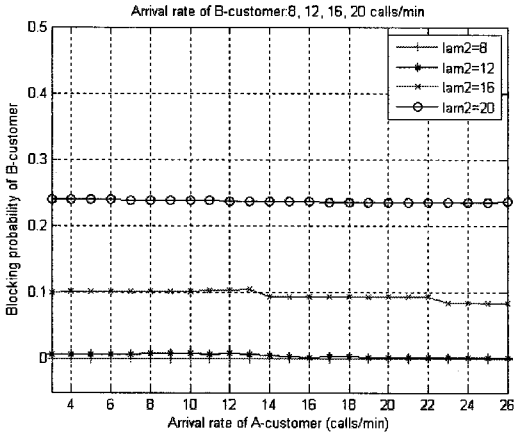

그림 4. A고객 도착율 대 블록킹율 P_B

미치지 못한다는 것을 알 수 있다. 이것은 B고객의 도착률이 고정되어 있을 때 A고객의 도착률이 커짐에 따라 일반원을 점유하는 A고객의 평균수는 증가하게 되고 이 증가분만큼 B큐의 평균 크기는 커지면서 B큐에서 B고객의 블록킹 확률은 상대적으로 작아지기 때문이다.

그림 5는 A고객의 도착률이 변할 때 A고객의 A큐에서 평균 대기시간의 변화를 나타내는 그림이다. B고객의 도착률이 낮을 때(lam2=4, 8)는 A고객의 도착률이 증가 할 때 A큐에서 A고객의 대기 시간은 지속적으로 늘어난다. 그러나 B고객의 도착률이 높을 때(lam2=12, 16)는 어느 한 정점('lam2=12'일 경우 A고객의 도착률이 24명인 점) 이후에는 오히려 대기시간이 줄어드는 것을 볼 수 있다. 이것은 정점을 경계로 대기 큐의 한계에 도달하고 블록킹이 발생하는 시점이며 대기 인내 시간 한계에 도달할 기회가 더 많아지는 것으로 볼 수 있기 때문이다. 여기서 부터는 4가지 다른 A고객의 도착률은 고정하고 B고객의 도착률을 변경하는 그림들을 다루기로 한다. 그림 6부터 그림 9까지에서 $\mu_1^{-1}=2$, $\mu_2^{-1}=3$, $\theta_1^{-1}=2$, $\theta_2^{-1}=4$ 분으로 택하기로 한다.

그림 6은 B고객의 도착률이 3부터 20까지 변할 때 A고객의 블록킹 확률의 변화를 나타내는 그림이다. A고객의 도착률이 낮을 때(lam1=10, 15), A고객의 블록킹 확률은 B고객의 도착률에 크게 영향을 받지 않는 다는 것을 확인할 수 있다. 그러나 A고객의 도착률이 높을 때(lam1=20, 25)는 A고객의 블록킹 확률은 특정한(B고객의 도착률이 13인) 값까지 높아지고 이후 변화가 거의 없음을 알 수 있다. 이유는 B고객의 도착률이 클 경우 일반원을 점유하는 A고객의 평균수가 적어짐에 따라 A큐의 평균 크기가 커지면서 A고객의 블록킹 확률이 작아지게 되기 때문이다.

그림 7은 B고객의 도착률이 변할 때 B고객의 블록킹 확률을 나타내는 그림이다. B고객의 도착률이 가로축을 따라 커질 때 곡선 각각에 대한 B고객의 블록킹 확률은 지수적으로 증가하는 것을 알 수 있다. 또한, A고객의 도착률은 B고객의 블록킹 확률에 영향을 미치지 않는 다는 것을 볼 수 있다.

그림 8은 B고객의 도착률이 변할 때 A큐에서 대기하는 A고객의 평균 대기시간의 변화를 나타내는 그림이다. B고객의 도착률이 증가할 때 A고객의 A큐에서 대기시간도 증가한다. 특히 B고객의 도착률이 14보다 클 때 A고객의 대기시간은 큰 변화가 없다

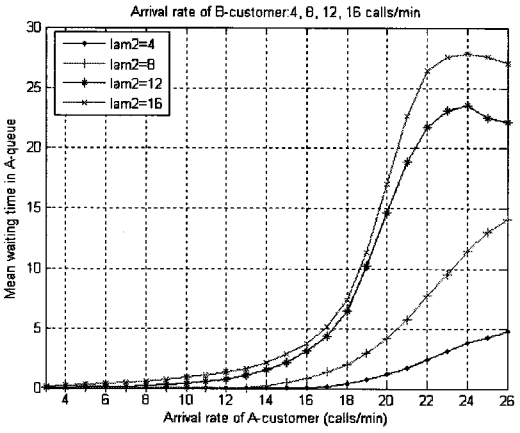

그림 5. A고객 도착율 대 대기시간 W_{qA}

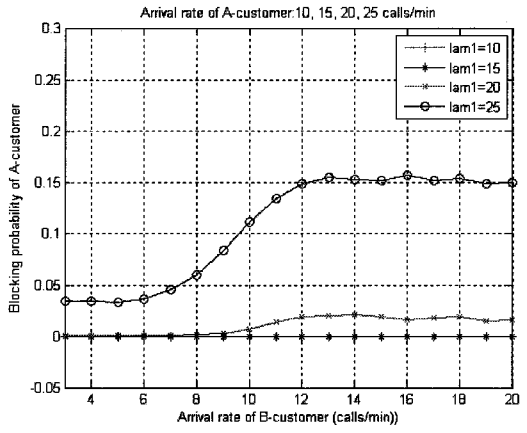

그림 6. B고객 도착율 대 블록킹율 P_A

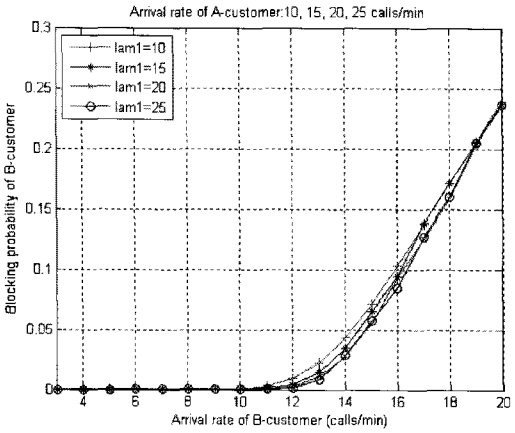

그림 7. B고객 도착율 대 블록킹율 P_B

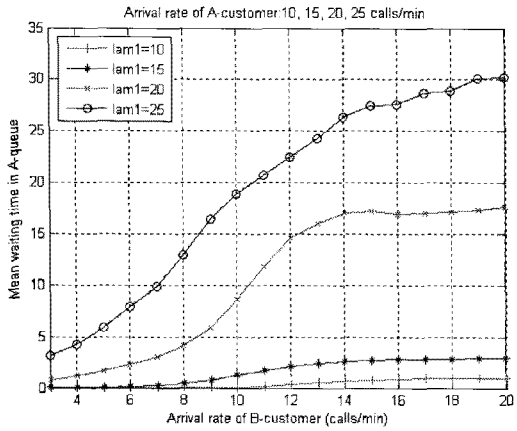

그림 8. B고객 도착율 대 대기시간 W_{qA}

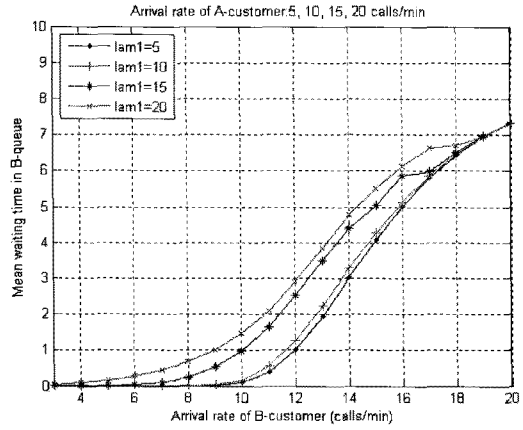

그림 9. B고객 도착율 대 대기시간 W_{qB}

는 것을 볼 수 있다. 이에 대한 이유는 그림 6에서와 같다.

그림 9는 B고객의 도착률이 변할 때 B고객의 B큐에서 평균 대기시간의 변화를 나타내는 그림이다. B고객의 도착률이 증가할 때 B고객의 대기 시간이 증가한다. A고객의 입력률이 낮을 때 B고객의 대기시간에 영향이 거의 없음을 알 수 있다.

4. 결 론

본 논문에서는 N-설계 모델에 대한 시스템의 상태 공간을 여러 개의 하위 공간으로 나누는 소위 분할방법을 이용한 근사적 분석 기법을 다루었다. 본 논문에서 다룬 N-설계 모델은 분석 결과가 비교적 쉽게 얻어지는 무한대기 공간을 갖는 기존의 모델과는 달리 두 개의 큐 모두 유한 대기 공간을 가지며 또한 큐에서 대기고객은 지수분포 인내시간을 가진 후 중도포기 하는 모델이다. 정확한 확률과정론적 접근법으로 N-설계 콜센터를 분석한다면 상태공간의 폭발적 증가로 수치계산이 불가능하다는 사실은 이미 알려져 있다. 본 논문에서 논의한 근사 시스템의 상태 공간의 크기는 서버의 수가 증가함에 따라서 천천히 증가하는 근사기법으로 이 문제를 해결하고 성능척도들을 구했다. 수치계산 결과로 고객의 도착률이 변할 때 일반원을 점유하는 A고객의 평균수와 두 유형의 고객호의 블록킹 확률과 대기시간 등의 변화 경향을 살펴보았다. 근사방법의 정확도와 고객의 인내시간에 대한 시스템의 영향 및 상담원의 최적 배치전략은 추후과제로 남겨둔다.

참 고 문 헌

[1] A. Mandelbaum and S. Zeltyn, "Service Engineering in Action: The Palm/ Erlang-A Queue, with Applications to Call Centers," *Teaching note to Service Engineering course*, Technion-Israel Institute of Technology, 2005.

[2] S. Borst, A. Mandelbaum, M. and I. Reiman, "Dimensioning Large Call Centers," *Operations Research*, Vol.52, No.1, pp. 17-34, 2004.

[3] R. Stolletz and S. Helber, "Performance analysis of an inbound call center with skills-based routing," *OR Spectrum*, Vol.26, pp. 331-352, 2004.

[4] M. Shimkin, A. and Mandelbaum, "Rational Abandonment from Tele-Queues: Nonlinear Waiting Costs with Heterogeneous Preferences," *Queueing Systems*, Vol.47, pp. 117-146, 2004.

[5] A. Mandelbaum and S. Zeltyn, "The impact of customer's patience on delay and abandon ment: some empirically-driven experiments with the M/M/n+G queue," *OR Spectrum*, Vol.26, pp. 377-411, 2004.

[6] N. Gans, G. Koole and A. Mandelbaum, "Commissioned Paper, Telephone Call Centers: Tutorial, Review, and Research Prospect," *Manufacturing & Science Operations Management*, Vol.5, No.2, pp. 79-141, 2003.

[7] R. A. Shumsky, "Appoximation and analysis of a call center with flexible and specialized servers," *OR Spectrum,* Vol.26, pp. 307-330, 2004.

[8] S. Helber and K.-H. Waldmann, *Call Center Management, Vol.26,* OR Spectrum, Springer-Verlag, 2004.

[9] D. Gross and C. H. Harris, *Fundamentals of Queueing Theory*, John Wiley & Sons, Inc. 1985.

[10] A. Kukzura, "The interrupted poisson process as an overflow process," *Bell System Technical Journal*, Vol.52, No,3, pp. 437-448, 1973.

[11] R. O. Onvural, *Asychronous Transfer Mode Networks: Performance Issues, Second edition*, Artech House, 1995.

[12] H. Heffes and D. M. Lucantony, "A Markov modulated characterization of packetized voice, data traffic and related statistical multiplexer performance," *IEEE J. Selected Areas Comm.,*. Vol.4, No.6, pp. 856-868, 1986.

[13] K. S. Meier-Hellstern, "The Analysis of a Queue Arising in Overflow Model," *IEEE Trans. on Comm.,* Vol.37, pp. 367-372, 1989.

[14] W. Stadje, "The busy periods of some queueing systems," *Stochastic processes and their Applications*, Vol.55, pp. 159-167, 1995.

[15] H. Tagaki, *Queueing Analysis, Vol. 2: Finite Systems*, IBM Japan, Ltd., North-Holland, 1993.

박 철 근

1983년 2월 부산대학교 수학과 이학사
1986년 2월 한국과학기술원 응용수학과 이학석사
1992년 3월~1995년 8월 한국과학기술원 수학과 박사
1996년 4월~1997년 2월 한국통신 통신망연구소, 교환기술연구소 선임연구원(팀장)
1997년 3월~현재 선문대학교 정보통신공학부, 교수
관심분야 : 트래픽공학, 통신망해석, 큐잉이론

성 수 학

1982년 경북대학교 수학과 학사
1985년 한국과학기술원 응용수학과 이학석사
1988년 한국과학기술원 응용수학과 이학박사
1991년 한국전자통신연구원 연구원
1991년 9월~현재 배재대학교 전산수학과, 교수
관심분야 : 확률극한, 암호이론

정 해

1987년 2월 한양대학교 전자통신공학과 학사
1991년 2월 한국과학기술원 전기 및 전자공학과 석사
1986년 2월 한국과학기술원 전기 및 전자공학과 박사
1991년 3월~1998년 7월 LG정보통신 선임연구원
1998년 8월~현재 금오공과대학교 전자공학부 부교수
2004년 1월~2005년 1월 Univ. of Texas at Dallas 방문교수
관심분야 : 가입자 액세스망, BcN, PON, Wibro