韓國經營科學會誌
第32권 第3號
2007年 9月

이중작업능력의 서버로 구성된 서비스시스템 설계*

김 성 철**

A Design Problem of a Service System with Bi-functional Servers*

Sung Chul Kim**

■ Abstract ■

In this paper, we consider a service system with bi-functional servers, which can switch between the primary service room and the secondary room. A service policy is characterized by the switching points which depend on the queue length in the primary service room and the service level requirement constraint of the secondary room. The primary service room is modeled as a Markovian queueing system and the throughput of the primary service room is function of the total number of bi-functional servers, the buffer capacity of the primary service room, and the service policy. There is a revenue obtained from throughput and costs due to servers and buffers. We study the problem of simultaneously determining the optimal number of servers, buffer capacity, and service policy to maximize profit of the service system, and develop an algorithm which can be successfully applied with the small number of computations.

Keyword : Optimization Problem, Bi-functional Worker, Queueing System, Nonlinear Interger Programming, Second Order Property, Marginal Analysis

1. 서 론

본 논문은 이중작업능력(bi-functional)을 갖는 서버(server)들로 구성된 서비스시스템의 설계에 관한 문제를 다룬다. 이중작업능력을 갖는 서버란 주영역과 보조영역에서 서로 다른 두 종류의 작업

논문접수일 : 2007년 05월 24일 논문게재확정일 : 2007년 07월 02일
* 본 연구는 2007학년도 덕성여자대학교 연구비 지원으로 이루어 졌음.
** 덕성여자대학교 경영학과

을 수행할 수 있는 교차훈련된(cross-trained) 서버를 의미한다. 이러한 이중작업능력을 갖는 시스템은 서비스시스템, 제조시스템, 통신시스템, 그리고 설비보전시스템 등 다양하게 제시될 수 있다. 그 중 한 예로서 소매상의 예를 보기로 하자. 소매상은 주영역으로 고객을 맞이하는 매장과 보조영역으로 재고를 유지하는 창고로 구성되어 매장과 창고에서 주어진 일을 모두 담당할 수 있는 교차훈련된 종업원은 주영역인 매장에서는 고객을 맞이하고 제품을 판매하며 보조영역인 창고에서는 상품을 입출고하고 분류 정리한다. 매장은 도착하는 고객에 대하여 서비스를 효율적으로 제공할 수 있도록 설계하는 것이 매우 중요하며 매장에는 고객이 많은 경우에는 많은 종업원이 배치되고 고객이 적은 경우에는 적은 수의 종업원이 배치되며 매장에 배치되지 않은 종업원은 창고에서 재고와 관련된 작업을 수행한다. 제조시스템의 경우에 주영역은 시장이나 상위단계(upstream)의 수요를 처리하는 제조공간(shop)을 보조공간은 제조에 소요되는 재료나 부품의 재고를 담당하는 창고를, 통신시스템의 예로서 저장전송(store & forward)의 경우를 보면 주영역은 통신업무를 보조업무는 여타의 정보처리를, 그리고 설비보존시스템의 경우에는 주영역은 고장난 설비의 교정보전을 보조영역은 예방보전의 경우로 설명할 수 있다. 이러한 다양한 서비스 메커니즘(mechanism)을 갖는 서비스시스템, 제조시스템, 통신시스템 등의 모든 시스템을 총칭하여 본 논문에서는 서비스시스템으로 명명하기로 한다. 그러므로 이러한 서비스시스템의 주영역은 수요종속적인 서비스능력을 갖는 대기행렬시스템(queueing system)으로 모형화되고 수요의 도착과정과 서비스과정에 의하여 수행도가 결정되며 최적화 모형이 수립될 수 있다. 이에 반하여 보조영역은 시간적으로 매장보다 훨씬 민감하지 않은 일상적인 업무가 수행되며 주어진 업무를 수행할 수 있는 일정수준의 서버는 주어진 최적화 모형에서 제약조건으로 모형화 될 수 있을 것이다.

만약 주어진 서비스시스템에 s개의 이중작업능력의 서버가 존재하고 서비스를 요하는 고객의 수가 $k_{n-1}+1$과 k_n사이에 있는 경우에는 주영역에 n $(1 \le n \le s)$개의 서버를 배치하는 서비스정책을 실행한다고 하자. k_s는 시스템 내에 존재 가능한 최대 고객수로 이를 c, $s \le c$,라고 하면 이는 주어진 서비스시스템에 있어서 주영역의 대기능력(buffer capacity)이 된다. 그러므로 본 논문에서는 서비스시스템의 총 서버의 수 s, 주영역의 대기능력 c, 그리고 주영역에서의 서비스정책 $K=(k_0, k_1, \cdots, k_s)$, 즉 서버의 교체점(switching point) k_n, $n=0, \cdots, s$,를 동시에 결정하는 대기행렬시스템의 최적화 문제를 제시하고자 한다. 이를 부연설명하면 서버나 대기능력을 확보하기 위한 비용(cost)이 소요되며 반면에 확보된 총 서버의 수와 주영역의 대기능력, 그리고 서비스정책의 함수로서 생산율(throughput)과 생산율에 따른 수익(revenue)이 발생하며 수익과 비용이 통합적으로 고려되어 주어진 서비스시스템의 이익(profit)이 산정된다. 반면에 보조영역에서 필요로 하는 작업을 수행하기 위하여 보조영역에서 요구되는 서버는 그 수의 기대치(expected value)의 하한(lower bound), B_L,이 존재하며 이는 주어진 최적화문제에 있어서 제약조건(constraint)이 된다. 그러므로 주어진 최적화문제는 보조영역의 서비스능력과 관련된 제약식을 만족시키면서 이익이 최대가 되도록 서비스시스템을 설계하는 것이다.

이중작업능력을 갖는 서비스시스템에 있어서 최적 서비스정책을 결정하는 최적화문제로서 Berman, et al.[3]이 본 논문과 가장 관련이 되는 논문이다. Berman, et al.[3]은 일정 수로 주어진 총 서버와 주영역의 대기능력을 갖는 서비스시스템에서 보조영역의 서비스능력과 관련된 제약식을 만족시키면서 주영역에서 고객의 평균대기시간(waiting time)을 최소화하는 서비스정책을 결정하는 최적화문제를 다루었다. Berman and Larson[1]은 또한 주영역과 보조영역사이에서 서버가 교체됨에 따른 비용을 수반하는 경우로 주어진 최적화문제를 연장하였으며 Berman and Sapna[2]는 주어진 최적화문제를 주영역과 보조영역의 작업들 간에 상관관계가

존재하는 경우로 연장하였다.

서버의 수나 대기능력을 설계하는 문제는 제조시스템이나 컴퓨터시스템의 최적화를 위하여 하나의 작업장(station)으로 구성된 대기시스템보다는 대기네트워크의 모형화를 통하여 다양하게 제시되었다. 그 중 본 논문과 가장 관련이 되는 논문은 Shanthikumar and Yao[9]를 들 수 있다. Shanthikumar and Yao[9]에서는 각 작업장이 제한된 대기능력을 갖는 폐쇄대기네트워크(closed queueing network)에 있어서 생산율의 증가하는 오목성(increasing concavity)을 이용하여 서버와 대기능력을 배분하는 문제를 다루었다. Wein[10]은 서비스능력의 설계에 따른 예산 제약이 존재할 때 일반화된 Jackson 대기네트워크에 존재하는 고객의 대기시간을 최소화하도록 주어진 서비스능력을 각 작업장에 배분하는 최적화문제를 다루었다. 제조시스템에 있어서 서비스능력이나 대기능력의 설계에 관한 최적화문제는 Buzacott and Shanthikumar[4]에 다양하게 제시되고 있다.

제 2장에서는 주영역과 보조영역으로 구성된 서비스시스템에 있어서 주어진 보조영역의 서비스능력을 만족시키면서 최적의 서비스정책, 서버의 수, 그리고 대기능력을 설계하는 최적화문제를 모형화한다. 이를 위하여 주어진 대기시스템의 상태확률과 설계모수들의 수행도가 정의된다. 제 3장에서는 주어진 최적화문제의 최적화절차를 도출하는데 요구되는 설계모수들에 대한 수행도의 일계특성과 이계특성이 도출된다. 제 4장에서는 주어진 최적화문제의 최적화절차가 유도되며 제 2장에서 논의된 설계모수들에 대한 수행도의 일계특성과 이계특성이 중요한 기반이 된다. 제 5장에서는 주어진 최적화문제의 이해를 돕기 위하여 최적화 단계에 따른 수치예가 제시된다. 제 6장에서는 결어로서 마감한다.

2. 모형화

먼저 이중작업능력을 갖는 총 서버의 수가 s, 주영역의 대기능력이 c, 그리고 서비스정책 $K=(k_0, k_1, \cdots, k_s)$가 주어져 있을 때 주어진 대기시스템의 수행도를 산정하는 모형을 다루기로 한다. 고객이 주영역에 도착하는 과정은 기대치 λ인 포아송(Poisson)분포에 의하며 주영역에서의 서버의 서비스시간은 서비스율(service rate)이 μ인 지수분포를 갖는다. 이제 $\rho=\lambda/\mu$라 하면 ρ는 주영역에 제공된 부하(offered load)로서 부여된 일의 양을 의미한다.

총 서버의 수 s와 주영역의 대기능력 c가 주어졌을 때 서비스정책 $K=(k_0, k_1, \cdots, k_s)$, $|k_n - k_{n-1}| \geq 1$, $k_0 \geq 0$, 가 주어지면 주어진 대기시스템의 상태확률(state probability) $p_{s,c,K}(x)$, $x=0, 1, \cdots, c$,는 다음과 같다.

$$p_{s,c,K}(x)=\beta_{s,c,K}(x)p_{s,c,K}(k_0),\ k_{n-1}<x\leq k_n,\quad 1\leq n\leq s,$$
$$p_{s,c,K}(k_0) = \frac{1}{\sum_{x=k_0}^{k_s}\beta_{s,c,K}(x)}. \tag{2.1}$$

여기에서 $\beta_{s,c,K}(x) = (\frac{\rho}{n})^{x-k_{n-1}}(\frac{\rho}{n-1})^{k_{n-1}-k_{n-2}}\cdots(\frac{\rho}{1})^{k_1-k_0}$, $k_{n-1}<x\leq k_n$, 를 의미한다. $c=k_s$이므로 $p_{s,c,K}(k_s)=p_{s,c,K}(c)$는 봉쇄(blocking)확률로서 주어진 대기시스템의 생산율을 결정하는 중요한 모수(parameter)가 된다.

주영역에 있는 서버의 수의 기대치를 $F_{s,c,K}$라 하면 보조영역에 있는 서버의 기대치 $B_{s,c,K}=s-F_{s,c,K}$가 된다.

$$F_{s,c,K} = \sum_{n=1}^{s}\sum_{x=k_{n-1}+1}^{k_n} np_{s,c,K}(x). \tag{2.2}$$

주어진 대기시스템의 생산율(throughput)을 $TH_{s,c,K}$라고 하면 이는 다음과 같이 정리될 수 있다.

$$TH_{s,c,K} = \sum_{n=1}^{s}\sum_{x=k_{n-1}+1}^{k_n} n\mu p_{s,c,K}(x) = \lambda\{1-p_{s,c,K}(c)\}. \tag{2.3}$$

이제 주어진 서비스시스템의 생산율, $TH_{s,c,K}$의 함수로 표시되는 수익함수를 $f(TH_{s,c,K})$, 총 서버 s를 확보하는데 소요되는 비용함수를 $h(s)$, 주영역의 대기능력 c를 확보하는데 소요되는 비용함수를 $g(c)$라고 정의하자. 여기에서 서버와 대기능력 확보와 관련되는 비용은 주어진 대기시스템 생산율과 일치하도록 단위기간 당 비용으로 치환된 금액을 의미한다. 이제 의사결정 변수로서 총 서버의 수 s, 주영역의 대기능력 c, 그리고 서비스정책 K에 의하여 정의되는 이익함수를 $\Phi(s,c,K)$라고 정의하면 $\Phi(s,c,K)$를 최대화시키는 최적화문제는 다음과 같은 비선형(nonlinear)의 정수계획문제(integer programming)로 정식화될 수 있다.

$$\underset{\substack{1\le s\le c\\ K}}{Max.}\ \Phi(s,c,K)=f\{\lambda[1-p_{s,c,K}(c)]\}-h(s)-g(c)$$

$$s.t. \qquad B_{s,c,K}\ge B_L. \tag{2.4}$$

식 (2.4)의 첫 줄은 서버 및 대기능력의 확보로 산정되는 수익함수와 서버 및 대기능력의 배분에 따른 비용함수의 차로 산정되는 이익함수를 정식화한 것이며 둘째 줄은 보조영역에서 요구되는 서비스능력의 하한 B_L을 제약식으로 나타낸 것이다.

이제 수익함수 $f\{TH_{s,c,K}\}$가 생산율 $TH_{s,c,K}=\lambda\{1-p_{s,c,K}(c)\}$에 대하여 증가하는 오목(concave)함수이고 비용함수 $h(s)$와 $g(c)$가 각각 총 서버 s와 대기능력 c에 대하여 증가하는 볼록(convex)함수라고 하자. 만약 총 서버 s와 대기능력 c에 대하여 생산율 $TH_{s,c,K}$가 증가하는 오목함수이면 수익함수 $f(TH_{s,c,K})$는 총 서버 s와 대기능력 c에 대하여 증가하는 오목함수가 되며 비용함수 $h(s)$와 $g(c)$가 각각 총 서버 s와 대기능력 c에 대하여 증가하는 볼록함수이므로 이익함수 $\Phi(s,c,K)$는 오목함수가 되어 주어진 최적화문제는 한계분석법(marginal analysis)[6]을 통하여 용이하게 해결될 수 있을 것이다. 그러므로 설계모수에 대하여 수익함수의 일계특성과 이계특성을 도출하는 일은 매우 중요한 일이라 할 수 있다.

3. 수행도

서비스정책 $K=(k_0, k_1,\cdots,k_s)$는 총 서버의 수 s와 주영역의 대기능력 c에 따라서 다양하게 제시되며 서로 다른 수행도를 갖는다. 그러므로 먼저 총 서버의 수 s와 주영역의 대기능력 c가 일정하게 주어진 경우 가장 특수하게 주어지는 두 서비스정책 $K_L=(c-s, c-s+1,\cdots,c)$와 $K_U=(0,1,\cdots,s-1,c)$의 경우를 보기로 하자. 서비스정책 K_L의 경우에는 주영역의 서비스시스템의 상태(state)가 $c-s$에서 c사이에서 정의되고 $M/M/s/s$ 대기시스템으로 모형화되어 결과적으로 주영역에 존재하는 서버의 수의 기대치는 가장 적고 봉쇄확률은 가장 크며 생산율이 가장 작은 서비스정책이 된다. 반면에 서비스정책 K_U의 경우는 상태가 0에서 c사이에서 정의되어 $M/M/s/c$ 대기시스템으로 모형화되며 주영역에 존재하는 서버의 수의 기대치는 가장 크며 봉쇄확률은 가장 적으나 생산율이 가장 큰 서비스정책이 된다. 그러므로 서비스정책 K_L은 주어진 최적화문제의 제약식과 관련하여 실행가능성(feasibility)을 확인하는데 서비스정책 K_U는 수익함수를 최대화하는 서비스정책으로 주어진 최적화문제의 최적해를 도출하는데 유용하게 활용될 수 있다.

최적화문제에 있어서 가장 중요하게 고려되어야 할 내용은 언급된 바와 같이 주어진 설계모수에 대하여 수행도의 특성을 도출하는 일이라 하겠다. 특히 일계모멘트(the first moment)와 이계모멘트(the second moment)는 해의 공간을 현저히 감소시키고 최적화의 과정을 용이하게 하여 매우 유용한 결과를 제시한다. 그러므로 먼저 주어진 서비스시스템의 중요한 수행도인 생산율의 특성들에 대하여 살펴보기로 한다.

먼저 서비스정책이 $K_L=(c-s, c-s+1,\cdots,c)$인 경우 생산율 $TH_{s,c,K_L}=\lambda\{1-p_{s,c,K_L}(c)\}$이며 봉쇄확률 $p_{s,c,K_L}(c)$는 다음과 같이 정의된다.

$$p_{s,c,K_L}(c) = \frac{\rho^s/s!}{\sum_{x=0}^{s}(\rho^x/x!)}. \tag{3.1}$$

생산율 TH_{s,c,K_L}은 주영역의 대기능력 c에 대하여는 독립적으로 수행도에 변화가 없으나 총 서버의 수 $s(\leq c)$에 대하여는 다음이 성립한다.

정리 1. 주영역의 대기용량 c가 주어져 있을 때 생산율 TH_{s,c,K_L}은 총 서버의 수 s, $s=0,\cdots,c$,에 대하여 증가하는 오목함수이다.

증명 : 생산율 TH_{s,c,K_L}의 정의에 의하여 $Max.\, TH_{s,c,K_L} = Min.\, p_{s,c,K_L}(c)$와 같으므로 생산율 TH_{s,c,K_L}가 총 서버의 수 s, $s=0,\cdots,c$에 대하여 증가하는 오목함수임을 증명하는 것은 봉쇄확률 $p_{s,c,K_L}(c)$가 총 서버의 수 s에 대하여 감소하는 볼록함수임을 증명하는 것과 같다. 총 서버의 수 s에 대한 생산율 TH_{s,c,K_L}의 증가성과 오목성에 대한 증명은 약간의 수치적 계산 후에 얻어지는 다음의 결과에 의한다.

$$p_{s+1,c,K_L}(c)-p_{s,c,K_L}(c) = \frac{1}{(\sum_{x=0}^{s}\frac{\rho^x}{x!})(\sum_{x=0}^{s+1}\frac{\rho^x}{x!})}\{-\frac{\rho^s}{s!}(1-\frac{\rho}{s+1})\sum_{x=0}^{s}\frac{\rho^x}{x!} - \frac{\rho^{2s+1}}{s!(s+1)!}\} \leq 0. \tag{3.2}$$

$$\{p_{s,c,K_L}(c)-p_{s+1,c,K_L}(c)\}-\{p_{s+1,c,K_L}(c)-p_{s+2,c,K_L}(c)\} = \frac{1}{(\sum_{x=0}^{s}\frac{\rho^x}{x!})(\sum_{x=0}^{s+1}\frac{\rho^x}{x!})(\sum_{x=0}^{s+2}\frac{\rho^x}{x!})} \times\{(\sum_{x=0}^{s}\frac{\rho^x}{x!})(\sum_{x=0}^{s+1}\frac{\rho^x}{x!}) [(\frac{\rho^s}{s!}-\frac{\rho^{s+1}}{(s+1)!})-(\frac{\rho^{s+1}}{(s+1)!}-\frac{\rho^{s+2}}{(s+2)!})]+\frac{\rho^{s+2}}{(s+2)!} (\sum_{x=0}^{s}\frac{\rho^x}{x!})(\frac{\rho^s}{s!}-\frac{\rho^{s+1}}{(s+1)!})+\frac{\rho^{s+1}}{(s+1)!}(\sum_{x=0}^{s}\frac{\rho^x}{x!}) [\frac{\rho^s}{s!}-\frac{\rho^{s+2}}{(s+2)!}] + \frac{\rho^{2s+1}}{s!(s+1)!}[\frac{\rho^{s+1}}{(s+1)!}+\frac{\rho^{s+2}}{(s+2)!}]\} \geq 0. \tag{3.3}$$

그러므로 생산율 TH_{s,c,K_L}가 총 서버의 수 s, $s=0,\cdots,c$에 대하여 증가하는 오목함수임을 결론지을 수 있다. 생산율 TH_{s,c,K_L}은 대기능력 $c(c\geq s)$에 독립적으로 주영역의 대기능력 c가 증가하여도 생산율 TH_{s,c,K_L}은 변하지 않으며 결과적으로 $c=s$인 경우에 대기능력 비용이 가장 적다. 주어진 결과는 주어진 최적화문제의 실행가능영역(feasible region)의 경계에서의 특성을 제시하며 만약 서비스정책 K_L만을 적용하는 최적화문제에 있어서는 주어진 특성들은 매우 용이하게 활용될 수 있다.

이제 서비스정책 $K_U=(0,1,\cdots,s-1,c)$인 경우를 보자. 이 경우에는 $b=c-s$라고 하자. b는 주영역에서 총 서버를 제외한 여분의 대기능력을 의미하며 c와 b는 서비스정책 K_U가 적용되는 경우에는 대기능력을 나타내는 의미로 필요에 따라 구분 없이 사용된다. 생산율 $TH_{s,c,K_U}=\lambda\{1-p_{s,c,K_U}(c)\}$이 되며 봉쇄확률 $p_{s,c,K_U}(c)$는 일반화된(generalized) Erlang loss 함수로 불리는 다음의 함수를 갖는다.

$$p_{s,c,K_U}(c) = \frac{\frac{\rho^c}{s!s^b}}{\sum_{k=0}^{s}\frac{\rho^k}{k!}+\frac{\rho^s}{s!}\sum_{l=1}^{b}(\frac{\rho}{s})^l} \tag{3.4}$$

정리 2. 총 서버의 수 s가 일정할 때 생산율 $TH_{s,c,K_U}=\lambda\{1-p_{s,c,K_U}(c)\}$는 주영역의 대기능력 $c(c\geq s)$에 대하여 증가하는 오목함수이다.

증명 : $Max.\, TH_{s,c,K_U} = Min.\, p_{s,c,K_U}(c)$이므로 생산율 TH_{s,c,K_U}가 주영역의 대기능력 c에 대하여 증가하는 오목함수임을 증명하는 것은 봉쇄확률 $p_{s,c,K_U}(c)$가 대기능력 c에 대하여 감소하는 볼록함수임을 증명하는 것과 같으며 약간의 수치적 계산 후에 다음이 성립한다.

$$p_{s,c,K_U}(c)-p_{s,c+1,K_U}(c+1) = \frac{\rho^c}{s!s^b}\left\{\sum_{l=0}^{s}\frac{\rho^l}{l!}+\frac{\rho^{s+1}(s^{b+1}-\rho^{b+1})}{s!s^{b+1}(s-\rho)}\right\} - \frac{\rho^{c+1}}{s!s^{b+1}}\left\{\sum_{l=0}^{s}\frac{\rho^l}{l!}+\frac{\rho^{b+1}(s^b-\rho^b)}{s!s^b(s-\rho)}\right\} = \frac{\rho^{s+1}}{s!s^b}\left(1-\frac{\rho}{s}\right)$$

$$\left(\sum_{l=0}^{s}\frac{\rho^l}{l!}\right)+\frac{\rho^{s+c+1}}{(s!)^2 s^{b+1}} \geq 0. \tag{3.5}$$

$$\begin{aligned}&\{p_{s,c,K_U}(c)-p_{s,c+1,K_U}(c+1)\}\\&-\{p_{s,c+1,K_U}(c+1)-p_{s,c+2,K_U}(c+2)\}\\&=\frac{\rho^c}{s!s^b}\left\{\sum_{l=0}^{s}\frac{\rho^l}{l!}(1-\frac{\rho}{s})+\frac{\rho^{s+1}}{s!s}\right\}\left\{\sum_{l=0}^{s}\frac{\rho^l}{l!}(1-\frac{\rho}{s})\right.\\&\left.+\frac{\rho^{s+1}}{s!s^{b+2}}(s^{b+1}+\rho^{b+1})\right\} \geq 0.\end{aligned} \tag{3.6}$$

정리 3. 주영역의 대기능력 c가 주어져 있을 때 생산율 TH_{s,c,K_U}는 총 서버의 수 s, $s=1,\cdots,c$에 대하여 증가하는 오목함수이다.

증명 : 정리 2에서와 마찬가지로 이의 증명은 봉쇄확률 $p_{s,c,K_U}(c)$가 총 서버의 수 s, $s=1,\cdots,c$에 대하여 감소하는 볼록함수임을 증명하는 것과 같으므로 다음과 같이 정리될 수 있다.

$$\begin{aligned}&p_{s,c,K_U}(c)-p_{s+1,c,K_U}(c)=\frac{\rho^c}{s!}\left\{(\frac{1}{s})^b-(\frac{1}{s+1})^b\right\}\\&\sum_{k=0}^{s}\frac{\rho^k}{k!}+\frac{1}{s!(s+1)!s^b(s+1)^{b-1}}\sum_{k=0}^{b-1}\rho^k\\&\{(s+1)^{b-1-k}-s^{b-1-k}\}\geq 0.\end{aligned} \tag{3.7}$$

$$\begin{aligned}&\{p_{s,c,K_U}(c)-p_{s+1,c,K_U}(c)\}-\{p_{s+1,c,K_U}(c)-\\&p_{s+2,c,K_U}(c)\}\geq 0.\end{aligned} \tag{3.8}$$

식 (3.8)은 얼마간의 대수적 전개 후에 다음과 같이 정리된다.

$$\begin{aligned}&\left\{\frac{\rho^c[(s+1)^b-s^b]}{(s+1)!s^b(s+1)^{b-1}}\sum_{k=0}^{s}\frac{\rho^k}{k!}+\frac{\rho^c}{s!(s+1)!s^b(s+1)^{b-1}}\right.\\&\left.\sum_{k=s+1}^{c}[(s+1)^{c-k}-s^{c-k}]\rho^k\right\}\\&\left\{\sum_{l=0}^{s}\frac{\rho^l}{l!}+\frac{\rho^{s+1}[(s+2)^b-\rho^b]}{(s+2)!(s+2)^{b-2}(s+2-\rho)}\right\}\\&-\left\{\frac{\rho^c[(s+2)^{b-1}-(s+1)^{b-1}]}{(s+2)!(s+1)^{b-1}(s+2)^{b-2}}\sum_{k=0}^{s}\frac{\rho^k}{k!}\right.\\&+\frac{\rho^c}{(s+1)!(s+2)!(s+1)^{b-1}(s+2)^{b-2}}\\&\left.\sum_{k=s+1}^{c}[(s+2)^{c-k}-(s+1)^{c-k}]\rho^k\right\}\left\{\sum_{l=0}^{s}\frac{\rho^l}{l!}\right.\\&\left.+\frac{\rho^{s+1}(s^b-\rho^b)}{s!s^b(s-\rho)}\right\}\geq 0.\end{aligned} \tag{3.9}$$

식 (3.9)의 증명을 위하여 k가 $0,\cdots,s$인 경우와 $s+1,\cdots,c$인 경우로 구분한다. 먼저 $k=0,\cdots,s$인 경우는 다음과 같이 정리된다.

$$\begin{aligned}&\left(\sum_{l=0}^{s}\frac{\rho^l}{l!}\right)^2\frac{\rho^c}{(s+2)!s^b(s+1)^{b-1}(s+2)^{b-2}}\\&\{(s+2)^{b-1}[(s+1)^b-s^b]-s^b[(s+2)^{b-1}\\&-(s+1)^{b-1}]\}+\left(\sum_{l=0}^{s}\frac{\rho^l}{l!}\right)\rho^{s+c+1}\\&\left\{\frac{[(s+1)^b-s^b][(s+2)^b-\rho^b]}{(s+1)!(s+2)!s^b(s+1)^{b-1}(s+2)^{b-2}(s+2-\rho)}\right.\\&\left.-\frac{[(s+2)^{b-1}-(s+1)^{b-1}][s^b-\rho^b]}{s!(s+2)!s^b(s+1)^{b-1}(s+2)^{b-2}(s-\rho)}\right\}\geq 0.\end{aligned} \tag{3.10}$$

여기에서

$$\begin{aligned}&(s+2)^{b-1}\{(s+1)^b-s^b\}-s^b\{(s+2)^{b-1}-(s+1)^{b-1}\}\\&\geq(s+2)^b\{(s+1)^b-s^b\}-s^b\{(s+2)^b-(s+1)^b\}\geq 0.\end{aligned} \tag{3.11}$$

$$\begin{aligned}&(s-\rho)\{(s+1)^b-s^b\}\{(s+2)^b-\rho^b\}-(s+2-\rho)(s+1)\\&\{(s+2)^{b-1}-(s+1)^{b-1}\}(s^b-\rho^b)\geq\\&\{(s+1)^b-s^b\}\{(s+2)^b-\rho^b\}-\{(s+2)^b-(s+1)^b\}\\&(s^b-\rho^b)\geq 0.\end{aligned} \tag{3.12}$$

이제 $k=s+1,\cdots,c$인 부분을 정리하면 다음과 같다.

$$\begin{aligned}&\frac{\rho^c}{(s+1)!(s+2)!s^b(s+1)^{b-1}(s+2)^{b-2}}\left(\sum_{l=0}^{s}\frac{\rho^l}{l!}\right)\\&\sum_{k=s+1}^{c}\rho^k\{(s+1)(s+2)^{b-1}[(s+1)^{c-k}-s^{c-k}]\\&-s^b[(s+2)^{c-k}-(s+1)^{c-k}]\}+\end{aligned}$$

$$\frac{\rho^{s+c+1}}{s!(s+1)!(s+2)!s^b(s+1)^{b-1}(s+2)^{b-2}}$$

$$\left(\sum_{k=s+1}^{c}\rho^k\right)\times\left\{\frac{[(s+1)^{c-k}-s^{c-k}][(s+2)^b-\rho^b]}{s+2-\rho}\right.$$

$$\left.-\frac{[(s+2)^{c-k}-(s+1)^{c-k}](s^b-\rho^b)}{s-\rho}\right\}\geq 0. \qquad (3.13)$$

식 (3.13)의 첫 번째 항(term)은 다음의 관계식에 의하여 성립한다.

$$(s+1)(s+2)^{b-1}\{(s+1)^{c-k}-s^{c-k}\}$$
$$-s^b\{(s+2)^{c-k}-(s+1)^{c-k}\}$$
$$\geq(s+2)^{b-1}\{(s+1)^{c-k}-s^{c-k}\}-s^b$$
$$\{(s+2)^{c-1-k}-(s+1)^{c-1-k}\}\geq 0. \qquad (3.14)$$

식 (3.13)의 두 번째 항은 b가 짝수인 경우 다음과 같이 전개되어 성립한다. b가 홀수인 경우도 유사하게 증명된다.

$$\frac{\rho^{s+c+1}}{s!(s+1)!(s+2)!s^b(s+1)^{b-1}(s+2)^{b-2}}\left(\sum_{k=s+1}^{c-1}\rho^k\right)$$
$$\left\{\sum_{l=0}^{b/2-1}\rho^l\sum_{n=0}^{c-1-k}[(s+2)^{b-1-l}(s+1)^{b-2-n}s^n\right.$$
$$-(s+2)^{b-2-n}(s+1)^n s^{b-1-l}]-\sum_{l=0}^{b/2-1}\rho^{b-1-l}$$
$$\left.\sum_{n=0}^{c-1-k}[(s+2)^{b-2-n}(s+1)^n s^l-(s+2)^l\right.$$
$$\left.(s+1)^{b-2-n}s^n]\right\}\geq 0. \qquad (3.15)$$

정리 3의 결과는 Shanthikumar and Yao[9]에 제시된 Theorem 1의 결과에 있어서 주어진 폐쇄대기네트워크(closed queueing network)의 작업장이 작업장 1(station 1) 하나로 구성된 특수한 경우의 결과에 해당된다. $M/M/s/c$ 대기시스템의 일반화된 Erlang loss 함수의 서버의 수에 대한 볼록성은 Chang, et. al.[5]이 추계적 비교에 의하여 Pacheco[7]는 대수적으로 증명하였다. Pacheco[8]는 일반화된 Erlang loss 함수의 대기용량에 대한 볼록성을 증명하였다. Pacheco[7, 8]의 증명은 일반화된 Erlang loss 함수를 연속적인 함수로 보았으며 여기에서는 설계모수를 실제와 같이 이산적으로 다루었다.

지금까지의 결과로부터 생산율 TH_{s,c,K_U}는 총 서버의 수 s가 일정하면 주영역의 대기능력 $c(c\geq s)$에 대하여 증가하는 오목함수이며 대기능력 c가 일정하면 총 서버의 수 $s(s\leq c)$에 대하여 증가하는 오목함수임을 제시하고 있다. 그 결과로 제조시스템의 생산율 TH_{s,c,K_U}는 서버배분 s와 대기능력 c의 독립적인 설계모수에 대하여는 증가하는 오목함수임을 알 수 있다. 그러나 주어진 결과는 생산율 TH_{s,c,K_U}가 서버배분 s와 대기능력 c에 대한 공동의(joint) 오목성을 만족시키는 함수라는 결과는 제시하지 못한다. 이는 주어진 최적화문제의 해법을 도출하는데 어려움을 제시한다.

지금까지는 총 서버의 수와 주영역의 대기능력과 관련하여 대기시스템의 수행도의 특성들에 대하여 알아보았다. 여기에서는 서비스정책과 관련하여 다음의 수행도의 특성을 보기로 한다. 만약 총 서버의 수 s와 주영역의 대기능력 c가 주어진 상태에서 두 종류의 서비스정책 $K^1=(k_0, k_1, \cdots, k_i, \cdots, k_s)$와 $K^2=(k_0, k_1, \cdots, k_i-1, \cdots, k_s)$을 보자. 이는 $n\neq i$, $i\in(0, \cdots, s-1)$,에 대하여 $k_n^1=k_n^2$이며 $|k_i^1-k_{i-1}^1|\geq 2$, $k_i^2=k_i^1-1$임을 의미한다.

정리 4. $p_{s,c,K^1}(c)\geq p_{s,c,K^2}(c)$, $TH_{s,c,K^1}\leq TH_{s,c,K^2}$, $F_{s,c,K^1}\leq F_{s,c,K^2}$, 그리고 $B_{s,c,K^1}\geq B_{s,c,K^2}$이다.

증명 : 봉쇄확률 $p_{s,c,K}(c)$에 대하여는 다음이 성립한다.

$$p_{s,c,K^1}(c)-p_{s,c,K^2}(c)=\frac{\beta_{s,c,K^1}(c)}{\sum_{x=k_0}^{c}\beta_{s,c,K^1}}-\frac{\beta_{s,c,K^2}(c)}{\sum_{x=k_0}^{c}\beta_{s,c,K^2}(x)}$$

$$=\frac{1}{\sum_{x=k_0}^{c}\beta_{s,c,K^1}(x)\sum_{x=k_0}^{c}\beta_{s,c,K^2}(x)}\{\beta_{s,c,K^1}(c)-\beta_{s,c,K^2}(c)\}$$

$$\sum_{x=k_0}^{k_i-1} \beta_{s,c,K^1}(x) \geq 0. \quad (3.16)$$

여기에서 $\beta_{s,c,K^1}(x)=\beta_{s,c,K^2}(x)$, $x\in(k_0,\cdots,k_i-1)$이며 $\beta_{s,c,K^1}(c)\geq\beta_{s,c,K^2}(c)$이다. 생산율 $TH_{s,c,K}$에 대한 부등식의 결과는 봉쇄확률 $p_{s,c,K}(c)$에 대한 결과에 의하며 부등식 $F_{s,c,K^1}\leq F_{s,c,K^2}$와 $B_{s,c,K^1}\geq B_{s,c,K^2}$에 대한 증명은 Berman et. al.(2005)에 의한다.

4. 최적화

본 장에서는 서비스시스템의 설계에 있어서 이의 최적화과정을 다룬다. 이는 주어진 서비스시스템에 있어서 보조영역에서 요구하는 서비스 능력에 대한 제약조건을 만족시키면서 총 서버의 수, 주영역의 대기능력, 그리고 서비스정책의 결과로 산출되는 생산율과 생산율의 함수로서 표시되는 수익(revenue), 그리고 서버와 대기능력의 확보에 따른 비용의 차로 표시되는 서비스시스템의 이익을 최대화시키는 것이다.

먼저 총 서버의 수가 s이고 주영역의 대기능력이 c인 경우를 보자. 가능한 서비스정책 중에서 서비스정책 $K_U=(0,1,\cdots,s-1,c)$가 가장 적은 봉쇄확률과 가장 큰 생산율을 갖는다. 그러므로 만약 주어진 서비스정책 K_U에서 보조영역의 서비스능력과 관련된 제약식이 만족된다면 서비스정책 K_U가 가장 많은 수익을 제공하는 서비스정책이 될 것이다. 그러므로 주어진 최적화 과정에 있어서 보조영역의 서비스능력과 관련된 제약식을 만족한다고 가정하면 서비스정책 K_U하에서 최적 대기시스템을 설계할 수 있을 것이다. 그러므로 먼저 보조영영의 서비스능력과 관련된 제약식을 고려하지 않고 서비스정책 K_U를 적용하여 총 서버의 수 s와 주영역의 대기능력 c를 결정하는 최적화절차를 고려한다.

그러므로 최적화문제는 다음과 같다.

$$\begin{aligned} &Max._{\substack{1\leq c\\ 1\leq s\leq c}} \Phi(s,c,K_U) \\ &=f\{\lambda[1-p_{s,c,K_U}(c)]\}-h(s)-g(c). \end{aligned} \quad (4.1)$$

이의 최적화를 위하여 다음을 정의한다.

$$\begin{aligned} \theta(s_c,c,K_U) &= Max._{1\leq s\leq c}\ \theta(s,c,K_U) \\ &= Max._{1\leq s\leq c}\ f\{\lambda[1-p_{s,c,K_U}(c)]\}-h(s). \end{aligned} \quad (4.2)$$

식 (4.2)는 주영역의 대기능력 c가 주어진 경우 최적 총 서버의 수 s_c와 최적 수익 $\theta(s,c,K_U)$를 결정하는 문제이다. 설계모수 총 서버의 수 s에 대하여 수익함수 $f\{\lambda[1-p_{s,c,K_U}(c)]\}$가 증가하는 오목함수이고 비용함수 $h(s)$는 증가하는 볼록함수이므로 수익함수 $\theta(s,c,K_U)$는 오목함수가 되어 최적 총 서버의 수 s_c는 한계분석법으로 쉽게 산정될 수 있다. 이는 서버의 수 s를 하나씩 순차적으로 증가시키면서 최대점을 도출하는 방법으로 $\nabla(s)=\theta(s+1,c,K_U)-\theta(s,c,K_U)$, $1\leq s\leq c-1$,라고 정의하면 처음으로 $\nabla(s)\leq 0$이 되는 s가 최적 s_c가 된다. 이는 의사결정변수 s를 한 단위씩 증가시키면서 이익함수 $\theta(s,c,K_U)$가 처음으로 감소하는 총 서버의 수 s에서 알고리즘을 종료함을 말한다.

그러므로 식 (4.1)의 총 서버의 수와 주영역의 대기용량을 동시에 결정하는 원문제는 다음과 같다.

$$Max._{1\leq c}\ \theta(s_c,c,K_U)-g(c). \quad (4.3)$$

만약 생산율함수 TH_{s,c,K_U}가 총 서버의 수 s와 주영역의 대기능력 c에 대하여 공동으로 증가하는 오목함수이면 주어진 최적화문제는 한계분석법으로 매우 쉽게 해결될 수 있다. 즉 최초로 $\nabla(c)=\{\theta(s_{c+1},c+1,K_U)-g(c+1)\}-\{\theta(s_c,c,K_U)-g(c)\}\leq 0$ 이 되는 대기능력 c에서 알고리즘을 종료하며 최적해 (s_{c^*},c^*)를 구할 수 있다. 그러나 생산율함수 TH_{s,c,K_U}가 총 서버의 수 s와 주영역의 대기능력 c의 각각에 대하여는 증가하는 오목함수임이 증명되었으나 공동으로 증가하는 오목함수임은 증명하지 못하고 있다. 그러므로 전술된 한계분석법을 적용하여 최적해를 구하기는 곤란하다. 그럼에도 불구

하고 주어진 문제는 생산율함수 TH_{s,c,K_U}가 서버의 수 s와 대기능력 c 각각에 대하여 증가하는 오목성을 만족시키므로 Shanthikumar and Yao[9]에서와 같이 효율적으로 해결될 수 있다.

정리 5. 식 (4.3)로 정의되는 총 서버의 수와 주영역의 대기능력을 결정하는 최적화문제에 있어서 최적 대기능력 c^*를 결정하는 절차는 다음과 같다.

$$\theta(s_{c^*}, c^*, K_U) - g(c^*) = Max._{1 \le c \le c_L} \theta(s_c, c, K_U) - g(c). \quad (4.4)$$

여기에서 c_L은 대기능력을 결정하는데 고려되는 한계치로서 다음의 관계식을 만족시키는 가장 적은 정수 c이다.

$$\frac{\theta(s_c, c, K_U)}{c} \le g(c) - g(c-1). \quad (4.5)$$

증명 : $c \le c_L$의 경우에는 정의에 의하여 최적해가 보장된다. $c > c_L$인 경우에는 다음의 일련의 부등식이 유도된다.

$$\theta(s_c, c, K_U) - \theta(s_{c_L}, c_L, K_U) \le \theta(s_c, c, K_U) - \theta(s_c, c_L, K_U)$$
$$\le (c - c_L)\frac{\theta(s_c, c_L, K_U)}{c_L} \le (c - c_L)\frac{\theta(s_{c_L}, c_L, K_U)}{c_L}$$
$$\le (c - c_L)\{g(c_L) - g(c_L - 1)\} \le g(c) - g(c_L). \quad (4.6)$$

그러므로 $c > c_L$에 대하여 다음이 성립하고 위의 절차에 의하여 얻어진 해는 최적해가 보장된다.

$$\theta(s_{c_L}, c_L, K_U) - g(c_L) \ge \theta(s_c, c, K_U) - g(c). \quad (4.7)$$

위의 증명에서 첫 번째와 세 번째 부등식은 식 (4.2)의 정의에 의하며 두 번째 부등식은 수익함수 $\theta(s, c, K_U)$의 대기능력 c에 대한 오목성과 오목성은 sublinearity(비음정수 m, $n(m \le n)$, 이산적 함수 $\theta(\cdot): R \to R$, 그리고 $\theta(0) = 0$에 대하여 $\theta(n) - \theta(m) \le \frac{n-m}{m}\theta(m)$)를 만족시킴을 적용하였고 네 번째 부등식은 식 (4.5)의 정의에 의하며 다섯 번째 부등식은 $g(c)$의 오목성(비음정수 l, m, $n(l \le m \le n)$, 이산적 함수 $\theta(\cdot): R \to R$, 그리고 $\theta(0) = 0$에 대하여 $(n-m)\{\theta(m) - \theta(l)\} \le (m-l)\{\theta(n) - \theta(m)\}$)에 의한다. 수익함수 $\theta(s_c, c, K_U)$가 대기능력 c에 대하여 오목성을 만족시킴은 증명하지 못하였으나 위의 첫 번째 항과 네 번째 항의 관계로부터 sublinearity를 만족시킴을 알 수 있다. 주어진 최적화 절차는 한계분석법과 비교하여 얼마간의 복잡성이 추가됨에도 불구하고 매우 편리하게 적용될 수 있음을 알 수 있다.

이제 본 문제인 보조영역의 서비스능력의 제약식을 직접 고려하는 식 (2.4)의 최적화 절차를 구하는 문제로 돌아가기로 한다. 본 문제인 식 (2.4)로 표시되는 최적화문제의 최적화 절차는 식 (4.2)에 선행하여 보조영역의 서비스능력을 만족시키면서 주어진 총 서버의 수 s에 대하여 수익(생산율)을 최대화시키는 서비스정책 K를 결정하는 단계를 보완하고 그 결과로 식 (4.2)가 수정되고 정리 5의 절차에 또한 이를 고려함으로써 쉽게 해결될 수 있다.

식 (4.2)에 선행하여 주어진 총 서버의 수 s, $s = 1, \cdots, c$에서 보조영역의 서비스능력을 고려하는 최적의 서비스정책을 채택하는 과정은 다음과 같다.

$$Max._K \theta(s, c, K) = Max._K f\{\lambda[1 - p_{s,c,K}(c)]\} - h(s),$$
$$s.t. \quad B_{s,c,K} \ge B_L. \quad (4.8)$$

주어진 총 서버의 수 s에 대하여 비용함수 $h(s)$는 상수이며 수익함수 $f(TH_{s,c,K})$는 생산율 $TH_{s,c,K}$에 대하여 증가함수이므로 식 (4.8)의 목적함수는 $Min._K\ p_{s,c,K}(c)$와 동일하다. 주어진 서버배분 s, $s = 1, \cdots, c$,에 대하여 다음의 세 가지 경우를 고려한다. 첫 번째, 서비스정책 K_L이 보조영역의 서비스

능력 제약식을 만족시키지 못하면 주어진 서버배분 s는 실행불가능(infeasible)하다. 이 경우에는 다음 $s+1$로 간다. 두 번째, 서비스정책 K_U가 보조영역의 서비스능력의 제약식을 만족시키면 서비스정책 K_U가 주어진 총 서버의 수 s에서 채택될 서비스정책이 된다. 그러므로 제시된 최적화 절차를 진행한다. 세 번째, 서비스정책 K_L은 보조영역의 서비스능력의 제약식을 만족시키나 서비스정책 K_U는 만족시키지 못하면 주어진 총 서버의 수 s에서 최적 서비스정책을 결정하는 과정이 필요하다. 정리4와 Berman, et. al. (2005)에 의하여 최적 서비스정책을 결정하는 휴리스틱(heuristic)은 다양하게 제시된다.

이를 위하여 다음을 정의한다. 주어진 서비스정책 K에 대하여 만약 $|k_n - k_{n-1}|>1$, $k_{n-1}\geq 0$, $n<s$, 이면 k_n을 타입(type)1 원소라고 하고 $|k_{n+1}-k_n|>1$, $n<s$,이면 k_n을 타입 2 원소라고 정의한다. 타입1 원소를 한 단위 감소시키면 봉쇄확률이 감소하여 생산율은 증가하나 보조영역의 서비스능력이 감소하고 타입 2 원소를 한 단위 증가시키면 생산율은 감소하나 보조영역의 서비스능력이 증가하여 실행가능성(feasibility)이 증가한다. 그러므로 이를 종합하여 다음을 제시한다.

step 1 : 만약 $B_{s,c,K_L}<B_L$이면 주어진 s는 실행불가능하다.

step 2 : 만약 $B_{s,c,K_U}\geq B_L$이면 $K_{s,c}=K_U$가 되며 $\theta(s, c, K_{s,c})$를 계산한다.

step 3 : 만약 $B_{s,c,K_L}\geq B_L$이고 $B_{s,c,K_U}<B_L$이면 $K=K_L$로 놓고 $\theta(s,c,K)$를 계산하고 $\theta(s,c,K_{s,c})=\theta(s,c,K)$로 놓는다. $J=s$로 놓는다.

step 4 : $0\leq j^*<J$이고 k_{j^*}는 타입 1 원소인 가장 작은 j^*를 찾는다. 만약 이런 j^*가 없으면 step 6로 간다. $k_{j^*}=k_{j^*}-1$로 놓는다. 만약 $B_{s,c,K}<B_L$이면 $J=j^*$로 놓고 step 6로 간다.

step 5 : 만약 $\theta(s,c,K)>\theta(s,c,K_{s,c})$이면 $k_{s,c}=K$로 놓고 $\theta(s,c,K_{s,c})$를 계산한다. step 4로 간다.

step 6 : $0\leq j^*<J$이고 k_{j^*}가 타입 2 원소인 가장 작은 j^*를 찾는다. 만약 이러한 j^*가 없으면 step 7로 간다. $k_{j^*}=k_{j^*}+1$로 놓는다. 만약 $B_{s,c,K}<B_L$이면 step 6를 반복한다. 그렇지 않으면 step 5로 간다.

step 7 : 멈춘다. 현재의 $K_{s,c}$를 최적 서비스정책으로 채택한다.

그 결과로 식 (4.2)와 정리 5는 다음과 같이 수정됨으로써 주어진 최적화 절차는 완료된다. 먼저 식(4.8)의 경우는 다음과 같다. 이 경우에도 식 (4.2)와 같이 한계분석법을 휴리스틱으로 적용한다. 이 경우에는 서비스정책 K_U를 적용했던 식 (4.2)의 경우와는 달리 제약식에 의하여 결정되는 최적 서비스정책 $K_{s,c}$를 직접 고려하여 수행도의 이계특성을 도출하기 어렵기 때문이다.

$$\theta(s_c, c, K_{s_c,c}) = Max_{\,1\leq s\leq c}\,\theta(s,c,K_{s,c})$$
$$= Max_{\,1\leq s\leq c}\; f\{\lambda[1-p_{s,c,K_{s,c}}(c)]\}-h(s). \qquad (4.9)$$

정리 5의 경우에는 식 (4.5)가 동일하게 적용될 수 있다. 식 (4.5)에 의하여 결정되는 한계 대기능력 c_L은 주영역에서의 대기능력에 대한 평균수익보다 한계비용이 더 커지는 특정 대기능력을 의미하므로 수익함수의 대기능력에 대한 오목성과 비용함수의 증가하는 볼록성에 의하여 알고리즘을 종료하는 한계 대기능력이 된다. 그러므로 보조영역에서의 서비스능력에 대한 제약조건이 추가되면 수익이 감소되어 그 결과는 식 (4.5)는 여전히 유효하게 적용될 수 있다. 그러므로 정리 5의 최적화절차는 다음과 같이 정리된다.

$$\theta(s_{c^*}, c^*, K_{s_{c^*},c^*}) - g(c^*)$$
$$= Max_{\,1\leq c\leq c_L}\;\theta(s_c, c, K_{s_c c}) - g(c), \qquad (4.10)$$

주어진 알고리즘에서 주영역의 대기능력의 한계치 c_L은 다음의 조건을 만족시키는 가장 작은 정수 c이다.

$$\frac{\theta(s_c, c, K_{s_c,c})}{c} \le g(c) - g(c-1). \quad (4.11)$$

5. 수치 예

본 장에서는 주어진 결과에 대한 이해를 높이기 위하여 수치적 예를 제시한다. 고객이 주영역에 도착하는 과정은 기대치 $\lambda=8$인 포아송분포에 의하며 주영역에서의 서버의 서비스시간은 서비스율 $\mu=2$인 지수분포를 갖는다. 그러므로 주영역의 주어진 부하 $\rho=8/2=4$이다. 수익함수 $f\{TH_{s,c,K}\}=2TH_{s,c,K}$로 정의되며 수익함수$f(TH_{s,c,K})$는 생산율 $TH_{s,c,K}$에 대하여 선형(linear)의 함수이며 선형의 함수는 광의의 의미에서 오목함수이다. 총 서버의 비용함수 $h(s)=s^{8/7}$로서 비용함수 $h(s)$는 총 서버의 수 s에 대하여 볼록함수이다. 또한 주영역의 대기능력에 대한 비용함수 $g(c)=0.22c^{3/2}$도 비용함수 $g(c)$의 대기능력 c에 대하여 볼록함수이다.

<표 1>에는 주영역의 특정 대기능력 c에 대하여 식 (4.2)와 식 (4.9)로 주어지는 최적 총 서버의 수 s_c를 구하는 최적화절차에 대한 수치 예가 제시되어 있다. 주영역의 대기능력 $c=10$으로 주어졌으며 다양한 보조영역의 서비스능력과 관련된 제약식에 대하여 식 (4.9)로 주어진 최적 서버의 수 s_{10}와 이에 상응하는 최적 서비스정책에 대한 수치 예가 제시되어 있다. <표 1>에서 적용된 대기능력 $c=10$은 식 (4.11)에 의하여 결정되는 한계대기능력으로 처음으로 한계비용(1.017)이 처음으로 평균수익(0.939)보다 커지는 대기능력 c_L을 의미한다. '불가'는 제약식을 만족시키지 못하여 실행불가능 해를 갖는 경우를 말한다. 최대 수익을 제공하는 최적 총 서버의 수 s_{10}은 굵게 표시되었다.

<표 1>에서 서비스능력의 제약식이 없는 경우에는 설계모수 총 서버의 수 s에 대하여 수익함수 $\theta(s,c,K_U)$가 오목함수임을 수치적 결과로부터 확인할 수 있으며 그 결과로 한계분석법 또한 적절한

〈표 1〉 서비스정책과 서버배분($c=10$인 경우)

제약식	서버의 수(s)	1	2	3	4	5	6	7	8	9	10
$B\ge 0$ (없음)	$\theta(s,10,K_{s,10})$	2.999	5,786	8.237	9.388	9.028	7.938	6.581	5.114	3.587	2.020
	$B_{s,10,K_{s,10}}$	0.000	0.001	0.063	0.434	1.170	2.078	3.044	4.030	5.024	6.095
	$K_{s,10}$	K_U	K_U	K_U	K_U	K_U	K_U	K_U	K_U	K_U	K_U
$B\ge 1.0$	$\theta(s,10,K_{s,10})$	불가	불가	불가	7.122	9.028	7.938	6.581	5.114	3.587	2.202
	$B_{s,10,K_{s,10}}$				1.000	1.170	2.078	3.044	4.030	5.024	6.095
	$K_{s,10}$				(4, 5, 6, 9, 10)	K_U	K_U	K_U	K_U	K_U	K_U
$B\ge 1.5$	$\theta(s,10,K_{s,10})$	불가	불가	불가	불가	7.701	7.938	6.581	5.114	3.587	2.202
	$B_{s,10,K_{s,10}}$					1.502	2.078	3.044	4.030	5.024	6.095
	$K_{s,10}$					(3, 4, 5, 7, 8, 10)	K_U	K_U	K_U	K_U	K_U
$B\ge 2.0$	$\theta(s,10,K_{s,10})$	불가	불가	불가	불가	불가	7.938	6.581	5.114	3.587	2.202
	$B_{s,10,K_{s,10}}$						2.078	3.044	4.030	5.023	6.095
	$K_{s,10}$						K_U	K_U	K_U	K_U	K_U
$B\ge 2.5$	$\theta(s,10,K_{s,10})$	불가	불가	불가	불가	불가	불가	6.581	5.113	3.587	2.020
	$B_{s,10,K_{s,10}}$							3.044	4.030	5.024	6.095
	$K_{s,10}$							K_U	K_U	K_U	K_U

최적화 알고리즘임을 알 수 있다. 총 서버의 수 s가 증가하여 어느 한계를 넘어가면 오목함수인 수익함수 $f(TH_{s,c,K_U})$의 한계수익보다 증가하는 볼록함수인 비용함수 $h(s)$의 한계비용이 더 커져 수익함수 $\theta(s,c,K_U)$가 감소한다. 총 서버의 수 s가 증가함에 따라 보조영역의 서비스능력도 같이 증가한다.

서비스능력과 관련된 제약식을 갖는 경우에도 수치적 결과는 한계분석법이 최적화절차로 여전히 유효함을 알 수 있으며 제약조건이 까다로워질수록 서비스정책 K_U를 시작으로 생산율이 높은 서비스정책들이 최적 서비스정책에서 배제되어 최적 총 서버의 수 s_c도 동시에 증가한다.

<표 2>에는 식 (4.1)와 식 (4.10)의 최적화절차에 요구되는 주영역의 대기능력 $c(1 \le c \le c_L)$에 대하여 최적 수익함수 $\theta(s_c, c, K_{s_c,c})$에 관련된 수치적 결과가 제시되어 있다.

<표 2>에서 서비스능력의 제약식이 없는 경우에는 그 수치적 결과로부터 주영역의 대기능력 c가 증가하면 최적 총 서버의 수 s_c가 증가하며 수익함수 $\theta(s_c, c, K_{s_c,c})$가 sublinearity(실제적으로는 오목성)을 만족시킴을 알 수 있다. 또한 서버의 수가 증가하면 보조영역의 서비스능력도 같이 증가함을 알 수 있다.

〈표 2〉 대기능력과 서버배분

제약식	대기능력(c)	1	2	3	4	5	6	7	8	9	10
$B \ge 0$ (없음)	s_c	1	2	3	3	4	4	4	4	4	4
	$K_{s_c,c}$	K_U	K_U	K_U	K_U	K_U	K_U	K_U	K_U	K_U	K_U
	$\theta(s_c, c, K_{s_c,c})$	2.200	3.946	5.279	6.484	7.331	8.058	8.551	8.907	9.177	9.388
	$B_{s_c,c,K_{s_c,c}}$	0.200	0.462	0.803	0.501	0.948	0.766	0.643	0.554	0.487	0.434
$B \ge 1.0$	s_c	불가	불가	불가	4	4	5	5	5	5	5
	$K_{s_c,c}$				K_U	(0, 1, 2, 4, 5)	K_U	K_U	K_U	K_U	K_U
	$\theta(s_c, c, K_{s_c,c})$				6.153	6.757	7.509	8.123	8.533	8.820	9.028
	$B_{s_c,c,K_{s_c,c}}$				1.243	1.092	1.550	1.396	1.294	1.222	1.170
$B \ge 1.5$	s_c	불가	불가	불가	불가	5	5	5	5	6	6
	$K_{s_c,c}$					K_U	K_U	(0, 1, 2, 4, 5, 7)	(1, 2, 3, 5, 6, 8)	K_U	K_U
	$\theta(s_c, c, K_{s_c,c})$					6.522	7.509	7.701	7.701	7.773	7.938
	$B_{s_c,c,K_{s_c,c}}$					1.796	1.550	1.502	1.502	2.119	2.078
$B \ge 2.0$	s_c	불가	불가	불가	불가	불가	6	6	6	6	6
	$K_{s_c,c}$						K_U	K_U	K_U	K_U	K_U
	$\theta(s_c, c, K_{s_c,c})$						6.375	7.091	7.513	7.773	7.938
	$B_{s_c,c,K_{s_c,c}}$						2.469	2.290	2.184	2.119	2.078
$B \ge 2.5$	s_c	불가	불가	불가	불가	불가	불가	7	7	7	7
	$K_{s_c,c}$							K_U	K_U	K_U	K_U
	$\theta(s_c, c, K_{s_c,c})$							5.753	6.203	6.446	6.581
	$B_{s_c,c,K_{s_c,c}}$							3.251	3.138	3.078	3.044

<표 2>의 결과로부터 서비스능력과 관련된 제약식이 있는 경우에도 한계분석법이 여전히 유효함을 알 수 있으며 제약식의 조건이 까다로워질수록 최적 총 서버의 수 s_c도 증가하며 이는 서비스정책 K_U와 같은 생산율이 높은 서비스정책이 실행불가능 하여지기 때문에 결과적으로 생산율이 더 낮은 서비스정책이 채택될 수밖에 없기 때문이다. 예를 들어 주영역의 대기능력 $c=7$인 경우를 보자. 제약식이 없는 경우에는 $s_7=4$, 서비스정책은 $K_{4,7}=K_U=(0, 1, 2, 3, 7)$, 그리고 $\theta(4, 7, K_U)=8.551$이다. 그러나 제약식 $B_{s,c,K}\geq 1$이 도입되면 $s=4$에서 서비스정책 K_U가 보조영역 서비스능력 제약식을 만족시키지 못하므로 최적 서비스 정책 $K_{4,7}=(1, 2, 3, 6, 7)$이 되어 $\theta(4,7,K_{4,7})=7.122$로 감소된다. 그러므로 보조영역의 서비스능력을 만족시키는 서버의 수 $s_7=5$로 증가하고 최적 서비스정책 $K_{5,7}=K_U$가 된다. 그 결과로 보조영역의 서비스능력 제약식이 있는 경우에는 수익함수 $\theta(s_c, c, K_{s_c,c})$가 대기능력 c에 대하여 오목성뿐만이 아니라 sublinearity도 만족시키지 못함을 알 수 있다.

<표 3>에는 최적 대기능력을 결정하는 식 (4.4)와 식 (4.10)으로 주어지는 최적화 절차의 수치 예를 제시한다. 최대이익을 주는 이익함수 $\Phi(s_{c^*}, c^*, K_{s_{c^*},c^*})$는 굵게 표시되었다.

〈표 3〉 대기능력과 총수익

제약식	대기능력(c)	1	2	3	4	5	6	7	8	9	10
$B\geq 0$ (없음)	s_c	1	2	3	3	4	4	4	4	4	4
	$\theta(s_c, c, K_{s_c,c})$	2.200	3.946	5.278	6.484	7.331	8.058	8.551	8.907	9.177	9.388
	$g(c)$	0.220	0.622	1.143	1.760	2.460	3.233	4.074	4.978	5.940	6.957
	$\Phi(s_c, c, K_{s_c,c})$	1,980	3.323	4.136	4.724	4.872	4.825	4.477	3.929	3.237	2.438
$B\geq 1.0$	s_c	불가	불가	불가	4	4	5	5	5	5	5
	$\theta(s_c, c, K_{s_c,c})$				6.153	6.757	7.509	8.123	8.533	8.820	9.028
	$g(c)$				1.760	2.460	3.233	4.074	4.978	5.940	6.957
	$\Phi(s_c, c, K_{s_c,c})$				4.393	4.298	4.276	4.049	3.553	2.880	2.058
$B\geq 1.5$	s_c	불가	불가	불가	불가	5	5	5	5	6	6
	$\theta(s_c, c, K_{s_c,c})$					6.522	7.509	7.701	7.701	7.773	7.938
	$g(c)$					2.460	3.233	4.074	4.978	5.940	6.957
	$\Phi(s_c, c, K_{s_c,c})$					4.063	4.276	3.633	2.723	1.833	0.981
$B\geq 2.0$	s_c	불가	불가	불가	불가	불가	6	6	6	6	6
	$\theta(s_c, c, K_{s_c,c})$						6.375	7.091	7.513	7.773	7.938
	$g(c)$						3.233	4.074	4.978	5.940	6,957
	$\Phi(s_c, c, K_{s_c,c})$						3.142	3.016	2.535	1.833	0.981
$B\geq 2.5$	s_c	불가	불가	불가	불가	불가	불가	7	7	7	7
	$\theta(s_c, c, K_{s_c,c})$							5.753	6.203	6.446	6.581
	$g(c)$							4.074	4,978	5,940	6.957
	$\Phi(s_c, c, K_{s_c,c})$							1.678	1.225	0.506	-0.38

<표 3>의 결과로부터 최적 총 서버의 수 s_{c^*}와 주영역의 최적 대기능력 c^*는 다음과 같음을 알 수 있다. 먼저 보조영역의 서비스능력과 관련된 제약식이 없는 경우에는 주영역의 최적 대기능력 $c^*=5$, 최적 총 서버의 수 $s_{c^*}=4$가 되며 보조영역의 서비스능력이 1보다 커야하는 경우에는, 즉 $B_{s,c,K} \geq 1$인 제약식을 갖는 경우에는 $c^*=4$, $s_{c^*}=4$가 되며 $B_{s,c,K} \geq 1.5$인 경우에는 $c^*=6$, $s_{c^*}=5$, $B_{s,c,K} \geq 2$인 경우에는 $c^*=6$, $s_{c^*}=6$, 그리고 $B_{s,c,K} \geq 2.5$인 경우에는 $c^*=7$, $s_{c^*}=7$이 됨을 알 수 있다. 한계분석법에 의한 최적화 절차를 적용한 수치적결과는 결과적으로 주어진 최적화절차가 매우 유용하게 적용될 수 있음을 보여주고 있다.

6. 결 론

본 논문에서는 주영역과 보조영역에서 서로 다른 두 종류의 작업을 수행할 수 있는 이중작업능력을 갖는 서버들로 구성된 서비스시스템의 설계에 관한 문제를 다루었다. 주영역은 고객에 종속적인 서비스정책을 수행하는 대기시스템으로 모형화되고 총 서버의 수, 주영역의 대기능력, 그리고 주영역의 서비스정책의 함수로 산정되는 생산율과 생산율의 함수로 표시되는 수익이 존재하며 반면에 서버와 대기능력의 확보에 따른 비용함수가 존재하여 이들의 차로 표시되는 이익을 최대화시키는 비선형의 정수계획 최적화문제가 모형화하였다. 반면에 보조영역에서 필요로 하는 작업을 수행하기 위하여 요구되는 서비스능력의 하한이 설정되며 주어진 최적화문제의 제약조건으로 적용되었다. 주어진 서비스시스템의 설계모수인 총 서버의 수, 주영역의 대기능력, 그리고 서비스정책에 대한 수행도의 일계특성과 이계특성들이 도출되었고 도출된 특성들에 근거하여 주어진 최적화문제에 유용하게 적용될 수 있는 최적화절차가 수립되었다. 주어진 최적화절차를 적용하여 그 결과로 산정된 수치 예가 제시되어 최적화절차에 대한 이해를 도모하였으며 주어진 최적화절차는 매우 유용하게 활용될 수 있음을 알 수 있다.

참 고 문 헌

[1] Berman, O. and R.C. Larson, "A Queueing Control Model for Retail Services Having Back Room Operations and Cross Trained Workers," Computers and Operations Research, Vol.31(2004), pp.201-222.

[2] Berman, O. and K.P. Sapna, "Optimal Control of Servers in Front and Back Rooms with Correlated Work," IEE Transactions(to appear).

[3] Berman, O., J. Wang, K.P. Sapna, "Optimal Management of Cross-trained Workers in Services with Negligible Switching Costs," European *Journal of Operational Research*, Vol.167(2005), pp.349-369.

[4] Buzacott, J.A. and J.G. Shanthikumar, Stochastic Models of Manufacturing Systems, Prentice Hall, 1993.

[5] Chang, C., X. Cao, M. Pinedo and J.G. Shanthikumar, "Stochastic Convexity for Multidimensional Processes and its Applications," IEEE Transactions. Automatic Control, Vol.36(1991), pp.1347-1355.

[6] Fox, B., "Discrete Optimization via Marginal Analysis," Management Science, Vol.13 (1966), pp.210-216.

[7] Pacheco, A., "Second Order Properties of the Loss Probability in M/M/s/s+c Systems," Technical Report No.1011, School of OR&IE, Cornell University, New York(1992).

[8] Pacheco, A., "Second-Order Properties of the Loss Probability in M/M/s/s+c Systems," Queueing Systems, Vol.15(1994), pp. 289-308.

[9] Shanthikumar, J.G. and D.D, Yao, "Optimal

Server Allocation in a System of Multi-Server Stations," Management Science, Vol.3(1987), pp.1173-1180.

[10] Wein, L.M., "Capacity Allocation in Generalized JacksonNetworks," Operations Research Letters, Vol.8(1989), pp.143-146.