韓國經營科學會誌
第31권 第1號
2006年 3月

두 단계로 구성된 순환대기네트워크의 설계*

김 성 철**

A Design Problem of a Two-Stage Cyclic Queueing Network*

Sung-Chul Kim**

■ Abstract ■

In this paper we consider a design problem of a cyclic queueing network with two stages, each with a local buffer of limited capacity. Based on the theory of reversibility and product-form solution, we derive the throughput function of the network as a key performance measure to maximize. Two cases are considered. In case each stage consists of a single server, an optimal allocation policy of a given buffer capacity and work load between stages as well as the optimal number of customers is identified by exploiting the properties of the throughput function. In case each stage consists of multiple servers, the optimal policy developed for the single server case doesn't hold any more and an algorithm is developed to allocate with a small number of computations a given number of servers, buffer capacity as well as total work load and the total number of customers. The differences of the optimal policies between two cases and the implications of the results are also discussed. The results can be applied to support the design of certain manufacturing and computer/communication systems.

Keyword : Two-stage Cyclic Queueing Network, Simultaneous Optimization, Reversible

1. 서 론

본 논문에서는 두 단계(stage)로 구성된 순환대기네트워크(cyclic queueing network)의 설계에 관한 문제를 다룬다. 각 단계는 하나 또는 복수의 서버(server)로 구성되고 제한된 대기능력(buffer capacity)의 저장소(local storage)에 의하여 각 단계에서의 고객(customer)의 수는 제한된다. 네트워크

논문접수일 : 2005년 5월 9일 논문게재확정일 : 2005년 12월 5일
* 본 논문은 2005학년도 덕성여자대학교 연구비지원으로 이루어졌음.
** 덕성여자대학교 경영학과

내에 존재하는 총 고객의 수는 통제되어 일정하며 각각의 고객은 두 단계를 순차적으로 방문하여 각 단계에서 요구되는 서비스를 제공받는다. 각 단계에서의 고객의 서비스시간은 지수분포(exponential distribution)에 의하며 상호 독립적이다. 이러한 순환대기네트워크와 관련하여 본 논문에서는 네트워크의 생산율함수(throughput function)를 최대화하도록 네 종류의 의사결정변수를 동시에 설계하는 문제를 다룬다. 이는 주어진 총 서버(server)를 두 단계에 배분하는 문제, 주어진 총 대기능력을 두 단계에 배분하는 문제, 네트워크에 존재하는 총 작업물의 수를 결정하는 문제, 그리고 주어진 총 부하(work load)를 두 단계에 할당하는 문제를 동시에 설계하는 것이다. 특히 각 단계가 하나의 서버로 구성된 경우와 복수의 서버로 구성된 경우가 구분되고 두 경우가 각각 독립적으로 설계된다.

Sparaggis and Gong[12]은 하나의 서버로 구성된 두 단계의 순환대기네트워크에 있어서 대기능력을 배분하는 문제를 다루었으며 총 대기능력을 균등분배(even-as-possible split)하는 것이 최적임을 보였다. 이에 반하여 본 논문에서는 복수의 서버로 구성된 경우에는 균등분배가 더 이상 최적이지 못하며 더 많은 부하가 할당된 단계에 더 많은 서버와 대기능력을 할당하는 것이 더 효율적임을 보인다. Sparaggis and Gong[12]은 주어진 최적화문제를 generalized semi-Markov processes(GSMP)[5-7]로 모형화하여 서로 다른 대기능력의 배분에 대하여 수행도를 확률적으로(stocastically) 비교하였다. GSMP는 각 단계의 시간적 요소인 서비스시간과는 무관하게 구조적이며 논리적 특성들(structural and logical properties)에 의하여 시간에 따른(temporal) 수행도를 직접 산정하지 않고 수행도에 대한 결론을 도출한다. 이에 대비하여 본 논문에서는 의사결정 변수들과 주어진 제약조건 하에서 네트워크의 시간적 요소를 직접 고려하여 확률적 변화에 의한 네트워크의 양적인 수행도를 산정함으로써 네트워크의 생산율 함수를 최대화시키는 문제를 다룬다.

두 단계 순환대기네트워크는 제조시스템이나 컴퓨터시스템의 분석 및 설계 등 다양하게 활용될 수 있다[1, 3, 4]. 그러나 제한된 대기능력의 저장소를 갖는 순환대기네트워크는 일반적으로 승법형 해(product form solution)를 적용하여 정확한 해를 산정할 수 있는 용이한 방법들이 존재하지 않으며[8] 결과적으로 주어진 네트워크의 수행도를 산정하거나 설계한다는 것은 매우 어려운 문제가 된다. 그러나 특수한 경우로서 경로변환마코브과정(routing markov process)이 시간적으로 역류가능(time reversible)하면 주어진 네트워크도 또한 역류가능(reversible)하고 승법형 해가 유지되어 이러한 역류성은 정확한 해를 산정하는데 기본이 된다. 예를 들어 중앙서버(central server)를 갖는 네트워크에 있어서는 이러한 역류성이 만족됨을 쉽게 증명할 수 있다[18].

제2장에서는 본 논문과 관련되는 기본적인 내용을 기술한다. 제3장에서는 주어진 두 단계 순환대기네트워크와 수행도에 대하여 간략히 기술한다. 제4장은 하나의 서버로 구성된 네트워크를 다루며 주어진 문제에 대한 최적 배분이 제시된다. 제5장에서는 복수의 서버로 구성된 네트워크가 고려되어 최적배분정책이 도출되고 하나의 서버로 구성된 경우와 비교된다. 그리고 주어진 동시 최적화문제에 대한 최적배분알고리즘이 개발된다. 제6장에서는 수치적 예가 제시되고 얼마간의 논의로써 논문을 마감한다.

2. 기본정의

만약 상태공간(state space) S에서 정의되는 확률과정 $[X(t)]$가 시간 $t_0, t_1, \cdots, t_n \in T$에 있어서 $[X(t_1), \cdots, X(t_n)]$와 $[X(t_0 - t_1), \cdots, X(t_0 - t_n)]$이 동일한 분포를 이루면 역류가능하다고 하고 역류가능한 과정 $[X(t)]$는 정상과정(stationary process)을 이룬다. 정상마코브과정(stationary Markov process)은 다음의 상세균형방정식(detailed balance equations)을 만족시키는 합이 1인 양수(positive number), $\pi(j)$,

$j \in S$,의 집합이 존재할 때 역류가능하다.

$$\pi(j)q(j,k)=\pi(k)q(k,j), \quad \forall j,k \in S. \tag{2.1}$$

여기에서 $q(j,k)$는 상태 j에서 상태 k로의 전이율(transition rate)을 의미한다. 만약 이러한 $\pi(j)$, $j \in S$,의 모임이 존재한다면 이는 주어진 마코브과정의 균형분포(equilibrium distribution)를 이룬다.

역류가능한 확률과정 $\{X(t)\}$의 매우 중요한 특성은 상태공간 S가 부공간 $\Omega(\subset S)$로 절단(truncated)된 경우에도 균형상태(equilibrium state)의 마코브과정은 역시 역류가능하며 균형상태에서 $\pi(j)$, $j \in S$, 를 원과정의 상태확률이라 하면 부공간으로 절단된 경우에 있어서의 상태확률은 $\pi(j)/\sum_{k \in \Omega}\pi(k)$, $j \in \Omega$,가 성립된다는 것이다. 이는 역류가능한 확률과정 $\{X(t)\}$는 그 균등분포에 있어서 상태를 절단하는 것은 확률을 절단하는 것과 동일함을 의미한다.

R을 일련의 실수들의 집합이라고 하자. 두 벡터(vector) x, $y \in R^n$에 대하여 $x_{(1)} \geq x_{(2)} \geq \cdots \geq x_{(n)}$, $y_{(1)} \geq y_{(2)} \geq \cdots \geq y_{(n)}$이라고 하자. Majorization ordering, $x \leq_m y$은 다음과 같이 정의된다[10].

$$\sum_{i=1}^{k} x_{(i)} \leq \sum_{i=1}^{k} y_{(i)}, \quad k=1,\cdots,n-1,$$

그리고

$$\sum_{i=1}^{n} x_{(i)} = \sum_{i=1}^{n} y_{(i)}. \tag{2.2}$$

함수, $f : \Im^n \to R(\Im \subset R)$,가 majorization ordering을 만족시키면, f는 Schur 함수로 정의된다. 즉 $x \leq_m y$이면 $f(x) \leq (\geq) f(y)$가 성립되면 f는 Schur convex(concave) 함수이다.

만약 함수 f가 연속적이고 미분가능하고 $\Im^n$에서 상호대칭(symmetric)이며 다음을 만족시키면 함수 f는 Schur convex(concave)이다.

$$(x_1 - x_2)(\partial/\partial x_1 - \partial/\partial x_2)f(x) \geq (\leq) 0, \forall x \in \Im^n. \tag{2.3}$$

또한 만약 함수 f가 상호대칭인 볼록(convex)/오목(concave) 함수이면 함수 f는 Schur convex/Schur concave 함수이다([10], Chap.3, C.2).

3. 네트워크와 수행도

두 단계로 구성된 제한된 대기능력을 갖는 순환대기네트워크를 보자. 네트워크를 순환하는 총 고객의 수를 N이라 하고 단계 i, $i=1,2$,에 대하여 s_i는 서버의 수, b_i는 서버를 포함하는 대기능력으로 주어진 단계에서 서비스가 완료되었으나 다음 단계의 대기공간에 여유공간이 없는 경우에는 서비스를 완료한 고객은 다음 단계에 진입하지 못하고 주어진 단계에서 서버를 점유하고 서버는 작업을 멈추게 된다. $\sum_{i=1}^{2} b_i > N$를 가정하며 만약 $N \leq \min(b_1, b_2)$이면 주어진 네트워크는 무한의 대기능력을 가짐을 의미한다. s를 서버벡터, b를 대기능력 벡터라고 정의하면 $s=(s_1,s_2)$, 그리고 $b=(b_1,b_2)$를 의미한다. 단계 i에서 고객의 서비스시간은 기대치 $1/\mu_i$인 지수분포에 의하며 여기에서 μ_i는 단계 i의 서버의 서비스율(service rate)을 나타낸다. 네트워크에서의 작업물의 경로변환은 한정되고(finite) 규칙적(regular)인 마코브과정으로[9] 경로변환확률(routing probability)행렬 $[\gamma_{ij}]$에 의한다. γ_{ij}는 단계 i에서 단계 j로의 경로변환확률로서 $\gamma_{ij}=1$, $i,j=1,2$, 그리고 $i \neq j$이다. 단계 i에 대하여 v_i를 방문빈도(visit frequency), ρ_i를 부하(relative workload)라고 정의하면 $v_i=1$, 그리고 $\rho_i=1/\mu_i$임을 쉽게 알 수 있다.

$n_i(t)$, $i=1,2$,를 시간 t에 있어서 단계 i에 존재하는 고객의 수 즉 대기길이(queue length)라고 하자. $n_i(t) \leq b_i$이며 $n(t)=(n_1(t),n_2(t))$는 시간 t에 있어서 대기길이 벡터로 정의된다. 대기길이과정 $n(t)$의 역류성은 경로변환연쇄(routing chain)에 의하여 증명될 수 있으며[8] 경로변환연쇄 Π가 비음의 원소로 구성된 대칭행렬(symmetric matrix) A와 양수의 대각(diagonal)원소로 구성된 대각행렬 D의 곱으로 표시될 때 즉 $\Pi=AD$인 경우에 역류성이 성립한다. 그러므로 주어진 순환대기네트워크의 경로변환의 역류성은 쉽게 증명될 수 있으며 대기길이 과정의 역류성이 성립한다.

단계 i, $i=1,2$,의 대기길이가 n_i일 때 단계 i의 서

비스율을 $\mu_i(n_i)$라고 하면 만약 $n_1 \le s_i$이면 $\mu_i(n_i) = n_i\mu_i$, $n_i > s_i$이면 $\mu_i(n_i) = s_i\mu_i$가 된다. $\pi_i(n_i)$를 단계 i에서의 대기길이가 n_i일 주변확률(marginal probability)이라 하면 주어진 네트워크의 생산율함수는 $\Sigma_{n_i}\mu_i(n_i)\pi_i(n_i)$로 산정될 수 있다. 주어진 생산율함수는 네트워크의 수행도를 표시하는 측정치로 네트워크를 최적화하는데 기본이 된다.

4. 각 단계가 하나의 서버로 구성된 경우

먼저 가장 간단한 경우로서 각 단계가 하나의 서버를 갖는 경우 즉 $s=(1,1)$인 경우에 있어서 주어진 제약조건하에서 네트워크의 생산율함수를 최대화하는 최적 총 고객의 수 N^*, 최적 대기능력배분 $b^*=(b_1^*,b_2^*)$, 그리고 최적 부하할당 $\rho^*=(\rho_1^*,\rho_2^*)$를 설계하는 문제를 보기로 한다. 일반성을 해함이 없이 $b_1 = \min(b_1,b_2)$라고 하면 네트워크의 생산율함수를 $TH_N(s,b)$이라 정의하면 생산율함수 $TH_N(s,b)$는 3장에서 언급된 역류성에 의하여 다음과 같이 유도될 수 있다.

$$TH_N(s,b) =$$
$$\sum_{k=0}^{N-1}\rho_1^{N-1-k}\rho_2^k \Big/ \sum_{k=0}^{N}\rho_1^{N-k}\rho_2^k, \qquad N \le b_1,$$
$$\sum_{k=N-b_1}^{N-1}\rho_1^{N-1-k}\rho_2^k \Big/ \sum_{k=N-b_1}^{N}\rho_1^{N-k}\rho_2^k, \qquad b_1 < N < b_2,$$
$$\sum_{k=N-b_1}^{b_2-1}\rho_1^{N-1-k}\rho_2^k \Big/ \sum_{k=N-b_1}^{b_2}\rho_1^{N-k}\rho_2^k, \qquad b_2 \le N < b_1+b_2, \tag{4.1}$$

이며 $\rho_1 \neq \rho_2$과 $\rho_1 = \rho_2$인 경우 각각에 대하여 $TH_N(s,b)$은 다음과 같이 간단히 표현될 수 있다. 먼저 $\rho_1 \neq \rho_2$인 경우에는 다음과 같다.

$$TH_N(s,b) =$$
$$(\rho_1^N - \rho_2^N)/(\rho_1^{N+1} - \rho_2^{N+1}), \quad N \le b_1,$$
$$(\rho_1^{b_1} - \rho_2^{b_1})/(\rho_1^{b_1+1} - \rho_2^{b_1+1}), \quad b_1 < N < b_2,$$
$$(\rho_1^{b_1+b_2-N} - \rho_2^{b_1+b_2-N})/(\rho_1^{b_1+b_2+1-N} - \rho_2^{b_1+b_2+1-N}),$$
$$b_2 \le N < b_1+b_2. \tag{4.2}$$

$\rho_1 = \rho_2$인 경우에는 다음과 같다.

$$TH_N(s,b) =$$
$$N\rho_1^{N-1}/((N+1)\rho_1^N), \quad N \le b_1,$$
$$b_1\rho_1^{b_1-1}/((b_1+1)\rho_1^{b_1}), \quad b_1 < N < b_2,$$
$$(b_1+b_2-N)\rho_1^{b_1+b_2-N-1}/((b_1+b_2+1-N)\rho_1^{b_1+b_2-N}),$$
$$b_2 \le N < b_1+b_2. \tag{4.3}$$

다음의 결론들이 유도될 수 있다.

[정리 1] 대기능력벡터 $b=(b_1,b_2)$이 주어지면 생산율함수 $TH_N(s,b)$은 총 고객수 N에 대하여 $1 \le N \le b_1+b_2-1$에서 상호대칭(symmetric)인 오목함수이다. 부연하면 생산율은 총 고객수 N에 대하여 $N \le b_1$에서는 증가하여 $b_1 \le N \le b_2$에서는 동일하게 최대값을 갖고 $b_2 \le N < b_1+b_2$에서는 감소함을 의미한다.

<증명> 주어진 결과는 $N < b_1$에서는 다음이 성립됨을 보임으로써 쉽게 증명될 수 있다.

$$TH_N(s,b) - TH_{N-1}(s,b) \ge TH_{N+1}(s,b) - TH_N(s,b) \ge 0. \tag{4.4}$$

먼저 $\rho_1 \neq \rho_2$인 경우를 보기로 한다. 여기에서

$$TH_N(s,b) - TH_{N-1}(s,b) = (\rho_1\rho_2)^{N-1}(\rho_1-\rho_2)^2 \div [(\rho_1^N - \rho_2^N)(\rho_1^{N+1} - \rho_2^{N+1})] \ge 0, \tag{4.5}$$

이므로 다음이 성립한다.

$$(TH_N(s,b) - TH_{N-1}(s,b)) - (TH_{N+1}(s,b) - TH_N(s,b)) = (\rho_1\rho_2)^{N-1}(\rho_1-\rho_2)^3(\rho_1^{N+1} + \rho_2^{N+1}) \div [(\rho_1^N - \rho_2^N)(\rho_1^{N+1} - \rho_2^{N+1})(\rho_1^{N+2} - \rho_2^{N+2})] \ge 0. \tag{4.6}$$

$N = b_1$인 경우에는 식 (4.5)에 의하여 $TH_{b_1-1}(s,b) \le TH_{b_1}(s,b)$가 성립하며 식 (4.1)또는 식 (4.2)에 의하여 $TH_{b_1}(s,b) = TH_{b_1+1}(s,b)$가 되어 식 (4.4)가 성립한다. $b_2 \le N < b_1+b_2$인 경우에는 식 (4.5)과 식 (4.6)의 부등식 $\ge$이 부등식 $\le$으로 변경되고 마찬가지로 성립한다. 또한 $\rho_1 = \rho_2$인 경우는 식 (4.3)에서와 같

이 생산율함수 $TH_N(s,b)$이 단일변수로 치환되어 식 (4.4)는 증명방법은 유사하나 더 간단하게 증명될 수 있다. Q.E.D.

[정리 1]은 다음과 같이 설명될 수 있다. 생산율함수 $TH_N(s,b)$은 총 고객수 N이 증가함에 따라 $N \le b_1$에서는 이용률(utilization)이 증대되어 생산율이 증가하며 $b_2 \le N < b_1 + b_2$에서는 네트워크의 혼잡도(congestion)가 증대되어 생산율이 감소된다.

위의 직접적인 결과로 만약 대기능력벡터 $b=(b_1, b_2)$이 주어지면 생산율함수 $TH_N(s,b)$을 최대로 하는 총 고객수 N^*는 대기능력 b_1과 b_2 사이에서 설계하는 것이 최적임을 알 수 있다.

다음으로는 주어진 총 대기능력 $b_1+b_2=B$를 배분하여 최적 대기능력배분 $b^*=(b_1^*, b_2^*)$을 구하는 문제를 보기로 한다.

[정리 2] 총 대기능력 B가 주어지면 생산율함수 $TH_{N^*}(s,b)$는 대기능력벡터 $b=(b_1,b_2)$.에 대하여 Schur concave함수이다.

<증명> 생산율함수 $TH_{N^*}(s,b)$도 대기능력벡터 $b=(b_1,b_2)$에 대하여 상호대칭함수이므로 $TH_{N^*}(s,b)$가 $b=(b_1,b_2)$에 대하여 오목함수임을 증명함으로써 $TH_{N^*}(s,b)$가 $b=(b_1,b_2)$에 대하여 Schur concave함수임을 증명하고자 한다([10], Chap.3, C.2). 생산율함수 $TH_{N^*}(s,b)$이 대기능력벡터 $b=(b_1,b_2)$에 대하여 오목함수는 다음에 의한다.

$$2TH_{N^*}(s,(b_1,b_2)) \ge TH_{N^*}(s,(b_1-1,b_2+1)) + TH_{N^*}(s,(b_1+1,b_2-1)). \quad (4.7)$$

식 (4.7)은 다음과 같이 정리될 수 있다.

$$TH_{N^*}(s,(b_1,b_2)) - TH_{N^*}(s,(b_1-1,b_2+1)) \ge TH_{N^*}(s,(b_1+1,b_2-1)) - TH_{N^*}(s,(b_1,b_2)). \quad (4.8)$$

[정리 1]에 의하여 일반성을 저해함이 없이 $N^* = \min(b_1,b_2)$으로 산정할 수 있으므로 식 (4.8)은 다음과 같이 정리될 수 있다.

$$TH_{\min(b_1,b_2)}(s,(b_1,b_2)) - TH_{\min(b_1-1,b_2+1)}(s,(b_1-1,b_2+1)) \ge TH_{\min(b_1+1,b_2-1)}(s,(b_1+1,b_2-1)) - TH_{\min(b_1,b_2)}(s,(b_1,b_2)). \quad (4.9)$$

식 (4.9)는 총 대기능력 B가 짝수인 경우와 홀수인 경우에 있어서 다음과 같이 정리될 수 있다. 먼저 총대기능력 B가 짝수인 경우를 정리하면 다음과 같다.

$b_1+1 \le b_2-1$인 경우,

$$TH_{b_1}(s,(b_1,b_2)) - TH_{b_1-1}(s,(b_1-1,b_2+1)) \ge TH_{b_1+1}(s,(b_1+1,b_2-1)) - TH_{b_1}(s,(b_1,b_2)), \quad (4.10)$$

$b_1=b_2$인 경우,

$$TH_{b_1}(s,(b_1,b_2)) - TH_{b_1-1}(s,(b_1-1,b_2+1)) \ge TH_{b_2-1}(s,(b_1+1,b_2-1)) - TH_{b_2}(s,(b_1,b_2)), \quad (4.11)$$

$b_1-1 \ge b_2+1$인 경우,

$$TH_{b_2}(s,(b_1,b_2)) - TH_{b_2+1}(s,(b_1-1,b_2+1)) \ge TH_{b_2-1}(s,(b_1+1,b_2-1)) - TH_{b_2}(s,(b_1,b_2)). \quad (4.12)$$

먼저 $b_1+1 \le b_2-1$인 경우를 보자. 생산율 함수 $TH_N(s,b)$가 $b_1 \le N \le b_2$에서는 N과 b_2의 값에 관계없이 모두 동일하므로 $b_1+1 \le b_2-1$인 경우의 식 (4.10)은 [정리 1]에 주어진 식 (4.6)에서 N을 b_1으로 치환한 것과 같다. 그러므로 식 (4.8)이 성립한다. 또한 마찬가지로 생산율 함수 $TH_{b_1}(s,b)$가 $b_1+1 \le b_2-1$의 범위에서 b_1에 대하여 증가하는 특성, 즉

$$TH_{b_1+1}(s,(b_1+1,b_2-1)) - TH_{b_1}(s,(b_1,b_2)) \ge 0, \quad (4.13)$$

도 앞에 언급된 이유로 식 (4.5)의 N을 b_1으로 치환하여 동일하게 증명될 수 있다. $b_1=b_2$인 경우의 식 (4.11)은 $TH_{b_1}(s,b)$가 b_1에 대하여 증가하는 특성과 상호대칭성에 의하여 $b_1 > b_2$가 되면 $N^*=b_2$임을 이용하여 식 (4.11)의 좌변은 ≥ 0, 부등식의 우변은 ≤ 0이 되어 또한 성립된다. 마지막으로 $b_1-1 \ge b_2+1$인 경우에는 식 (4.12)의 좌변과 우변 모두에 -1을 곱하면 다음과 같이 정리된다.

$$TH_{b_2}(s,(b_1,b_2))-TH_{b_2-1}(s,(b_1+1,b_2-1))$$
$$\geq TH_{b_2+1}(s,(b_1-1,b_2+1))-TH_{b_2}(s,(b_1,b_2)). \quad (4.14)$$

식 (4.14)은 [정리 1]에 주어진 생산율 함수 $TH_N(s,b)$의 총 고객수 N에 대한 상호대칭성에 의하여 $b_1+1\leq b_2-1$인 경우에서와 마찬가지로 성립됨을 알 수 있다. 이제 총 대기능력 B가 홀수인 경우에는 $b_1+1<b_2-1$, $b_1=b_2-1$, $b_1-1=b_2$, 그리고 $b_1-1>b_2$로 구분되어 총 대기능력 B가 짝수인 경우와 유사한 방법으로 쉽게 증명될 수 있다. 그러므로 생산율함수 $TH_{N^*}(s,b)$의 대기능력벡터 $b=(b_1,b_2)$에 대한 Schur concavity는 생산율함수 $TH_{N^*}(s,b)$가 $b=(b_1,b_2)$에 대하여 대칭함수이므로 쉽게 결론지어질 수 있다. 또한 $\rho_1=\rho_2$인 경우도 [정리 1]에서와 같이 단일변수로서 매우 수월하게 증명될 수 있다. Q.E.D.

그러므로 다양한 대기능력벡터가 주어지면 majorization ordering을 적용하여 생산율을 쉽게 비교할 수 있다. 이는 두 대기능력벡터 b^1과 b^2에 있어서 $b^1\leq_m b^2$이 성립되면 또한 $TH_{N^*}(s,b^1)\geq TH_{N^*}(s,b^2)$이 성립되어 생산율을 직접 산정하지 않고도 생산율을 확률적으로 비교할 수 있음을 의미한다. 결과적으로 균등한 배분 즉 $b_1=b_2$이나 $b_1+1=b_2$가 majorization 하에서 가장 적은 벡터이므로 생산율 $TH_N(s,b)$을 최대화한다.

이제 주어진 총 부하를 각 단계에 할당하는 문제를 보기로 하자. 부호의 편의를 위하여 주어진 서버벡터 $s=(s_1,s_2)$와 대기능력벡터 $b=(b_1,b_2)$에 있어서 $TH_N(\rho)$을 부하벡터 $\rho=(\rho_1,\rho_2)$에 대한 생산율함수라고 정의한다.

[정리 3] 생산율함수 $TH_N(\rho)$은 부하벡터 $\rho=(\rho_1,\rho_2)$에 대하여 Schur concave함수이다.

<증명> 식 (2.3)으로부터 다음이 성립함을 보임으로써 쉽게 증명될 수 있다.

$$(\rho_1-\rho_2)(\partial/\partial\rho_1-\partial/\partial\rho_2)TH_N(\rho)\leq 0. \quad (4.15)$$

먼저 $N\leq b_1$인 경우를 증명하기로 한다. 이를 위하여 다음을 정의한다.

$$\Phi_N(\rho)=\sum_{n=0}^{N}\rho_1^n\rho_2^{N-n}. \quad (4.16)$$

그러므로,

$$\begin{aligned}(\partial/\partial\rho_1)\Phi_N(\rho)&=\sum_{n=1}^{N}n\rho_1^{n-1}\rho_2^{N-n}\\&=\sum_{n=0}^{N-1}(n+1)\rho_1^n\rho_2^{N-1-n}\\&=\rho_1\sum_{n=1}^{N-1}n\rho_1^{n-1}\rho_2^{N-1-n}+\Phi_{N-1}(\rho)\\&=\rho_1(\partial/\partial\rho_1\Phi_{N-1}(\rho))+\Phi_{N-1}(\rho). \quad (4.17)\end{aligned}$$

식 (4.17)을 총 고객수 N에 대하여 반복적으로 적용하면 생산율함수 $TH_N(\rho)$의 부하벡터 ρ에 대한 대칭성에 의하여 다음을 유도할 수 있다.

$$(\partial/\partial\rho_i)\Phi_N(\rho)=\sum_{n=0}^{N-1}\rho_i^n\Phi_{N-1-n}(\rho),\ i=1,2. \quad (4.18)$$

그러므로,

$$\begin{aligned}&(\partial/\partial\rho_i)TH_N(\rho)\\&=\Phi_N^{-2}(\rho)\{\sum_{n=0}^{N-2}\rho_i^n[\Phi_{N-2-n}(\rho)\Phi_N(\rho)\\&\quad-\Phi_{N-1}(\rho)\Phi_{N-1-n}(\rho)]-\rho_i^{N-1}\Phi_{N-1}(\rho)\},\\&\quad i=1,2, \quad (4.19)\\&(\partial/\partial\rho_1-\partial/\partial\rho_2)TH_N(\rho)\\&=\Phi_N^{-2}(\rho)\{\sum_{n=0}^{N-2}(\rho_1^n-\rho_2^n)[\Phi_{N-2-n}(\rho)\Phi_N(\rho)\\&\quad-\Phi_{N-1}(\rho)\Phi_{N-1-n}(\rho)]\\&\quad-(\rho_1^{N-1}-\rho_2^{N-1})\Phi_{N-1}(\rho)\}. \quad (4.20)\end{aligned}$$

[정리 1]에 의하여 $N\leq b_1$에 대하여 $TH_N(\rho)$가 N에 대하여 증가함수이므로, 즉

$$\Phi_{N-2-n}(\rho)\Phi_N(\rho)-\Phi_{N-1}(\rho)\Phi_{N-1-n}(\rho)\leq 0, \quad (4.21)$$

의 부등식이 만족되므로 식 (4.15)가 성립한다. $b_1<N<b_2$인 경우는 같은 방법으로 증명되며 $b_2\leq N<b_1+b_2$인 경우는 $N\leq b_1$의 경우와 상호대칭으로 주어진 결과가 성립된다. 주어진 결과는 Yao[17]의 특수한 경우로 볼 수 있다. Q.E.D.

그 결과로 만약 총 부하 즉 $\rho_1+\rho_2=L$이 주어져 있다면 서로 다른 부하벡터 또한 majorization ordering에 의하여 쉽게 비교될 수 있다. 즉 부하벡터 ρ^1과 ρ^2에 있어서 $\rho^1 \le_m \rho^2$은 $TH_N(\rho^1) \ge TH_N(\rho^2)$임을 의미한다. 그러므로 majorization하에서 가장 적은 부하벡터 즉 균등부하 $\rho=(L/2, L/2)$가 생산율함수 $TH_N(s,b)$을 최대화한다.

그러므로 각 단계가 하나의 서버로 구성되어 있는 경우에는 majorization ordering이 적용되어 다양한 대기능력과 부하벡터에 대한 설계를 비교할 수 있는 방법을 제시한다. 또한 결과적으로 하나의 서버로 구성된 네트워크에서 이러한 변수들을 동시에 최적화시키기 위해서는 대기능력은 두 단계에 가능한 한 균등하게 배분하고 총 고객수는 배분된 대기능력과 동일하게 그리고 부하는 두 단계에 균등하게 할당함이 최적임을 알 수 있다.

5. 각 단계가 복수의 서버로 구성된 경우

각 단계가 복수의 서버로 구성된 경우에는 서버벡터 $s=(s_1,s_2)$, 대기능력벡터 $b=(b_1,b_2)$, 부하벡터 $\rho=(\rho_1,\rho_2)$, 그리고 총 고객수 N의 함수로서 정의되는 생산율함수 $TH_N(s,b)$은 다음과 같다.

$$TH_N(s,b) =$$
$$G_{N-1}^{0,N-1}(s,b)/G_N^{0,N}(s,b), \qquad N \le b_1,$$
$$G_{N-1}^{N-b_1,N-1}(s,b)/G_N^{N-b_1,N}(s,b), \quad b_1 < N < b_2,$$
$$G_{N-1}^{N-b_1,b_2-1}(s,b)/G_N^{N-b_1,b_2}(s,b), \quad b_2 \le N < b_1+b_2.$$

여기에서

$$G_N^{m,n}(s,b) = \sum_{k=m}^{n} \rho_1^{N-k}\rho_2^k/[\theta_1(N-k)\theta_2(k)]$$

그리고

$$\theta_i(n_i) = \begin{cases} n_i!, & n_i \le s_i, \\ s_i! s_i^{n_i-s_i}, & n_i > s_i. \end{cases} \qquad (5.1)$$

각 단계가 복수의 서버로 구성된 경우에는 하나로 서버로 구성된 경우와는 달리 4장에서의 생산율에 관련된 결과나 특성들이 더 이상 일반적으로 성립되지 않으며 역의 예(counterexample)들이 쉽게 보여 질 수 있다. 여러 다른 모수에 대한 생산율의 수치예가 6장에 제시되며 이러한 수치적 결과에 의하여 다음의 특성들을 유추할 수 있다.

하나의 서버로 구성된 경우에는 주어진 대기능력벡터 $b=(b_1,b_2)$에 대하여 생산율함수를 최대화시키는 최적 총 고객수 N^*는 b_1과 b_2 범위 내에 존재하였으나 복수의 서버로 구성된 경우에는 최적 총 고객수 N^*는 b_1과 b_2사이에 존재하나 경우에 따라서는 b_2를 초과하여 존재할 수 있다.

하나의 서버로 구성된 경우에는 대기능력의 균등한 배분이 생산율을 최대화하였으나 복수의 서버로 구성된 경우에는 불균등 부하 하에서는 균등한 대기능력의 배분뿐만이 아니라 균등한 서버의 배분도 생산율을 최대화하지 못한다. 뿐만이 아니라 하나의 서버의 경우에는 균등부하가 생산율을 최대화시키는 최적 부하정책이었으나 복수의 서버로 구성된 경우에 불균등부하에서도 네트워크의 생산율이 최대화됨을 알 수 있다.

하나의 서버로 구성된 경우에는 생산율은 대기능력에 대하여 상호대칭이었으나 복수의 서버로 구성된 경우에는 생산율함수가 대기능력뿐만이 아니라 서버에 대하여도 더 이상 상호대칭이 아님을 알 수 있다.

이제 네트워크 내에 배분될 총 서버와 총 대기능력 즉 $s_1+s_2=C$ 그리고 $b_1+b_2=B$가 주어져 있다고 하자. $s=(s_1,s_2)$, $s^R=(s_2,s_1)$, $b=(b_1,b_2)$, 그리고 $b^R=(b_2,b_1)$라 하고 $s_1 \le s_2$, $b_1 \le b_2$라고 하자. 또한 정의된 생산율함수 $TH_N(s,b)$에 있어서 $s=(s_1,s_2)$ 그리고 $b=(b_1,b_2)$이 주어진 경우 총 고객수 N에 대한 최대 생산율을 $TH_{N^*}(s,b)$라고 정의하고 $s=(s_1,s_2)$가 주어졌을 때 대기능력벡터 $b=(b_1,b_2)$와 총 고객수 N에 대한 최대 생산율을 $TH_{N^*}(s,b^*)$이라고 하자.

[정리 4] 만약 $\rho_1 \le \rho_2$이면 $TH_N(s^R,b^R) \le TH_N(s,b)$.

<증명> [정리 4]를 증명하기 위해서 $N \le b_1$에 대하여 다음을 증명하고자 한다.

$$G_{N-1}^{0,N-1}(s^R,b^R)\, G_N^{0,N}(s,b) \le G_{N-1}^{0,N-1}(s,b)\, G_N^{0,N}(s^R,b^R). \quad (5.2)$$

식 (5.2)을 대수적 전개함으로써 LHS(left-hand side)는

$$\text{LHS} = \xi_0\rho_1^{2N-1} + \xi_1\rho_1^{2N-2}\rho_2 + \cdots + \psi_{2N-2}\rho_1\rho_2^{2N-2} + \psi_{2N-1}\rho_2^{2N-1},$$

그리고 RHS(right-hand side)는 다음과 같이 표현될 수 있다.

$$\text{RHS} = \psi_{2N-1}\rho_1^{2N-1} + \psi_{2N-2}\rho_1^{2N-2}\rho_2 + \cdots + \xi_1\rho_1\rho_2^{2N-2} + \xi_0\rho_2^{2N-1}$$

여기에서 $k=0,\cdots,N-1$에 대하여

$$\xi_k = \sum_{l=0}^{k} 1/[\theta_1(l)\theta_1(N-k+l)\theta_2(k-l)\theta_2(N-1-l)],$$

$$\psi_{2N-1-k} = \sum_{l=0}^{k} 1/[\theta_1(l)\theta_1(N-1-k+l)\theta_2(k-l)\theta_2(N-l)] \quad (5.3)$$

이며 모두 양수이다. ρ_i, $i=1,2$,에 대한 대칭성에 의하여 다음이 성립한다.

$$\text{RHS} - \text{LHS} = \sum_{k=0}^{N-1} (\rho_1\rho_2)^k (\rho_2^{2N-1-2k} - \rho_1^{2N-1-2k}) \times (\xi_k - \psi_{2N-1-k}),$$

여기에서

$$\xi_k - \psi_{2N-1-k} = \sum_{l=0}^{k} [1/(\theta_1(l)\theta_1(N-k+l)\theta_2(k-l)\theta_2(N-l)] \times [\min(N-l,s_2) - \min(N-k+l,s_1)]. \quad (5.4)$$

$s_1=1$인 경우에는 모든 k와 l에 대하여 $\min(N-k+l,s_1)=1$이며 식 (5.4)의 두 번째 각괄호(brackets) 안의 값은 음수 값을 가질 수 없으며 식 (5.2)는 엄격한(strict) 부등식으로 만족된다. 이제 s_1의 값이 증가하여 $\min(s_2,N)$과 동일하여지면 $l>k/2$에 대하여 두 번째 각괄호 안의 모든 값들이 양수가 될 수 없으나 식 (5.1)에 의하여 정의되는 생산율함수가 상호대칭함수가 되어 결과적으로 $TH_N(s,b)=TH_N(s^R,b^R)$이 성립된다. 그러므로 식 (5.2)의 부등식은 등식으로 만족된다. 이 경우에는 대기능력이 무한한 것과 같으며 Shanthikumar and Yao[13]는 폐쇄 대기네트워크에서 각 작업장의 서비스율이 대기길이에 대하여 증가하는 concave이면 각 작업장의 서버의 수에 대하여 생산율이 증가하는 concave임을 보이고 있다. 그러므로 $1 \le s_1 < s_2$에 의하여 식 (5.2)가 성립된다.

위의 결과를 대수적으로 설명하면 만약 $s_1' > s_1$이라면 $\min(N-k+l,s_1) \le \min(N-k+l,s_1')$이고 $\min(N-l,s_2) - \min(N-k+l,s_1) \ge \min(N-l,s_2) - \min(N-k+l,s_1')$이므로 $s_1=s_2$에서 식 (5.4)가 0이 되는 것을 증명하면 $s_1 \le s_2$에 대하여 식 (5.4)가 0보다 크거나 0이 되어 식 (5.2)가 성립한다. 이제 $s_1=s_2$라 하면 주어진 k, $k=0,1,\cdots,N-1$,에 있어서 $l=c$와 $l=k-c$, $c=0,\cdots,\lfloor k/2 \rfloor$,에 대하여 두 번째 각 괄호 값은 절대 값이 같고 부호가 다르며 첫 번째 각 괄호 값은 서로 동일하여 $\nu(c)=-\nu(k-c)$가 성립하고 식 (5.4)는 0이 된다. 여기에서 $\nu(l)=[1/\theta_1(l)\theta_1(N-k+l)\theta_2(k-l)\theta_2(N-l)][\min(N-l,s_2)-\min(N-k+l,s_1)]$, $l=0,\cdots,k$,를 $\lfloor k/2 \rfloor$는 $k/2$보다 크지 않은 가장 큰 정수를 의미한다. 그러므로 $s_1 \le s_2$에 의하여 식 (5.2)가 만족된다.

이를 좀 더 부연설명하면 식 (5.4)에 있어서 만약 s_1이 증가하면 가장 큰 값을 갖는 $k(=N-1)$와 그에 상응하는 가장 큰 값의 $l(=k)$로부터 두 번째 각괄호 안의 값은 하나하나씩 음수 값을 가지기 시작한다. 결과적으로 $s_1=1$에서 엄격한 부등식으로 만족된 식 (5.2)는 $s_1=\min(s_2,N)$에서는 등식으로 만족되어 생산율함수가 서버 수 s_1에 대하여 증가함수임을 알 수 있다. $b_2 \le N \le b_1+b_2$인 경우는 $N \le b_1$인 경우와 상호 대칭되어 마찬가지로 증명될 수 있다. $b_1 < N < b_2$의 경우에는 $N=b_1+p$라 하고 다음을 정

의한다.

$$TH_N(s^R,b^R) = G_{N-1}^{0,b_1-1}(s^R,b^R)/G_N^{0,b_1}(s^R,b^R),$$

여기에서

$$G_N^{0,b_1}(s^R,b^R)=\sum_{k=0}^{b_1}\rho_1^{N-k}\rho_2^k/[\theta_2(N-k)\theta_1(k)]. \quad (5.5)$$

다음을 증명하고자 한다.

$$G_{N-1}^{0,b_1-1}(s^R,b^R)\,G_N^{N-b_1,N}(s,b) \le G_{N-1}^{N-b_1,N-1}(s,b)\,G_N^{0,b_1}(s^R,b^R). \quad (5.6)$$

식 (5.6)의 각 항들을 식 (5.3)에서와 같이 명시적(explicitly)으로 표현하여 유도하면

$$\text{RHS - LHS} = \sum_{k=p}^{N-1}(\rho_1\rho_2)^k(\rho_2^{2N-1-2k}-\rho_1^{2N-1-2k}) \times(\xi_k-\psi_{2N-1-k})$$

가 되며 여기에서

$$\xi_k-\psi_{2N-1-k} =\sum_{l=0}^{k-p}[1/(\theta_1(l)\theta_1(N-k+l)\theta_2(k-l)\theta_2(N-l)] \times[\min(N-l,s_2)-\min(N-k+l,s_1)]. \quad (5.7)$$

식 (5.7)에서도 $s_1=1$인 경우에는 두 번째 각괄호안의 양들이 모두 비음이며 식 (5.6)은 엄격한 부등식으로 성립되며 $N\le b_1$인 경우와 마찬가지로 $b_1=N$, $s_1=s_2$인 경우 $TH_N(s,b) = TH_N(s^R,b^R)$이 성립되며 식 (5.6)은 등식으로 만족되며 위의 결과가 증명된다. 이제 Shanthikumar and Yao[15]에 의하면 제한된 대기능력의 순환대기네트워크에 있어서 생산율은 서비스율, 대기능력에 대하여 증가하는 함수임을 보이고 있으며 결과적으로 $1\le s_1<s_2$에 의하여 식 (5.2)가 성립된다.

식 (5.7)도 식 (5.4)에서와 마찬가지로 $s_1=s_2$에 대하여 식 (5.7)이 0이 됨을 보이고 $s_1<s_2$이므로 식 (5.6)이 만족됨을 보일 수 있다. 즉 s_1이 증가하면 가장 큰 k와 그에 상응하는 가장 큰 l에서부터 두 번째 각괄호의 값들이 하나하나씩 양수 값을 갖지 못하게 된다. 그러나 식 (5.7)에서 주어진 p, $p=1,2,\cdots,$와 k, $k=p,\cdots,N-1$에 대하여 가장 큰 값을 갖는 l, $l=k-p+1,\cdots,k,$의 항들이 제거되어 가는데 이들은 음수 값을 가질 가능성이 가장 큰 항들이다. Q.E.D.

그러므로 더 많은 부하가 할당된 단계에 더 많은 서버와 대기능력을 할당하는 것이 생산율을 증대시킴을 결론지을 수 있다. 주어진 결과는 서버와 대기능력의 배분공간(allocation space)을 현저히 제한하여 주어진 동시 최적화문제의 복잡성(complexity)을 줄이는데 성공적으로 적용될 수 있을 것이다.

복수의 서버로 구성된 경우에는 서버배분, 대기능력배분, 그리고 부하할당에 대하여 상호대칭성이 더 이상 성립되지 않으나 이러한 의사결정변수들에 대하여 광의의 오목성과 관련된 특성들이 성립함을 알거나 추론(conjecture)할 수 있다. 폐쇄대기네트워크(closed queueing network)에 있어서 고려되는 다양한 의사결정변수에 대한 이러한 이계특성(second order property)들에 대한 결과는 그 중에서도 Stecke and Solberg[16], Shaked and Shanthikumar[11], Shanthikumar and Yao[13-15], 등에서 찾아 볼 수 있다. 이는 생산율 함수 $TH_N(s,b)$은 $s_1+s_2=C$, $b_1+b_2=B$가 주어져 있을 때 N에 대하여 증가하였다가 감소하고, $TH_{N^*}(s,b)$은 $s_1+s_2=C$가 주어져 있을 때 $b=(b_1,b_2)$, $b_1=s_1,\cdots,B-s_2$에 대하여 증가하였다가 감소하며 $TH_{N^*}(s,b^*)$은s_1, $s_1=1,\cdots,C-s_2$에 대하여 증가하였다가 감소함을 의미한다. 주어진 특성은 추구되는 동시 최적화문제에 Fox[2]의 한계배분알고리즘(marginal allocation algorithm)이 적용할 수 있게 한다.

이제 부하벡터 $\rho=(\rho_1,\rho_2)$가 주어지고 $\rho_1\le\rho_2$라고 하자. 총 서버의 수$s_1+s_2=C$와 총 대기능력 $b_1+b_2=B$ 또한 주어졌다고 하자. 여기에서는 의사결정변수에 대한 최적해 즉 $N^*,s^*=(s_1^*,s_2^*)$, 그리고$b^*=(b_1^*,b_2^*)$를 동시에 도출할 수 있는 간단한 알고리즘을

제시하고자 한다. $\lceil x \rceil$는 x보다 적지 않은 가장 적은 정수를 의미한다.

주어진 알고리즘은 다음과 같이 적용될 수 있다. 전체 알고리즘은 세 단계의 순환과정(loop)으로 구성된다. 전체(global)순환은 서버의 배분을 중간(middle)순환은 대기능력의 배분을 그리고 지역(local)순환은 총 고객수를 결정한다.

전체순환에 있어서는 먼저 초기화를 위하여 $s^0=(s_1^0,s_2^0)=(C-\lceil C/2\rceil,\lceil C/2\rceil)$, $s^k=(s_1^0-k,s_2^0+k)$를 설정한다. $k=0$으로부터 k를 하나씩 증가시켜 가면서 처음으로 $\triangle_s^k TH=TH_{N^*}(s^k,b^{k^*})-TH_{N^*}(s^{k+1},b^{k+1^*})>0$이 될 때까지 $\triangle_s^k TH$을 산정한다. 여기에서 $TH_{N^*}(s^k,b^{k^*})$은 주어진 s^k에 대하여 N과 b에 대한 최대 생산율을 의미한다. 처음으로 $\triangle_s^k TH>0$을 만족시키는 s^k가 최적 서버벡터이다. 즉 $s^*=s^k=(C-\lceil C/2\rceil-k,\lceil C/2\rceil+k)$임을 의미한다.

중간순환에 있어서는 주어진 서버벡터 s^k에 대하여 최적 대기능력의 배분 b^{k^*}가 도출된다. 먼저 $b^0=(b_1^0,b_2^0)=(B-\lceil B/2\rceil,\lceil B/2\rceil)$, $b^l=(b_1^0-l,b_2^0+l)$을 설정한다. $l=0$으로부터 l을 한 번에 하나씩 증가하면서 처음으로 $\triangle_b^l TH=TH_{N^*}(s^k,b^l)-TH_{N^*}(s^k,b^{l+1})>0$이 될 때까지 $\triangle_b^l TH$을 산정한다. 여기에서 $TH_{N^*}(s^k,b^l)$은 s^k와 b^l이 주어져 있을 때 N에 대한 생산율의 최대값을 의미한다. 처음으로 $\triangle_b^l TH>0$을 만족시키는 b^l이 최적 대기능력벡터가 된다. 즉 $b^{k^*}=b^l=(B-\lceil B/2\rceil-l,\lceil B/2\rceil+l)$임을 의미한다.

지역순환에서는 주어진 서버벡터 s^k와 대기능력벡터 b^l에 대하여 최적 총 고객수 N^*를 도출한다. 먼저 초기화를 위하여 $N^0=b_1^l, N^p=N^0+p$를 설정한다. 그리고 처음으로 $\triangle_N^p TH=TH_{N^p}(s^k,b^l)-TH_{N^{p-1}}(s^k,b^l)>0$이 될 때까지 p를 하나씩 증가시키면서 $\triangle_N^p TH$를 산정한다. 처음으로 $\triangle_N^p TH>0$를 만족시키는 N^p가 네트워크 내에 존재하는 최적 총 고객수가 된다. 즉 $N^*=b_1^l+p$임을 의미한다.

전체순환과 중간순환에서의 초기화의 설정은 더 많은 부하가 할당된 단계에 더 많은 서버와 대기능력이 배분되어야 한다는 특성에 의하여 제안되었다. 지역순환은 각 단계가 복수의 서버를 갖는 경우에는 최적 고객수가 b_1과 b_2 사이에 존재하거나 경우에 따라서는 b_2보다도 클 수 있다는 내용에 기초하였다. 네트워크의 생산율이 의사결정변수에 대하여 증가하였다가 감소한다는 추론에 근거하여 한계배분알고리즘을 적용할 수 있었다.

6. 수치 예와 논의

본 장에서는 의사결정변수에 대한 네트워크의 수행도에 대하여 5장의 결과를 보완하고 일반적인 이해를 높이기 위하여 몇 가지 수치예를 제시하고자 한다. $\rho=(0.3,0.9)$가 주어지고 8개의 서버와 16개의 대기능력을 배분하는 문제를 보기로 한다. <표 1>은 서버배분 $s=(k,8-k)$, $k=1,\cdots,7$와 주어진 서버배분 $s=(s_1,s_2)$에 대하여는 대기능력배분 $b=(l,16-l), l=s_1,\cdots,16-s_2$에서의 최대 생산율 $TH_{N^*}(s,b)$을 제시한다.

<표 1>로부터 주어진 동시 최적화문제의 최대 생산율 $TH_{N^*}(s^*,b^*)$은 6.778이며 이에 상응하는 최적 서버배분 $s^*=(3,5)$, 최적 대기능력배분 $b^*=(7,9)$, 그리고 최적 고객수 $N^*=10$임을 알 수 있다. 그러므로 이미 언급된 바와 같이 복수의 서버로 구성된 경우에는 하나의 서버로 구성된 경우의 결과가 더 이상 유지되지 않음을 알 수 있다.

<표 1>은 최적 고객수 N^*가 b_1과 b_2사이에 존재할 뿐만이 아니라 b_2보다 클 수도 있음을 제시하고 있다. 또한 예를 들어 $s=(4,4)$와 같은 균등한 서버의 배분이나 $b=(8,8)$과 같은 균등한 대기능력의 배분이 더 이상 생산율을 최대화하지 않으며 적절한 불균등 배분이 생산율을 최대화함을 제시한다. 또한 생산율이 서버배분 $s=(s_1,s_2)$이나 대기능력배분 $b=(b_1,b_2)$에 대하여 더 이상 상호대칭을 성립하지 않음을 제시하고 있다. 예를 들어, $TH_{10}((5,3),(9,7))=4.2835$, 그리고 $TH_{10}((3,5),(7,9))=6.778$은 전혀 다른 생산율을 제시한다.

〈표 1〉 서버와 대기능력 배분에 대한 생산율

Buffer \ Server Allocation	(1, 7)	(2, 6)	(3, 5)	(4, 4)	(5, 3)	(6, 2)	(7, 1)
(1, 15) (N^*)	2.5 (7-15)						
(2, 14)	3.0769 (8-14)	4.451 (7-14)					
(3, 13)	3.25 (10-13)	5.2974 (8-13)	5.4881 (7-13)				
(4, 12)	3.3058 (11-12)	5.7483 (9-12)	6.1287 (8-12)	5.3323 (7-12)			
(5, 11)	3.3242 (11)	6.0215 (10-11)	6.4852 (9-11)	5.5551 (8-11)	4.251 (7-11)		
(6, 10)	3.329 (10)	6.1775 (10)	6.7021 (10)	5.6469 (9-10)	4.2768 (8-10)	2.8565 (7-10)	
(7, 9)	3.3294 (9)	6.1983 (10)	6.778 (10)	5.6793 (10)	4.2834 (9)	2.857 (8-9)	1.4286 (7-9)
(8, 8)	3.3253 (8)	6.1151 (9)	6.7339 (10)	5.6827 (10)	4.2841 (9-10)	2.8571 (9)	1.4286 (8-9)
(9, 7)	3.3091 (7)	5.9014 (8-9)	6.5701 (9)	5.6611 (10)	4.2835 (10)	2.8571 (10)	1.4286 (8)
(10, 6)		5.5552 (7-10)	6.2711 (8-10)	5.5916 (9-10)	4.2793 (10)	2.857 (10)	1.4286 (9-10)
(11, 5)			5.7528 (7-11)	5.4217 (8-11)	4.2606 (9-11)	2.8566 (10-11)	1.4286 (9-11)
(12, 4)				4.9948 (7-12)	4.1876 (8-12)	2.8531 (9-12)	1.4286 (10-12)
(13, 3)					3.8953 (7-13)	2.8285 (8-13)	1.4286 (9-13)
(14, 2)						2.6549 (7-14)	1.4296 (8-14)
(15, 1)							1.4286 (7-15)

반면에 <표 1>의 결과들은 더 많은 부하를 갖는 단계에 더 많은 서버와 더 많은 대기능력을 배분하는 것이 생산율을 증대시킴을 보여준다. 또한 의사결정변수인 s, b, 그리고 N에 대하여 생산율이 개별적으로 또는 공동으로(jointly) 증대되었다가 감소함을 쉽게 확인할 수 있다.

더욱 흥미를 유도하는 결과는 균등부하, 균등한 서버와 대기능력의 배분이 생산율을 최대화하지 않는다는 것이다. 예를 들어 균등한 부하 $\rho=(0.5,0.5)$에서 균등하게 서버와 대기능력을 배분하는 $s=(4,4)$와 $b=(8,8)$에서 최적 고객수 $N^*=10$이며 생산율은 6.769이나 불균등부하인 $\rho=(0.3,0.7)$에서 불균등하게 배분된 $s=(3,5)$, $b=(7,9)$, 그리고 $N^*=10$에서의 생산율이 6.778임을 알 수 있다.

지금까지의 동시최적화문제에 부하문제를 추가로 포함하기 위해서는 최적 부하할당을 결정하기 위한 또 하나의 외곽(outermost)순환이 요구되며 주어진 연속적 변수의 최적화를 위하여 간단한 비

선형기법들이 적용될 수 있을 것이다.

마지막으로 주어진 의사결정변수들에 대한 다양한 수치예들은 매우 제한된 계산(iteration)으로 항상 최적해를 제시하였다. 복수의 서버로 구성된 경우에는 더 많은 부하가 할당된 단계에 더 많은 서버와 대기능력을 배분함으로써 생산율을 증대시킬 수 있었으며 최적배분에 있어서 예를 들어 균등배분과 같은 특성들이 존재하지 않음을 알 수 있다. 제시된 알고리즘은 주어진 동시 최적화문제에 매우 효율적으로 적용될 수 있을 것이다.

참 고 문 헌

[1] Bozer, Y.A. and J.A. White, "A Generalized Design and Performance Analysis Model for End-of-Aisle Order-Picking Systems," *IIE Trans.* Vol.28(1996), pp.271-280.

[2] Fox, B., "Discrete Optimization via Marginal Analysis," *Management Science*, Vol. 13(1969), pp.210-216.

[3] Gaver, D.P. and G.S. Shelder, "Processor Utilization in Multiprogramming Systems via Diffusion Approximation," *Operations Research*, Vol.21(1973), pp.569-576.

[4] Gelenbe, E., "On Approximate Computer System Models," *Associated Computing Machinery*, Vol.22(1975), pp.261-269.

[5] Glasserman, P.G. and D.D. Yao, "Monotonicity in Generalized Semi-Markov Processes," *Mathematics of Operations Research*, Vol.17(1992), pp.1-21.

[6] Glasserman, P.G. and D.D. Yao, *Monotone Structure in Discrete Event Systems*, Wiley, New York, 1994.

[7] Glasserman, P.G. and D.D. Yao, "Structured Buffer-Allocation Problems," *Discrete Event Dynamic Systems : Theory and Applications*, Vol.6(1996), pp.9-41.

[8] Kelly, F.P., *Reversibility and Stochastic Networks*, Wiley, New York, 1979.

[9] Kemeny, J.G. and J.L. Snell, *Finite Markov Chains*, Springer-Verlag, New York, 1976.

[10] Marshall, A.W. and I. Olkin, *Inequalities ; Theory of Majorization and It's Applications*, Academic Press, New York, 1979.

[11] Shaked, M. and J.G. Shanthikumar, "Stochastic Convexity and It's Applications," *Advances of Applied Probability*, Vol.20 (1988), pp.427-446.

[12] Sparaggis, P.D. and W. Gong, "Optimal Buffer Allocation in a Two-Stage Queueing System," *Journal of Applied Probability*, Vol.30(1993), pp.478-482.

[13] Shanthikumar, J.G. and D.D. Yao, "Optimal Server Allocation in A System of Multi-server Stations," *Management Science*, Vol.33, No.9(1987), pp.1173-1180.

[14] Shanthikuma, J.G. and D.D. Yao, "Second-Order Properties of the Throughput of a Closed Queueing Network," *Mathematics of Operations Research*, Vol.13(1988), pp.524-534.

[15] Shanthikumar, J.G. and D.D. Yao, "Monotonicity and Concavity Properties in Cyclic Queueing Networks with Finite Buffers," in H.G. Perros, and T. Altiok eds. : *Queueing Networks with Blocking*, North-Holland, pp.325-344, 1989.

[16] Stecke, K.E. and J.J. Solberg, "The Optimality of Unbalancing Both Workloads and Machine Group Sizes in Closed Queueing Networks of Multi-server Queues," *Operations Research*, Vol.33(1985), pp.882-920.

[17] Yao, D.D., "Some Properties of the Throughput Function of Closed Networks of Queues,"

Operations Research Letters, Vol.3, No.6 (1985), pp.313-317.

[18] Yao, D.D. and J.A. Buzacott, "Modeling a Class of Flexible Manufacturing Systems with Reversible Routing," *Operations Research*, Vol.35(1987), pp.87-93.