Journal of the Korean Institute
of Industrial Engineers
Vol.21, No.2, June, 1995

복수의 서버를 갖는 작업장으로 구성된 일반대기네트워크에 있어서 균형과 불균형부하

Balancing/Unbalancing in General Queueing Networks with Multi-Server Stations

김 성 철*

Kim, Sung Chul

Abstract

We consider a general queueing network with multi-server stations. The stations are under heavy traffics or moderate variable conditions. We develope an algorithm to determine the optimal loading policy, which minimizes the congestion in a network. Under more specified condition, majorization and arrangement orderings are established to compare, respectively, various loading and assignment policies. Implications of results are also discussed.

1. 서론

본 논문에서는 복수의 서버(server)를 갖는 M개의 작업장(station)으로 구성된 개방 대기네트워크(open queueing network)에 있어서 최적 부하(optimal loading)에 관한 문제를 다룬다. 주어진 네트워크에 있어서 도착 과정과 서비스 과정은 일반 분포에 의한다. 작업물(job)들의 작업장 i(i=1, … , M)에의 도착은 독립적이며 동일한 분포를 갖는 재생과정(renewal process)에 의하며 교통 방정식(traffic equation)에 의하여 유도되는 도착률 λ_i와 도착 사이 시간의 squared coefficient of variance (scv) ca_i가 존재한다. 또한 작업장 i (i=1, … ,M)는 s_i개의 동일한 병렬(parallel)의 서버로 구성되어 있으며 각 서버에서의 서비스 시간도 독립적이며 동일한 분포에 의하며 기대치 $1/\mu_i$와 scv cs_i를 갖는다. 각 작업장은 무한한 저장 능력의 작업장 저장소(local buffer)를 가지며 작업물의 작업은 선

* 덕성여자대학교 경영학과

입선출(first come first served)의 서비스 규칙에 의한다.

λ를 네트워크 외부로 부터 네트워크에로의 작업물의 총 도착률이라고 하고 $\nu_i=\lambda_i/\lambda$(i=1, … ,M)라고 하면 ν_i는 하나의 작업물의 작업장 i에의 기대 방문 횟수로 해석 될 수 있으며 이를 작업장 i에의 방문 빈도수(visit frequency)로 언급하기로 한다. 또한 $\rho_i=\lambda_i/\mu_i=\lambda\nu_i/\mu_i$, $\chi_i=\rho_i/s_i$(i=1, … ,M)라고 하면 ρ_i는 작업장 i에 할당된 작업량, 즉 부하(loading)를 의미하며 χ_i는 작업장 i의 서비스 강도(service intensity) 또는 이용률(utilization)을 나타낸다. 그러므로 $\underline{\rho}=(\rho_1, \cdots ,\rho_M)$은 부하 벡터(loading vector)로 명명하고 안정 상태를 보장하기 위하여는 모든 i(i=1, … ,M)에 대하여 $\rho_i < s_i$가 만족되어야 한다. 만약 임의의 $\underline{\rho}$에 대하여 $L=|\underline{\rho}|=\rho_1+ \cdots +\rho_M$이라 하면 L은 주어진 네트워크에 부과된 작업물들의 평균 총 작업량으로 해석 될 수 있다. 주어진 L에 대하여 하나의 부하 벡터는 주어진 총 작업량 L의 각 작업장에의 할당을 의미한다.

주어진 최적 부하에 관한 문제는 제조 시스템의 효율성과 관련지어서 많은 문헌에서 다루어져 왔다. 승법형(product form) 해를 갖는 개방 또는 폐쇄 대기 네트워크(Jackson 1957, Gorden과 Newell 1967)에 있어서 각 작업장에서의 서버의 수가 동일한 경우에는 균등 부하(balanced loading)가 최적 부하가 됨을 많은 문헌에서 찾아 볼 수 있다(Shanthikumar와 Stecke 1986, Yao와 Kim 1987a,b). 이에 반하여 각 작업장에서의 서버의 수가 동일하지 않는 경우에는 최적 부하의 결정에 관한 문제는 비선형 최적화 문제로 모형화되어 더 많은 수의 서버를 갖는 작업장에 상대적으로 더 많은 부하가 할당되는 불균형 부하(unbalanced loading)가 최적임이 보여질 수 있다(Stecke와 Solberg 1985, Bitran과 Tirupati 1989). 또한 도착 과정과 서비스 과정이 지수 분포를 갖지 않으며 하나의 서버를 갖는 작업장으로 구성된 비승법형 개방 대기 네트워크에 있어서는 각 작업장의 분산도가 동일한 경우에는 균등 부하가 최적이며 분산도가 동일하지 않는 경우에는 균등 부하가 최적이 아님이 분해법에 의한 근사화 방법(decomposition approximation)에 의하여 보여질 수 있다(Kim 1989). 본 논문에서는 Kim(1989)의 연장 선상에서 복수의 서버를 갖는 작업장으로 구성된 일반 대기 네트워크에서의 최적 부하 문제를 다루고자 한다.

도착 과정과 서비스 과정이 지수 분포를 갖지 않은 비승법형 대기 네트워크의 수행도를 측정하기 위해서는 분해법을 이용한 근사화 방법(Shanthikumar와 Buzacott 1981, Whitt 1983, Bitran과 Tirupati 1988)을 적용 할 수 있다. 주어진 분해법은 먼저 각 작업장에의 도착 사이 시간의 기대치와 scv를 유도하므로써 각 작업장의 도착 과정을 특징 지운다. 다음은 주어진 기대치와 scv를 이용하여 각 작업장에서의 수행도를 측정 할 수 있다. ω_i(상대적으로 T_i)(i=1, … , M)를 작업장 i에서의 임의의 작업물의 작업 개시 전까지의 (상대적으로 작업 소요 시간을 포함한) 평균 대기 시간이라고 하고 T를 임의의 작업물의 네트워크에서의 평균 체재 시간이라고 하자.

주어진 분해법에서는 각 작업장은 추계적으로 독립적으로 분석되며 $T_I = \nu_i(1/\mu_i+\omega_i)$ 그리고 $T = T_1+ \cdots +T_M$이 성립된다. 또한 N_I를 서버에서 작업중인 작업물을 포함하는 작업장 i에서의 평균 대기 길이 그리고 N을 네트워크에 존재하는 평균 총 작업물의 수라고 하면 Little 방정식(Heyman과 Sobel 1982, Frankel et al. 1981)에 의하여 $N_i = \lambda_i(1/\mu_i+\omega_i) = \rho_i+\lambda_i\omega_i$가 성립되며 $N = N_1+ \cdots +N_M$이 된다. 그러므로 주어진 총 작업량을 네트워크에 존재하는 평균 총 작업물의 수를 최소화하도록 할당하므로써 Little 방정식에 의하여 결과적으로 각 작업물의 평균 대기 시간 또한 최소화하게 된다.

2. 분해법

대기 네트워크에 있어서 정확한 해를 구할 수 있는 경우는 승법형 해가 존재하는 경우로 제한되어 있다(Jackson 1957, 1963, Gorden과 Newell 1967, Baskett et al. 1975, 그리고 Kelly 1979)로 제한되어 있다. 승법형 해가 존재하지 않는 일반 대기 네트워크에 있어서는 하나의 근사치 산정 방법으로 분해법이 있으며 이의 적용은 좋은 결과를 제시하는 것으로 알려지고 있다. 이러한 분해법에 의한 근사치 산정에서는 Jackson 네트워크에 있어서의 각 작업장의 독립성과 승법형 해를 일반화 시키려는 시도가 적용된다. 그러므로 각 작업장은 하나의 독립적인 대기 시스템으로 분석되며 일계 모멘트(the first moment)와 이계 모멘트(the second moment), 즉 기대치와 분산에 의한 수행도의 근사치가 산정된다.

분해법에 의한 간단한 설명은 다음과 같이 요약 될 수 있다. 첫번째 단계로써 각 작업장의 상호 작용을 분석한다. 이로써는

i) 각 작업장의 실제 도착률은 기본적인 교통 방정식에 의하여 산정된다. 즉

$$\lambda_i = a_I+\Sigma_{j=1}^{M}\lambda_j\gamma_{ji},\ i=1, \cdots ,M, \qquad (2.1)$$

ii) 도착 사이 시간의 scv를 다음의 일련의 방정식에 의하여 계산한다.

$$\lambda_i ca_i - \Sigma_{j=1}^{M}\{\lambda_j(1-\rho_i^2)\gamma_{ji}^2 ca_j\} = a_i ca_i^0+\Sigma_{j=1}^{M}\{\lambda_j\gamma_{ji}(\rho_i^2\gamma_{ji}cs_j+1-\gamma_{ji})\},\quad i=1, \cdots, M. \qquad (2.2)$$

여기에서

γ_{ji} : 작업장 j에서 작업장 i로의 작업물의 경로 변환 확률,

a_I : 외부로 부터 작업장 i로의 작업물의 도착률,

그리고

ca_i^0 : 외부로 부터 작업장 i로의 작업물의 도착 사이 시간의 scv

를 의미한다. 식(2.2)는 각 작업장에서의 작업물의 분할(splitting), 합병(merging), 그리고 관계식 $cd_i = \rho_i^2 cs_i+(1-\rho_i^2)ca_i$를 순차적으로 적용하므로써 유도 될 수 있다. 여기에서 cd_i는 작업장 i에서의 이탈 과정의 scv를 의미하며 이는 다음 도착되는 작업장에서의 도착 과정의 scv가 된다.

두번째 단계에서는 전 단계에서 얻어진 처음 두 모멘트를 이용하여 각각의 작업장은 독립적인 대기 시스템, 즉 $GI/G/s_i$ 시스템으

로 분석되어 각 작업장의 수행도는 근사화 된다. 즉 작업장 i(i=1, ⋯, M)에서의 ρ_i, ca_i, cs_i의 함수로써 작업 개시 전의 평균 대기 시간 $\omega_i(\rho_i,ca_i,cs_i)$은 교통 강도 극한 이론(heavy traffic limit theorem)(Halfin과 Whitt 1981, Kollerstrom 1974)에 의하여 다음과 같이 정리 될 수 있다.

$$\omega_i(\rho_i,ca_i,cs_i) = \{(ca_i+cs_i)/2\}\{C(s_i,\rho_i)/(s_i-\rho_i)\mu_i\}, \quad i=1, \cdots, M, \quad (2.3)$$

여기에서

$$C(s_i,\rho_i) = \{\rho_i^{si}/(s_i-1)!(s_i-\rho_i)\}/\{\Sigma_{k=0}^{s_i-1}\rho_i^k/k! + \rho_i^{si}/(s_i-1)!(s_i-\rho_i)\} \quad (2.4)$$

이다. 그러므로 주어진 s_i에 대하여 $\rho_i \to s_i$로 접근할 때, 즉 극한 교통 강도하에서 정확성이 증대되며 $ca_i \geqq 0.9$, $cs_i \geqq 0.9$ 또는 $ca_i \leqq 1.1$, $cs_i \leqq 1.1$인 경우에도 상대적으로 좋은 결과를 얻는 것으로 제시되고 있다(Whitt 1983). 여기에서 $C(s_i,\rho_i)$는 Erlang delay 공식으로 명명 될 수 있다. 그러므로

$$T_i(\rho_i,ca_i,cs_i) = \nu_i(1/\mu_i+\omega_i)$$
$$= (\rho_i/\lambda)\{1+[(ca_i+cs_i)/2][C(s_i,\rho_i)/(s_i-\rho_i)]\}, \quad i=1, \cdots, M, \quad (2.5)$$

그리고

$$N_i(\rho_i,ca_i,cs_i) = \lambda_i(1/\mu_i + \omega_i)$$
$$= \rho_i\{1+[(ca_i+cs_i)/2][C(s_i,\rho_i)/(s_i-\rho_i)]\}, \quad i=1, \cdots, M. \quad (2.6)$$

이 된다.

마지막 단계는 각 작업장의 수행도를 네트워크의 수행도로 종합하는 과정으로써 네트워크에서의 평균 총 작업물의 수,

$$N = \Sigma_{i=1}^{M} N_i(\rho_i,ca_i,cs_i) \quad (2.7)$$

이 된다.

3. 부하 문제

전술된 분해법에 기초하여 부하 문제는 다음과 같은 비선형 문제로 정식화 될 수 있다.

$$\text{Min } N = \Sigma_{i=1}^{M} N_i(\rho_i,ca_i,cs_i) \quad (3.1)$$

$$\text{s.t. } \Sigma_{i=1}^{M} \rho_i = L \quad (3.2)$$

식(2.2).

주어진 비선형 문제를 풀기 위하여는 부하 벡터 $\underline{\rho}$와 이에 대한 수행도 그리고 분산도에 대한 영향을 반복적인 관계에 의하여 계속적으로 산정하므로써 최적 부하 $\underline{\rho}^*$에 도달할 수 있다. 그러나 일반적으로 분산 계수 ca_i와 cs_i(i=1, ⋯, M)는 ρ_i에 대하여 독립적인 것으로 문헌에서 가정되고 있다(Kleinrock 1964, Wein 1989, Shanthikumar와 Yao 1988, Bitran과 Tirupati 1989). 실제적으로 주어진 제조 시스템에 있어서 다양한 제품이 생산되는 경우 등에는 분산도는 ρ_i의 적은 변화에 대하여는 민감하지 않으며 예상되는 반복적 관계에 있어서 초기 부하 벡터 $\underline{\rho}^0$를 적절히 선택하므로써 얻어지는 최적해는 분산도를 반복적으로 산정하여 얻어지는 최적해와 비교할 때 적당한 근사치 범위 안에 존재한다. 분산 계수의 ρ_i에 대한 독립성 하에서는 주어진 최적화 문제에 있어서 제약 조건 식(2.2)는 불필요하게 되어 제거 된다.

Lemma 1 : 만약 분산 계수 ca_i와 cs_i(i=1, ⋯, M)가 ρ_i에 대하여 독립적이라고 가정한다면 $N_i(\rho_i,ca_i,cs_i)$는 ρ_i에 대하여 비감소 convex 함수이다.

증명 : $C(s_i,\rho_i)/(s_i-\rho_i)\mu_i$의 convexity(Lee와 Cohen 1983, Grassman 1983)로 부터 주어진

목적 함수는 convex 함수의 합으로 표시되며 결과적으로 convex 함수이다.

그러므로 주어진 비선형 문제는 목적 함수가 convex이고 제약 조건이 convex set을 형성하므로써 하나의 최적해 $\underline{\rho}^*$가 존재한다.

부하 문제의 최적해를 구하기 위하여는 목적 함수의 convexity를 이용한 Greedy 휴리스틱 등 다양한 방법이 제시 될 수 있다. 여기에서는 문제의 최적성의 조건을 이용하여 다음과 같은 반복적 알고리즘을 제시하고자 한다.

단계 1 : 초기화 : k=0.

$|\underline{\rho}^0| = L$인 초기 부하 벡터 $\underline{\rho}^0$를 할당하고 네트워크의 수행도 추정치 평균 총 작업물의 수 N^0를 산정한다. 단계 3으로.

단계 2 : 만약 $N^k > N^{k-1}$이면 알고리즘을 끝내고 그렇지 않으면 단계 3으로.

단계 3 : 모든 i(i=1, ⋯, M)에 대하여

$$\partial N/\partial \rho_i = 1+\{(ca_1+cs_1)/2\}C(s_1,\rho_i)\{1+[s_1+\rho_i-\rho_iC(s_1,\rho_i)]/(s_1-\rho_i)^2\} \quad (3.3)$$

을 산정하고

$$\rho_j^k = \rho_j^{k-1}+\varepsilon,\ \rho_n^k = \rho_n^{k-1}-\varepsilon,\ \rho_i^k = \rho_i^{k-1} \quad (i\neq j,n) \quad (3.4)$$

로 부하를 조정한다. 여기에서

$j = \text{argmin}\{\partial N/\partial \rho_i \mid i=1, \cdots, M\}$,

$n = \text{argmax}\{\partial N/\partial \rho_i \mid i=1, \cdots, M\}$.

단계 4 : 평균 총 작업물의 수 N^k를 산정한다. 단계 2로.

주어진 알고리즘은 원 문제의 최적성의 조건 하에서 제약 조건 (3.2)의 라그랑쥐 승수(Lagrange multiflier) π가 모든 i(i=1, ⋯ ,M)에 대하여 $\partial N/\partial \rho_i$으로써 모든 i에 대하여 동일하다는 조건에 의한다. 또한 이를 알고리즘에 적용할 경우 극한 강도 하에서는 ρ_i에 대한 수행도의 편도함수 값 $\partial N/\partial \rho_i$가 적은 ε에 대하여 매우 민감하므로 이에 따라 발생 할 수 있는 알고리즘 적용상의 순환 과정(cycling)을 회피하고 convexity를 고려하였다. 그러므로 주어진 알고리즘은 원(primal) 문제의 실행 가능성과 상보 여유성(complementary slackness)을 만족 시키므로써 ε를 충분히 적게 하므로써 즉 $\varepsilon\to 0$인 경우에 최적해에 도달한다.

4. 특수한 경우

각 작업장이 동일한 갯수의 서버로 구성되어 있는 경우에는 즉 모든 i(i=1, ⋯ ,M)에 대하여 $s_i=s$인 경우에는 다음의 결과들을 유도 할 수 있다. 먼저 각 작업장의 분산 계수가 동일한 경우 즉 $ca_1+cs_1=c$인 경우에는 균등 부하가 최적임은 3장의 일반적인 부하 문제의 해법 과정을 통하여 쉽게 보여질 수 있다. 여기에서는 여러가지 부하 벡터를 서로 비교 할 수 있는 기준을 제시 하고자 한다.

Theorem 2 : 분산 계수가 동일한 경우에는 부하 벡터 $\underline{\rho}$의 함수로써 평균 총 작업물의 수 $N(\underline{\rho})$는 $\underline{\rho}$에 대하여 Schur convex 함수(부록 참조)이다. 즉 두개의 부하 벡터 $\underline{\rho}^1$과 $\underline{\rho}^2$에 대하여 $\underline{\rho}^1 \leq_m \underline{\rho}^2$이면 $N(\underline{\rho}^1)\leq N(\underline{\rho}^2)$이다.

Theorem 2를 증명하기 위하여는 다음의 부호들을 필요로 하며 모든 i(i=1, ⋯ ,M)에 대하여 $s_i=s$이므로 부호에서 s는 생략하기로

한다.

$$A(\rho_i) = \Sigma_{k=0}^{s-1}\rho_i^k/k!,$$
$$B(\rho_i) = \rho_i^s/[(s-1)!(s-\rho_i)]. \quad (4.1)$$

그러므로

$$C(\rho_i) = B(\rho_i)/[A(\rho_i)+B(\rho_i)], \quad (4.2)$$

$$\partial N(\underline{\rho})/\partial \rho_i = 1+\alpha C(\rho_i)\{1+[s+\rho_i-\rho_i C(\rho_i)]/(s-\rho_i)\}, \quad (4.3)$$

Lemma 3 : i) $\rho_1 \geq \rho_2 \rightarrow \rho_1/(s-\rho_1) \geq \rho_2/(s-\rho_2)$, $1/(s-\rho_1)^2 \geq 1/(s-\rho_2)^2$. (4.4)

ii) $\rho_1 \geq \rho_2 \rightarrow A(\rho_2)B(\rho_1) \geq A(\rho_1)B(\rho_2)$. (4.5)

iii) $\rho_1 \geq \rho_2 \rightarrow A(\rho_1)/\{(s-\rho_1)[A(\rho_1)+B(\rho_1)]\} \geq A(\rho_2)/\{(s-\rho_2)[A(\rho_2)+B(\rho_2)]\}$. (4.6)

iv) $\rho_1 \geq \rho_2 \rightarrow C(\rho_1) \geq C(\rho_2)$. (4.7)

v) $\rho_1 \geq \rho_2 \rightarrow [s+\rho_1-\rho_1 C(\rho_1)]/(s-\rho_1)^2 \geq [s+\rho_2-\rho_2 C(\rho_2)]/(s-\rho_2)^2$. (4.8)

증명 : i) 대수적 조작에 의하여 쉽게 유도 됨.

ii) $\rho_1^s A(\rho_2) \geq \rho_2^s A(\rho_1)$과 i)에 의함.

iii) Yao와 Kim(1989b)에 주어짐.

iv) 간단한 대수적 조작에 의하여 ii)로 치환 가능.

v) $[s+\rho_1-\rho_1 C(\rho_1)]/(s-\rho_1)^2 = s/(s-\rho_1)^2 + \rho_1 A(\rho_1)/(s-\rho_1)^2[A(\rho_1)+B(\rho_1)]$이므로 i)과 ii)에 의하여 증명됨.

Theorem 2의 증명 : $N(\underline{\rho})$의 $\underline{\rho}$에 대한 Schur convexity를 증명하기 위하여는 $N(\underline{\rho})$의 대칭성에 의하여 $\rho_1 \geq \rho_2$이면 $\partial N(\underline{\rho})/\partial \rho_1 - \partial N(\underline{\rho})/\partial \rho_2 \geq 0$임을 보이면 된다. 그러므로 주어진 결과는 Lemma 3의 iv)와 v)에 의한다.

Majorization ordering 하에서 부하 벡터의 감소는 네트워크 내의 총 작업물의 수의 추계적 감소를 의미한다. $\Sigma_{i=1}^{M}\rho_i = L$에서 L이 주어진 상수이면 Majorization ordering하에서 작은 부하 벡터는 균등 부하 즉 $\underline{\rho} = (L/M, \cdots, L/M)$에 더 근접한 경우를 의미하며 결과적으로 균등 부하가 최적 부하가 되며 네트워크 내의 총 작업물의 수를 최소화 한다.

Theorem 4 : 만약 $\underline{\rho}^1 \leq_{wm} \underline{\rho}^2$이면 $N(\underline{\rho}^1) \leq N(\underline{\rho}^2)$이다.

증명 : i)과 iii)에 의하여 증명됨.

위의 부하 문제는 다음의 할당 문제(assignment problem)로 쉽게 연장 될 수 있다. 만약 M그룹의 서버들이 있고 각 그룹은 s개의 서버로 구성되어 있으며 그룹 i($i=1, \cdots, M$)의 각 서버의 서비스율은 μ_i라 하고 서버 그룹 i는 작업장 i에 꼭 배치되지 않는 반면에 작업장 i에의 작업물의 도착률은 λ_i라고 하자. 또한 $\lambda_1 \geq \cdots \geq \lambda_M$이면 $\underline{\lambda}\downarrow = (\lambda_1, \cdots, \lambda_M)$으로 표시되고 부호 $\downarrow$는 벡터 성분들을 감소하는 순으로 배열하는 것을 의미한다. 그러면 M개의 서버 그룹의 각 작업장에의 할당은 $\{\mu_i \mid i=1, \cdots, M\}$의 순열 벡터로 표시될 수 있다. 그러므로 2M 벡터 $(\underline{\lambda}, \underline{\mu})$를 할당 벡터(assignment vector)로 언급 할 수 있으며 주어진 할당 벡터들 간에는 arrangement ordering $\leq_a$이 성립 될 수 있다. 주어진 ordering 하에서 가장 큰 할당 벡터는 $\underline{\lambda}$의 성분의 배열 순서대로 $\underline{\mu}$의 성분을 할당하는 것 즉 $(\underline{\lambda}\downarrow, \underline{\mu}\downarrow)$이고 가장 작은 할당 벡터는

반대로 할당하는 것 즉 ($\underline{\lambda}\downarrow$, $\underline{\mu}\uparrow$)가 된다.

Lemma 5 : ($\underline{\lambda}\downarrow$, $\underline{\mu}^1$)과 ($\underline{\lambda}\downarrow$, $\underline{\mu}^2$)가 두개의 할당 벡터라고 하고 $\underline{\rho}^1$과 $\underline{\rho}^2$가 각각 상응하는 부하 벡터라고 하면

($\underline{\lambda}\downarrow$, $\underline{\mu}^2$)$\leq_a$($\underline{\lambda}\downarrow$, $\underline{\mu}^1$) $\rightarrow$ $\underline{\rho}^1\leq_{wm}\underline{\rho}^2$. (4.9)

증명 : Yao와 Kim(1989a)에 주어짐.

Theorem 6 : ($\underline{\lambda}\downarrow$, $\underline{\mu}^2$)$\leq_a$($\underline{\lambda}\downarrow$, $\underline{\mu}^1$)이면 N($\underline{\lambda}\downarrow$, $\underline{\mu}^1$)$\leq$N($\underline{\lambda}\downarrow$, $\underline{\mu}^2$).

증명 : Theorem 4와 Lemma 5에 의함.

각 작업장의 서버의 수가 모두 같은 경우에도 분산 계수가 서로 다른 경우에는 위의 결과는 성립되지 않는다. 각 작업장에서의 도착 과정이나 서비스 과정의 분산도의 증가는 이탈 과정(departure process)의 분산도의 증가를 가져오므로써 다음 작업장에의 도착 과정의 분산도를 증대시키고 결과적으로 분산도의 증대는 네트워크의 수행도의 감소를 가져온다고 할 수 있다(Surech와 Whitt 1990). 이러한 관점에서 각 작업장은 분산 계수와 부하와의 관계에 있어서 다음이 성립된다.

Theorem 7 : 만약 M개의 작업장이 분산도의 순서로 번호가 주어진다면 즉 $ca_1+cs_1 \leq \cdots \leq ca_M+cs_M$이라 하면 최적 부하는 $\rho_1^* \geq \cdots \geq \rho_M^*$를 만족시킨다.

증명 : Lemma 1, 최적성의 조건 그리고 식 (3.3)에 의한다.

5. 실행 예

본 절에서는 세개의 작업장으로 구성된 네트워크에 있어서 부하 문제에 대한 실행 예를 제시하고자 한다. 예제에서 사용된 모수와 결과는 Table 1에 요약되어 있으며 적용된 스탭 길이(step length) 즉 $\varepsilon=0.001$이다.

각 작업장이 상당히 높은 교통 강도를 가지고 있는 주어진 예제에 있어서 각 작업장의 부하에 대한 편도함수가 매우 민감하여 작은 스탭 길이의 변화에 대하여도 편도함수의 변화는 매우 크게 되어 결과적으로 제시된 최적 부하에 있어서 각 작업장의 편도함수 값은 동일하지 않고 약간의 차이가 존재한다. 또한 주어진 결과는 동일한 분산도 하에서는 서버 배분에 있어서 불균형이 커질수록 수행도가 증가하며 결과적으로 네트워크 내의 총 대기 길이가 짧아진다는 것을 알 수 있으며 이는 서버 모음(server pooling)에 관한 기존의 결과(Smith와 Whitt 1981)들과 같은 방향의 결과로 볼 수 있으나 비교에 주의가 필요하다. 불균형의 서버 배분 하에서는 서버의 수가 많은 작업장의 분산도가 상대적으로 적을수록 평균 대기 길이가 작아지며 수행도가 커진다고 할 수 있다. 동일한 서버 배분 하에서는 분산도가 같은 경우에만 균등 부하가 최적이며 그 외에는 불균형 부하가 최적이나 균등 부하의 수행도가 상대적으로 낮음을 알 수 있다.

6. 결론

본 논문은 복수의 서버를 갖는 작업장으로 구성된 일반 대기 네트워크에 있어서 최적 부하 문제를 비선형의 convex program으로 모형화하여 부하 벡터에 대한 각 작업장에서의 편도함수를 이용한 반복적 방법을 제시하였

Table 1. 실행 예

L	서버배분			분 산 도 ca_i+cs_i			결 과						총대기 길 이
							부 하			편도함수			
	1	2	3	1	2	3	1	2	3	1	2	3	
5.8	1	2	3	2.6	2.0	1.4	0.9427	1.9293	2.9280	395.2	400.6	405.6	78.9975
	1	2	3	2.4	2.0	1.6	0.9547	1.9303	2.9240	406.3	412.2	415.8	81.137
	1	2	3	2.1	2.0	1.9	0.9507	1.9313	2.9180	431.4	424.3	424.1	83.7966
	1	2	3	2.0	2.0	2.0	0.9517	1.9313	2.9170	428.1	436.9	435.6	84.54725
	1	2	3	1.9	2.0	2.1	0.9527	1.9323	2.9150	424.1	450.1	436.1	85.22919
	2	2	2	2.0	2.0	2.0	1.9333	1.9333	1.9333	450.1	450.1	450.1	87.47466
	1	2	3	1.4	2.0	2.6	0.9807	1.9333	2.9060	452.8	450.1	441.3	87.61548
	2	2	2	2.4	2.0	1.6	1.9263	1.9333	1.9403	442.2	450.1	449.7	87.87192
	2	2	2	1.9	2.0	2.1	1.9353	1.9333	1.9313	454.5	450.1	445.5	88.4388
	2	2	2	2.1	2.1	2.1	1.9333	1.9333	1.9333	472.6	472.6	472.6	92.60839

다. 주어진 결과는 서버 배분과 분산도가 네트워크의 수행도에 미치는 영향을 보였으며 서버 배분에 있어서는 불균형 배분이 불균형 배분에 있어서는 서버의 수가 많은 작업장이 더 낮은 분산도를 갖는 경우에 수행도가 높아짐을 보였다. 특수한 경우로써 각 작업장의 서버의 수와 분산도가 같은 경우에는 Yao와 Kim(1989b)이 연장되어 다양한 부하 정책이 majorization ordering과 arrangement ordering 하에서 비교 될 수 있음을 보였다.

복수의 서버를 갖는 작업장으로 구성된 일반 대기 네트워크에 있어서 최적 부하에 관한 주어진 결과는 새로운 것으로 볼 수 있다. 그러나 극한 교통 강도 또는 제한된 분산도 등의 가정들은 주어진 결과의 일반화에 대한 앞으로의 연구의 필요성을 제시한다.

참 고 문 헌

[1] Baskett, F., Chandy, K.M., Muntz, R.R., and F.G. Palacios, "Open, Closed, and Mixed Networks of Queues with Different Classes of Customers," J. Associ. Comput. Machi., 22 (1975).

[2] Bitran, G.R. and D. Tirupati, "Multiproduct Queueing networks with Deterministic Routing : Decomposition Approach and the Notion of Interface," Magt Sci. 34 (1988).

[3] ———— and ————, "Tradeoff Curves, Targeting and Balancing in Queueing Networks," Oper. Res., 37 (1989).

[4] Frankel, P., Konig, D., Arndt, U. and Schmidt, V., Queues and Point Processes Akademie Verlag, Berlin (1981).

[5] Gorden, W.J. and Newell, G.F., "Closed

Queueing Networks with Exponential Servers," Oper. Res., 15 (1967).

[6] Grassman, W., "The Convexity of the Mean Queue Size of the M/M/c Queue with Respect to the Traffic Intensity," J. Appl. Proba., 20 (1983).

[7] Halfin, S. and W. Whitt, "Heavy-Traffic Limits for Queues with Many Exponential Servers," Oper. Res., 29 (1981).

[8] Heyman, D.P. and M.J. Sobel, Stochastic Models in Operations Research, Vol.1, Macgraw Hill, New York (1982).

[9] Jackson, J.R., "Networks of Waiting Lines," Oper. Res., 5 (1957).

[10] ————, "Jobshop like Queueing Systems," Magt Sci., 10 (1963).

[11] Kelly, F., Reversibility and Stochastic Networks, Willey, New York (1979).

[13] Kim, S.C., "Reducing Congestion in General Queueing Networks," J. Korean OR/MS Soc., 14 (1989).

[14] Kleinrock, L., Communication Nets : Stochastic Message Flow and Delay, Dover Publications, Inc., New York (1979).

[15] Kollerstrom, J., "Heavy Traffic Theory for Queues with Several Servers. I.," J. Appl. Prob., 11 (1974).

[16] Lee, H.L. and M.A.Cohen, "A Note on the Convexity of Performance Measures of M/M/c Queueing Systems," J. Appl. Proba., 20 (1983).

[17] Marshall, A.W. and I. Olkin, Inequalities : Theory of Majorization and It's Applications, Academic Press, New York (1979).

[18] Shanthikumar, J.G. and J.A. Buzacott, "Open Queueing network Models of Dynamic Job Shops," Inter. J. Prod. Res., 19 (1981).

[19] ———————— and K.E. Stecke, "Reducing Work-In-Process Inventory in Certain Classes of Flexible Manufacturing Systems," Europ. J. Oper. Res., 26 (1986).

[20] ———————— and D.D.Yao, "On Server Allocation in Multiple Center Manufacturing Systems," Oper. Res., 36 (1988).

[21] Smith, D.R. and W. Whitt, "Resource Sharing for Efficiency in Traffic Systems," Bell System Tech. J., 60 (1981).

[22] 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," Oper. Res., 33, 4 (1985).

[23] Suresh, S. and W. Whitt, "Arranging Queues in Series : A Simulation Experiments," Magt Sci., 36 (1990).

[24] Wein, L.M., "Capacity Allocation in Generalized Jackson Networks," Oper. Res. Letters, 8 (1989).

[25] Whitt, W., "The Queueing Networks Analyzer," The Bell System Tech. J., 62 (1983).

[26] Yao, D.D. and S.C.Kim, "Some Order Relations in Closed Networks of Queues

with Multiserver Stations," Naval Res. Logist., Vol. 34 (1987a).

[27] ——— and ———, "Reducing the Congestion in a Class of Job Shops," Magt Sci., 33, 9 (1987b).

부록

본 부록에서는 본 논문에서 적용된 majorization ordering과 arrangement ordering에 대한 결과를 요약하여 제시하고자 한다. 여기에서 사용되는 M-O는 참고 문헌 Marshall과 Olkin(1979)를 의미한다.

1. $\underline{x}$와 $\underline{y}$는 실수의 성분으로 구성된 n차원의 벡터이며 $x_{[i]}$와 $y_{[i]}$는 각 벡터에 있어서 i번째로 큰 성분을 의미한다고 하면 majorization ordering $\underline{x} \leq_m \underline{y}$는 다음과 같이 정의된다.

$$\Sigma_{i=1}^{k} x_{[i]} \leq \Sigma_{i=1}^{k} y_{[i]},\ k=1, \cdots, n-1,$$

$$\Sigma_{i=1}^{n} x_{[i]} = \Sigma_{i=1}^{n} y_{[i]}. \qquad (A.1)$$

2. 만약 (A.1)의 마지막 방정식이 $\leq$로 변경되면 주어진 ordering은 weak majorization ordering $\leq_{wm}$으로 정의된다(M-O의 1A).

3. 만약 $\underline{x} \leq_{wm} \underline{y}$이면 각 성분에 있어서 $\underline{x} \leq \underline{z}$, 그리고 $\underline{z} \leq_m \underline{y}$인 벡터 $\underline{z}$가 존재한다 (M-O의 5A).

4. 실수의 값으로 정의되는 함수 $f(\cdot)$에 있어서 $\underline{x} \leq_m \underline{y}$일때 $f(\underline{x}) \leq (\geq) f(\underline{y})$이면 주어진 함수 $f(\cdot)$는 Schur convex(concave) 함수로 정의된다. 만약 함수 $f(\underline{x})$가 연속적이고 미분 가능하면 다음이 성립되는 경우에만 Schur convex(concave) 함수이다.

$$x_i \leq x_j \rightarrow (\partial / \partial x_i) f(\underline{x}) \leq (\geq) (\partial / \partial x_j) f(\underline{y})$$

(M-O의 3A). (A.2)

5. 두개의 할당 벡터의 arrangement ordering $(\underline{\lambda}\downarrow, \underline{\mu}^2) \leq_a (\underline{\lambda}\downarrow, \underline{\mu}^1)$은 벡터 $\underline{\mu}^1$으로 부터 $\underline{\mu}^2$의 근접한 두개의 성분을 그들이 감소하는 순서로 계속적으로 교환하므로써 얻어지는 경우로 정의 할 수 있다(M-O의 6F).