Naumov, Valeriy; Martikainen, Olli

Working Paper

Optimal resource allocation in multiclass networks

ETLA Discussion Papers, No. 1262

Provided in Cooperation with:
The Research Institute of the Finnish Economy (ETLA), Helsinki

Suggested Citation: Naumov, Valeriy; Martikainen, Olli (2011) : Optimal resource allocation in multiclass networks, ETLA Discussion Papers, No. 1262, The Research Institute of the Finnish Economy (ETLA), Helsinki

This Version is available at:
https://hdl.handle.net/10419/87783

ETLA ELINKEINOELÄMÄN TUTKIMUSLAITOS
THE RESEARCH INSTITUTE OF THE FINNISH ECONOMY

Keskusteluaiheita
Discussion Papers
14 December 2011

No **1262**

Optimal Resource Allocation in Multiclass Networks

Valeriy Naumov* – Olli Martikainen**

* ETLA – The Research Institute of the Finnish Economy, valeriy.naumov@etla.fi
** ETLA – The Research Institute of the Finnish Economy, olli.martikainen@etla.fi

The authors thank the Technology Industries of Finland Centennial Foundation and Tekes – the Finnish Funding Agency for Technology and Innovation for research funding.

ISSN 0781–6847

Contents

Abstract

In this paper, we study resource allocation in multiclass networks having several types of flexible servers and general constraints on the number of servers at each station. Each job class is characterized by the station where the job is processed and by the amount of work allocated to that station upon job arrival. Servers may have different skills, working efficiency and resource requirements. We propose a simple method to calculate an upper bound for the maximum network throughput achievable with static resource allocation.

Key words: Resource allocation, flexible server, multiclass network, throughput, bottleneck

JEL: C61, C62, C68

Tiivistelmä

Tässä artikkelissa tutkitaan joustavia erityyppisiä palvelimia käyttävän moniluokkaverkon resurssiallokaatiota, kun verkon solmuissa on erityyppisiä palvelimia ja palvelinten kokonaismäärät noudattavat yleisiä rajoitusehtoja. Jokaiseen asiakasluokkaan liittyy verkon solmu, jossa asiakkaita palvellaan, sekä työmäärä, jonka asiakkaat tuovat solmuun. Palvelimissa sallitaan eri osaamisluokkia, palvelutehokkuuksia ja resurssitarpeita. Esitämme yksinkertaisen tavan verkon maksimiläpäisyn ylärajan laskemiseksi staattisella resurssiallokaatiolla.

Asiasanat: Resurssiallokaatio, joustava palvelin, moniluokkaverkko, läpäisy, pullonkaula

1 Introduction

Stochastic network models with flexible servers are widely used for the analysis and optimization of computer and communication networks [12], manufacturing systems [13], and health care services [14]. In such models, "server flexibility" denotes a server's job processing capacity at different network stations. There is extensive literature analyzing the throughput of systems with flexible servers, as exemplified by [1]-[4]. Andradottir et al. [1] propose linear programming (LP) to analyze the optimal allocation of a given number of flexible servers in a multiclass network. Al-Azzoni and Down [7] use the same allocation LP model for mapping tasks onto flexible servers. Down and Karakostas [2] extend the LP model in [1] and study server allocation under a constraint on the number of servers at each station.

In this paper, we are interested in the optimization of networks having flexible servers and using non-sharable resources, such as health care teams consisting of people and equipment. This optimization problem can be considered as a multiagent resource allocation problem [10], where servers are modeled as agents that receive resources. We use an approach similar to activity analysis ([11]) for the optimization of a multiclass network having limited, continuous, non-sharable resources and having different types of flexible servers that can operate at stations only after the allocation of a group of resources. Server skills and their efficiency at different stations are represented in the form of a productivity matrix. Our analysis is based on a few assumptions: 1) Upon arrival, each job inputs some quantity of work into a station; 2) The aggregate productivity of servers allocated to a particular station is an additive function of the individual servers' productivities; 3) The expected rate of work assigned to each station must not exceed the total productivity of servers allocated to that station. We propose a solution for the problem of optimal static resource allocation to servers that maximizes network throughput while satisfying constraints on the number of servers at each station.

We generalize the results of [15] by introducing resource requirements for servers. In particular, we consider servers as teams that have different types of resources. For example, in a hospital operation room, the server might consist of the following resources: surgeon, anesthesiologist, nurses and any necessary special equipment. This concept of a server, which has both complementary and non-complementary resources, represents a new idea and (to our knowledge) has not yet been introduced elsewhere.

2 Description of the system

We consider a network composed of N stations, K classes of jobs and M types of servers. The stations represent the job processing stage, and each station consists of an infinite buffer and several servers that work in parallel. A job's class uniquely identifies the job's station, and a given job can change class after each processing stage. We use $\mathcal{J}(n)$ to denote the set of job classes processed at station n. Figure 1 shows the relation between network components and parameters. Upon the completion of service, a job in class i is either routed to class j with probability p_{ij} or leaves the network with probability $1-\sum_{j=1}^{K} p_{ij}$. The transition from class i to class j may correspond to a transition of the job from one station to another, or it may represent the job's transition to another class within the same station. We assume that the matrix $\mathbf{I}-\mathbf{P}$ is invertible, which implies that each job class has a finite expected time to leave the

network. A network is open if jobs are allowed to both enter and leave the system. We will call a network clopen if there are no arrivals, so that jobs can only leave the system.

We assume that a job transmits to a station some volume of work that must be performed by servers, which are defined as independent, exponentially distributed random variables. We denote by v_j the expected volume of work that each job of class j brings to a station. We consider networks having multi-skilled servers that are split into *M* types according to their functionality. Each server type is characterized by server productivity and its resource quantities required for operation. The sets of stations where servers operate may overlap, and each server can be allocated to any station where the server is operational. Server allocation can be described by an integer $N \times M$ − matrix $\mathbf{X} = [x_{nm}]$, where x_{nm} is the number of servers of type *m* allocated to station n.

A server's productivity is represented by a constant value that describes the volume of work that the server is able to process at a station per unit of time. We denote the productivity of type *m* servers at station n by π_{nm}. Servers of type *m* having zero productivity (i.e., $\pi_{nm} = 0$) implies that these servers are not operational at station n. Thus, a server can receive jobs at a given station only if its productivity at the station is positive. For that matter, a server can be allocated to any station where it is operational. For example, servers of type 1 may be operational only at stations 1, 2 and 3, whereas servers of type 2 may be operational only at stations 2, 3 and 4. The $N \times M$ -matrix $\mathbf{\Pi} = [\pi_{nm}]$ will be called the productivity matrix corresponding to a given set of servers and stations. The resulting skill matrix $\mathbf{\Sigma} = [\sigma_{nm}]$ is defined by setting $\sigma_{nm} = 1$ if $\pi_{nm} > 0$ and by setting $\sigma_{nm} = 0$ if $\pi_{nm} = 0$. If the elements of the productivity matrix are required to be either 0 or 1, the productivity matrix coincides with the skill matrix.

Say that there are L types of resources and that the resources of each type are limited and fixed in quantity. We denote the quantity of type l resources required for a type m server to operate by r_{ml}. For each resource of type l, the server allocation matrix $\mathbf{X}$ must satisfy the following constraints:

$$\sum_{n=1}^{N} \sum_{m=1}^{M} x_{nm} r_{ml} \le R_l, \quad l = 1, \ldots, L, \tag{1}$$

where R_l is the total quantity of type l resources.

The quantity of untapped resources of type l, u_l, is given by the difference between the total quantity and allocated quantity of type l resources, i.e.,

$$u_l = R_l - \sum_{n=1}^{N} \sum_{k=1}^{K} x_{nk} r_{kl}.$$

Let $U = \{l \mid u_l = 0\}$ be the set containing the types of missing resources. We say that a nonnegative vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_K)$ is the composition vector for the server of type k if $\boldsymbol{\theta}$ satisfies the following properties:

1. $r_{kl} \le \sum_{j=1}^{K} \theta_j r_{jl}$, for all $l \in U$;
2. If $0 \le \xi_j \le \theta_j$ for all $j = 1,\ldots,K$, and $r_{kl} \le \sum_{j=1}^{K} \xi_j r_{jl}$ for all $l \in U$, then $\boldsymbol{\xi} = \boldsymbol{\theta}$.

Note that the set of composition vectors depends on both the server resource requirements and the types of missing resources, which, in turn, depend of the server allocation $\mathbf{X}$.

We assume that servers cooperate in the sense that if there are multiple servers allocated to a station, they pool their efforts, so that the aggregated productivity of all servers allocated to station n can be calculated as

$$\eta_n(\mathbf{X}) = \sum_{m=1}^{M} \pi_{nm} x_{nm}, \quad n = 1,2,\ldots,N. \tag{2}$$

Let $\mathcal{N} = \{1,2,\ldots,N\}$ represent the set of all stations, and $\mathcal{G} = \{S_1, S_2, \ldots S_K\}$ denote a collection of non-empty subsets of $\mathcal{N}$. We assume that for each set of stations $S_i \in \mathcal{G}$ a positive number B_i is specified serving as an upper limit for the number of servers allocated to stations belonging to the set $S_i \in \mathcal{G}$. Therefore, only those server allocations are feasible that satisfy the constraints

$$\sum_{n \in S_i} \sum_{m=1}^{M} x_{nm} \le B_i, \quad i = 1,2,\ldots,K. \tag{3}$$

For example, if $\mathcal{G} = \{\{1\},\{2\},\ldots,\{N\},\mathcal{N}\}$, then the number of servers at station *n* has upper bound B_n, $n = 1,2,\ldots,N$, whereas the total number of servers in the network cannot exceed B_{N+1}.

3 Resource allocation

3.1 Open networks

Assume that arriving jobs are routed to class j with probability α_j, where $\sum_{j=1}^{K} \alpha_j = 1$. The expected number of visits γ_j to class *j*, called the visiting ratio, can be uniquely determined by solving the following linear system [9]:

$$\gamma_j = \alpha_j + \sum_{i=1}^{K} \gamma_i p_{ij}, \quad j = 1,2,\ldots,K.$$

The total expected workload at station *n*, which is the expected volume of work that each job places in station *n* during the job's lifetime, is given by

$$w_n = \sum_{j \in \mathcal{J}(n)} \gamma_j v_j. \tag{4}$$

We assume that w_n is positive for each station n, i.e., that the system does not have superfluous stations.

The total expected service time required for processing a job at a station n over all that job's visits to the station can be calculated as

$$\tau_n(\mathbf{X}) = \frac{w_n}{\eta_n(\mathbf{X})}. \tag{5}$$

and the saturation rate of station n can be calculated as

$$\mu_n(\mathbf{X}) = \frac{\eta_n(\mathbf{X})}{w_n}. \tag{6}$$

At each station of a stable network, the job arrival rate a cannot exceed the saturation rate [8]-[9]. Therefore, a job arrival rate a is feasible only if it satisfies the inequality $a \le \lambda(\mathbf{X})$, where the network throughput $\lambda(\mathbf{X})$ is defined as

$$\lambda(\mathbf{X}) = \min_{1 \le n \le N} \mu_n(\mathbf{X}).$$

Equivalently, the job arrival rate a must satisfy the following constraints:

$$aw_n \le \eta_n(\mathbf{X}), \quad n = 1,2,\ldots,N,$$

These constraints demand that the total expected amount of work placed into each station per unit of time does not exceed the total productivity of the servers allocated to the station. These inequalities compose a necessary but not a sufficient condition for network stability. Non-stable networks have long been observed in practice; see [16] for an example of such a communications network. Necessary and sufficient conditions for the stability of various stochastic networks can be found in [17].

For any arrival rate satisfying $a \le \lambda(\mathbf{X})$, the utilization of station n can be calculated as

$$\rho_n(\mathbf{X}) = \frac{a}{\mu_n(\mathbf{X})}. \tag{7}$$

A bottleneck station is defined to be any station n for which the saturation rate $\mu_n(\mathbf{X})$ attains its minimum value $\lambda(\mathbf{X})$ [8]. It follows from (7) that bottleneck stations have the highest utilization in the system. We define the utilization of type m servers, $\alpha_m(\mathbf{X})$, and type l resources, $\beta_l(\mathbf{X})$, as

$$\alpha_m(\mathbf{X}) = \begin{cases} \frac{1}{x_m} \sum_{n=1}^{N} \rho_n(\mathbf{X}) x_{nm}, & x_m > 0, \\ 0, & x_m = 0, \end{cases}, \tag{8}$$

$$\beta_l(\mathbf{X}) = \frac{1}{R_l} \sum_{m=1}^{M} \alpha_m(\mathbf{X}) x_m r_{ml} = \frac{1}{R_l} \sum_{n=1}^{N} \sum_{m=1}^{M} \rho_n(\mathbf{X}) x_{nm} r_{ml}. \tag{9}$$

where $x_m = \sum_{n=1}^{N} x_{nm}$ is the total number of type m servers. It is easy to see that server and resource utilizations cannot exceed the utilization of a bottleneck station. In a balanced network, all stations are equally utilized, and as a result, in (8) and (9) we have that $\rho_n(\mathbf{X}) = \rho(\mathbf{X})$ for all $n = 1,2,\ldots,N$. In this case, all servers are also equally utilized:

$$\alpha_m(\mathbf{X}) = \rho(\mathbf{X}), \; m = 1,2,\ldots,M,$$

As for resource utilization, we have the following equation:

$$\beta_l(\mathbf{X}) = \frac{\rho(\mathbf{X})}{R_l} \sum_{m=1}^{M} x_m r_{ml} \ , \ l = 1,\ldots,L \ .$$

We will therefore formulate the resource allocation problem for an open network (RAO) as the following integer programming problem:

(RAO)

maximize

$$\lambda(\mathbf{X}) = \min_{1 \le n \le N} \frac{1}{w_n} \sum_{m=1}^{M} \pi_{nm} x_{nm} \tag{10}$$

subject to:

$$\sum_{n=1}^{N} \sum_{m=1}^{M} x_{nm} r_{ml} \le R_l \ , \ \ l = 1,\ldots,L \ , \tag{11}$$

$$\sum_{n \in S_i} \sum_{m=1}^{M} x_{nm} \le B_i \ , \quad i = 1,2,\ldots,K \ , \tag{12}$$

$$x_{nm} \in \mathbb{N} \ , \ \ n = 1,2,\ldots,N \ , \ m = 1,2,\ldots,M \ . \tag{13}$$

Here $\mathbb{N}$ denotes the set of natural numbers, and expected workloads w_n are calculated using . For the solution $\widetilde{\mathbf{X}}$ of an RAO, the value of $\lambda(\widetilde{\mathbf{X}})$ gives the maximum throughput that can be achieved with static server allocation.

3.2 Clopen networks

Now consider a clopen network, for which at time $t = 0$ there are q_i class i jobs for $j = 1,2,\ldots,K$. Let $Q(t)$ represent the number of jobs in service within the network at time t and $\Theta = \inf_{t \ge 0} \{Q(t) = 0\}$ denote the time until the network is empty, or time-to-empty. We want to find an allocation of servers that minimizes the expected such time-to-empty, $\theta = \mathbb{E}\Theta$.

The expected number of visits of servers entering class j from class i, γ_{ij}, can be uniquely determined by solving the linear system

$$\gamma_{ij} = \delta_{ij} + \sum_{k=1}^{K} \gamma_{ik} p_{kj} \ , \ \ i,j = 1,2,\ldots,K \ .$$

Taking into account the network's initial state of the network, the expected number of class j jobs processed by the network before it becomes empty can be calculated as

$$Q_j = \sum_{k=1}^{K} q_k \gamma_{kj} \ , \tag{14}$$

This value also can be determined directly from the linear system

$$Q_j = q_j + \sum_{k=1}^{K} Q_k p_{kj}\,, \quad j = 1,2,\ldots,K\,.$$

Therefore, the expected volume of work placed to the station n can be calculated as

$$W_n = \sum_{j \in \mathcal{J}(n)} Q_j v_j\,, \tag{15}$$

and the total expected service time over all visits of all jobs at station n can be calculated as

$$T_n(\mathbf{X}) = \frac{W_n}{\eta_n(\mathbf{X})}\,. \tag{16}$$

Because the network time-to-empty cannot be less than any given station's time-to-empty, we deduce the following bound for the expected network time-to-empty:

$$\frac{1}{\theta} \le \min_{1 \le n \le N} \frac{1}{W_n} \sum_{m=1}^{M} \pi_{nm} x_{nm}\,. \tag{17}$$

Therefore, we can formulate the resource allocation problem for a clopen network (RAC) as the problem of maximizing the right side of inequality (17), as follows:

(RAC)

maximize

$$\Lambda(\mathbf{X}) = \min_{1 \le n \le N} \frac{1}{W_n} \sum_{m=1}^{M} \pi_{nm} x_{nm} \tag{18}$$

subject to:

$$\sum_{n=1}^{N} \sum_{m=1}^{M} x_{nm} r_{ml} \le R_l\,, \quad l = 1,\ldots,L\,, \tag{19}$$

$$\sum_{n \in S_i} \sum_{m=1}^{M} x_{nm} \le B_i\,, \quad i = 1,2,\ldots,K\,, \tag{20}$$

$$x_{nm} \in \mathbb{N}\,, \quad n = 1,2,\ldots,N\,,\ m = 1,2,\ldots,M\,. \tag{21}$$

Note that the formulations of the server allocation problem for open and clopen networks are similar. However, the implied meaning of the corresponding objective functions is different. For the optimal allocation of servers specified by the solution $\widetilde{\mathbf{X}}$ to the RAC, the value of $\theta = \Lambda(\widetilde{\mathbf{X}})^{-1}$ gives the minimum expected network time-to-empty that can be achieved with static server allocation.

4 Upper bounds for the maximum throughput

The solution of the RAO and RAC, both of which rely on integer programming, presents a difficult task, but some simplifications may help to estimate the optimal solution [5]. Below we give an upper bound for the throughput of open networks. Similar results can be obtained for clopen networks by substituting $\lambda(\mathbf{X})$ and w_n with $\Lambda(\mathbf{X})$ and W_n.

4.1 The case $\pi_{nm} = \pi_n$ for all *n* and *m*

Let $\mathbf{X}^*$ be a solution to an RAO, and assume that server productivities are independent of the server type, i.e., assume that

$$\pi_{nm} = \pi_n > 0 \tag{22}$$

for all *n* and *m*. Then for all stations *n* we have that

$$\lambda(\mathbf{X}^*)\frac{w_n}{\pi_n} \le \sum_{m=1}^{M} x^*_{nm}\ , \tag{23}$$

and it follows from (23) that for each set $S_i \in \mathcal{G}$ the following inequality is valid:

$$\lambda(\mathbf{X}^*)\sum_{n\in S_i}\frac{w_n}{\pi_n} \le \sum_{n\in S_i}\sum_{m=1}^{M} x^*_{nm} \le B_i\ .$$

Therefore, the maximum throughput of the network is given by the following upper bound:

$$\lambda(\mathbf{X}^*) \le \min_{1\le i\le K}\frac{B_i}{\sum\limits_{n\in S_i}\frac{w_n}{\pi_n}}. \tag{24}$$

Note that the denominator in (24) is the total expected service time required for processing a job over all its visits to the set of stations S_i.

In addition to assumption (22), let resource requirements also be independent of the server type, i.e., suppose that for all *m* and *l*,

$$r_{ml} = r_l > 0 \tag{25}$$

In this case, due to (11), the total quantity of allocated servers satisfies the following inequality:

$$\sum_{n=1}^{N}\sum_{m=1}^{M} x^*_{nm} \le \frac{R_l}{r_l}. \tag{26}$$

Furthermore, another bound for the network throughput follows from (23):

$$\lambda(\mathbf{X}^*) \le \frac{\min\limits_{1\le l\le L}\frac{R_l}{r_l}}{\sum\limits_{n=1}^{N}\frac{w_n}{\pi_n}}. \tag{27}$$

4.2 General case

In the general case, a relaxation of the constraints in (21) can be used to evaluate the maximum throughput $\lambda(\mathbf{X})$. Consider the following Relaxed RAO (RRAO):

(RGAO)

maximize

$$\lambda(\mathbf{X}) = \min_{1 \le n \le N} \left(\frac{1}{w_n} \sum_{m=1}^{M} \pi_{nm} x_{nm} \right) \tag{28}$$

subject to:

$$\sum_{n=1}^{N} \sum_{m=1}^{M} x_{nm} r_{ml} \le R_l, \quad l = 1, \ldots, L, \tag{29}$$

$$\sum_{n \in S_i} \sum_{m=1}^{M} x_{nm} \le B_i, \qquad i = 1,2,\ldots,K, \tag{30}$$

$$x_{nm} \ge 0, \quad n = 1,2,\ldots,N, \; m = 1,2,\ldots,M. \tag{31}$$

The only difference between RRAO and GAO is that here, the unknown variables x_{nm} may take on any nonnegative real values. Because the search space in RAO is larger than in RAO, the value of the objective function $\lambda(\widetilde{\mathbf{X}})$ for a solution $\widetilde{\mathbf{X}}$ to RRAO yields an upper bound for the maximum network throughput $\lambda(\mathbf{X}^*)$ provided by the solution $\mathbf{X}^*$ from RAO.

For any solution $\mathbf{Y}$ to the RRAO, there exists a balanced solution $\mathbf{X}$ giving the same throughput as $\mathbf{Y}$ and satisfying the following equalities:

$$\mu_n(\mathbf{X}) = \lambda(\mathbf{Y}), \quad n = 1,2,\ldots,N. \tag{32}$$

For example, the matrix $\mathbf{X}$ defined by

$$x_{nm} = y_{nm} \frac{\lambda(\mathbf{Y})}{\mu_n(\mathbf{Y})}$$

satisfies conditions (29)-(32).

The RRAO problem has an equivalent linear programming formulation:

(LRAO)

maximize λ

subject to:

$$\lambda w_n \le \sum_{m=1}^{M} \pi_{nm} x_{nm}, \quad n = 1,2,\ldots,N,$$

$$\sum_{n=1}^{N} \sum_{m=1}^{M} x_{nm} r_{ml} \le R_l, \quad l = 1, \ldots, L,$$

$$\sum_{n \in S_i} \sum_{m=1}^{M} x_{nm} \le B_i, \qquad i = 1,2,\ldots,K,$$

$$x_{nm} \ge 0, \quad n = 1,2,\ldots,N, \; m = 1,2,\ldots,M.$$

We prefer the formulation given in (28)-(31) because it helps us understand the idea underlying the server allocation procedure presented in the next section. This procedure uses server composition vectors to determine stations with an excess of resources and balances the network by allocating it to the bottleneck station.

5 Solution of the relaxed resource allocation problem

Starting with some initial server allocation, we can consistently increase network throughput by moving servers from non-bottleneck stations to bottleneck stations. Figure 2 illustrates the resource allocation procedure for a simplified RRAO in which the constraints from (30) have been omitted. This procedure is iterative and converges to a solution with required accuracy ε, where $0 < \varepsilon << 1$, while undertaking the following steps:

1. For each resource of type l, compute the quantity of untapped resources of type l, u_l.
2. Compute the saturation flow of each station and determine a permutation $q(1), q(2), \ldots, q(N)$ of the indices 1,2,..., N that will place the saturation flows into non-decreasing order, $\mu_{q(1)} \le \mu_{q(2)} \le \ldots \le \mu_{q(N)}$.
3. For each server of type j, compute v_j as follows. If $x_{q(i)j} = 0$ for all i, then set $v_j = 0$. Otherwise, compute the maximal index i such that $x_{q(i)j} > 0$, and set $v_j = q(i)$.
4. For each server of type k having $\pi_{q(1)k} > 0$ and each composition vector $\boldsymbol{\theta}$ corresponding to this server type, perform the following steps.
   a) For each station i having corresponding to a nonempty set $Z_i = \{j \mid \theta_j > 0, v_j = i\}$, compute parameters a_i, b_i and c_i as
   $$a_i = \min_{j \in Z_i} \left(\frac{x_{q(i)j}}{\theta_j} \right), \; b_i = \frac{\mu_{q(i)} - \mu_{q(1)}}{\dfrac{\pi_{q(1)k}}{w_{q(1)}} + \dfrac{1}{w_{q(i)}} \displaystyle\sum_{j \in Z_i} \pi_{q(i)j} \theta_j}, \; c_i = \min(a_i, b_i).$$
   b) Compute parameter $\delta_k(\boldsymbol{\theta})$ as
   $$\delta_k(\boldsymbol{\theta}) = \min_{i : Z_i \neq \varnothing} c_i. \tag{33}$$
   c) Compute the quantity of type k servers that can be obtained by resource composition as specified by vector $\boldsymbol{\theta}$:
   $$\Delta_k(\boldsymbol{\theta}) = \min_{l : r_{kl} > 0} \frac{1}{r_{kl}} \left(u_l + \delta_k(\boldsymbol{\theta}) \sum_{j=1}^{K} \theta_j r_{jl} \right). \tag{34}$$
   d) Compute the throughput gain achieved after addition of $\Delta_k(\boldsymbol{\theta})$ units of type k servers to the bottleneck station as
   $$g_k(\boldsymbol{\theta}) = \Delta_k(\boldsymbol{\theta}) \frac{\pi_{q(1)k}}{w_{q(1)}}. \tag{35}$$

5. Select servers of type $k = \hat{k}$ and take the composition vector $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ for which the throughput gain $g_k(\boldsymbol{\theta})$, achieved after the addition of $\Delta_k(\boldsymbol{\theta})$ units of type k servers to the bottleneck station, attains its maximum value.
6. Adjust the allocated quantities of servers by adding $\Delta_{\hat{k}}(\hat{\boldsymbol{\theta}})$ servers of type $\hat{k}$ to station $q(1)$ and removing $\hat{\theta}_j \delta_{\hat{k}}(\hat{\boldsymbol{\theta}})$ servers of type j from station v_j; i.e., set

$$\begin{aligned} x_{q(1),\hat{k}} &:= x_{q(1),\hat{k}} + \Delta_{\hat{k}}(\hat{\boldsymbol{\theta}}), \\ x_{v_j,j} &:= x_{v_j,j} - \hat{\theta}_j \delta_{\hat{k}}(\hat{\boldsymbol{\theta}}), \qquad j = 1,2,\ldots,K. \end{aligned} \tag{36}$$

7. Compute the throughput $\lambda(\mathbf{X})$.
8. If $\Delta_{i^*k^*} < \varepsilon\lambda(\mathbf{X})$, then convergence has been obtained, and terminate the server allocation procedure; otherwise, return to step 1.

Before each iteration of this method, there is a group of bottleneck stations having the same saturation rate. In each iteration the saturation rate of a bottleneck station in the group increases, and after increasing the saturation rate of the last bottleneck station in the group, a new group of bottleneck stations arises. The saturation rate in the new group of bottleneck stations is higher than that of the previous group of bottleneck stations. Therefore, the sequence of network throughputs calculated by each iteration is a non-decreasing, bounded and convergent sequence. When convergence is reached, $\lambda(\mathbf{X})$ gives the highest throughput that can be achieved with available resources.

6 Examples

Consider an open two-station network processing three classes of jobs. External arrivals belong to either class 1, with probability 0.8, or class 2, with probability 0.2. The first station serves class 1 and 2 jobs alone. After service, class 1 jobs leave the network, while class 2 jobs arrive at the second station as class 3 jobs. The quantity of resources and the number of servers at the second stations are limited by $R_1 = 5$, $R_2 = 3$ and $b_2 = 2$. Server productivity and resource requirements are presented in Figure 3, which also depicts external arrivals and job class distribution. Table 1 gives network throughput and server allocation for different values of the value b_1 limiting the maximum number of servers at the first station. Utilization of stations, servers and resources for the arrival rate $a = 0.6$ are given in Table 2.

Table 1. Network throughput and resource allocation

Capacity of station 1	Network throughput	Server allocation			
		Station	Quantities of type 1 servers	Quantities of type 2 servers	Quantities of type 3 servers
$b_1 = 1$	0.800	1	0.000	1.000	0.000
		2	0.000	0.000	0.364
$b_1 = 2$	1.600	1	0.000	2.000	0.000
		2	0.000	0.000	0.727
$b_1 = 3$	2.250	1	1.125	1.875	0.000
		2	0.000	0.000	1.023
$b_1 = 4$	2.883	1	2.375	1.625	0.000
		2	0.000	0.000	1.311
$b_1 = 5$	3.500	1	3.750	1.250	0.000
		2	0.000	0.000	1.591
$b_1 = 6$	3.500	1	3.750	1.250	0.000
		2	0.000	0.000	1.591

Table 1. Utilization of network components for arrival rate $a = 0.6$

Capacity of station 1	Station		Server type			Resource type	
	1	2	1	2	3	1	2
$b_1 = 1$	0.750	0.750	0.000	0.750	0.750	0.150	0.341
$b_1 = 2$	0.375	0.375	0.000	0.375	0.375	0.150	0.341
$b_1 = 3$	0.267	0.267	0.267	0.267	0.267	0.160	0.258
$b_1 = 4$	0.208	0.208	0.208	0.208	0.208	0.166	0.207
$b_1 = 5$	0.171	0.171	0.171	0.171	0.171	0.171	0.162
$b_1 = 6$	0.171	0.171	0.171	0.171	0.171	0.171	0.162

7 Conclusions

In this paper, we have studied resource allocation in multiclass networks having several types of flexible servers that operate only if certain required resources are available. We have proposed a method for the calculation of an upper bound for the maximum network throughput of open networks and a lower bound for the minimum time-to-empty for clopen networks; this method can be achieved with static allocation of resources to servers. Our results for servers representing teams of resources are new and, to our knowledge, have not been published elsewhere.

References

[1] Andradottir, S., Ayhan, H. and D.G. Down, Dynamic server allocation for queueing networks with flexible servers, *Operations Research*, Vol. 51 (6), pp. 952–968, 2003.

[2] Down, D.G. and Karakostas, G.. Maximizing throughput in queueing networks with limited flexibility, *European J. of Operational Research*, Vol. 187(1), pp. 98–112, 2008.

[3] Hopp, W.J. and Van Oyen, M.P. Agile workforce evaluation: a framework for cross-training and coordination. *IIE Transactions*, Vol. 36 (10), pp. 919–940, 2004.

[4] Gurumurthi, S. and S. Benjaafar, Modeling and Analysis of Flexible Queueing Systems, *Naval Research Logistics*, Vol. 51, pp. 755–782, 2004.

[5] Yves Pochet and Laurence A. Wolsey. *Production Planning by Mixed Integer Programming*, Springer, 2006.

[6] O.Z. Aksin et.al. A review of workforce cross-training in call centers from an operations management perspective, in *Workforce Cross Training Handbook* (D. Nembhard ed.), CRC Press, pp. 211–240, 2007.

[7] Issam Al-Azzoni and D. G. Down, Linear programming based affinity scheduling for heterogeneous computing systems. IEEE Transactions on Parallel and Distributed Systems, Vol. 19(2), pp. 1671–1682, 2008.

[8] Denning, P.J. Throughput. In *Wiley Encyclopedia of Computer Science and Engineering* (B.W. Wah ed.), Wiley-Interscience, 2008.

[9] Richard J. Boucherie and Nico M. van Dijk. *Queueing Networks: A Fundamental Approach*, Springer, 2011.

[10] Y. Chevaleyre et.al. Issues in Multiagent Resource Allocation. *Informatica*, Vol. 30, No. 1, pp.3–31, 2006.

[11] Alan S. Mann. Basic concepts of activity analysis. In *Studies in Process Analysis* (Alan S. Manne and Harry M. Markowitz eds). John Wiley & Sons, New York, pp. 417–422, 1963.

[12] Micha Pióro and Deepankar Medhi. Routing, Flow, and Capacity Design in Communication and Computer Networks, Elsevier, 2004.

[13] George Shanthikumar, David D. Yao and W. H. M. Zijm. Stochastic modeling and optimization of manufacturing systems and supply chains, Springer, 2003.

[14] François Sainfort, John Blake, Diwakar Gupta and Ronald L. Rardin. Operations Research for Health Care Delivery Systems, WTEC, 2005.

[15] Valeriy Naumov and Olli Martikainen. Method for throughput maximization of multiclass network with flexible servers. ETLA Discussion papers 1261, 1-27, 2011.

[16] O. Martikainen and M. Lahti. Performance analysis of OSI TP4/CLNP on FDDI, in *Proc. 2nd MultiG Workshop*, Stockholm, 17.6.1991, 1 – 14, 1991.

[17] Maury Bramson. Stability of queuing networks. *Probability Surveys*, Vol. 15, pp. 169-345, 2008.

Figure 1 Network components and parameters

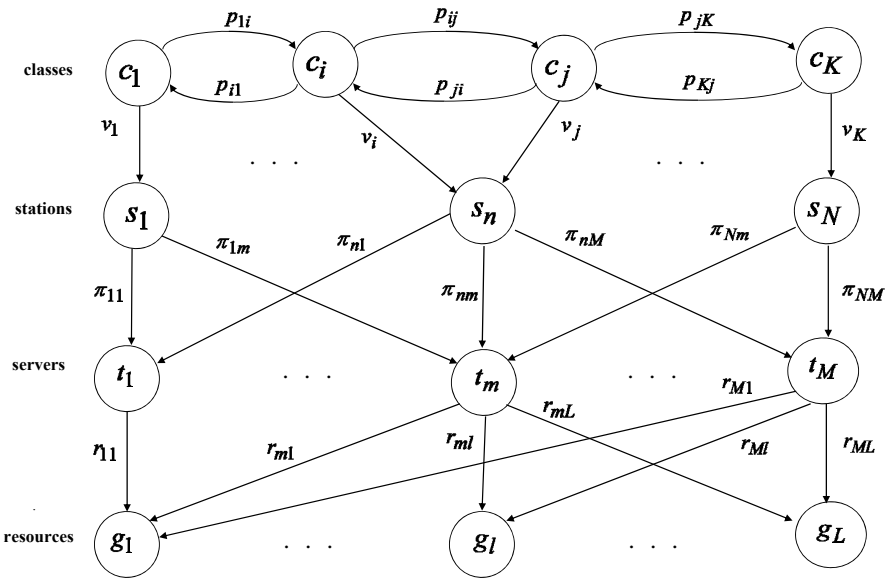


Figure 2 Resource allocation procedure

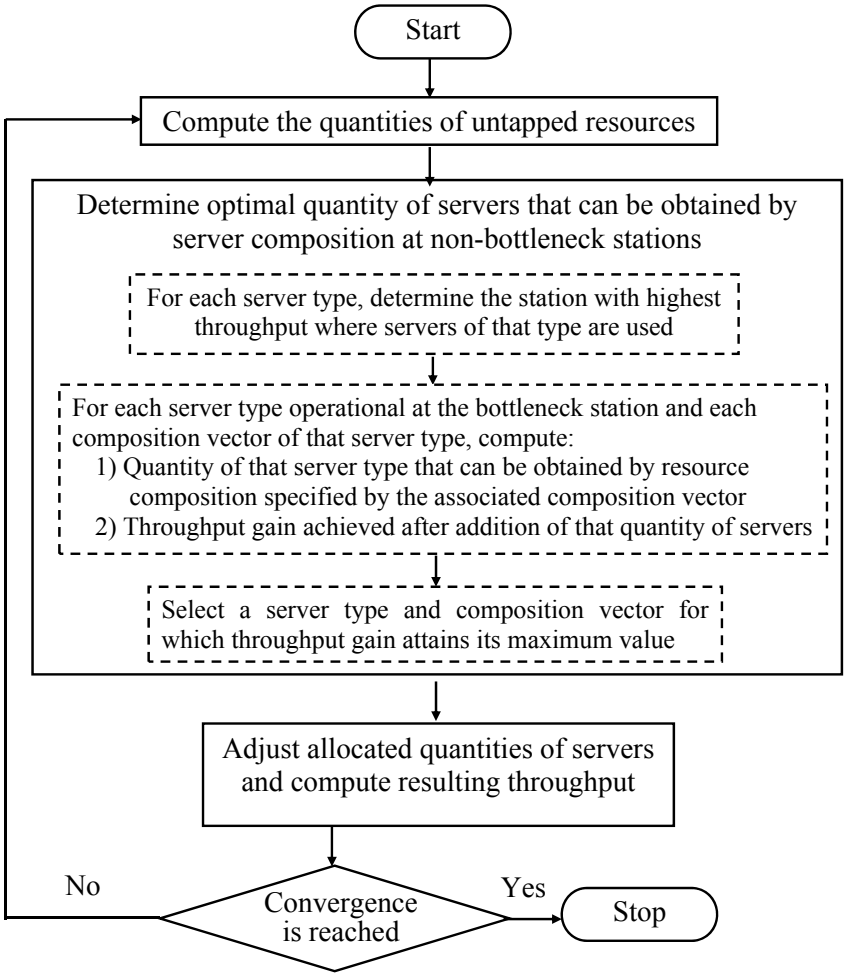
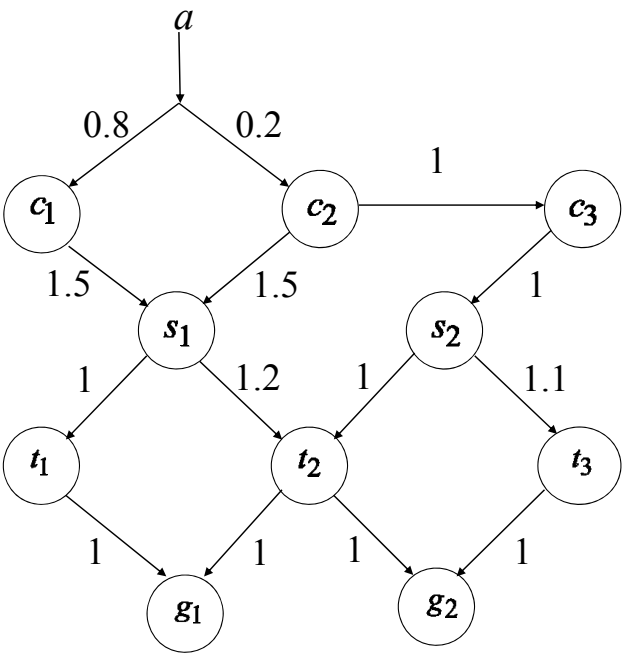


Figure 3 Components and parameters of example network



Aikaisemmin ilmestynyt ETLAn Keskusteluaiheita-sarjassa

Previously published in the ETLA Discussion Papers Series

No 1247 *Antti Kauhanen*, The Perils of Altering Incentive Plans. A Case Study. 08.04.2011. 22 p.

No 1248 *Rita Asplund – Sami Napari*, Intangible Capital and Wages. An Analysis of Wage Gaps Across Occupations and Genders in Czech Republic, Finland and Norway. 11.04.2011. 18 p.

No 1249 *Mari Kangasniemi – Antti Kauhanen*, Performance-related Pay and Gender Wage Differences. 21.04.2011. 19 p.

No 1250 *Ye Zhang*, Wireless Acquisition of Process Data. 24.05.2011. 52 p.

No 1251 *Rita Asplund – Erling Barth – Per Lundborg – Kjersti Misje Nilsen*, Challenges of Nordic Labour Markets: A Polarization of Working Life? 08.06.2011. 21 p.

No 1252 *Jari Hyvärinen*, Innovaatiotoiminta: Näkemyksiä ympäristö- ja energia-alaan. 1.6.2011. 39 s.

No 1253 *Ari Hyytinen – Mika Maliranta*, Firm Lifecycles and External Restructuring. 17.06.2011. 34 p.

No 1254 *Timo Seppälä – Olli Martikainen*, Europe Lagging Behind in ICT Evolution: Patenting Trends of Leading ICT Companies. 22.06.2011. 18 p.

No 1255 *Paavo Suni – Pekka Ylä-Anttila*, Kilpailukyky ja globaalin toimintaympäristön muutos. Suomen koneteollisuus maailmantaloudessa. 19.08.2011. 39 s.

No 1256 *Jari Hyvärinen*, Innovaatiotoiminta: Näkemyksiä hyvinvointialaan ja työelämän kehittämiseen. 31.8.2011. 28 s.

No 1257 *Terttu Luukkonen – Matthias Deschryvere – Fabio Bertoni – Tuomo Nikulainen*, Importance of the Non-financial Value Added of Government and Independent Venture Capitalists. 2.9.2011. 28 p.

No 1258 *Ari Hyytinen – Mika Pajarinen – Pekka Ylä-Anttila*, Finpron vaikuttavuus – Finpron palveluiden käytön vaikutukset yritysten kansainvälistymiseen ja menestymiseen. 15.9.2011. 32 p.

No 1259 *Kari E.O. Alho*, How to Restore Sustainability of the Euro? 19.9.2011. 27 p.

No 1260 *Heli Koski*, Does Marginal Cost Pricing of Public Sector Information Spur Firm Growth? 28.9.2011. 15 p.

No 1261 *Valeriy Naumov – Olli Martikainen*, Method for Throughput Maximization of Multiclass Networks with Flexible Servers. 13.12.2011. 19 p.

Elinkeinoelämän Tutkimuslaitoksen julkaisemat "Keskusteluaiheita" ovat raportteja alustavista tutkimustuloksista ja väliraportteja tekeillä olevista tutkimuksista. Tässä sarjassa julkaistuja monisteita on mahdollista ostaa Taloustieto Oy:stä kopiointi- ja toimituskuluja vastaavaan hintaan.

Papers in this series are reports on preliminary research results and on studies in progress. They are sold by Taloustieto Oy for a nominal fee covering copying and postage costs.

Julkaisut ovat ladattavissa pdf-muodossa osoitteessa: www.etla.fi/julkaisuhaku.php
Publications in pdf can be downloaded at www.etla.fi/eng/julkaisuhaku.php

ETLA
Elinkeinoelämän Tutkimuslaitos
The Research Institute of the Finnish Economy
Lönnrotinkatu 4 B
00120 Helsinki

Puh. 09-609 900
Fax 09-601 753
www.etla.fi
etunimi.sukunimi@etla.fi

ISSN 0781–6847