Performance of Distributed and Parallel Systems
T. Hasegawa, H. Takagi, Y. Takahashi (Editors)
Elsevier Science Publishers B.V. (North-Holland)
IFIP, 1989

Throughput and Response Time Optimization in Queueing Network Models of Computer Systems

I. F. Akyildiz and G. Bolch***

*School of Information and Computer Science
Georgia Institute of Technology
Atlanta, Georgia 30332
U. S. A.

**Informatik IV
University of Erlangen-Nuernberg
8520 Erlangen
F. R. G.

ABSTRACT

Queueing network models are being used to analyze various optimization problems such as server allocation, design and capacity issues, optimal routing and workload allocation in computer systems, communication networks and flexible manufacturing systems. This paper presents procedures for optimizing the performance and cost parameters for closed queueing network models with arbitrarily connected single and infinite server stations. The throughputs are maximized for fixed costs. The mean response times are minimized for given fixed throughput values. In both cases the linear and nonlinear cost functions are considered. The method of Lagrange multipliers is used to solve the optimization problems. Numerical examples are given to illustrate and to discuss the solutions.

Key Words: Computer System Design, Performance Evaluation, Optimization, Queueing Networks, Lagrange Multipliers.

1. Introduction

Cost evaluation and performance prediction processes are an important step in the planning and design of computer systems, communication networks and flexible manufacturing systems. Queueing network models have received special interest for performance analysis in the last two decades. Several analytical methods have been derived for the analysis of queueing network models in recent years [6, 11, 24]. In addition to the computation of performance meas-

Akyildiz's work was supported by the School of Information and Computer Science, ICS, and by the Airforce Office of the Scientific Research, AFOSR, under Grant No. AFOSR-88-0028.

ures, queueing network models can also be used for optimization of performance measures. In the optimization procedure an objective function, such as the costs, the throughput, the utilization or response time, can be obtained by appropriate selection of certain system input parameters which are called as decision variables. Within the modeling by queueing networks, the quantities such as service rates and the number of jobs can be assumed as decision variables by which the objective function is to be reached. Decison variables are selected subject to certain constraints. For example, we optimize the throughput of a given queueing network model by proper selection of service rates at each station subject to the cost constraints which restrict the range of the decision variables. Queueing network models represent the relationship between the objective function and the system input parameters. Within these models it is possible to have an optimal design of computer systems, communication networks and flexible manufacturing systems by using the mathematical optimization techniques.

Several authors have discussed the issue of optimization in recent years. Trivedi and Wagner [21] consider a computer configuration design problem where the computer system is modeled by a closed central server model. The system throughput is the objective function to be maximized by proper choice of device speeds subject to a cost constraint. A non-linear cost function is considered in the analysis. Trivedi, Wagner and Sigmon [22], Trivedi and Sigmon [23] analyze a computer system configuration problem in which the objective is to select the CPU speed, the capacities of secondary storage devices and the allocation of a set of files across the secondary storage devices so as to maximize the system throughput subject to a cost constraint. Kleinrock [10], Chandy, Hogarth and Sauer [5] consider a similar decison model as Trivedi and Kinicki [20] but their model is an open queueing network model. Chandy, Hogarth and Sauer [5] extend Kleinrock's model [10] in two different directions. First they allow locally balanced open queueing networks with multiple jobs classes as well as the open networks analyzed by diffusion approximation [12]. Second they consider a rich class of nonlinear cost functions. Ferrari [6] uses a cyclic queueing model to optimize throughput subject to a nonlinear cost constraint. He solves the problem graphically and hence his method is restricted to problems with only a few devices.

Von Mayrhauser and Trivedi [25] consider a configuration design problem where the computer system is modeled as a closed queueing network. The mean response time to an interactive user request is minimized and the speeds of the devices are the decision variables. Geist and Trivedi [7] developed an optimization model for assigning a fixed set of files across an assemblage of memory devices so as to maximize system throughput. Trivedi and Kinicki [20] and Trivedi and Wagner [21] consider single server queueing networks and show by using the results of Price [15] that the optimization problem is global. They use the convolution algorithm as the base for optimization and maximize the throughput and minimize the costs. Note also that they do not give explicit closed form solutions for the optimization problem.

Kenevan and von Mayrhauser [9] show that the throughput is a log convex function of the number of items in a closed, single class, network of an arbitrary number of single and infinite servers. They also prove that reciprocal throughput is a convex function of the relative utilization of the servers which is the generalization of Price's proof [15]. Kobayashi and Gerla [13] determine the optimal routing in central server models with single server stations and multiple classes of jobs. Heiss and Totzauer [8] consider open BCMP queueing networks with load independent service rates. They determine the throughputs to optimize a linear objective function which is chosen to be a weighted sum of station utilizations. As restrictions, minimal and maximal throughputs and maximal response times are allowed per job class. Stecke [19] investigates BCMP networks in which she imposes a constraint on the total workload in the system. She shows that throughput as a function of the ratio of the service rate at a server to the sum of the workloads, not purely concave but rather quasi-concave.

This paper is organized as follows: In section 2 the optimization problems are presented. In section 3 we briefly describe the mean value analysis, henceforth in short form MVA, and derive the constraint which will be used in application of Lagrange multiplier technique for optimization of performance measures. In section 4 we obtain an iterative formula for maximum throughput. The formula has a closed form solution in case of linear cost constraints. Response time minization procedure is given in section 5. Additonally, numerical examples are given to illustrate and to discuss the solutions.

2. Optimization

In the optimization procedure of performance measures such as response times, throughputs, utilizations, the objective is to determine the optimal service rates as decision variables subject to certain constraints. The cost constraint is the most common case and has the following form in the linear case:

$$\sum_{i=1}^{N} c_i \mu_i = C \tag{1}$$

where N is the number of stations in the queueing network. The parameter c_i is the cost constraint for station i. C denotes the total cost of the queueing network.

The nonlinear cost function is the more realistic case where the total costs are computed as follows:

$$\sum_{i=1}^{N} c_i \mu_i^{\alpha_i} = C \qquad \alpha_i \geq 1 \tag{2}$$

where α_i controls the increase of the costs for each station i in the network. For the special case, $\alpha_i = 1$ for $i = 1,2,..,N$ we obtain the linear cost function, equation (1).

In the following we summarize the different optimization problems which we will investigate in this work.

i) *Throughput Maximization under Fixed Costs*

The total cost for the queueing network model is assumed to be known. The total cost is the budget available for purchasing a given number of stations with specific service rates. The total throughput is maximized by determining optimal service rates under cost constraints. We investigate this problem in section 4.

ii) *Response Time Minimization for Each Station*

The response time of each station for a given queueing network is minimized where a given fixed throughput value is controlled. The service rates are also, in this case, under cost constraints. In section 5 a solution is given to this problem.

3. Mean Value Analysis

We consider BCMP [3] queueing networks with N stations and K jobs. We denote μ_i as the service rate of the i-th station. As generally known, BCMP networks contain the following station types: *Type I: • /M/m - FCFS, Type II: • /G/1 - PS, Type III: • /G/IS (Infinite Servers), Type IV: • /G/1-LCFS-PR.*

The Bard/Schweitzer [2,18] algorithm provides the following solution for the analysis of BCMP networks containing single and infinite server stations.

The *mean response time* of the *i-th station (for $i = 1,\dots,N$) is:*

$$\bar{t}_i = \begin{cases} \dfrac{1}{\mu_i}\left(1 + \dfrac{K-1}{K}\ \bar{k}_i \right) & Type(i) \neq IS \\ \dfrac{1}{\mu_i} & Type(i) = IS \end{cases} \tag{3}$$

The *throughput* of the network is obtained by Little's law:

$$\lambda = \frac{K}{\sum_{i=1}^{N} e_i\ \bar{t}_i} \tag{4}$$

The *mean number of jobs* at the i-th station (for $i=1,..,N$) is also obtained by Little's law:

$$\overline{k_i} = \lambda \cdot e_i \cdot \overline{t_i} \tag{5}$$

where e_i is the mean number of visits that a job makes to station i:

$$e_i = \sum_{j=1}^{N} e_j \cdot p_{ji} \qquad \text{for } i=1,..,N$$

p_{ji} is the transition probability that a job after completing service at station j proceeds to station i.

In the following we derive new formulas which provide a different perspective on performance measures like mean response times and mean number of jobs.

By substituting equation (5) in equation (3) we obtain:

$$\overline{t_i} = \begin{cases} \frac{1}{\mu_i}\left(1 + \frac{K-1}{K} \lambda \, e_i \, \overline{t_i} \right) & Type(i) \neq IS \\ \frac{1}{\mu_i} & Type(i) = IS \end{cases} \tag{6}$$

Solving equation (6) for $\overline{t_i}$ we get

$$\overline{t_i} = \begin{cases} \dfrac{\frac{1}{\mu_i}}{\left(1 - \frac{K-1}{K} \frac{\lambda \, e_i}{\mu_i} \right)} & Type(i) \neq IS \\ \frac{1}{\mu_i} & Type(i) = IS \end{cases} \tag{7}$$

From equation (4) we derive

$$\lambda \cdot \sum_{i=1}^{N} e_i \; \overline{t_i} = K \tag{8}$$

From equations (7 and 8) we obtain

$$\sum_{Type(i) \neq IS}^{N} \frac{\frac{\lambda \, e_i}{\mu_i}}{1 - \frac{K-1}{K} \lambda \, \frac{e_i}{\mu_i}} + \sum_{Type(i) = IS}^{N} \frac{\lambda \, e_i}{\mu_i} = K \tag{9}$$

Equation (9) is the constraint used in application of Lagrange multipliers technique for optimization of performance measures. In [1] we assumed $\frac{K-1}{K}$ is approximately equal to one and simplified equation (9). Using the simplified equations we then solved the optimization problems accordingly. However, the accuracy of the results is violated by this assumption. Similar way of attacking the optimization problems as in [1] was also utilized by [4,17].

4. Throughput Optimization for Fixed Costs

In the planning phase of computer systems, communication networks and flexible manufacturing systems we have to consider the fact that only a certain amount of money is available. The objective is to optimize the performance measures within the available budget. In this section we show how the throughput can be maximized by selecting the service rates of each station within the available budget.

The objective here is to find the optimal service rates μ_i^* which provide the maximum throughput λ^* subject to the nonlinear costs. The linear cost is a special case of the nonlinear case. Note that another solution was given by Trivedi and Wagner [21] to this problem. They derive an explicit formula for total throughput λ from the normalization constant which is then maximized subject to the linear costs, equation (1), dependent on service rates. However, queueing networks studied by [21] may contain only single server stations. We include infinite server stations to the model considered for optimization problems. The solution suggested by [21] is further simplified here by using the MVA as the base for optimization.

The optimal service rates μ_i^* which determine the maximum throughput λ for a given cost constraint cannot be computed by a closed form solution. However, they are determined iteratively as follows:

- *Initialize* the auxilary quantities

$$\mu_i^{(0)} = 1 \qquad \text{for } i = 1, ..., N$$

- *Iterate for* $n = 1,2,\ldots$ until the deviation between the iterations is small:

$$\lambda^{(n)} = \frac{C\ K}{\left[\sum_{j=1}^{N} \sqrt{c_i\ e_i\ \mu_i^{\alpha-1^{(n)}}}\right]^2 + (K-1) \sum_{\neq IS} c_i\ e_i\ \mu_i^{\alpha-1^{(n)}}} \tag{10}$$

$$\mu_i^{(n+1)} = \begin{cases} \dfrac{\lambda^{(n)}\ e_i}{K} \left[\dfrac{\sum_{j=1}^{N} \sqrt{c_j\ e_j\ \mu_j^{\alpha-1^{(n)}}}}{\sqrt{c_i\ e_i\ \mu_i^{\alpha-1^{(n)}}}} + (K-1) \right] & \text{for} \quad Type(i) \neq IS \\ \lambda^{(n)}\ e_i \dfrac{\sum_{j=1}^{N} \sqrt{c_j\ e_j\ \mu_j^{\alpha-1^{(n)}}}}{K\ \sqrt{c_i\ e_i\ \mu_i^{\alpha-1^{(n)}}}} & \text{for} \quad Type(i) = IS \end{cases} \tag{11}$$

Note that based on our test studies we found out that the iteration converges for $0 < \alpha < 2$. This is also based on the fact that the values for μ_i are not initiated appropriately.

Derivation.

First we rewrite equation (9) as follows:

$$\sum_{Type(i) \neq IS}^{N} \frac{\lambda\ e_i}{\mu_i - \dfrac{K-1}{K}\ \lambda\ e_i} + \sum_{Type(i) = IS}^{N} \frac{\lambda\ e_i}{\mu_i} = K \tag{12}$$

Equation (12) implicitly defines λ as a function of $\underline{\mu}$, i.e., $\lambda = \lambda(\underline{\mu})$

By differentiating equation (12) by μ_i we obtain

$$\frac{\partial \lambda}{\partial \mu_i} = \begin{cases} \dfrac{\lambda\ e_i}{(\mu_i - \dfrac{K-1}{K}\ \lambda\ e_i)^2} * A & \text{for} \quad \neq IS \\ \dfrac{\lambda\ e_i}{\mu_i^2} * A & \text{for} \quad IS \end{cases} \tag{13}$$

where

$$A = \left[\sum_{\substack{j \neq IS \\ \forall j=1}}^{N} \frac{e_j\ \mu_j}{(\mu_j - \dfrac{K-1}{K} \lambda\ \ e_j\ )^2} + \sum_{\substack{j = IS \\ \forall j=1}}^{N} \frac{e_j}{\mu_j} \right]^{-1} \tag{14}$$

The Lagrange function $L(\underline{\mu}, y)$ with objective function $\lambda(\underline{\mu})$ is written as follows:

$$L(\underline{\mu},\ y) = \lambda(\underline{\mu}) + y\ (\ \sum_{i=1}^{N} c_i\ \mu_i^{\alpha} - \ C\ ) \tag{15}$$

where $\lambda(\underline{\mu})$ is defined by equation (12).

By differentiating $L(\underline{\mu},y)$ by μ_i and y we obtain the following nonlinear system equation for determining the maximum throughput and the optimal service rates:

$$\frac{\partial L}{\partial \mu_i} : \frac{\partial \lambda}{\partial \mu_i} + y\,\alpha\, c_i\, \mu_i^{\alpha-1} = 0 \tag{16}$$

$$\frac{\partial L}{\partial y} : \quad \sum_{i=1}^{N} c_i\, \mu_i^{\alpha} = C \tag{17}$$

By substituting equation (13) into (16) we obtain

$$\frac{\partial L}{\partial \mu_i} = \begin{cases} \dfrac{\lambda\, e_i}{(\mu_i - \frac{K-1}{K}\lambda\, e_i)^2} * A \;+\; y\,\alpha\, c_i\, \mu_i^{\alpha-1} = 0 & \text{for} \quad \neq IS \\[2ex] \dfrac{\lambda\, e_i}{\mu_i^2} * A \;+\; y\,\alpha\, c_i\, \mu_i^{\alpha-1} = 0 & \text{for} \quad IS \end{cases} \tag{18}$$

Rewriting equation (18) provides

$$\frac{\partial L}{\partial \mu_i} = \begin{cases} \dfrac{\lambda\, e_i}{\mu_i - \frac{K-1}{K}\lambda\, e_i} * A \;+\; y\,\alpha\, e_i\, \mu_i^{\alpha-1}\,(\mu_i - \frac{K-1}{K}\lambda\, e_i) = 0 & \text{for} \quad \neq IS \\[2ex] \dfrac{\lambda\, e_i}{\mu_i} * A \;+\; y\,\alpha\, c_i \mu_i^{\alpha} = 0 & \text{for} \quad IS \end{cases} \tag{19}$$

By summing equation (19) over all stations i we obtain

$$K * A \;+\; y\,\alpha\,(\; C - \frac{K-1}{K}\lambda \sum_{\neq IS} c_i\, e_i\, \mu_i^{\alpha-1}\;) = 0 \tag{20}$$

Then it follows that

$$A = \frac{y\,\alpha\,(\;\frac{K-1}{K}\lambda \sum_{\neq IS} c_i\, e_i\, \mu_i^{\alpha-1} - C)}{K} \tag{21}$$

By substituting equation (21) into equation (18) we derive

$$\left[\frac{K\,\lambda\, c_i\, e_i\, \mu_i^{\alpha-1}}{C - \frac{K-1}{K}\lambda \sum_{\neq IS} c_i\, e_i\, \mu_i^{\alpha-1}}\right]^{1/2} = \begin{cases} \dfrac{\lambda\, e_i}{\mu_i - \frac{K-1}{K}\lambda\, e_i} & \text{for} \quad \neq IS \\[2ex] \dfrac{\lambda\, e_i}{\mu_i} & \text{for} \quad IS \end{cases} \tag{22}$$

By summing over all stations i we obtain

$$\left[\frac{K\,\lambda}{C - \frac{K-1}{K}\lambda \sum_{\neq IS} c_i\, e_i\, \mu_i^{\alpha-1}}\right]^{1/2} \sum_{i=1}^{N} \sqrt{c_i\, e_i\, \mu_i^{\alpha-1}} = K \tag{23}$$

By solving equation (23) for λ equation (10) is derived.

From equation (23) we obtain

$$C - \frac{K-1}{K}\lambda \sum_{\neq IS} c_i e_i \mu_i^{\alpha-1} = \frac{\lambda}{K}\Big(\sum_{i=1}^{N} \sqrt{c_i e_i \mu_i^{\alpha-1}}\Big)^2 \tag{24}$$

By substituting equation (24) into equation (22) and by rewriting we obtain equation (11).

Note that for linear costs, i. e., $\alpha=1$, equation (10) and (11) become the following closed forms, i.e., there is no need for iterations in linear cost case.

$$\lambda^* = \frac{C \cdot K}{\Big(\sum_{i=1}^{N}\sqrt{c_i e_i}\Big)^2 + (K-1)\sum_{i \neq IS} c_i e_i} \tag{25}$$

$$\mu_i^* = \begin{cases} \dfrac{\lambda e_i}{K}\left[\dfrac{\sum_{j=1}^{N}\sqrt{c_j e_j}}{\sqrt{c_i e_i}} + (K-1)\right] & \text{for} \quad Type(i) \neq IS \\ \lambda e_i \dfrac{\sum_{j=1}^{N}\sqrt{c_j e_j}}{K\sqrt{c_i e_i}} & \text{for} \quad Type(i) = IS \end{cases} \tag{26}$$

Example.

We consider the following queueing network model:

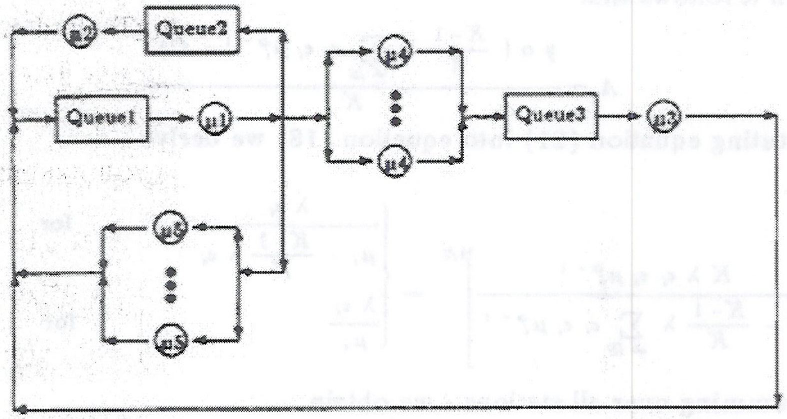

Figure 1.

There are $N=5$ stations, $K=20$ jobs. Stations 4 and 5 are of Type 3 where other stations have Types of 1,2 or 4. Transition probabilities are given by:

$$p_{12}=0.2;\ p_{14}=0.5;\ p_{16}=0.3;\ p_{43}=p_{21}=p_{31}=p_{61}=1.$$

The mean number of visits e_i that a job makes to station i is:

$$e_1=1;\ e_2=0.2;\ e_3=0.5;\ e_4=0.5;\ e_5=0.3.$$

The service rate costs c_i are:

$$c_1=10;\ c_2=5;\ c_3=5;\ c_4=2;\ c_5=1.$$

We investigate the effect of different parameters α, of the cost function, equation (2), on the maximum throughput λ. In Table 1 we show different service rates μ_i^* computed using the iterative procedure, equations (10 and 11), for different α values. We also give the throughput values (λ) which are obtained by mean value analysis using the optimal service rates μ_i^*.

α	0.5	0.75	1	1.25	1.5	1.75	1.95
μ_1^*	37.153	12.688	7.242	5.105	4.011	3.358	3.330
μ_2^*	7.810	2.689	1.550	1.105	0.801	0.748	0.678
μ_3^*	19.154	6.642	3.738	2.639	2.078	1.745	1.563
μ_4^*	1.085	0.496	0.370	0.331	0.320	0.321	0.326
μ_5^*	1.226	0.551	0.405	0.359	0.344	0.343	0.347
λ^*	36.007	12.395	7.109	5.028	3.960	3.322	2.793
λ	38.318	12.479	7.151	5.048	3.912	3.444	2.987

Table 1. ($K=70$, $C=100$)

The values for $\alpha=1$ are computed from equations (25 and 26). From Table 1 it is clear that the service rates μ_i decrease at the CPU with increasing α since the cost factor is the greatest ($c_1=10$) at the CPU. The following graph shows the relationship between the maximal throughput and the quantity α. The throughput decreases exponentially with increasing α.

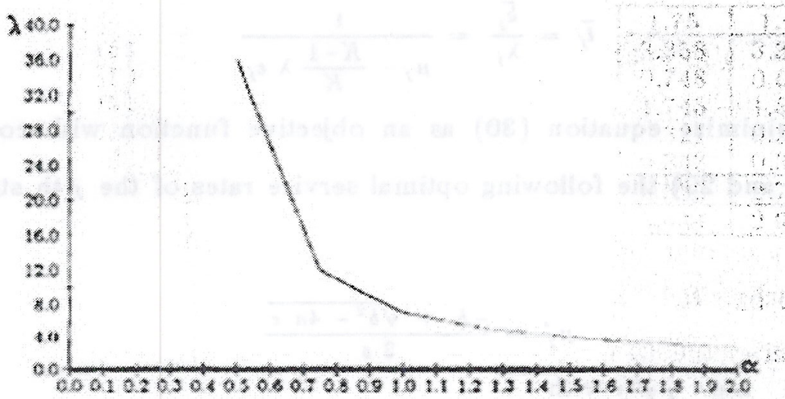

Figure 2. Maximum Throughput dependent on α

5. Response Time Minimization

The optimization question here is to minimize the objective function, the total response time T subject to the fixed cost constraints. This is equivalent to the question to maximize the throughput since there is a dependency between the throughput and response time which can be seen in Little's law:

$$T = \frac{K}{\lambda} \tag{27}$$

From the equations

$$\lambda_i \; \overline{t_i} = \overline{k_i} \qquad \lambda_i = \lambda \; e_i \qquad \text{for} \quad i = 1,\ldots,N \tag{28}$$

it is easy to show that the total response time T can be computed by:

$$T = \sum_{i=1}^{N} e_i \; \overline{t_i} \tag{29}$$

We concentrate our investigation on maximizing the response times for different types of stations individually in the queueing network.

5.1. Single Server Stations

The objective is to minimize the response time $\overline{t_j}$ of a specific station of Type 1,2,4 where a given throughput value λ is maintained. The service rates have a linear cost constraint, equation (1). The total cost C is known. The response time $\overline{t_j}$ in a Type 1,2,4 station is given by Little's law:

$$\overline{t_j} = \frac{\overline{k_j}}{\lambda_j} = \frac{1}{\mu_j - \frac{K-1}{K} \lambda \; e_j} \tag{30}$$

If we minimize equation (30) as an objective function with conditions, equations (9 and 29) the following optimal service rates of the j-th station are obtained:

$$\mu_j^* = \frac{-b \; + \; \sqrt{b^2 - 4a \; c}}{2 \; a} \tag{31}$$

for $j = 1,\ldots,N$ and $j \neq i$ with

$$a = K \; c_j \tag{32}$$

$$b = \lambda \left[\sum_{i \neq j}^{N} \sqrt{e_i \; c_i} \right]^2 - K \; \lambda \; e_j \; c_j - K \; C + (K-1) \; \lambda \sum_{i \neq j \; \delta \neq IS}^{N} e_i \; c_i \tag{33}$$

$$c = K \; \lambda \; e_j \; C - (K-1) \; \lambda^2 \; e_j \sum_{i \neq j \; \delta \neq IS}^{N} e_i \; c_i \; - \; \frac{K-1}{K} \lambda^2 \; e_j \left[\sum_{i \neq j}^{N} \sqrt{e_i \; c_i} \right] \tag{34}$$

The service rates of the remaining stations i for $i=1,...,N$ and $i \neq j$ are obtained by:

$$\mu_i^* = \begin{cases} \Phi \sqrt{\dfrac{e_i}{c_i}} + \dfrac{K-1}{K} \lambda e_i & \text{for} \quad Type(i) \neq IS \\ \Phi \sqrt{\dfrac{e_i}{c_i}} & \text{for} \quad Type(i) = IS \end{cases} \tag{35}$$

where

$$\Phi = \frac{C - \lambda \dfrac{K-1}{K} \displaystyle\sum_{i \neq j \, \delta \neq IS}^{N} e_i c_i - c_j \mu_j^*}{\displaystyle\sum_{i \neq j}^{N} \sqrt{e_i c_i}} \tag{36}$$

Derivation.

The Lagrange function is derived as follows:

$$L(\mu, y_1, y_2) = \frac{1}{\mu_j - \dfrac{K-1}{K} \lambda e_j} + y_1 \left(\sum_{\neq IS} \frac{\lambda e_i}{\mu_i - \dfrac{K-1}{K} \lambda e_i} + \sum_{IS} \frac{\lambda e_i}{\mu_i} - K \right) + y_2 \left[\sum_{IS} \mu_i c_i - C \right] \tag{37}$$

Differentiating by μ_i, μ_j, y_1 and y_2 we obtain the following system of equations:

$$\frac{\partial L}{\partial \mu_j} = \frac{-1}{(\mu_j - \dfrac{K-1}{K} \lambda e_j)^2} + y_1 \frac{-\lambda e_j}{(\mu_j - \dfrac{K-1}{K} \lambda e_j)^2} + y_2 c_j = 0 \tag{38}$$

for i, $j = 1,2,...,N$ and $i \neq j$.

$$\frac{\partial L}{\partial \mu_i} = \begin{cases} y_1 \dfrac{-\lambda e_i}{(\mu_i - \dfrac{K-1}{K} \lambda e_i)^2} + y_2 c_i = 0 & \text{for} \quad \neq IS \\ y_1 \dfrac{-\lambda e_i}{\mu_i^2} + y_2 c_i = 0 & \text{for} \quad IS \end{cases} \tag{39}$$

for $i,j = 1,...,N$ and $i \neq j$.

$$\frac{\partial L}{\partial y_1} : \quad \sum_{\neq IS} \frac{\lambda e_i}{\mu_i - \dfrac{K-1}{K} \lambda e_i} + \sum_{IS} \frac{\lambda e_i}{\mu_i} = K \tag{40}$$

$$\frac{\partial L}{\partial y_2} : \quad \sum_{i=1}^{N} c_i \mu_i = C \tag{41}$$

We can solve equation (39) as follows:

$$\sqrt{\frac{y_2}{y_1}\, c_i\, \lambda\, e_i} = \begin{cases} \dfrac{\lambda\, e_i}{\mu_i - \dfrac{K-1}{K}\lambda e_i} & \text{for} \quad \neq IS \\ \dfrac{\lambda\, e_i}{\mu_i} & \text{for} \quad IS \end{cases} \qquad (42)$$

for $i,j = 1,\ldots,N$ and $i \neq j$.

Rewriting equation (42) we get

$$\mu_i = \begin{cases} \sqrt{\dfrac{y_1\, \lambda\, e_i}{y_2\, c_i}} + \dfrac{K-1}{K}\lambda\, e_i & \text{for} \quad \neq IS \\ \sqrt{\dfrac{y_1\, \lambda\, e_i}{y_2\, c_i}} & \text{for} \quad IS \end{cases} \qquad (43)$$

for $i,j = 1,\ldots,N$ and $i \neq j$.

By substituting equation (42) in equation (40) we obtain

$$\frac{\lambda\, e_j}{\mu_j - \dfrac{K-1}{K}\lambda\, e_j} + \sqrt{\frac{y_2}{y_1}}\,\lambda \sum_{i \neq j} \sqrt{e_i\, c_i} = K \qquad (44)$$

for $i,j = 1,\ldots,N \quad i \neq j$

We substitute equation (43) in equation (41) and get → $\frac{K-1}{K}$

$$c_j\, \mu_j + \sqrt{\frac{y_1}{y_2}\lambda} \sum_{i \neq j}^{N} \sqrt{e_i\, c_i} + \lambda \sum_{i \neq j\ 8 \neq IS}^{N} e_i\, c_s = C \qquad (45)$$

For the sake of simplicity we introduce the following auxiliary quantities:

$$\Phi = \sqrt{\frac{y_1}{y_2}\,\lambda} \qquad (46)$$

$$w = \sum_{i \neq j}^{N} \sqrt{e_i\, c_i} \qquad (47)$$

$$s = \sum_{i \neq j\ 8 \neq IS}^{N} e_i\, c_i \longrightarrow \frac{K-1}{K} \qquad (48)$$

From (45) we obtain then equation (36) which has the following rewritten form

$$\Phi = \frac{C - \lambda\, s - c_j\, \mu_j}{w} \qquad (49)$$

We substitute (49) into (44) and obtain the following equation for determining the optimal service rates μ_j^* in the j-th station:

$$K\ c_j\ \mu_j^2 + [\lambda w^2 - K\lambda e_j c_j - KC + (K-1)\lambda s]\mu_j + \lambda e_j [KC - (K-1)\lambda s] - \frac{(K-1)}{K}\lambda w^2 = 0 \quad (50)$$

Note that this equation generally possesses two solutions. The mean response times $\overline{t}_j$ are minimized when the optimal service rates μ_j^* are chosen to be the greater values of the solution. The lesser values maximize the response times. Using μ_j^* we determine the auxiliary variable Φ from equation (49) and the service rates μ_i^* of the remaining stations $i=1,..,N$ and $j \neq i$ from equation (35).

Remark. Equation (35) may not always admit a unique solution. This is the case when the desired throughput value λ cannot be obtained under the given total cost C. In this case the solution procedure, equations (35 and 36), provides negative service rates which will be demonstrated in the following example. In the following example we minimize the response time at the CPU.

Example.

There are $N=4$ stations and $K=10$ total number of jobs. The total cost is $C=10$.

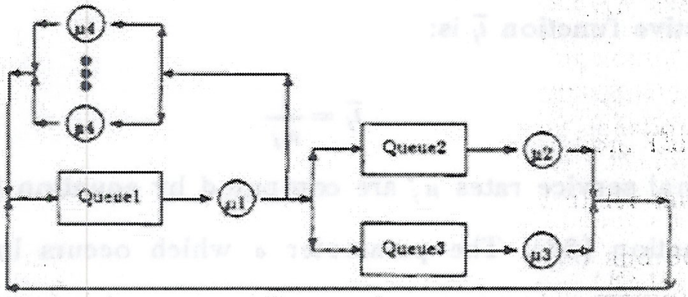

Figure 3.

The mean number of visits e_i is computed:

$$e_1 = 1\ ;\ e_2 = 0.4\ ;\ e_3 = e_4 = 0.3$$

Single costs for the service rates are $c_1 = c_2 = c_3 = c_4 = 1$

The objective is to minimize the response time $\overline{t}_1$ of the first station subject to the total cost $C=10$ while keeping the total throughput fixed at $\lambda=100$.

From equations (31 and 35) we obtain the following optimal result for service rates:

$$\mu_1^* = 108.49 \ ; \ \mu_2^* = -21.7 \ ; \ \mu_3^* = -23.41 \ ; \ \mu_4^* = -53.41$$

It can easily be seen that the total throughput of $\lambda = 100$ cannot be reached for the given cost constraint of $C = 10$.

Now we select new total throughput as $\lambda = 1$ and obtain as optimal service rates

$$\overline{t}_1 = 0.122 \ ; \ \mu_1^* = 0.068 \ ; \ \mu_2^* = 0.471 \ ; \ \mu_3^* = 0.366 \ ; \ \mu_4^* = 0.096$$

Using MVA with these optimal service rates we obtain the throughput and the mean response time of the first station:

$$\lambda = 1.029 \qquad \overline{t}_1 = 0.124$$

5.2. Infinite Server Stations

Since terminals can be modeled as an infinite server station we investigate the case where the response time of Type 3 stations is minimized. The objective is to determine the optimal service rates μ_i^* such that the response time $\overline{t}_j$ of a station j of Type 3 is minimized while reaching the value of given total throughput value λ.

The objective function $\overline{t}_j$ is:

$$\overline{t}_j = \frac{1}{\mu_j} \tag{51}$$

The optimal service rates μ_j^* are computed by equation (31) with Φ determined by equation (36). The parameter a which occurs in equation (31) is obtained from equation (32). However, the parameters b and c have a slightly different form here than in equations (33) and (34):

$$b = \lambda \left(\sum_{i \neq j}^{N} \sqrt{e_i \, c_i} \right)^2 - \lambda \, e_j \, c_j - K \; C + (K-1) \, \lambda \left| \sum_{i \neq j \, \delta \neq IS}^{N} e_i \, c_i \right| \tag{52}$$

$$c = \lambda \; e_j \left(C - \frac{K-1}{K} \lambda \sum_{i \neq j \, \delta \neq IS}^{N} e_i \, c_i \right) \tag{53}$$

The Lagrange function L also has a different form than in equation (37):

$$L(\underline{\mu}, y_1, y_2) = \frac{1}{\mu_j} + y_1 \left[\sum_{\neq IS} \frac{\lambda e_i}{\mu_i - \frac{K-1}{K}\lambda e_i} + \sum_{IS} \frac{\lambda e_i}{\mu_i} - K \right] + y_2 \left[\sum_{IS} \mu_i c_i - C \right] \quad (54)$$

Differentiating by μ_i, μ_j, y_1 and y_2 we obtain the following system of equations:

$$\frac{\partial L}{\partial \mu_j} = \frac{-1}{(\mu_j)^2} + y_1 \frac{-\lambda e_j}{(\mu_j)^2} + y_2 c_j = 0 \quad (55)$$

$\frac{\partial L}{\partial \mu_i}$, $\frac{\partial L}{\partial y_1}$ and $\frac{\partial L}{\partial y_2}$ are given by equations (39, 40 and 41), respectively.

Equation (44) becomes

$$\frac{\lambda e_j}{\mu_j} + \sqrt{\frac{y_2}{y_1}}\, \lambda \sum_{i \neq j}^{N} \sqrt{e_i c_i} = K \quad (56)$$

Substituting equation (49) into equation (56) we obtain the following system of equations by which the optimal service rates μ_j in the j-th station can be determined:

$$K\, c_j \mu_j^2 + [\lambda w^2 - K\, C + (K-1)\lambda s - \lambda e_j c_j]\, \mu_j + \lambda e_j (C - \frac{K-1}{K}\lambda s) = 0 \quad (57)$$

Example.

We investigate the same example as in Figure 3. We minimize the response time for fixed total cost of $C = 10$ while a total throughput $\lambda = 1$ is maintained. From equations (31 and 35) we obtain the optimal service rates,

$$\mu_1^* = 1.119\ ;\ \mu_2^* = 0.198\ ;\ \mu_3^* = 0.390\ ;\ \mu_4^* = 7.893$$

Using the optimal service rate μ_4^* we then compute the minimum response time for station 4 from equation (51)

$$\bar{t}_4 = 0.125$$

Using these optimal service rates we run MVA and obtain the following throughput and mean response time results:

$$\lambda = 1.004 \qquad \bar{t}_4 = 0.125$$

Let us also analyze the case where the service rates are modified and show the effect of this modification on throughput λ and response time $\bar{t}_4$. The total cost is still $C = 10$. Here we select a faster service time for terminal (station 4).

$$\mu_1 = 1 \;\; ; \;\; \mu_2 = 0.5 \;\; ; \;\; \mu_3 = 0.5 \;\; ; \;\; \mu_4 = 8$$

Using these service rates we compute by MVA

$$\lambda = 0.962 \qquad \bar{t}_4 = 0.125$$

In this case the mean response time is the same as the above case. However, the throughput value decreases below the given throughput value of $\lambda = 1$.

Now we select small values for service rates of stations 1,2 and 3 as

$$\mu_1 = 1.2 \;\; ; \;\; \mu_2 = 0.5 \;\; ; \;\; \mu_3 = 0.5 \;\; ; \;\; \mu_4 = 7.8$$

In this case we obtain with MVA

$$\lambda = 1.081 \qquad \bar{t}_4 = 0.128$$

Let us select the following values for service rates which deviate more from the optimal values

$$\mu_1 = 1.5 \;\; ; \;\; \mu_2 = 0.8 \;\; ; \;\; \mu_3 = 0.7 \;\; ; \;\; \mu_4 = 7$$

Using these service rates MVA provides

$$\lambda = 1.451 \qquad \bar{t}_4 = 0.142$$

As can easily be seen the total throughput is increasing for smaller values of service rates in stations 1, 2 and 3. Consequently, the mean response time $\bar{t}_4$ increases.

References

1. I. F. Akyildiz and G. Bolch, "Optimization of Performance Measures in Queueing Network Models of Computer Systems", *Technical Report, ICS-GIT-87-039, November 1987. Presented at "Analysis and Control of Large Scale Stochastic Systems" Conference in North Carolina in May 1988.*
2. Y. Bard, "Some Extension to Multiclass Queueing Network Analysis", *Proc. of Performance 79,* Feb. 1979, Vienna, Vol. 1.
3. F. Baskett, K. M. Chandy, R. R. Muntz and F. G. Palacios, "Open, Closed and Mixed Queues of Networks with Different Types of Customers", *Journal of the ACM,* Vol. 22, No. 2, April 1975, pp. 248-260.
4. G. Bolch, G. Fleischmann and R. Schreppel, "Ein Funktionales Konzept zur Analyse von Warteschlangennetzen und Optimierung von Leistungsgroessen", *Proc. of the GI/NTG Int. Conference, Springer Verlag, Vol. 154,* Oct. 1987, pp. 327-342.
5. K. M. Chandy, J. Hogarth and C. H. Sauer, "Selecting Capacities in Computer Communication Systems", *IEEE Transactions on Software Engineering,* Vol. 4, July 1977, pp. 290-295.

6. D. Ferrari, "Computer Systems Performance Evaluation", *Prentice Hall, NJ, 1978.*

7. R. Geist and K. Trivedi, "Optimal Design of Multilevel Storage Hierarchies", *IEEE Transactions on Computers,* Vol. C-31, No. 3, March 1982. pp. 249-260.

8. H. Heiss and G. Totzauer, "Optimizing Utilization under Response Time Constraints", *Computing Journal,* Springer Verlag, Vol. 35, pp. 1-12, 1985.

9. J. R. Kenevan and A. K. von Mayrhauser, "Convexity and Concavity Properties of Analytic Queueing Models for Computer Systems", *Performance'84,* North-Holland, 1984, pp. 361-375.

10. L. Kleinrock, "Analytic and Simulation Methods in Computer Network Design", *AFIPS Conf. Proc.* Vol. 36, June 1969, pp. 569-579.

11. L. Kleinrock, "Queueing Systems", "*Vol. II: Computer Applications*", *John Wiley and Sons, 1976.*

12. H. Kobayashi, "Application of Diffusion Approximation to Queueing Networks I: Equilibrium Queue Distributions", *Journal of the ACM,* Vol. 21, No. 2, April 1974, pp. 316-328

13. H. Kobayashi and M. Gerla, "Optimal Routing in Closed Queueing Networks", *ACM Transactions on Computer Systems,* Vol. 1, No. 4, 1983, pp. 294-310.

14. A. Lazar, "Optimal Flow Control of an M/M/m Queue", *Journal of the ACM,* Vol. 31, No. 1, January 1984, pp. 86-89.

15. T. G. Price, "Probability Models of Multiprogrammed Computer Systems" Ph.D. Diss., Department of EE, Stanford University, December 1974.

16. M. Reiser and S. Lavenberg, "Mean Value Analysis of Closed Multichain Queueing Networks", *Journal of the ACM,* April 1980, pp. 313-322.

17. R. Schreppel, "Ermittlung Optimaler Leistungsgroessen mit Hilfe Analytischer Warteschlangenmodelle", *Diplomarbeit at IMMDIV of the University of Erlangen-Nuernberg, 1986.*

18. P. J. Schweitzer, "Approximate Analysis of Multiclass Closed Networks of Queues", *Int. Conference Stochastic Control and Optimization,* Amsterdam, 1979.

19. K. E. Stecke, "On the Nonconcavity of Throughput in Certain Closed Queueing Networks", *Performance Evaluation,* Vol. 6, North-Holland, 1986, pp. 293-305.

20. K. S. Trivedi and R. Kinicki, "A Model for Computer Configuration Design", *IEEE Computer,* April 1980, pp. 47-54.

21. K. S. Trivedi and R. A. Wagner, "A Decision Model for Closed Queueing Networks", *IEEE Transactions on Software Engineering,* Vol. SE-5, No. 4, July 1979.

22. K. S. Trivedi, R. Wagner and T. Sigmon, "Optimal Selection of CPU Speed, Device Capacities and File Assignments", *Journal of the ACM,* Vol. 27, No. 3, July 1980, pp. 457-473.

23. K. S. Trivedi and T. M. Sigmon, "Optimal Design of Linear Storage Hierarchies", *Journal of the ACM,* Vol. 28, No. 2, 1981, pp. 270-288.

24. K. S. Trivedi, "Probability and Statistics with Reliability, Queueing and Computer Science Applications" , *Prentice Hall, 1982.*

25. A. K. von Mayrhauser and K. S. Trivedi, "Computer Configuration Design to Minimize Response Time", *Computer Performance journal,* Butterworth Co., Vol. 3, No. 1, March 1982, pp. 32-39.