ROUTING GAMES

by

A. A. Economides and J. A. Silvester

Technical Report CENG 89-38

September 1989

Electrical Engineering - Systems Department
University of Southern California
Los Angeles, CA 90089-0781

ROUTING GAMES

by

Anastasios A. Economides *and* John A. Silvester

Computer Engineering Division
Electrical Engineering - Systems Department
University of Southern California, SAL 300
Los Angeles, CA 90089-0781

ABSTRACT

Previous work on multi-objective routing takes a socialistic approach to minimize some global objective function. In this paper, we take a different approach using a game theoretic formulation. We focus on a simple example of two classes which minimize a delay objective. We present three cases. The first case (baseline) does global optimization where the routing policies for the two classes are forced to be equal. The second case is where the two classes cooperate to minimize the same objective function of global average delay. In general, this Pareto formulation will have a multiplicity of solutions which allow us to use secondary objectives to select the operating point. The third case is where each class optimizes its own objective function (which may or may not be identical)- this corresponds to the classical non-cooperative Nash game. This allows different objectives to be adopted by the different classes.

1. INTRODUCTION

The usual approach to distributed system design and control is the optimization of a single function, which may be the combination of multiple objectives as seen by the system administrator [2]. Thus, it is assumed that all customers in the system cooperate for the socially optimum, such as optimizing the average customer performance.

However, in a real distributed environment there is a diversity of customer classes, each one with possibly different objectives. These different classes of customers compete for the limited common resources of the distributed system in order to optimize their own objectives, ignoring the inconvenience that they cause to the other customer classes. For example, different telecommunication companies may share the same communication links and one of them may want to maximize the throughput of its customers, another may want to minimize its average customer delay and a third may want to minimize the blocking probability of its customers. Another example is when different users share a multiprocessor system and one group of users wants to maximize its throughput, similarly another group of users wants to maximize its own throughput, another group of users wants to minimize its average response time and finally another group of users wants to minimize the variance of its response time.

Customers of a given class arrive to the distributed system requiring transfer to a destination node. The problem of deciding through which path each customer will be routed is the routing problem (Fig. 1). Kobayashi & Gerla [8] consider the multiple class routing problem in closed queueing networks. Each closed chain corresponds to a different class of customers. However, the average delay is not convex, for closed chains routing, and therefore local minima can be found. de Souza e Silva & Gerla [3] similarly consider the load balancing problem in a product form queueing network with fixed closed chain routing. They minimize a measure of the average delay with respect to the open chains flows.

In this paper, for simplicity of presentation, we consider two classes of customers

which select between two links joining the entry point and the destination (the more general case is described in [5]). We formulate and solve the routing problem both as a team optimization problem and as a Nash non cooperative game [1] among the two competing classes of customers, where each class of customers tries to operate in the most beneficial way for its own customers.

In section 2, we define the notation for a simple two link network that is shared among two customer classes. In section 3, we study the routing problem when the two classes of customers are considered as one, using the traditional optimization approach. In section 4, we study the same problem when the two classes of customers are treated separately, however both want to minimize the average customer delay. In section 5, we introduce the alternative formulation and solution of the problem using the Nash game approach, where the two classes compete for sharing the two servers and each class wants to minimize the average delay of its own customers. Finally, in section 6, we conclude on this new approach for performance evaluation and optimization of distributed computing systems.

2. QUEUEING MODEL

In this section, we introduce a simple queueing model of two servers that are shared by customers of two classes. The problem is to assign these customers to the two servers (links) so as to minimize a delay objective. An application is routing, where two classes of packets may use two different links for transmission between the source and destination. Another application is load sharing for a multiprocessor system, where two classes of jobs may use two processors for execution.

Let class α customers arrive to the system with rate λ^α (Poisson arrivals) and class β customers arrive to the system with rate λ^β (Poisson arrivals). So, the total arrival rate is $\lambda = \lambda^\alpha + \lambda^\beta$. Customers of both classes may be served at either of the two servers, which have service rates C_1 and C_2. So, the total system capacity is $C = C_1 + C_2$. Let that the service requirement of each job be exponentially distributed

with mean 1. A class α customer is assigned to server 1 with probability P_1^α and to server 2 with probability P_2^α; and a class β customer is assigned to server 1 with probability P_1^β and to server 2 with probability P_2^β. Let also the superscript * at a variable denote the optimum value of that variable.

Furthermore, for stability reasons it is assumed that the total arrival rate is less than the total service rate :

$\lambda^\alpha + \lambda^\beta \leq C_1 + C_2$ *or* $\lambda \leq C$

In the following sections, we consider three formulations and solutions for sharing the two servers among customers of the two classes.

3. TRAFFIC AGGREGATION

In this section, we find the optimal routing policy, when the two classes are aggregated into a single class. Therefore, a class α customer is assigned to a server with the same probability as a class β customer, i.e. $P_1^\alpha = P_1^\beta = P_1$ and $P_2^\alpha = P_2^\beta = P_2$. If both classes want to minimize the average customer delay in the system, then we have the delay objective [7]:

$$J(P_1, P_2) = \frac{P_1}{C_1 - (\lambda^\alpha + \lambda^\beta) * P_1} + \frac{P_2}{C_2 - (\lambda^\alpha + \lambda^\beta) * P_2}$$

and we have the following optimization problem :

minimize $J(P_1, P_2)$

with respect to $P_1, \ P_2$

so that $P_1 + P_2 = 1, \ P_1 \geq 0, \ P_2 \geq 0.$

This is a simple problem and can easily be solved [2, 4]. Let first define the auxiliary variable

$$K_1 = \frac{C_1 + C_2 - \lambda^\alpha - \lambda^\beta}{\lambda^\alpha + \lambda^\beta} * \frac{\sqrt{C_1}}{\sqrt{C_1} + \sqrt{C_2}}$$

Then, the following policy optimally assigns the arriving customers to the two servers:

$$If \quad C_1 - \sqrt{C_1 C_2} \leq \lambda^\alpha + \lambda^\beta \quad and \quad C_2 - \sqrt{C_1 C_2} \leq \lambda^\alpha + \lambda^\beta \leq C_1 + C_2$$

$$then \quad P_1^* = \frac{C_1}{\lambda^\alpha + \lambda^\beta} - K_1$$

$$If \quad 0 \leq \lambda^\alpha + \lambda^\beta \leq C_1 - \sqrt{C_1 C_2}$$

$$then \quad P_1^* = 1$$

$$If \quad 0 \leq \lambda^\alpha + \lambda^\beta \leq C_2 - \sqrt{C_1 C_2}$$

$$then \quad P_1^* = 0$$

Of course, the optimum routing probability to the other server is $P_2^* = 1 - P_1^*$.

In Fig. 2, we show the optimum routing probability to server 1, P_1^*, versus the system utilization, λ/C, for fixed server 2 capacity, $C_2 = 1$, and different server 1 capacities, $C_1 = 1$, 2, 3, 5, 7 and 10. When the two servers have equal capacity, $C_1 = C_2 = 1$, then the flow is split half to each server ($P^* = 0.5$). As we increase the server 1 capacity, then this server tends to be exclusively used ($P_1^* = 1$) for a larger range of system utilization.

4. TEAM OPTIMIZATION

In this section, we find the optimum routing decisions, when each class is treated independently from the other. A class α customer is assigned to a server with possibly different probability than a class β customer. However, both classes minimize the same objective - the average customer delay. This problem can be considered as a cooperative team game [1] between the two classes, where each class solves the following problem :

minimize

$$J(P_1^\alpha, P_2^\alpha, P_1^\beta, P_2^\beta) = \frac{\lambda^\alpha * P_1^\alpha + \lambda^\beta * P_1^\beta}{\lambda^\alpha + \lambda^\beta} * \frac{1}{C_1 - \lambda^\alpha * P_1^\alpha - \lambda^\beta * P_1^\beta} +$$

$$+ \frac{\lambda^\alpha * P_2^\alpha + \lambda^\beta * P_2^\beta}{\lambda^\alpha + \lambda^\beta} * \frac{1}{C_2 - \lambda^\alpha * P_2^\alpha - \lambda^\beta * P_2^\beta}$$

with respect to $P_1^\alpha,\ P_2^\alpha,\ P_1^\beta,\ P_2^\beta$

so that $P_1^\alpha + P_2^\alpha = 1,\ \ P_1^\beta + P_2^\beta = 1,\ \ P_1^\alpha,\ P_2^\alpha,\ P_1^\beta,\ P_2^\beta \geq 0.$

The objective function $J(P_1^\alpha, P_2^\alpha, P_1^\beta, P_2^\beta)$ is convex since it is a sum of convex functions.

Let define the auxiliary variables

$$K_1^\alpha = \frac{C_1 + C_2 - \lambda^\alpha - \lambda^\beta}{\lambda^\alpha} * \frac{\sqrt{C_1}}{\sqrt{C_1} + \sqrt{C_2}}$$

$$K_1^\beta = \frac{C_1 + C_2 - \lambda^\alpha - \lambda^\beta}{\lambda^\beta} * \frac{\sqrt{C_1}}{\sqrt{C_1} + \sqrt{C_2}}$$

Then, the following policy [6] will optimally assign the arriving customers to the two servers:

Team routing :

If $\lambda^\alpha + \lambda^\beta \leq C_1 + C_2$,

then $$P_1^{\alpha*} = \frac{C_1 - \lambda^\beta P_1^{\beta*}}{\lambda^\alpha} - K_1^\alpha$$

$$P_1^{\beta*} = \frac{C_1 - \lambda^\alpha P_1^{\alpha*}}{\lambda^\beta} - K_1^\beta$$

accept the solution only if

$$C_1 - \lambda^\beta P_1^{\beta*} - \sqrt{\frac{C_1}{C_2}}(C_2 - \lambda^\beta P_2^{\beta*}) \leq \lambda^\alpha$$

$$C_2 - \lambda^\beta P_2^{\beta*} - \sqrt{\frac{C_2}{C_1}}(C_1 - \lambda^\beta P_1^{\beta*}) \leq \lambda^\alpha$$

$$C_1 - \lambda^\alpha P_1^{\alpha*} - \sqrt{\frac{C_1}{C_2}}(C_2 - \lambda^\alpha P_2^{\alpha*}) \leq \lambda^\beta$$

$$C_2 - \lambda^\alpha P_2^{\alpha*} - \sqrt{\frac{C_2}{C_1}}(C_1 - \lambda^\alpha P_1^{\alpha*}) \leq \lambda^\beta$$

If $\lambda^\alpha + \lambda^\beta \leq C_1 - \sqrt{C_1 C_2}$,

then $P_1^{\alpha*} = 1, \quad P_1^{\beta*} = 1$

If $\lambda^\alpha \sqrt{C_2} - \lambda^\beta \sqrt{C_1} = \sqrt{C_1 C_2}(\sqrt{C_1} - \sqrt{C_2})$,

then $P_1^{\alpha*} = 1, \quad P_1^{\beta*} = 0$

$If \quad \lambda^{\alpha}\sqrt{C_1} - \lambda^{\beta}\sqrt{C_2} = \sqrt{C_1 C_2}(\sqrt{C_2} - \sqrt{C_1}),$

$then \quad P_1^{\alpha *} = 0, \quad P_1^{\beta *} = 1$

$If \quad \lambda^{\alpha} + \lambda^{\beta} \leq C_2 - \sqrt{C_1 C_2},$

$then \quad P_1^{\alpha *} = 0, \quad P_1^{\beta *} = 0$

$If \quad \lambda^{\alpha} + \lambda^{\beta} \geq C_1 - \sqrt{C_1 C_2} \quad and \quad \lambda^{\alpha}\sqrt{C_2} - \lambda^{\beta}\sqrt{C_1} \leq \sqrt{C_1 C_2}(\sqrt{C_1} - \sqrt{C_2}),$

$then \quad P_1^{\alpha *} = 1$

$$P_1^{\beta *} = \frac{C_1 - \lambda^{\alpha}}{\lambda^{\beta}} - K_1^{\beta}$$

$If \quad \lambda^{\alpha} + \lambda^{\beta} \geq C_2 - \sqrt{C_1 C_2} \quad and \quad \lambda^{\alpha}\sqrt{C_1} - \lambda^{\beta}\sqrt{C_2} \leq \sqrt{C_1 C_2}(\sqrt{C_2} - \sqrt{C_1}),$

$then \quad P_1^{\alpha *} = 0$

$$P_1^{\beta *} = \frac{C_1}{\lambda^{\beta}} - K_1^{\beta}$$

$If \quad \lambda^{\alpha} + \lambda^{\beta} \geq C_1 - \sqrt{C_1 C_2} \quad and \quad \lambda^{\beta}\sqrt{C_2} - \lambda^{\alpha}\sqrt{C_1} \leq \sqrt{C_1 C_2}(\sqrt{C_1} - \sqrt{C_2}),$

$then \quad P_1^{\beta *} = 1$

$$P_1^{\alpha *} = \frac{C_1 - \lambda^{\beta}}{\lambda^{\alpha}} - K_1^{\alpha}$$

$$If \quad \lambda^\alpha + \lambda^\beta \geq C_2 - \sqrt{C_1 C_2} \quad and \quad \lambda^\beta \sqrt{C_1} - \lambda^\alpha \sqrt{C_2} \leq \sqrt{C_1 C_2}(\sqrt{C_2} - \sqrt{C_1}),$$

$$then \quad P_1^{\beta *} = 0$$

$$P_1^{\alpha *} = \frac{C_1}{\lambda^\alpha} - K_1^\alpha$$

Of course, the optimum routing probabilities to the other server are $P_2^{\alpha *} = 1 - P_1^{\alpha *}$ and $P_2^{\beta *} = 1 - P_1^{\beta *}$.

In order to find the optimum routing probabilities $(P_1^{\alpha *}, P_1^{\beta *})$ for the first case of the team routing policy, we give all possible values to $P_1^\beta \in [0, 1]$ and calculate the corresponding values for the P_1^α

$$P_1^\alpha = \frac{C_1 - \lambda^\beta P_1^\beta}{\lambda^\alpha} - N_1^\alpha(P_1^\beta)$$

Then we check if the conditions for the resulting routing probabilities $P_1^\alpha, P_2^\alpha, P_1^\beta, P_2^\beta$ are satisfied and if yes the we accept them as optimum routing probabilities.

In Fig. 3, we show the optimum routing probabilities $(P_1^{\alpha *}, P_1^{\beta *})$ for a simple homogeneous case $C_1 = C_2 = 1$ and $\lambda^\alpha = \lambda^\beta = 0.1, ..., 0.9$. We note that the solution pairs form a straight line which is general the case as stated in the Proposition 1. Thus we have a multiplicity of optimum routing probabilities and we can choose any pair of them with some other criterion. For example if we want $P_1^\alpha = P_1^\beta$, then the solution set reduces to a single point that is also the solution of section 3, where we treat the two classes as one.

Proposition 1 : The set of the optimum routing probabilities $(P_1^{\alpha *}, P_1^{\beta *})$ for fixed arrival rate λ^α and λ^β and fixed server capacities C_1 and C_2 forms a straight line.

Proof :

The general equation that gives the optimum routing probabilities is

$$P_1^{\alpha *} = \frac{C_1 - \lambda^\beta P_1^{\beta *}}{\lambda^\alpha} - K_1^\alpha$$

Obviously, this equation describes a straight line. □

In Fig. 4, we show the optimum routing probabilities $(P_1^{\alpha*}, P_1^{\beta*})$ for fixed server capacities, $C_1 = 2$, $C_2 = 1$, fixed class β arrival rate, $\lambda^\beta = 1$, and different class α arrival rates, $\lambda^\alpha = 0.1, ..., 1.9$. We notice something remarkable. The straight line solutions for different class α arrival rates intersect at a single intersection point. This means that there is a common pair of optimum routing probabilities $(P_1^{\alpha*}, P_1^{\beta*})$, where we can optimally operate for different class α arrival rates. So, we can use the optimum routing probabilities of the intersection point and operate optimally even if class α arrival rate varies. The next Proposition 2 describes this result more formally.

Proposition 2 : For a given system C_1, C_2, with fixed class β arrival rate λ^β,

If

$$0 \leq C_1 - (C_1 + C_2 - \lambda^\beta) * \frac{\sqrt{C_1}}{\sqrt{C_1} + \sqrt{C_2}} \leq \lambda^\beta$$

then the straight lines of the team optimum probabilities $(P_1^{\alpha*}, P_1^{\beta*})$, for different class α arrival rates λ^α ($\lambda^\alpha + \lambda^\beta \leq C_1 + C_2$), intersect at a single point

$$(P_1^{\alpha*}, P_1^{\beta*}) = (\frac{\sqrt{C_1}}{\sqrt{C_1} + \sqrt{C_2}}, \frac{C_1}{\lambda^\beta} - \frac{C_1 + C_2 - \lambda^\beta}{\lambda^\beta} * \frac{\sqrt{C_1}}{\sqrt{C_1} + \sqrt{C_2}}$$

i.e. this intersection point is independent of the class α arrival rate.

Proof :

Let the straight line of the optimum routing probabilities for a given class α arrival rate λ_1^α be

$$P_1^{\alpha*} = \frac{C_1 - \lambda^\beta P_1^{\beta*}}{\lambda_1^\alpha} - \frac{C_1 + C_2 - \lambda_1^\alpha - \lambda^\beta}{\lambda_1^\alpha} * \frac{\sqrt{C_1}}{\sqrt{C_1} + \sqrt{C_2}}$$

Let also the straight line of the optimum routing probabilities for another given class α arrival rate λ_2^α be

$$P_1^{\alpha**} = \frac{C_1 - \lambda^\beta P_1^{\beta**}}{\lambda_2^\alpha} - \frac{C_1 + C_2 - \lambda_2^\alpha - \lambda^\beta}{\lambda_2^\alpha} * \frac{\sqrt{C_1}}{\sqrt{C_1} + \sqrt{C_2}}$$

Since the slope of each line depends on the class α arrival rate, these lines will intersect each other. Now in order to prove that all lines intersect at a single point,

it is enough to prove that the intersection point is independent of the class α arrival rate λ^α.

Let $(P_1^{\alpha\#}, P_1^{\beta\#})$ be the intersection point of these two lines, i.e. $P_1^{\alpha*} = P_1^{\alpha**} = P_1^{\alpha\#}$ and $P_1^{\beta*} = P_1^{\beta**} = P_1^{\beta\#}$. Then

$$\frac{C_1 - \lambda^\beta P_1^{\beta\#}}{\lambda_1^\alpha} - \frac{C_1 + C_2 - \lambda_1^\alpha - \lambda^\beta}{\lambda_1^\alpha} * \frac{\sqrt{C_1}}{\sqrt{C_1} + \sqrt{C_2}} = \frac{C_1 - \lambda^\beta P_1^{\beta\#}}{\lambda_2^\alpha} - \frac{C_1 + C_2 - \lambda_2^\alpha - \lambda^\beta}{\lambda_2^\alpha} * \frac{\sqrt{C_1}}{\sqrt{C_1} + \sqrt{C_2}}$$

and finally

$$P_1^{\beta\#} = \frac{C_1}{\lambda^\beta} - \frac{C_1 + C_2 - \lambda^\beta}{\lambda^\beta} * \frac{\sqrt{C_1}}{\sqrt{C_1} + \sqrt{C_2}}$$

$$P_1^{\alpha\#} = \frac{\sqrt{C_1}}{\sqrt{C_1} + \sqrt{C_2}}$$

Thus the intersection point $(P_1^{\alpha\#}, P_1^{\beta\#})$ is independent from the class α arrival rate λ^α. In order for the intersection point be also a solution, it must be in the range $0 \leq P_1^{\beta\#} \leq 1$. Thus the result. □

In Fig. 5, we show the optimum routing probabilities $(P_1^{\alpha*}, P_1^{\beta*})$ for fixed arrival rates $\lambda^\alpha = 2$, $\lambda^\beta = 1$, fixed server 2 capacity $C_2 = 1$ and different server 1 capacities $C_1 = 2.1, ..., 3.8$. We see that the solution lines are parallel.

Proposition 3 : The straight lines of the optimum routing probabilities for fixed arrival rates $\lambda^\alpha, \lambda^\beta$, fixed server 2 capacity C_2 and different server 1 capacities C_1 are parallel.

Proof :

The optimum routing probabilities are described by the folowing straight line

$$P_1^{\alpha*} = \frac{C_1 - \lambda^\beta P_1^*}{\lambda^\alpha} - \frac{C_1 + C_2 - \lambda^\alpha - \lambda^\beta}{\lambda^\alpha} * \frac{\sqrt{C_1}}{\sqrt{C_1} + \sqrt{C_2}}$$

that has slope independent of the capacity of the servers. □

As we have seen we have a set of optimum routing probability pairs $(P^{\alpha}{}_{*1}, P_1^{\beta*})$ that all achieve the same global minimum delay. However, these optimum routing

probabilities will give different average delay for each class. So, we can choose the operating point using another delay objective. In Fig. 6, we show the difference in the average delay of class α and class β customers, $J^{\alpha*} - J^{\beta*}$, versus the class α optimum routing probability, $P_1^{\alpha*}$, for fixed server capacities, $C_1 = 2, \quad C_2 = 1$, fixed class β arrival rate, $\lambda^\beta = 1$, and different class α arrival rates, $\lambda^\alpha = 0.1, ..., 1.9$. An example is when it is desired that both classes have the same average delay. Then this point will be the intersection of the delay difference line and the zero delay difference line. The operating point for this case is the same as the solution of section 3, where we aggregate the two classes into a single class and therefore we treat them similarly. Another example is when a secondary objective is that class α should have better treatment than class β. Then the lowest point of the delay difference line $J^{\alpha*} - J^{\beta*}$ is chosen.

5. NASH EQUILIBRIUM

In this section, we find the optimum routing decisions, when each class chooses the best strategy for its customers given the decision of the other class. Class α assigns its customers to the two servers such that the average delay of its customers is minimized, given that class β assigns optimally its customers. Similarly, class β assigns its customers to the two servers such that the average delay of its customers is minimized, given that class α assign optimally its customers. Therefore customers of different classes do not have the same objective and they compete for sharing the two servers. We formulate and solve the above multiobjective optimization problem as a non cooperative Nash game [1] between the two classes. After reaching a Nash equilibrium, no class of customers will have a rational motive to unilaterally deviate from its equilibrium strategy.

Class α solves the following problem :

minimize

$$J^\alpha(P_1^\alpha, P_2^\alpha, P_1^{\beta*}, P_2^{\beta*}) = \frac{P_1^\alpha}{C_1 - \lambda^\alpha * P_1^\alpha - \lambda^\beta * P_1^{\beta*}} + \frac{P_2^\alpha}{C_2 - \lambda^\alpha * P_2^\alpha - \lambda^\beta * P_2^{\beta*}}$$

with respect to $P_1^\alpha,\ P_2^\alpha$

so that $P_1^\alpha + P_2^\alpha = 1,\quad P_1^\alpha,\ P_2^\alpha \geq 0.$

The objective function $J^\alpha(P_1^\alpha, P_2^\alpha, P_1^{\beta*}, P_2^{\beta*})$ is convex over the convex space $P_1^\alpha + P_2^\alpha = 1, P_1^\alpha, P_2^\alpha \geq 0$, since it is a sum of convex functions.

On the other hand, class β will solve the following problem :

minimize

$$J^\beta(P_1^{\alpha*}, P_2^{\alpha*}, P_1^\beta, P_2^\beta) = \frac{P_1^\beta}{C_1 - \lambda^\alpha * P_1^{\alpha*} - \lambda^\beta * P_1^\beta} + \frac{P_2^\beta}{C_2 - \lambda^\alpha * P_2^{\alpha*} - \lambda^\beta * P_2^\beta}$$

with respect to $P_1^\beta,\ P_2^\beta$

so that $P_1^\beta + P_2^\beta = 1,\quad P_1^\beta,\ P_2^\beta \geq 0.$

The objective function $J^\beta(P_1^{\alpha*}, P_2^{\alpha*}, P_1^\beta, P_2^\beta)$ is convex over the convex space $P_1^\beta + P_2^\beta = 1, P_1^\beta, P_2^\beta \geq 0$, since it is a sum of convex functions.

When the players are in a Nash equilibrium, no player can improve his cost by altering his decision unilaterally. Next, we give the definition of a Nash equilibrium [1] in our context:

Definition : A vector $[P_1^\alpha, P_2^\alpha, P_1^\beta, P_2^\beta]$ with $P_1^\alpha + P_2^\alpha = 1,\quad P_1^\beta + P_2^\beta = 1$, and $P_1^\alpha,\ P_2^\alpha,\ P_1^\beta,\ P_2^\beta \geq 0.$ is called Nash equilibrium for a two player nonzero-sum infinite game if

$$J^\alpha(P_1^{\alpha*}, P_2^{\alpha*}, P_1^{\beta*}, P_2^{\beta*}) \leq \inf_{\substack{P_1^\alpha + P_2^\alpha = 1,\\ P_1^\alpha, P_2^\alpha \geq 0}} J^\alpha(P_1^\alpha, P_2^\alpha, P_1^{\beta*}, P_2^{\beta*})$$

$$J^\beta(P_1^{\alpha*}, P_2^{\alpha*}, P_1^{\beta*}, P_2^{\beta*}) \leq \inf_{\substack{P_1^\beta + P_2^\beta = 1,\\ P_1^\beta, P_2^\beta \geq 0}} J^\beta(P_1^{\alpha*}, P_2^{\alpha*}, P_1^\beta, P_2^\beta)$$

Therefore each class will minimize its average customer delay given that the other class has minimized the average delay of its customers.

Let define the auxiliary variables

$$N_1^{\alpha*}(P_1^{\beta*}) = \frac{C_1 + C_2 - \lambda^\alpha - \lambda^\beta}{\lambda^\alpha} * \frac{\sqrt{C_1 - \lambda^\beta * P_1^{\beta*}}}{\sqrt{C_1 - \lambda^\beta * P_1^{\beta*}} + \sqrt{C_2 - \lambda^\beta * P_2^{\beta*}}}$$

$$N_1^{\beta*}(P_1^{\alpha*}) = \frac{C_1 + C_2 - \lambda^\alpha - \lambda^\beta}{\lambda^\beta} * \frac{\sqrt{C_1 - \lambda^\alpha * P_1^{\alpha*}}}{\sqrt{C_1 - \lambda^\alpha * P_1^{\alpha*}} + \sqrt{C_2 - \lambda^\alpha * P_2^{\alpha*}}}$$

Then, the following policy [6] will route the arriving customers to the two servers such that a Nash equilibrium is achieved:

Nash routing :

If $\lambda^\alpha + \lambda^\beta \leq C_1 + C_2$,

then $$P_1^{\alpha*} = \frac{C_1 - \lambda^\beta P_1^{\beta*}}{\lambda^\alpha} - N_1^\alpha(P_1^{\beta*})$$

$$P_1^{\beta*} = \frac{C_1 - \lambda^\alpha P_1^{\alpha*}}{\lambda^\beta} - N_1^\beta(P_1^{\alpha*})$$

accept the solution only if

$$C_1 - \lambda^\beta P_1^{\beta*} - \sqrt{(C_1 - \lambda^\beta P_1^{\beta*})(C_2 - \lambda^\beta P_2^{\beta*})} \leq \lambda^\alpha$$

$$C_2 - \lambda^\beta P_2^{\beta} - \sqrt{(C_1 - \lambda^\beta P_1^{\beta*})(C_2 - \lambda^\beta P_2^{\beta*})} \leq \lambda^\alpha$$

$$C_1 - \lambda^\alpha P_1^{\alpha*} - \sqrt{(C_1 - \lambda^\alpha P_1^{\alpha*})(C_2 - \lambda^\alpha P_2^{\alpha*})} \leq \lambda^\beta$$

$$C_2 - \lambda^\alpha P_2^{\alpha*} - \sqrt{(C_1 - \lambda^\alpha P_1^{\alpha*})(C_2 - \lambda^\alpha P_2^{\alpha*})} \leq \lambda^\beta$$

If $\lambda^\alpha + \lambda^\beta \leq C_1 - \sqrt{(C_1 - \lambda^\alpha)C_2}$ *and* $\lambda^\alpha + \lambda^\beta \leq C_1 - \sqrt{(C_1 - \lambda^\beta)C_2}$,

then $P_1^{\alpha*} = 1, \quad P_1^{\beta*} = 1$

If $0 \leq \lambda^{\alpha} \leq C_1 - \sqrt{(C_1(C_2 - \lambda^{\beta})}$ *and* $0 \leq \lambda^{\beta} \leq C_2 - \sqrt{(C_1 - \lambda^{\alpha})C_2}$,

then $P_1^{\alpha *} = 1, \quad P_1^{\beta *} = 0$

If $0 \leq \lambda^{\alpha} \leq C_2 - \sqrt{(C_1 - \lambda^{\beta})C_2}$ *and* $0 \leq \lambda^{\beta} \leq C_1 - \sqrt{C_1(C_2 - \lambda^{\alpha})}$,

then $P_1^{\alpha *} = 0, \quad P_1^{\beta *} = 1$

If $\lambda^{\alpha} + \lambda^{\beta} \leq C_2 - \sqrt{C_1(C_2 - \lambda^{\alpha})}$ *and* $\lambda^{\alpha} + \lambda^{\beta} \leq C_2 - \sqrt{C_1(C_2 - \lambda^{\beta})}$,

then $P_1^{\alpha *} = 0, \quad P_1^{\beta *} = 0$

If $\lambda^{\alpha} + \lambda^{\beta} \geq C_1 - \sqrt{(C_1 - \lambda^{\alpha})C_2}$ *and* $\lambda^{\beta} \geq C_2 - \sqrt{(C_1 - \lambda^{\alpha})C_2}$,

then $P_1^{\alpha *} = 1$

$$P_1^{\beta *} = \frac{C_1 - \lambda^{\alpha}}{\lambda^{\beta}} - N_1^{\beta}(1)$$

accept the solution only if

$$\lambda^{\alpha} \leq C_1 - \lambda^{\beta} P_1^{\beta *} - \sqrt{(C_1 - \lambda^{\beta} P_1^{\beta *})(C_2 - \lambda^{\beta} P_2^{\beta *})}$$

If $\lambda^\alpha + \lambda^\beta \geq C_2 - \sqrt{C_1(C_2 - \lambda^\alpha)}$ *and* $\lambda^\beta \geq C_1 - \sqrt{C_1(C_2 - \lambda^\alpha)}$,

then $P_1^{\alpha*} = 0$

$$P_1^{\beta*} = \frac{C_1}{\lambda^\beta} - N_1^\beta(0)$$

accept the solution only if

$$\lambda^\alpha \leq C_2 - \lambda^\beta P_2^{\beta*} - \sqrt{(C_1 - \lambda^\beta P_1^{\beta*})(C_2 - \lambda^\beta P_2^{\beta*})}$$

If $\lambda^\alpha + \lambda^\beta \geq C_1 - \sqrt{(C_1 - \lambda^\beta)C_2}$ *and* $\lambda^\alpha \geq C_2 - \sqrt{(C_1 - \lambda^\beta)C_2}$,

then $P_1^{\beta*} = 1$

$$P_1^{\alpha*} = \frac{C_1 - \lambda^\beta}{\lambda^\alpha} - N_1^\alpha(1)$$

accept the solution only if

$$\lambda^\beta \leq C_1 - \lambda^\alpha P_1^{\alpha*} - \sqrt{(C_1 - \lambda^\alpha P_1^{\alpha*})(C_2 - \lambda^\alpha P_2^{\alpha*})}$$

If $\lambda^\alpha + \lambda^\beta \geq C_2 - \sqrt{C_1(C_2 - \lambda^\beta)}$ *and* $\lambda^\alpha \geq C_1 - \sqrt{C_1(C_2 - \lambda^\beta)}$,

then $P_1^{\beta *} = 0$

$$P_1^{\alpha *} = \frac{C_1}{\lambda^\alpha} - N_1^\alpha(0)$$

accept the solution only if

$$\lambda^\beta \leq C_2 - \lambda^\alpha P_2^{\alpha *} - \sqrt{(C_1 - \lambda^\alpha P_1^{\alpha *})(C_2 - \lambda^\alpha P_2^{\alpha *})}$$

Of course, the Nash equilibrium routing probabilities to the other server are $P_2^{\alpha *} = 1 - P_1^{\alpha *}$ and $P_2^{\beta *} = 1 - P_1^{\beta *}$.

Theorem 1 : For the above routing game, there exists a Nash equilibrium.

Proof : It is known [1] that a two player non-zero sum game, in which the action spaces $P_1^\alpha + P_2^\alpha = 1, \quad P_1^\alpha, P_2^\alpha \geq 0$ and $P_1^\beta + P_2^\beta = 1, \quad P_1^\beta, P_2^\beta \geq 0$ are compact and the cost functionals J^α, J^β are continuous on the product of the action spaces admits a Nash equilibrium in mixed strategies. □

In order to find the Nash equilibrium routing probabilities $P_1^{\alpha *}, P_1^{\beta *})$ for the first case of the Nash routing, we need an iterative algorithm to calculate them. So, starting with $P_1^{\alpha *}(0) = P_1^{\beta *}(0) = 0$, we iterate according to the following algorithm:

$$P_1^\alpha(k+1) = \frac{C_1 - \lambda^\beta P_1^\beta(k)}{\lambda^\alpha} - N_1^\alpha(P_1^\beta(k))$$

$$P_1^\beta(k+1) = \frac{C_1 - \lambda^\alpha P_1^\alpha(k)}{\lambda^\beta} - N_1^\beta(P_1^\alpha(k))$$

In Fig. 7, we show the average delay difference between the two classes $J^{\alpha *} - J^{\beta *}$ for fixed server capacities $C_1 = 2, C_2 = 1$, fixed class β arrival rate $\lambda^\beta = 1$ and different class α arrival rates λ^α. When the class α arrival rate is equal to the class β arrival rate $\lambda^\alpha = \lambda^\beta = 1$, then both classes have the same average delay. When a

class has larger arrival rate then it also has larger aveage delay. For very small class α arrival rate λ^α, we notice something peculiar: the average delay difference curve is not monotonic with the arrival rate. This happens because for these values we hit the boundary, as we see in Fig. 8.

In Fig. 8, we show the Nash equilibrium routing probabilities of the two classes $P_1^{\alpha*}$ and $P_1^{\beta*}$, for fixed server capacities, $C_1 = 2, C_2 = 1$, fixed class β arrival rate, $\lambda^\beta = 1$ and different class α arrival rates, λ^α. We see that for very small class α arrival rate λ^α, class α uses exclusively the faster server 1 ($P_1^{\alpha*} = 1$). For equal arrival rates $\lambda^\alpha = \lambda^\beta = 1$, the Nash equilibrium routing probabilities intersect at the point $P_1^{\alpha*} = P_1^{\beta*}$. As we increase the arrival rate they depart each other to meet again when the arrival rate becomes large.

For comparison, we also show in Fig. 9 the team optimum routing probabilities of the two classes, $P_1^{\alpha*}$ and $P_1^{\beta*}$, for fixed server capacities, $C_1 = 2, C_2 = 1$, fixed class β arrival rate, $\lambda^\beta = 1$ and different class α arrival rates, λ^α. In this case we have a multiplicity of solutions. The set of the class α routing probabilities $P_1^{\alpha*}$ is in red, while the set of the class β routing probabilities $P_1^{\beta*}$ is in blue. Now there is a large range where these sets intersect each other.

6. CONCLUSIONS

In this paper, we formulated and solved a simple *two class routing* problem. When the two classes of customers cooperate for minimizing the average customer delay, then we formulate and solve the problem as a *team optimization problem.* When the two classes of customers compete among themselves and each class wants to minimize the average delay of its own customers, we introduce an alternative methodology for multiobjective performance optimization. In this case, we formulate and solve the problem as a *non cooperative Nash game.* Each class of customers chooses the best strategy for its customers given the decisions of the other class. A Nash equilibrium is achieved, where no class of customers has a rational motive to unilaterally depart

from its strategy.

Straightforward extensions are to consider multiple (> 2) classes, as well as more than two servers. Also the customer service times may be generally distributed. Details are left to our report [6]. The methodology also applies to the combined load sharing, routing and congestion control problem of multiple competing classes of customers in an arbitrary distributed computing system, see [5].

In summary, we have presented a novel approach which leads itself to multi-objective optimization problems. Applications to different real scenarios remain to be investigated.

References

[1] T. Basar and G. J. Olsder. *Dynamic Noncooperative Game theory.* Academic Press, 1982.

[2] D.P. Bertsekas and R. Gallager. *Data Networks.* Prentice Hall, 1987.

[3] E. de Souza e Silva and M. Gerla. Load balancing in distributed systems with multiple classes and site constraints. *Performance '84, E. Gelenbe (ed.), pp. 17-33,* North Holland 1984.

[4] A.A. Economides and J.A. Silvester. Optimal routing in a network with unreliable links. *Proc. of IEEE Computer Networking Symposium,* pp. 288-297, IEEE 1988.

[5] A.A. Economides and J.A. Silvester. Load sharing, routing and congestion control in distributed computing system as a nash game. *Electrical Engineering - Systems Dept.,* University of Southern California, 1989.

[6] A.A. Economides and J.A. Silvester. Routing of cooperating and competing multiple classes: Team optimization and nash equilibrium. *Electrical Engineering - Systems Dept.*, University of Southern California, 1989.

[7] L. Kleinrock. *Queueing Systems, Vol. 2: Applications.* J. Wiley & Sons, 1976.

[8] H. Kobayashi and M. Gerla. Optimal routing in closed queueuing networks. *ACM Transactions on Computer Systems, Vol. 1, No. 4, pp. 294-310*, Nov. 1983.

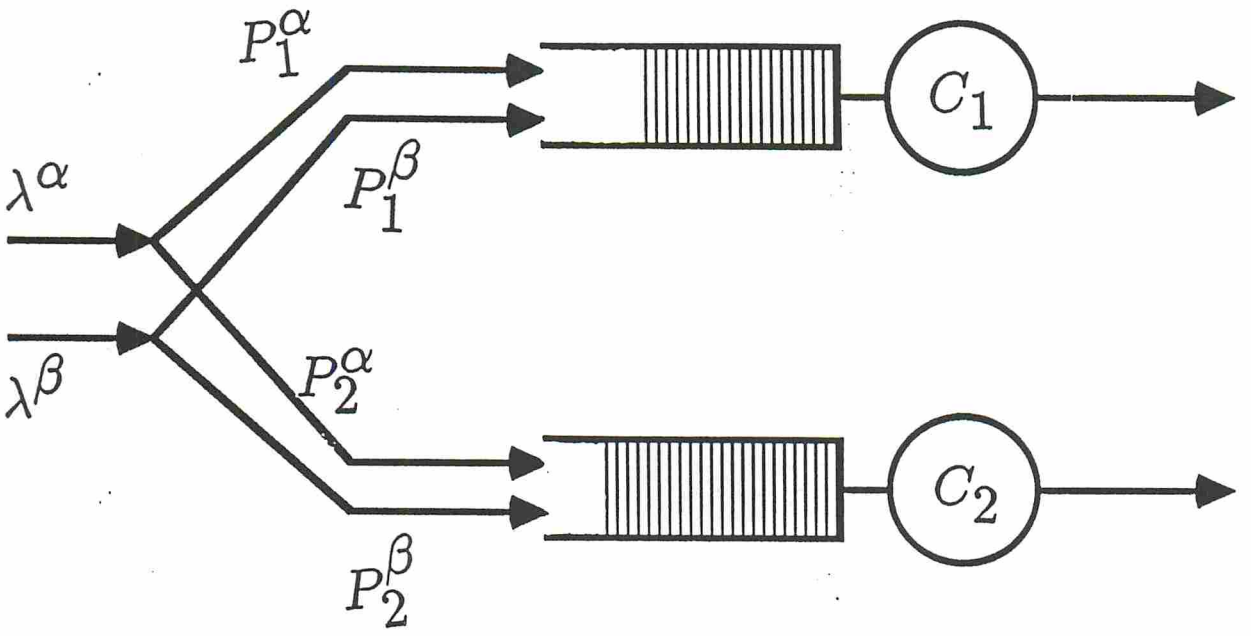

Fig. 1 A two class routing problem

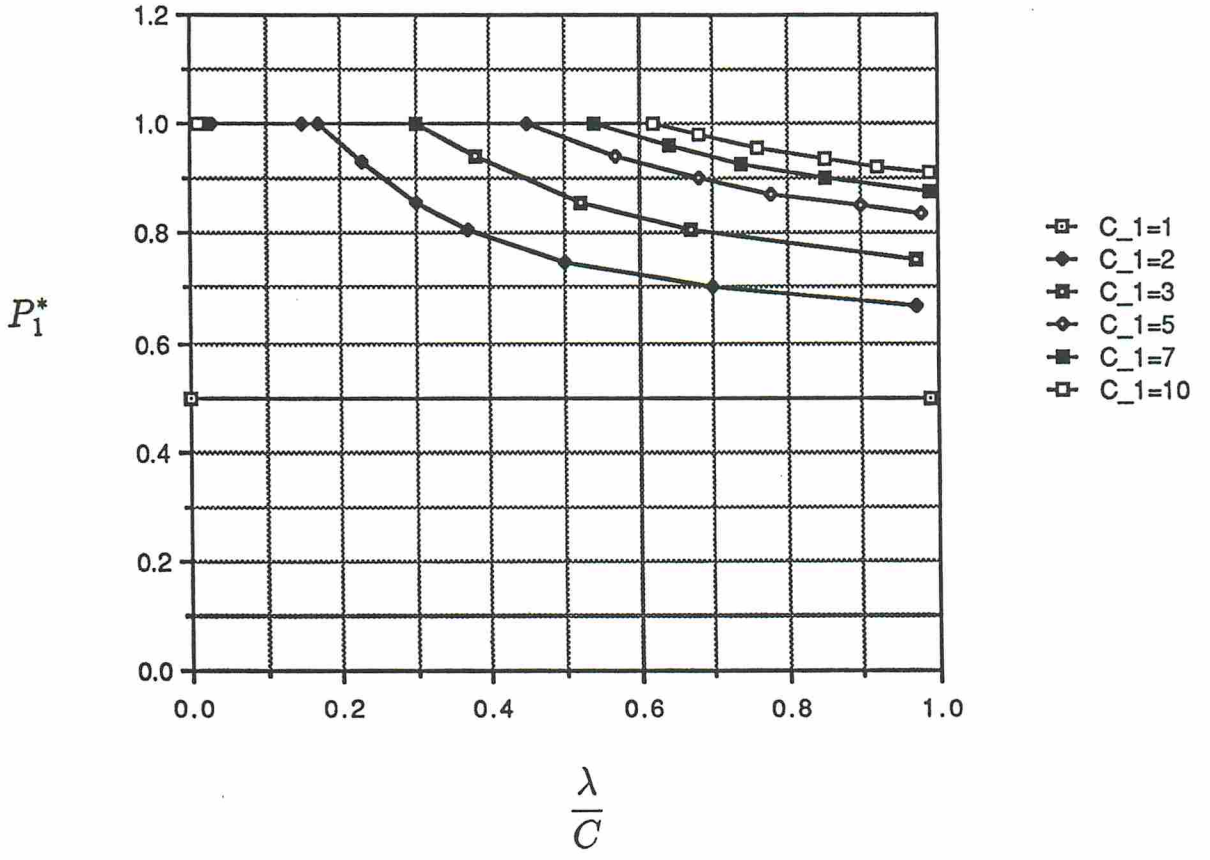

Fig. 2 The optimum routing probability to server 1, P_1^*, versus the system utilization $\frac{\lambda}{C}$, for fixed server 2 capacity $C_2 = 1$ and different server 1 capacities $C_1 = 1, 2, 3, 5, 7$ and 10.

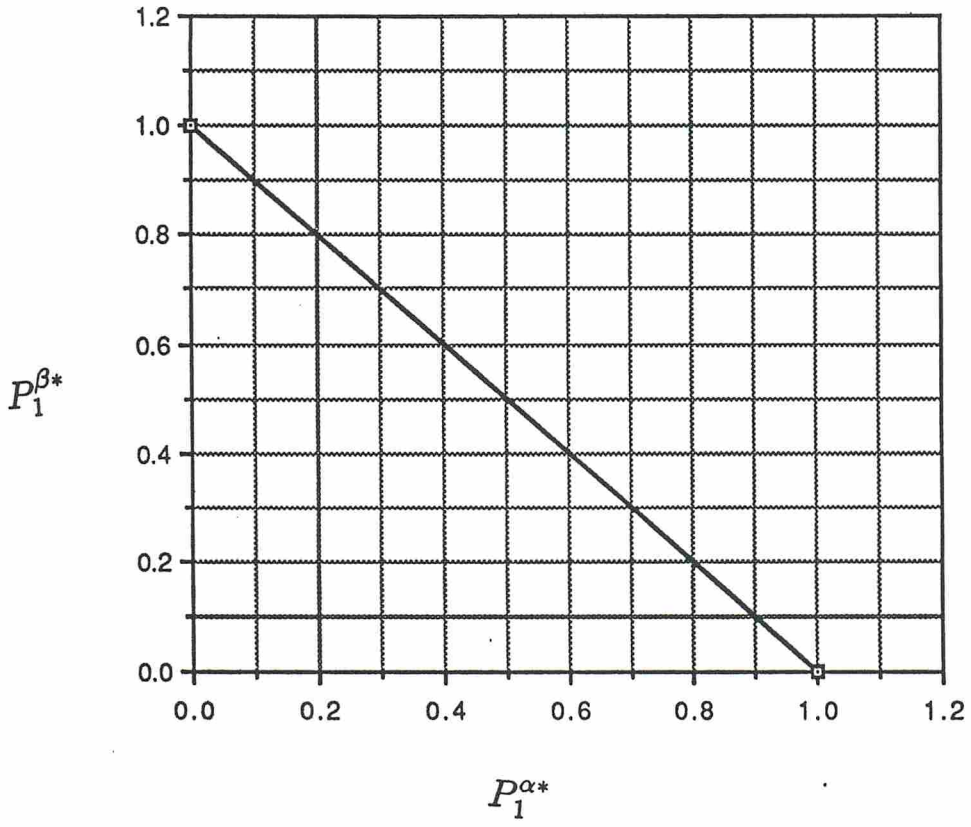

Fig. 3 The optimum routing probabilities $(P_1^{\alpha*}, P_1^{\beta*})$ for a homogeneous network $C_1 = C_2 = 1$ and $\lambda^\alpha = \lambda^\beta = 0.1, ..., 0.9$.

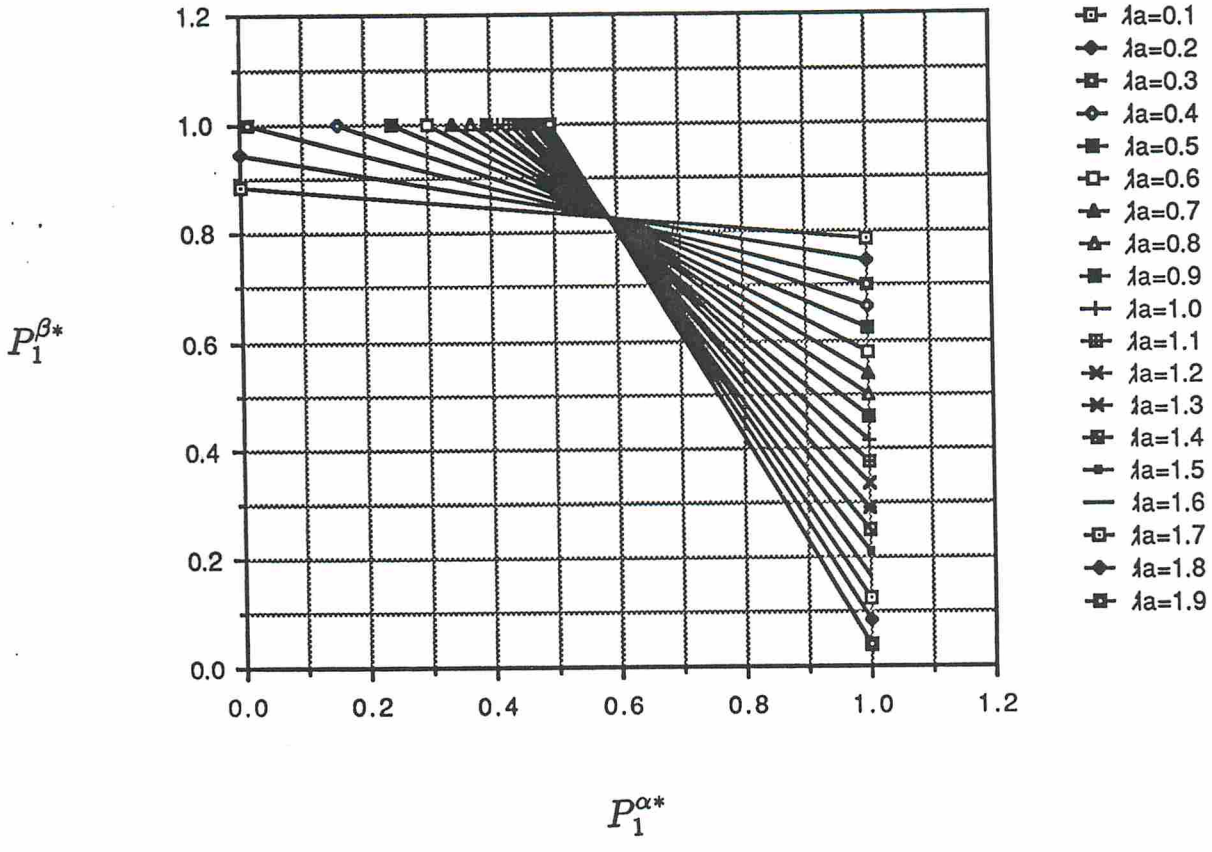

Fig. 4 The optimum routing probabilities $(P_1^{\alpha*}, P_1^{\beta*})$ for fixed server capacities $C_1 = 2$ and $C_2 = 1$, fixed class β arrival rate $\lambda^\beta = 1$ and different class α arrival rates $\lambda^\alpha = \lambda^\beta = 0.1, ..., 1.9$.

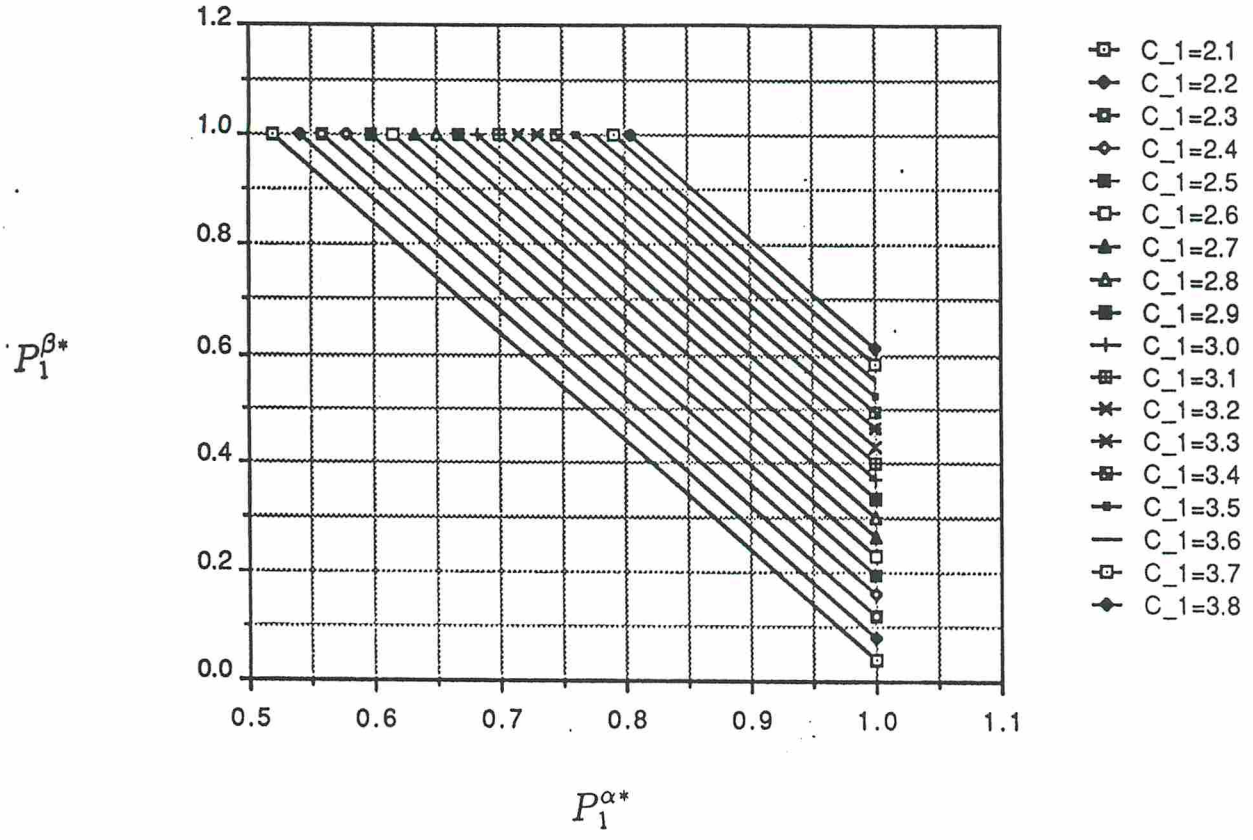

Fig. 5 The optimum routing probabilities $(P_1^{\alpha*}, P_1^{\beta*})$ for fixed arrival rates $\lambda^\alpha = 2$ and $\lambda^\beta = 1$, fixed server 2 capacity $C_2 = 1$, and server 1 capacities $C_1 = 2.1,...,3.8$.

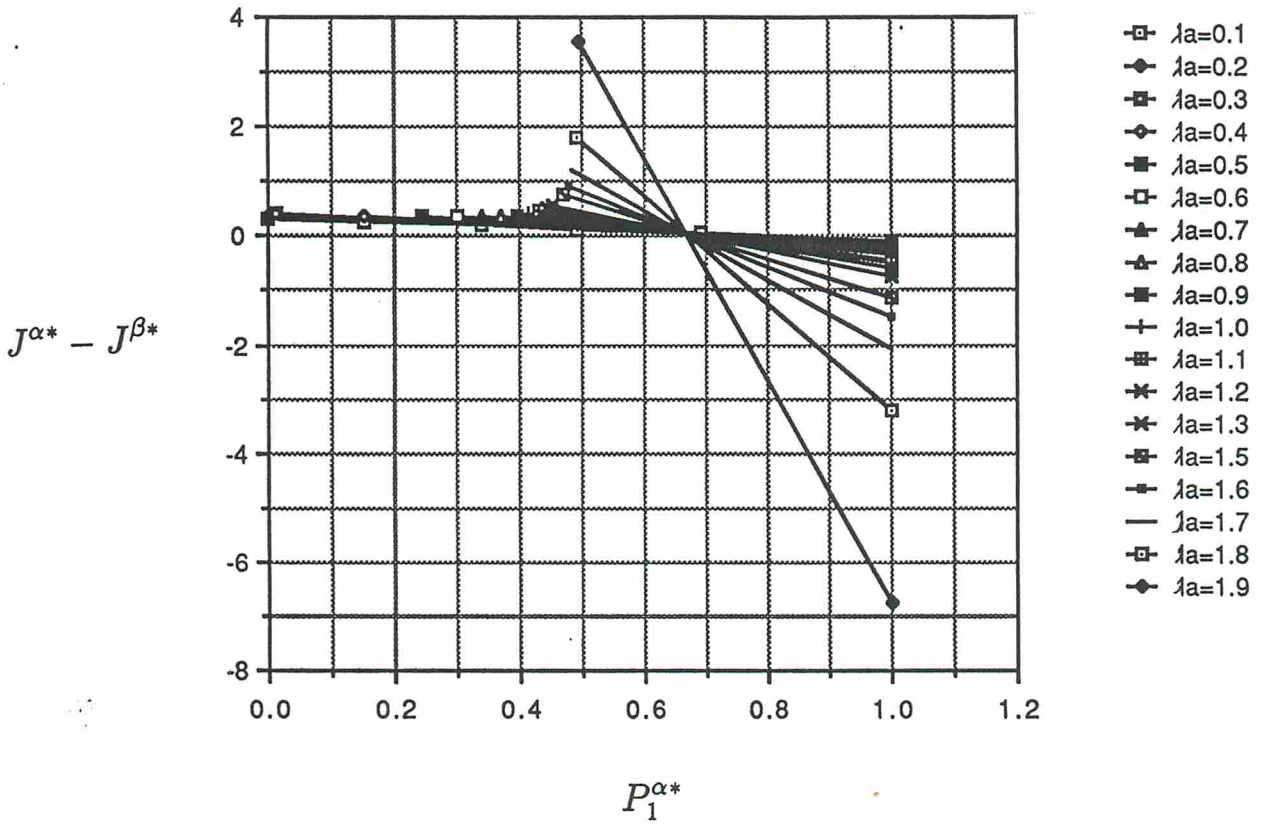

Fig. 6 The difference of the optimum average delay of class α minus the optimum average delay of class β, $J^{\alpha*} - J^{\beta*}$, for fixed server capacities $C_1 = 2$ and $C_1 = 1$, fixed class β arrival rate $\lambda^\beta = 1$ and different class α arrival rates $\lambda^\alpha = 0.1, ..., 1.9$.

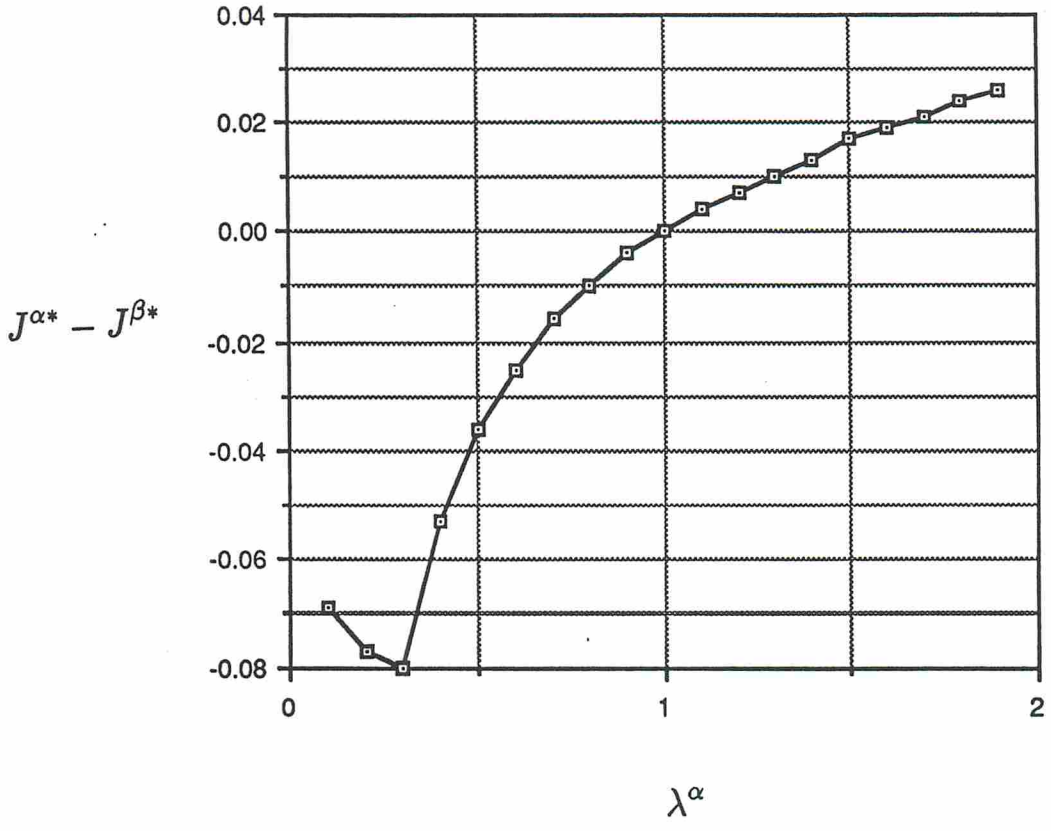

Fig. 7 The difference of the Nash equilibrium average delay of class α minus the Nash equilibrium average delay of class β, $J^{\alpha*} - J^{\beta*}$, for fixed server capacities $C_1 = 2$ and $C_1 = 1$, fixed class β arrival rate $\lambda^{\beta} = 1$ and different class α arrival rates.

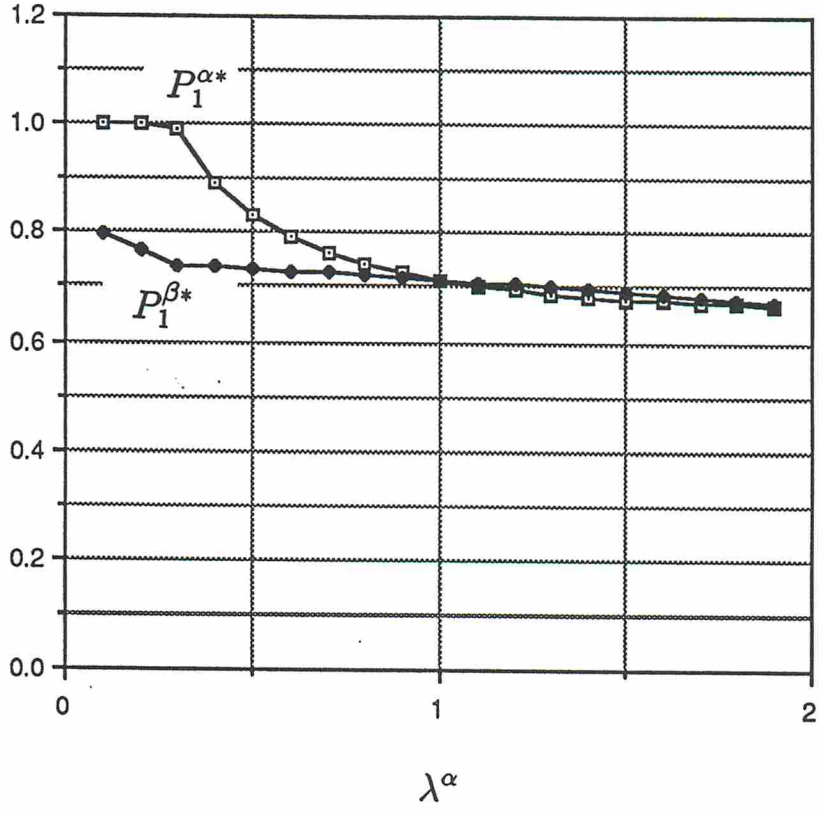

Fig. 8 The Nash equilibrium routing probabilities of class α, $P_1^{\alpha*}$ and class β, $P_1^{\beta*}$, for fixed servercapacities $C_1 = 2$ and $C_2 = 1$, fixed class β arrival rate $\lambda^\beta = 1$ and different class α arrival rates λ^α.

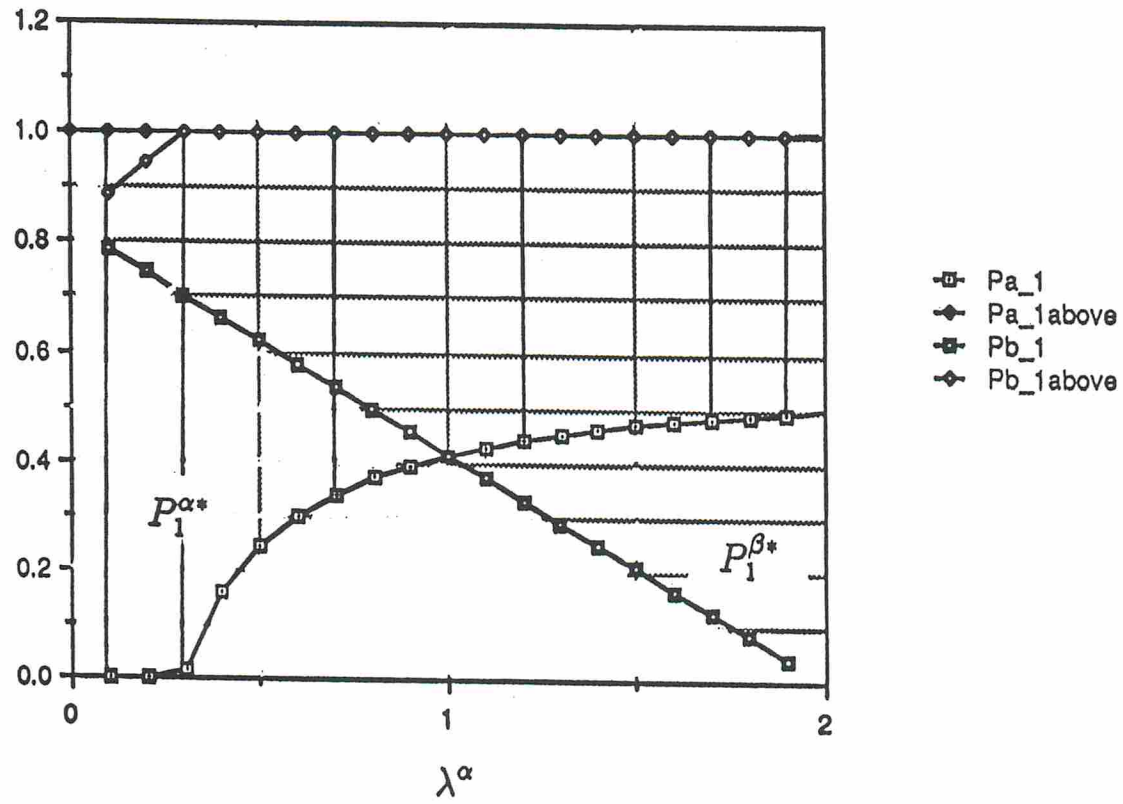

Fig. 9 The team optimum routing probabilities of class α, $P_1^{\alpha*}$ and class β, $P_1^{\beta*}$, for fixed servercapacities $C_1 = 2$ and $C_2 = 1$, fixed class β arrival rate $\lambda^\beta = 1$ and different class α arrival rates λ^α.