An Efficient Resource Allocation Scheme for Mobile Multimedia Networks

Vijay Kumar B.P. and P. Venkataram
Protocol Engineering and Technology (PET) unit. Electrical Communication Engineering Dept.,
Indian Institute of Science, Bangalore-560012,INDIA,
E-mail:{vijaybp, pallapa}@ece.iisc.ernet.in

Abstract—In mobile multimedia networks the traffic fluctuation is unpredictable and also due to limited resource availability, lhe resource allocation to multimedia applications of varying Quality of Service (QoS) requirement becomes a complex issue. This paper proposes an efficient resource allocation scheme based on resource reduction of running applications without hampering their QoS guarantee, in a single mobile cellular environment We propose a Linear Programming (LP) based resource reductiun far efficient Resource Allocation (RA). Artificial Neural Network model is used to solve lhe linear programming problem, which facilitates in real time control decision in the practical systems. The model is computationally less expensive and faster for resource allocation.

The suggested scheme along with the LP-based resource reduction has shown that it is appropriate for the reduction of assigned resources for running applications during over load conditions and allocation of resources to requesting applications. The simulation results for lhe proposed scheme yielded an improved resource utilization and lower percentage of rejection lo hand-off and new applications due to efficient resource allocation.

***Keywards:** Resource Allocation, Mobile networks, Quality of Service (QoS), Artificial Neural Network, Linear Programming.*

I. INTRODUCTION

With the advancement in wireless communication networks and portable computing technologies, the transport of real time multimedia traffic over the wireless channel is challenging due to the severe resource constraints of the wireless link and mobility. A key characteristics of multimedia type application services is that they require different quality of service (QoS) guarantees. Due to the limitations of the radio spectrum, the wireless systems use micro/pico-cellular architectures in order to provide a higher capacity. Because of the small coverage area of micro/pico-cells, the hand-off events in the systems will occur at a much higher rate as compared to macro-cellular systems. Frequent hand-offs in wirelesslmobile networks introduce a new paradigm in the area of optimal resource allocation to maintain the QoS needs of the running applications.

Most of the existing resource allocation schemes to guarantee QoS in mobile multimedia networks have considered bandwidth as one of the QoS parameters and their performance is evaluated for connection blocking and dropping probabilities of hand-off and new connection requests, respectively. The schemes provided in [1][2] considered the dynamic resource allocation scheme to provide QoS assurance to heterogeneous traffic in multimedia wireless networks. Several other methods [3][4] followed resource reservation strategies for admission control to deal with hand-off and new connection requests.

In this paper, an efficient resource allocation scheme is proposed for multimedia traffic carried in mobile networks. The proposed scheme is restricted to the applications within a cell without loss of generality, by considering the local information of unused resources and allocated resources to the running applications to determine whether it can meet the resource requirement of running and requesting applications. The scheme uses a Linear Programming (LP) based resource reduction model for resource allocation to the running and requesting applications and to maximize the resource utilization and minimize average rejection of requesting hand-off and new applications. The LP problem is solved by using Artificial Neural Networks (ANN) which are inherently parallel and distributed, which aids in real time information processing and decision making capability.

The rest of the paper is organized as follows. Some of the definitions that are followed in this paper, the proposed resource allocation scheme and LP-based Resource Reduction model are described in section II. The simulation model along with the simulation results are presented in section III. Finally, we provide concluding remarks in section IV.

II. LP-BASED RESOURCE ALLOCATION SCHEME

A. Definitions

We introduce some of the definitions that are used in this paper.

Definition 1: Network Resources are a set of resources, $\{1, 2, \ldots q)$ available for users in a mobile cell, and the set of parameters $\{P_1, P_2, \ldots P_q\}$ indicate their maximum capacities respectively.

For example, if j^{th} resource is the bandwidth having a maximum capacity of 100 Mbps, then $P_j = 100$.

0-7803-7605-6/02/$17.00 ©2002 IEEE.

Definition **2:** The ***UserApplications*** represent load in the system, which are categorized into ***Running, Hand-off*** and ***New Applications*** and each application request all the q resources, whose requirement depends on the type of application such as audio, video or data.

Running Application: An application for which the connection is already established and network resources have been allocated to it in the cell is called a Running Application. Each of the running application's resource requirement is shown as follows.

R_{ij}^{min} = Minimum value of j^{th} resource requirement of i^{th} running application.

R_{ij}^{max} = Maximum value of j^{th} ———— " ————.

R_{ij}^{alloc} =Allocated value of j^{th} ———— " ————.

For example, if an i^{th} running application request for bandwidth (say, j^{th} *resource)* ofmaximum value **64** Kbps and minimum value of **32** Kbps, and let allocated bandwidth is **48** Kbps which will be well within 64 Kbps and **32** Kbps.

That is: $R_{ij}^{min} = \mathbf{32} \leq R_{ij}^{alloc} = 48 \leq R_{ij}^{max} = \mathbf{64}.$

Hand-off Application: When a mobile user move from one cell to another cell then hand-off takes place. An application whose connection is already established and resources were allocated to it in the previous cell will request for resources in the new cell, such an application is called Hand-off application. Each of the Hand-off Application's resource requirement is shown as follows.

H_{ij}^{min} = Minimum value of j^{th} resource requirement of i^{th} hand-off application.

H_{ij}^{max} = Maximum value of j^{th} ———— " ————.

H_{ij}^{alloc} =Allocated value of j^{th} ———— " ————.

New Application: When a Mobile user initiate an application within a cell and request for resources is called new application, Each of the New Application's resource requirement is shown as follows.

N_{ij}^{min} = Minimum value of j^{th} resource requirement of i^{th} new application.

N_{ij}^{max} = Maximum value of j^{th} ———— " ————.

Allocated quantity for **all** q resources to new applications will be zero (i.e., $N_{ij}^{alloc} = 0$ *Vi).*

Definition *3:* The ***rr-tuple (Resource Requirement-tuple)*** of an application is expressed as a set of resource requirement and allocated for an application.

For example, the ***rr-tuple*** for an i^{th} hand-off application is as follows.

$$< (H_{i1}^{min}, H_{i1}^{max}, H_{i1}^{alloc}), (H_{i2}^{min}, H_{i2}^{max}, H_{i2}^{alloc}), \ldots (H_{iq}^{min}, H_{iq}^{max}, H_{iq}^{alloc}) >$$

and similarly for i^{th} new application, the ***rr-tuple*** is expressed as follows.

$$< (N_{i1}^{min}, N_{i1}^{max}, 0), (N_{i2}^{min}, N_{i2}^{max}, 0), \ldots (N_{iq}^{min}, N_{iq}^{max}, 0) > .$$

Definition **4**: ***Available resource*** is a vector $\Theta = [\Theta_1, \Theta_2, .. \Theta_q]$ indicating the amount of available quantities of the resources **1,** ***2,*** ..., q, respectively.

A. Resource Allocation Procedure

To discuss the proposed resource allocation scheme, consider a mobile cell environment in which ***r*** applications have been running (or scheduled) at an instant, whose resource requirement are given in rr – ***tuples*** for running applications, as in definition **3.** Let h hand-off and n new applications request for resources (or scheduling) at that instant of time (i.e., when r running applications are being scheduled). These requests are processed for resource with certain priority as follows.

One can use a variety of techniques to determine which of the applications should be given priority. In the proposed scheme we give two level priorities. Firstly,we give priority to hand-off applications over new applications. Secondly, the priority among the applications, which is based on their average requirement over all q resources. For hand-off applications the order is as specified in equation (I).

$$\frac{\left(\sum_{j=1}^{q} \frac{H_{1j}^{alloc}}{P_j}\right)}{q} < \frac{\left(\sum_{j=1}^{q} \frac{H_{2j}^{alloc}}{P_j}\right)}{q} < \ldots, \frac{\left(\sum_{j=1}^{q} \frac{H_{hj}^{alloc}}{P_j}\right)}{q} \qquad (1)$$

and, for new applications the order is:

$$\frac{\left(\sum_{j=1}^{q} \frac{N_{1j}^{min}}{P_j}\right)}{q} < \frac{\left(\sum_{j=1}^{q} \frac{N_{2j}^{min}}{P_j}\right)}{q} < \ldots \frac{\left(\sum_{j=1}^{q} \frac{N_{nj}^{min}}{P_j}\right)}{q} \qquad (2)$$

This arrangement make the applications with higher requirement to suffer more, but it makes way for more applications to schedule with increase in average resource utilization.

The process of Allocation of the resources for hand-off and new applications follows any one of the following two conditions.

Condition **1:** If the available resources **(O)**are more than or equal to the sum of total allocated resources of h hand-off and the total minimum requirement of n new applications then, allocate minimum resources to new applications and previously allocated resources to hand-off applications from the available resource ***(Θ)*** To allocate excess resources available ($i.e.$, $\Theta_j - (\sum_{i=1}^{h} H_{ij}^{alloc} + \sum_{i=1}^{n} N_{ij}^{min}) \forall j \in (1, q|,)$ to requesting applications well with in the requirement and to provide some fairness to the applications, we convert this problem to a Linear Programming problem, which is given as:

$$\begin{aligned} \textit{maximize} \quad & \textstyle\sum_{j=1}^{q} (\sum_{i=1}^{h} \varphi_{ij} X_{ij} + \sum_{i=1}^{n} \psi_{ij} Y_{ij}) \qquad (3) \\ \textit{subject to} \quad & \textstyle\sum_{i=1}^{h} X_{ij} + \sum_{i=1}^{n} Y_{ij} \leq \\ & \quad (\Theta_j - (\textstyle\sum_{i=1}^{h} H_{ij}^{alloc} + \sum_{i=1}^{n} N_{ij}^{min})), \\ & 0 \leq X_{ij} \leq (H_{ij}^{max} - H_{ij}^{alloc}), \end{aligned}$$

$$0 \leq Y_{ij} \leq (N^{max}_{ij} - N^{min}_{ij}),$$
$$X_{ij}, Y_{kj} \geq 0 \quad \forall i \in [1, h],\ \forall k \in [1, n], \forall j \in [1, q]$$

Where $0 \leq \psi_{ij} < 1$, could be the relative importance or weight given to i^{th} hand-off or new application for j^{th} resource. The relative importance is given in terms of their resource requirement quantity, i.e., $\varphi_{ij} = \frac{H^{max}_{ij} - H^{alloc}_{ij}}{\Theta_j}$ for hand-off applications and $\psi_{ij} = \frac{N^{max}_{ij} - N^{min}_{ij}}{\Theta_j}$ for new applications. X_{ij} and Y_{ij} are the decision variables to be solved for excess resources to get total allocated resources for hand-off and new applications respectively.

The total allocated resources for hand-off and new applications is as follows.

$R^{(r+i)j} = H^{alloc}_{ij} + X_{ij} \quad \forall i \in [1, h], \forall j \in [1, q]$ for hand-off applications,

$R''_{(r+h+i)} = N^{min}_{ij} + Y_{ij} \quad \forall i \in [1, n], \forall j \in [1, q]$ for new applications.

Here, we solve for $\Theta_j < (\sum_{i=1}^{h} H^{max}_{ij} + \sum_{i=1}^{n} N^{max}_{ij})$, otherwise the problem is trivial.

Condition 2: If the available resources (O) are less than the total allocated resources of ***h*** hand-off and the total minimum requirement of n new applications then the resource allocation is done in sequence: first all hand-off applications are considered and later the new applications using equation (4) and *(5)* respectively.

The resource allocation for hand-off applications is given by:

While $((\mathbf{O}, -\sum_{k=1}^{i-1} H^{alloc}_{kj}) > H^{alloc}_{ij}, \forall j \in [1, q])$
$$R^{alloc}_{(r+i)j} = H^{alloc}_{ij}; \qquad \forall i \in [1, h], \tag{4}$$
and, for new applications:

While $((\Theta_j - (\sum_{k=1}^{h} H^{alloc}_{kj} + \sum_{l=1}^{i-1} N^{min}_{lj})) > N^{min}_{ij}, \forall j \in [1, q])$
$$R^{alloc}_{(r+h+i)j} = N^{min}_{ij}; \qquad \forall i \in [1, n], \tag{5}$$

To accommodate possible number of remaining hand-off and new applications a Linear Programming based Resource Reduction with ANN model (described in the next section) is used to reduce the allocated quantity of resources for running applications. The LP-RR updates the available resource (Θ) and are allocated to possible hand-off and new applications. **An** algorithm I, describes the resource allocation scheme.

Algorithm 1: Resource Allocation Scheme

```
-----------------------------------
IF ( h H and n N request )
THEN   /* H = Hand-off and N = New appls. */
BEGIN
Step I:
Arrange H and N in the order of their resource
requirement;   /* equation (I) and (2) */
Step 2:
IF ( Θ ≥ Total requirement of h and n applications)
  THEN
    Allocate minimum resources for h and n, and
    add excess resources to h and n appls. Eq. (3);
  ELSE
    begin
     Allocate Θ to possible h and n appls. Eq. (4-5);
     Call LP-RR;
     Allocate O to possible number of remaining H and N;
    end;
END;
-----------------------------------
```

B. Linear Programming based Resource Reduction

In this subsection we discuss the principle of LP based Resource Reduction (LP-RR), which reduces the allocated resources of the running applications well within their resource requirement.

Let r applications have been running (or scheduled) in a mobile cell and the applications have been allocated a certain proportion of the q resource. We need to determine the fair reduction of resources from the applications without violating the resource requirement given in their respective $\tau\tau - tuple$. The reduction of resources is carried out only if the total allocated quantity of resource is greater than the total minimum requirement of running applications by certain threshold value (γ).

i.e., *IF* $(\sum_{i=1}^{r} R^{alloc}_{ij} - \sum_{i=1}^{r} R^{min}_{ij}) > \gamma_j \quad \forall j \in [1, q]$,

then the total allocated quantity of resource is reduced by certain quantity, which is fixed by the reduction parameter (Ω). (we have chosen Ω_j value to vary between 0.9 to 0.7 if γ_j value is 10% and above the maximum quantity of j^{th} resource). Now the problem reduces to the following in Linear Programming problem, which is given by :

$$\begin{aligned} &maximize && \textstyle\sum_{j=1}^{q} \sum_{i=1}^{r} c_{ij} R_{ij} \\ &subject\ to && \textstyle\sum_{i=1}^{r} R_{ij} \leq \Omega_j (P_j - \Theta_j), \\ & && R^{min}_{ij} \leq R_{ij} \leq R^{alloc}_{ij}, \\ & && R_{ij}, \geq 0 \qquad \forall i \in [1, r], \forall j \in [1, q]. \end{aligned} \tag{6}$$

where R_{ij} 's are decision variables to be solved, which gives the newly allocated quantity of resources for running applications **on** solving the above LP problem. c_{ij} 's are the weights chosen for j^{th} resource of i^{th} application based on the allocated quantity of resource with respect to maximum and minimum resource requirement. We define the weight function as follows

$$c_{ij} = \frac{R^{max}_{ij} - R^{alloc}_{ij}}{R^{max}_{ij} - R^{min}_{ij}}$$

Ω_j is a reduction parameter value for j^{th} resource to keep the resource reduction well within the requirement range.

Convert the given LP problem to minimization equivalent linear programming problem by adding slack and surplus variables.

For example, the obtained LP problem *(6)* is converted to minimization equivalent problem as follows:

$$\begin{aligned} &minimize && \textstyle -\sum_{j=1}^{q} \sum_{i=1}^{r} c_{ij} R,, \\ &subject\ to && \textstyle\sum_{i=1}^{r} R_{ij} + s_j = \Omega_j (P_j - \Theta_j), \end{aligned} \tag{7}$$

$$\begin{aligned} R_{ij} + slk_{ij} &= R_{ij}^{alloc}, \\ R_{ij} - slk_{ij}^{'} &= R_{ij}^{min}, \\ R_{ij} &\geq 0 \qquad \forall\, i \in [1, r], \forall\, j \in [1, q]. \end{aligned}$$

Where s_j, slk_{ij} and $slk_{ij}^{'}$ are the slack and surplus variables respectively.

The LP equation given in equation (7) can be represented in more compact matrix form equation, as follows.

$$\begin{aligned} &\textit{minimize} && c^T x \\ &\text{subject to} && Vx = b, \\ & && 0 \leq x \leq x_{max}. \end{aligned} \tag{8}$$

Where the matrix $V \in \Re^{(2(r*q)+q)\times(3(r*q)+q)}$ is called the constraint matrix of the LP, the vector c $\in \Re^{3(r*q)+q}$ is the coefficient ofobjective function variables, b $= [\Omega_i\,(P_i - \Theta_i)\,|\,i \in [1, q], R_{ij}^{alloc}, R_{ij}^{min}\,|\,i \in [1, r], j \in [1, q]]$ is the right handside vector in the constraint equation and $x = [R_{ij},\ \Theta_j,\ s_j\ slk_{ij},\ slk_{ij}^{'}\,|\,i \in [1, r], j \in [1, q],]$ is the decision variable vector. Above LP problem (8) is solved for decision variables by using Artificial Neural Network by designing its connection weight matrix $(W)$and biasing thresholds (θ) [5][6][7]. Where $W = -AV^T V$ and $\theta = AV^T$ b $- B \exp(-Ct)$ c for the above ANN model. The sigmoid activation function is used for neuron output. A, B, and C are positive scalar parameters. The values chosen for the parameters based on the maximum value of the decision variables (x_{max}) are eiven in section III.

III. Simulation

The simulation model composed of a single cell, which will keep contact with its six neighboring cells. Each cell contains a base station, which is responsible for the connection setup and tear-down of new applications and to serve hand-off applications. Also, base station controls the resource redistribution to running applications using LP-RR, as and when the request for resources from new and hand-off application arrives. For simulation purpose we have considered two of the resources, for example bandwidth and buffer having a maximum capacity of 45 Mbps and 500 Mbytes respectively, i.e., $P_1 = 45$ *Mbps* and $P_2 = 500$ Mbytes. Two types of applications traffic are considered in the simulation, a new application, which is initiated by a mobile user within a cell and a hand-off application, which is handed off from the neighboring cells.

We have considered the bandwidth and buffer requirement of each application in the range of 0.1 Mbps to 3 Mbps and 5 Mhytes to 30 Mbytes respectively. The minimum value of resource requirement is considered to about 10% to 50% of the maximum value. The reduction parameter (Ω) value in constraint equation *(6)* is carefully chosen in order to avoid drastic variation in the allocated values of resources to running applications, (we have chosen its value value to vary between 0.7 to 0.9). Based on the maximum value of the decision variables in the proposed LP-RR simulation model, the parameter values for the neural network to solve the LP equations are: $A = B = 300$ and $C = 0.01$.

The simulation is carried out on a Pentium III 550 MHz PC, for 10,000 arrival events with random number of application arrivals (maximum upto 50) with random resource requirement (with in the specified requirement) of each requesting applications to realize realistic network traffic. The performance measures obtained through the simulation are the percentage utilization of bandwidth and buffer, resources utilization by hand-off and new applications, average percentage rejection of new and hand-off applications due to lack of resources. The results are plotted as a function of number of applications arrival observed for all the 10,000 arrival events.

We have taken the simulation results with equal priority to hand-off and new applications, and with priority to hand-off applications over new applications. In order to evaluate the performance, the proposed scheme is compared with the scheme *without LP-RR* model.

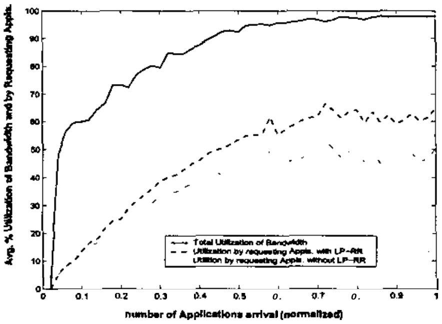

Figure I: Resource utilization (with equal priority to hand-off and new applications).

B. Results

Fig. 1 shows the graph for an average percentage utilization of bandwidth for the simulation results based on equal priority to hand-off and new applications plotted with respect to the number of applications arrival (normalized to peak arrivals). (similar results are taken for buffer utilization, because of space limitation we could not show the graph).

The results, for with and without LP-RR model, are plotted and obtained an improved average utilization of resources by the requesting applications (about 15% to 20%) for the scheme with LP-RR model. The total utilization of resources by running and requesting applications is also plotted to study the total performance of the system. The utilization of bandwidth and buffer has increased up to 92% and 97% gradually as the number of applications arrival increases to peak value respectively.

Fig. 2 shows the average percentage rejection of requesting applications with respect to the number of applications arrival. The graph shows that with LP-RR the average rejection of requesting applications has decreased to about *15%* to 20% for different number of applications arrival events (normalized to peak arrivals), as compared to the scheme without LP-RR. This indicates the increased resource utilization by requesting applications (shown in Fig. 1).

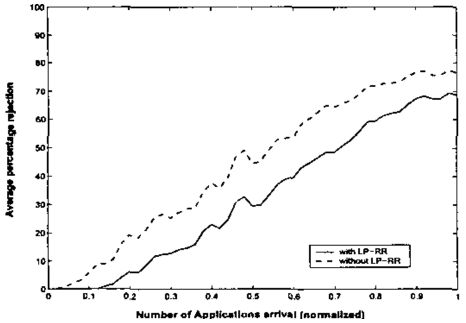

Figure 2: Percentage rejection of applications (with equal priority to hand-off and new applications).

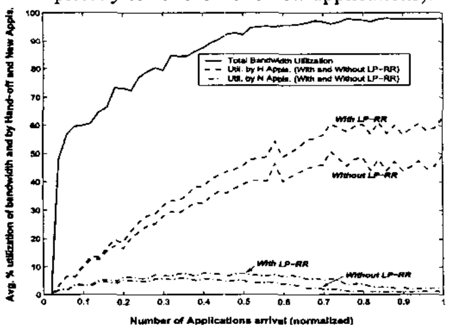

Figure 3: Resource utilization (with priority).

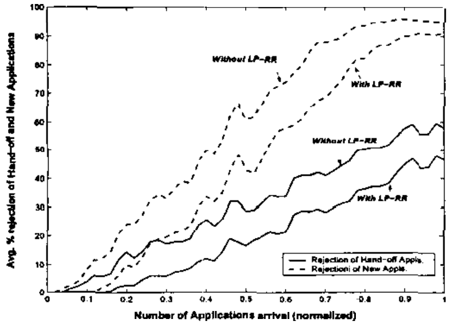

Figure 4: Percentage Rejection of hand-off and new applications (with priority).

Fig. 3 shows the graphs for an average percentage of bandwidth utilization as a function of applications arrival with priority to hand-off applications over new applications. The resource utilization by hand-off and new applications has increased up to **15%** at average and peak amvals, with LP-RR than with-out LP-RR. This shows that the proposed scheme has increased the admitted hand-off and new applications as compared to the scheme with-out LP-RR. The graph of percentage rejection to hand-off and new applications is shown in Fig. 4. Since the proposed scheme reduces the assigned resources for running applications to adjust for the requesting applications, the proposed scheme achieved lower percentage of rejection to new and hand-off applications (up to 20%) than the scheme without LP-RR. The priority given to hand-off applications has increased the rejection for new applications up to 90% at peak arrivals, where as for hand-off applications it is about **50%.**

The graph of percentage reduction of allocated resources for running applications with respect to their maximum resource requirement and their requirement range are ploned. The results show that the resource reduction is proportional to the applications requirement and requirement range, thus achieving some fairness among the applications of different resource requirement and range of requirement (because of space limitation we could **not** show the graphs for the same).

IV. CONCLUSION

In this paper, an efficient resource allocation scheme for mobile multimedia networks has been proposed. The scheme provides the resource allocation to hand-off and new applications by reducing and reassigning the allocated resources for running applications. The resource utilization by the applications has increased with LP-RR as compared to without LP-RR and gave an improved results by adjusting the resource reduction parameter value based on the resources requirement of requesting applications. The ANN model used for resource reduction has shown a good performance as a computational model. The Artificial neural network used is capable of generating feasible solution **to** linear programming problems.

It is shown through simulation results that the proposed scheme has improved the resource utilization and lower percentage of rejection to hand-off and new applications as compared to the scheme without LP-RR. LP-RR has maintained some fairness during the reduction of allocated resources for running applications.

REFERENCES

[1] P. Ramanathan, K. M. Shivalingam, P. Agrawal and Shalinee K, "Dynamic Resource Allocation Schemes During Hand-ff for Mobile Multimedia Wireless Networks", ***IEEE J. on Select Area Commun.,*** **Val., 17, no.7, pp. 1270-1283,July,** 1999.

[2] C. Oliveira, J. B. Kim and T. Suda, "An Adaptive Bandwidth Reservation Scheme **far** High-Speed Multimedia Wireless Networks", ***IEEE J. Select. Area Commun.,*** **Vol.,** 16,no. 8, pp. 858-874, August, 1998.

[3] M. Naghshineh and M. Schwartz, "Distributed **call** admission control in mobile/wireless networks". ***IEEE J. Select Areas Commun.***, Vol. 14, no. 4, pp. 711-717, May, 1996.

[4] **S.** H. Oh and D. W. Tcha, "Prioritized channel assignment in a **cellular** radio network", ***IEEE Trans. Communications,*** Vol. 40, **no.** 7, pp. 1259-1269, July, 1992.

[5] J. Wang and V. Chankong, "Recurrent neural networks for linear programming: Analysis and design principles" ***Computers Operations research,*** **vol.** 19, no. 3/4, pp 297-311. 1992.

[6] A. Cichocki and R. Unbehauen, Neural Networks **for** Optimization and Signal Processing, John-Wiley and sons, Stuttgart, 1993.

[7] J.J.Hopfield and D.W.Tank, "Neural computation **of** decisions in optimization problems", ***Biological Cybernetics,*****vol.** 52, no. 1, pp.141-152, 1985.