Proactive Resource Reservation in Next-Generation Wireless Networks

Arijit Ukil
Innovation Labs
Tata Consultancy Services
Kolkata, India
arijit.ukil@tcs.com

Jaydip Sen
Innovation Labs
Tat Consultancy Services
Kolkata, India
jaydip.sen@tcs.com

Abstract— **Resource reservation is a key management function in future wireless networks, particularly to provide high-end multimedia applications to mobile users with quality of service (QoS) guarantees and to optimize the overall system resource utilization. In prediction based resource reservation, the radio resource is reserved from an a priori estimation of the future traffic load. In this paper, we have presented a scheme for proactive resource reservation in next generation wireless networks. In this scheme, the future traffic estimation is based on Normalized Least Mean Square (NLMS) adaptive filtering method. Based on the estimated characteristic of future traffic load, network congestion notification and network utility factor are evaluated. We have applied tele-traffic theory based on these two parameters to optimize the resource reservation scheme. The novelty of the scheme is the integration of adaptive filtering and tele-traffic theory in a coupled manner in order to find an optimum solution for resource reservation in next generation wireless networks like WiMAX. To show the effectiveness of our proposed scheme and the efficacy of the different parameters for optimizing the overall resource reservation process, we have presented computer simulation results, which include traffic estimation and predictive resource reservation.**

Keywords -resource reservation;, tele-traffic theory; quality of service; WiMAX; Normalized Least Mean Square; congestion notification; network utilization;

I. INTRODUCTION

A wireless cell can only host a limited number of calls. New calls and incoming handoff calls should not compromise the quality of the ongoing calls in the cell. Due to limited and varying wireless channel condition and traffic characteristics, resource reservation is one of the most important components of the overall resource management process in wireless networks. As the next generation wireless networks cater for various kinds of applications from typical voice to high-end multimedia with varying degree of QoS parameters, resource management holds a key position in order to optimize the overall system performance. The performance of a wireless network is bounded by the efficiency of resource management. The aim of radio resource management in wireless cellular communication networks is to share the available and often rather limited radio spectrum between users as efficiently as possible. Here, 'efficiency' refers to use of capacity in the sense of maximizing the data traffic load with respect to satisfying the QoS requirements of as many mobile users as possible, and overcoming the fundamental difficulties of radio wave propagation in the wireless environment [1]. Resource management mostly consists of three broad parts, namely resource allocation, connection admission control and resource reservation. Resource reservation is responsible for participating in the two important activities of resource management, viz. to maximize the number of users in a cell at any instant of time and to ensure planned (low) blocking. It can be noted that the main purpose of resource manager is to allocate physical radio resources when requested by the *radio resource control* (RRC) layer from the knowledge of radio network configuration and state data. In fact, the allocation is a reservation of a proportion of the available radio resources according to the channel request from RRC layer for each radio connection. It may be noted that resource reservation, or more broadly, resource management is not a trivial task. In order to find optimal solution, we must consider a number of constraint parameters and some complex convex functions. For some time, researchers have put good amount of effort to find optimal solutions for this problem. Senarath et al have proposed prioritization and queuing-based solutions for lowering the connection dropping rate [2]. In prioritization, handoff connections are given higher priority compared to new connections. However, this approach results in the starvation of new connections in cells intersected by the highways. Since the number of handoff requests to these cells is very high, stationary users in such cells will not be able to get enough radio resources. Another method of increasing network capacity is to introduce some form of adaptation into the resource management of current cellular systems. A plethora of concepts attempting to introduce such adaptation have been proposed [3]. Yu have proposed a statistical strategy for resource reservation through the estimation of a user's transfer probabilities, which represent the possibilities of the user leaving the current cell and entering the neighboring cells and the resources reserved for a user in each base station are proportional to the user's transfer probabilities [4]. Genetic algorithm-based adaptive channel allocation and reservation policy utilizing channel borrowing from neighborhood cells is proposed in [10]. However, there has been not much work done so far for designing a proactive and QoS-aware resource reservation model for next generation wireless networks like WiMAX. In [9], a new dynamic resource reservation scheme to improve system connectivity is proposed, which incorporates pre-reservation technique to provide seamless handoff and path prediction. In this paper, we have attempted to develop a scheme and algorithm for this purpose, which has practical significance in the presence of large number real-time applications in a broadband wireless network. This method does not incur much overhead and it is also not computationally complex, so it is highly suitable in the presence of delay-constraint

applications like video conferencing, real-time multimedia streaming. The paper is organized as follows. In section II, the architecture of the proposed scheme is presented. In section III, we discuss the traffic load prediction scheme and in section IV, we present the resource reservation scheme based on tele-traffic theory. Simulation results are presented in section V, and the paper is concluded in section VI.

II. SYSTEM ARCHITECTURE

The proposed proactive resource system architecture is shown in Fig. 1. The *traffic prediction* module estimates future traffic (at some future point K_0) from the knowledge of the existing applications' QoS, QoS of the new call requesting for entry in the cell or network, the existing traffic load and the current network conditions. In the traffic prediction block, NLMS-based prediction algorithm is implemented, which predicts the future condition of the network in terms of stability and loading, and passes on the information to the *resource reservation* block. Resource reservation block reserves the amount of radio resource in terms of frequency bandwidth for future traffic based on the congestion notification sent by the traffic prediction block. The resource reservation block, which is a part of the resource management system, is located in the base station (BS). The BS acts as the central controller in order to properly look after the overall resource management process including resource reservation. For the purpose of simplicity, we have shown the resource reservation part only, with the assumption that proper resource allocation and connection admission control components are present.

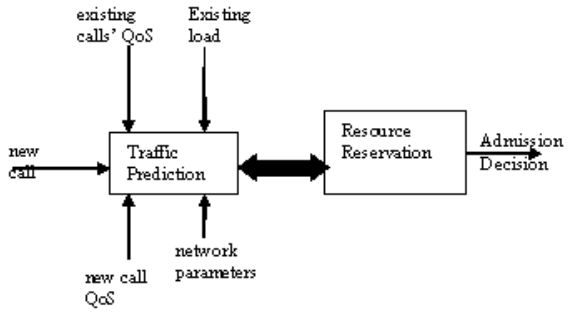

Figure 1. Architecture of the proactive resource reservation system

III. TRAFFIC PREDICTION

In this section, we present the traffic prediction scheme and an algorithm, which is based on NLMS method of adaptive filtering. NLMS is an LMS filter with time-varying step size. NLMS has the advantage of minimizing gradient noise amplification problem suffered by LMS filter [5]. The classical adaptive filtering problem can be stated in the following manner. Given an input signal *u(n)* and a desired signal *d(n)*, determine the filter (*w*) that minimizes the error, *e(n)*, between the output of the filter *y(n)*, and the desired signal *d(n)*. For the case of *finite impulse response* (FIR) filters, an algorithm that solves this problem is the well known LMS. This is given by the following equation:

$$w(n+1) = w(n) + \mu u(n) * e(n) \qquad (1)$$

The equation (1) updates the vector of the filter coefficients *w(n)*. The output of the filter is *y(n) = wT(n) u(n)* with *u(n) = [u(n)... u(n-N+1)]* were *N* is the filter length, and *e(n) = d(n) - y(n)*.

NLMS also exhibits faster rate of convergence than standard LMS for both correlated and uncorrelated input data. Estimation of instantaneous data rate of the future user is made, and is denoted by $C(k+K_0)$, which is predicted data rate of $k+K_0$ th user , where *k* is the current user and $C(k)$ is instantaneous data rate of k^{th} user. NLMS filter order is *M*. $\overline{C(k)}$ is $M \times 1$ vector of instantaneous data rate of the past *M* users and $\overline{\omega(k)}$ is tap-weight vector of dimension $M \times 1$, which are updated dynamically based on the error between predicted and actual data rate value. Prediction error is denoted by *e(n)*, which is the measured error between predicted value of the future user data rate and actual value. These tap-weights are updated dynamically based on the prediction error. The instantaneous data rate after K_0 users from the admission of k^{th} user is predicted based on $C(k+K_0)$. From NLMS adaptive filter theory [6], this predicted value is generated based on the following:

$$C(k+K_0) = \overline{\omega(k)}^T \times \overline{C(k)} \qquad (2)$$

The tap-weights are updated using the following:

$$\overline{\omega(k+1)} = \overline{\omega(k)} + \mu(k) \times \overline{C(k)} \times e(k) \qquad (3)$$

where,

$\mu(k)$ is the step-size parameter. It is defined as follows:

$$\mu(k) = \frac{\mu_0}{\left\|C(k)\right\|^2} \qquad (4)$$

The error in the prediction is computed using the following:

$$e(k) = C(k+K_0) - \overline{\omega(n)}^T \times \overline{C(k)} \qquad (5)$$

where, μ_0 is a positive real scaling factor, which controls the tap-weight vector $\overline{\omega(k)}$ from one iteration to another without changing the direction of the vector. The magnitude of $C(k+K_0)$ is the indicator of future traffic load. When this exceeds beyond the maximum tolerable load on the network, the network is considered as overloaded for the future iteration and is expressed by the notation m_o. The network is properly loaded when the value of $C(k+K_0)$ is in between the maximum tolerable load and minimum possible load, this condition is expressed by the notation m_j. When the value of $C(k+K_0)$ is less than threshold of minimum load on the network, the network is considered to be under-loaded. This condition of the network is denoted by the symbol m_u.

IV. PREDICTIVE RESOURCE RESERVATION

In order to optimize the resource utilization and minimization of call dropping (P_D) and new call blocking

(P_B) probabilities, a priori congestion notification in terms of $m_x|_{x=o,j,u}$ is sent to the resource allocation and reservation unit during each the decision unit. In fact, it involves very complex mathematical computation to find the optimal solution for dynamic resource reservation which minimizes (P_{cd}) and (P_{cb}) simultaneously while the QoS requirement of the users are satisfied. It is to be noted that for the overall system to be properly optimized, over-loading and under-loading of the network both are to be avoided to the maximum possible extent. Where as, under-loading inevitably brings forth the low resource utilization problem, over-loading rapidly degrades system performance. Such a situation will cause an unstable nervous system. So, the objective of the a priori congestion notification is to determine the optimal dynamic load balancing criteria. This is not feasible in real time systems, particularly in situations where mobile base stations are involved due to its high computational complexity.

With this background information, we now present the proposed resource reservation scheme and the algorithm. The salient features of the algorithm are as follows:

1. It optimizes the resource utilization (η) factor of the system.
2. It guarantees that the resources allocated to a real-time connection or specifically the *unsolicited grant service* (UGS) QoS class users to maintain the minimum resource or bandwidth requirement as per the agreement at the call setup time.
3. It is proportionally fair in the sense that the resource or the bandwidth borrowed from the existing connection for reservation purpose is proportional to the ratio of connection's current allocation and mean carried traffic of the cell.
4. It attempts to maintain stability of the resource allocation and connection admission control system by converging towards the optimal system condition, i.e. to prolong or achieve congestion notification state.

We have introduced a new and simplified resource reservation scheme by the modification of tele-traffic theory developed for wireless cellular networks [7]. We essentially follow the approach of slightly degrading or controlling the QoS of the users in the overloaded cell by lowering the data rate of one or several services that are primarily insensitive to increased delays and can gracefully tolerate lesser throughput until the time the phase of system overloading is over, i.e. between the transition period from m_o to m_j. For an optimally loaded system, we attempt to converge the system resource utilization to the maximum so that new call admission does not hamper the system stability or QoS degradation as well as the optimally loaded stage is sustained to the largest possible duration, i.e. the transition time from m_j to m_o or m_j to m_u should be as high as possible. For under-loaded system, we attempt to upgrade the QoS of the users to m_j as far as possible.

The model of the resource reservation system is now described below.

Consider the scenario where the arrival rate γ_k of new call *k*, is follows the Poisson distribution. So, the distribution function for arrival of new calls is $\rho_k(t) = \gamma_k e^{-\gamma_k t}$. We assume that each call in the originating cell may complete in the cell or may handover to one of the neighboring cells after certain time periods that are exponentially distributed with mean values $\frac{1}{\mu_c}, \frac{1}{\mu_h}$ respectively. This means that existing calls complete at rate μ_c and handover calls depart the cell at a rate μ_h. So, the average total call termination rate of the cell is $\mu_T = \mu_c + \mu_h$. From Little's law, the traffic intensity is $\Gamma = \frac{\gamma_k}{\mu_T}$. Considering M/M/c/c queuing model, the blocking probability P_B for the new incoming call *k*, using Erlang-B formula is:

$$P_B = \frac{\frac{(\Gamma)^N}{N!}}{\sum_{n=0}^{N} \frac{(\Gamma)^n}{n!}} \tag{6}$$

where *N* is the number of channels in the cell.

The total traffic carried by the cell is given by:

$$Z = \Gamma(1 - P_B) \tag{7}$$

Now, we have to consider the entry of new calls including calls generated due to handover. Because of the memory-less property of exponential distribution, one can assume that the handover process is also a Poisson process with mean intensity γ_H. Therefore, the total call arrival intensity may be computed using the following:

$$\gamma_T = \gamma_k + \gamma_H \tag{8}$$

The effective offered traffic intensity of the cell is computed as follows:

$$\Gamma_e = \frac{\gamma_T}{\mu_T} = \frac{\gamma_k + \gamma_H}{\mu_C + \mu_h} \tag{9}$$

Now we define some parameters: P_H = the probability that a new call that is not blocked would require at least one handoff, P_{HH} = the probability that a call that has already been handed off successfully would require another handoff, and P_f = the probability of handoff failure

The handovers are not independent processes. They are related to the new call arrivals in the cells of the mobile network. If we denote with P_{Be} be the overall effective blocking probability in a cell including new calls blocking and handover blocking, then P_{Be} can be computed as:

$$P_{Be} = \frac{\frac{(\Gamma_e)^N}{N!}}{\sum_{n=0}^{N} \frac{(\Gamma_e)^n}{n!}} \tag{10}$$

Hong and Rappaport have proposed a traffic model for a hexagonal cell (approximated by a circle) [8]. They assume that the vehicles are spread evenly over the service area; thus, the location of a vehicle when a call is initiated by the user is uniformly distributed in the cell. They also assume that a vehicle initiating a call moves from the current location in any direction with equal probability and that this direction does not change while the vehicle remains in the cell. These we feel are very much valid assumptions in today's mobile wireless scenario. From these assumptions they showed that the arrival rate of handoff calls is given by:

$$\gamma_H = \frac{P_H(1-P_{Be})}{1-P_{HH}(1-P_f)} \times \gamma_k \qquad (11)$$

From this we can deduce the effective offered carried traffic of the cell as:

$$\mathrm{Z}_e = \Gamma(1-P_{Be}) \qquad (12)$$

Now, we will develop the proposed resource reservation and allocation scheme. Let, $\Delta\lambda_k$ be the data rate tolerance value of the connection *k*, which is specified at the call setup as per QoS requirement agreement. So one can write

$$\Delta\lambda_k = \lambda_p^k - \lambda_s^k \qquad (13)$$

It may be noted that for CBR call, $\Delta\lambda_k \to 0$. At any instance, the connections are allocated data rate λ_k which lies between the maximum and sustainable data rate. Thus,

$$\lambda_k = \lambda_s^k + \phi_k \Delta\lambda_k \qquad (14)$$

From this, we calculate the aggregate data rate, which represents the amount data rate demanded by the connections of a cell. It may be noted that aggregate data rate is adjustable and network controller or base station has full control over it, i.e. Ψ is controllable and QoS dependent. Where as, effective offered carried traffic Z_e is mainly dependent on the traffic characteristics, amount of resources (channels) the cell has and Z_e is external to the network controller. In fact, our aim is to optimize Ψ so as to match it with Z_e. We calculate the aggregate data rate as

$$\Psi = \sum_{k=1}^{K} \lambda_k$$

$$= \sum_{k=1}^{K} \lambda_s^k + \sum_{k=1}^{K} \phi_k \Delta\lambda_k \qquad (15)$$

where ϕ_k represents the fraction of the data rate tolerance the connection *k* is allocated over and above its assigned minimum or sustainable data rate value. This fraction represents the elasticity of the connection. We term ϕ_k as elasticity factor. By accepting or attempting to accept a new call including handover calls, the overall equilibrium of the cell is disturbed. A priori congestion notification $m_x\big|_{x=o,j,u}$ helps network controller to take advance action for balancing the system equilibrium and tilting it to the most optimal and stable state. We consider the three congestion notifications sent by the decision unit on which the resource reservation and allocation scheme acts.

1. When congestion notification m_o is sent:

In this case, the system is overloaded, implying

a. the probability of successful entry of a new call is very less, i.e. $P_B \to 1$.
b. the probability of handoff failure is very high, i.e. $P_f \to 1$. Even when handoff calls are admitted, there is a chance of severe degradation of existing connections QoS.
c. the resource utilization factor (η) is more than 1, where resource utilization factor is defined as:

$$\eta = \frac{Z_e + C(k+K_0)}{\Psi} \qquad (16)$$

So, the objective is to minimize the aggregate data rate which in effect, maximizes the reserved bandwidth R(k), i.e. the objective is:

$$\min imize\{\Psi\} = \min inimize\left\{\sum_{k=1}^{K} \lambda_s^k + \sum_{k=1}^{K} \phi_k \Delta\lambda_k\right\}$$

$$= \quad Y + \Delta Y \min imize\left\{\sum_{k=1}^{K} \phi_k\right\} \qquad (17)$$

Since the minimum sustainable rate and maximum or peak rate of each connection is constant, and are negotiated at the time of call setup, the objective function is:

$\min imize\left\{\sum_{k=1}^{K} \phi_k\right\} = \min imize\{\phi_k\}$, as the data rate tolerance fraction values of individual connections are independent. The objective is to be satisfied under the constraint of maintaining the fairness criteria of:

$$\Delta\phi_k = \frac{\lambda_k}{Z_e}$$

where $\Delta\phi_k = \frac{m_k}{L}$, the idea is that bandwidth is to be borrowed gradually, so that QoS degradation is graceful, not abrupt. This optimization process will continue until m_o is changed to m_j or $\phi_k = 0, \forall k$, which ever is reached earlier.

2. When congestion notification m_j is sent:

In this case, the system is properly or optimally loaded and the system is stable even when a new connection is admitted and resource utilization factor tends to unity, $\eta \le 1, \eta \to 1$. The only issue is that when, delay-sensitive connections are assigned with their effective capacity, i.e. with data rate equal to $\min\{\lambda_p^k, (\lambda_s^k + \sigma_k)\}$ for long time, its delay violation may happen. In order to avoid this, the optimization objective is:

$$\max imize\{\phi_k\}$$

subject to:

$$Z_{enew} \ge Z_e + \lambda_p^k,$$

$$\min\{\lambda_p^k, (\lambda_s^k + \sigma_k)\} \le C(k+K_0) + R(k) \text{ and } \Delta\phi_k = \frac{\lambda_k}{Z_e}.$$

3. When congestion notification m_u is sent:

In this case, the system is underutilized and $0 < \eta < 1$, which means the system resource (bandwidth) is wasted and the reserved bandwidth is more than the required value. So, the objective is to push $\eta \to 1$, by providing the existing connections with QoS as maximum as possible while properly reserving resource for the new connection. So, the objective is:

$$\max imize\{\phi_k\}$$

subject to:

$$\min\{\lambda_p^k, (\lambda_s^k + \sigma_k)\} \le C(k+K_0) + R(k) \text{ and } \Delta\phi_k = \frac{\lambda_k}{Z_e}$$

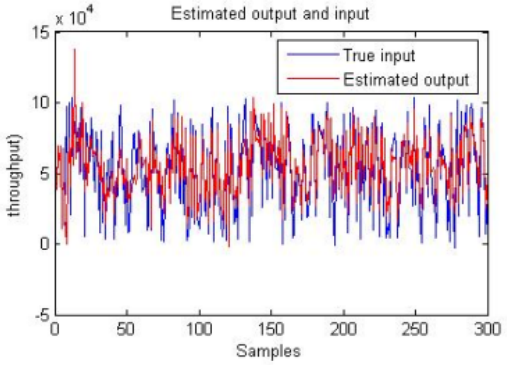

Figure 2. Traffic load estimation for 300 continuous admission requests from application with different QoS requirements

V. SIMULATION RESULTS

In this section, we present the simulation results of our proposed proactive resource reservation scheme. The simulation results are performed in MATLAB platform. In the simulation set up, we have considered random QoS characteristics of the available traffic, which consists of random number of UGS, NRTPS, RTPS, BE traffic. We have selected K_0 = 15, which we find optimum after several simulation runs (> 50) in different parameter settings as well as considering the channel coherence time. In Fig. 2, we show that our estimation of prediction of future traffic load characterization is close to the original one. We have considered 300 samples to validate our scheme. It may be pointed out K_0 plays an important role. We found that, for more magnitude of K_0, prediction accuracy becomes very less and with low value of K_0, NLMS has very little role to play. Proactive resource reservation value in terms of byte is shown in Fig. 3. Here we have considered 40 resource reservation instants. It can be noted that this reservation is done proactively, i.e., based on the value of K_0. Resource is reserved for the predicted traffic load after K_0 cycles. It may be observed that for some cases, no reservation is made, which indicates overloading condition and may result into call blocking and packet dropping.

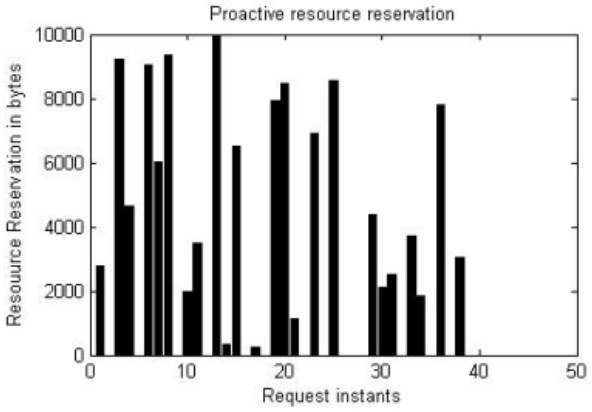

Figure3. Resource reservation up to 40 requests

VI. CONCLUSION

Design of a proactive and QoS-aware resource reservation scheme is an important issue for delivering real-time applications over next-generation wireless networks. In this paper, we have presented a resource reservation scheme and an algorithm for proactive resource reservation that is based on the concepts of adaptive filtering and tele-traffic theory. The proposed algorithm has been simulated in MATLAB platform and the results have demonstrated effectiveness of the proposed scheme.

REFERENCES

[1] S. Naghian, "Location-Sensitive Intelligent Radio Resource Management and its Application in WCDMA Mobile Systems", *Doctoral thesis*, Helsinki University of Technology, Applied Electronics Laboratory, Research report B8, 2001.

[2] G. Senarath and D. Everitt, "Performance of handover priority and queuing systems under different handover request strategies for microcellular mobile communication systems", in *Proc. of IEEE Vehicular Technology Conference*, pp .897-901, 1995.

[3] I. Katzela, M. Naghshineh, "Channel Assignment Schemes: A Comprehensive Survey", *IEEE Personal Communications*, pp. 10-31, June 1996.

[4] W.W.H. Changhua He Yu, "Resource reservation in wireless networks based on pattern recognition", In *Proc. of International Joint Conference on Neural Networks*, pp.2264-2269, 2001.

[5] Z. Dziong, M. Jia, and P. Mermelstein, "Adaptive Traffic Admission for Integrated Services in CDMA Wireless access Networks," *IEEE JSAC*, vol. 14, no. 9, Dec. 1996, pp.1737–47.

[6] Simon Haykin, *Adaptive Filter Theory*, Prentice Hall, 4th Edition, 2002.

[7] Haring, G., et al., "Loss Formulas and Their Application to Optimization for Cellular Networks," *IEEE Transactions on vehicular Technology*, Vol. 50, No. 3, May 2001, pp. 664–673.

[8] D. Hong and S. S. Rappaport, "Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures", *IEEE Trans. Vehicular. Technology,* Vol. 35, No. 3, pp. 448–461, August, 1986.

[9] H-W Feng, W-Y. Kao, J-Ji. Huang, and D. Shiung, "A dynamic resource reservation scheme designed for improving multicast protocols in HMIPv6-based networks", In Proceedings of VTC-Spring, 2006.

[10] S.S.M. Patra, K.Roy, S. Banerjee, and D.P. Viyarthi, "Improved generic algorithm for channel allocation with channel borrowing" IEEE Trans. on Mobile Computing, Vol 5, No 7, pp 884-892, Jul 2006.