Bearer Service Allocation and Pricing in Heterogeneous Wireless Networks

Wenhui Zhang
University of Stuttgart, Institute of Communication Networks and Computer Engineering
IKR, Pfaffenwaldring 47, D-70569 Stuttgart, Germany
Tel: +49 711 685 7964, E-mail: zhang@ikr.uni-stuttgart.de

***Abstract*–In next generation heterogeneous wireless networks, a user with a multi-mode terminal may have network access for different bearer services using various access technologies. In such a scenario, the user will select the best access network with good QoS and low cost. A profit-seeking network service provider, in order to maximize his revenue and be competitive in the market, has to efficiently utilize limited resources, and properly price the bearer services. Based on analysis and simulations, a simple allocation policy is proposed with the aim to maximize network capacity and user QoS, which can also simplify network selection process. In addition, the pricing policy for multi-services in heterogeneous wireless networks is also investigated in this paper.**

Keywords: heterogeneous networks, resource allocation, bandwidth degradation, pricing, QoS.

I. INTRODUCTION

Wireless communications have undergone fast growth in the last two decades. It is envisaged that in next generation wireless networks, heterogeneous wireless access technologies e.g. 2G, 3G networks and the Wireless LAN (WLAN) etc. will coexist, and provide multiple bearer services. In addition, seamless handover will be possible between these technologies, which may differ in coverage area, QoS and price. A user with a multi-mode terminal can choose the best access network any time and any where. This scenario is referred to as Always Best Connected [1], which invokes a lot of challenges from both the network and the user side. In such an interworking scenario, users will select the best access network with good QoS and low cost. A profit-seeking network service provider, in order to maximize his revenue and be competitive in the market, has to efficiently utilize limited resources in order to serve as many users as possible with high QoS. Therefore, in addition to improving the performance of each individual access technology, improving the joint performance of the networks is also an important task. Research in this area shows that interworking of heterogeneous wireless networks can reduce the call blocking and dropping probability [2], and properly allocating bearer services to heterogeneous networks can improve the total network capacity [3].

A proper price policy can never be overlooked because it directly influences user demand and network revenue. There are numerous papers on pricing for communication networks using the principles of Microeconomics [4]. Kelly et al. use congestion price for rate control in wired networks, where elastic traffic users value only the throughput [5] and dynamically change their data rate. Siris applies a similar approach in CDMA networks [6]. However, these approaches suffer from their complexity, and may not accepted by the users. A simple pricing scheme, Paris Metro Pricing (PMP) suggests to partition the network into two parts which only differ in price [7].

Pricing and resource allocation in heterogeneous networks are closely related to each other. On the one hand, allocation of bearer service should maximize the capacity and QoS of users, which has influence on price. One the other hand, price directly affects the user selection of networks and the traffic demand, and consequently, the load and QoS of the networks. The main task of this paper is to analyse combined pricing and allocation algorithms in heterogeneous networks, with the objective to jointly optimise multi-service provisioning and pricing. Section II discusses algorithms for bearer service allocation in heterogeneous networks. Section III presents the price scheme. Finally, Section IV concludes the paper.

II. BEARER SERVICE ALLOCATION

A. *Wireless networks and capacity*

The most widely used wireless communication system GSM is based on TDMA technology, and the third generation mobile communication system UMTS is based on WCDMA, which is optimised for multimedia services. Both GSM and UMTS can share the same core network, and provide ubiquitous coverage area with similar cell size. In addition, both systems provide multi-bearer services, and four QoS classes have been identified for different kinds of traffic: conversational, streaming, interactive, background classes [10]. The same QoS parameters have been standardized, which allows easy integration of both systems. Compared with cellular systems, the WLAN has smaller coverage area, and is suitable for best-effort high data rate services. It has been deployed in non-contiguous hot spot areas for wireless Internet access, and there is a trend to integrate the WLAN and cellular systems [11].

The capacity of a certain bearer service in a wireless network can be measured as the total throughput when certain ratio of users have satisfying QoS level prescribed by the network [12]. Comparing the capacities of different systems is inherently difficult due to technology differences. However, some preliminary results of service capacities in GSM and UMTS are available. Rysavy shows that GPRS is relative good at supporting low rate data services, and UMTS is relative efficient for high data rate services [13]. Comparable results are reported by evaluating the performance of UMTS and GPRS for WWW data service [14]. Furuskär shows similar results [3], and indicates that in a multi-service wireless network, linear capacity region between single bearer service capacity end points can be achieved as shown in Fig. 1, where the capacity for voice and high rate data service in GSM are C_1^1 and C_2^1 respectively; the capacity for voice and high rate data service for UMTS are C_1^2 and C_2^2 respectively, with

$$C_1^1/C_2^1 > C_1^2/C_2^2 . \qquad (1)$$

Mathematically, when a network supports N different bearer services with linear capacity region, its load ρ in percentage can be described by (2), where x_j and C_j are the throughput and total capacity of the service type j.

$$\rho = \sum_{j=1}^{N} \frac{x_j}{C_j} \qquad (2)$$

When capacity region is not linear, it can be approximated by several linear parts. For simplicity, linear capacity region is assumed in this paper.

B. Capacity based bearer service allocation

Since each network has different efficiency in supporing different bearer services, the maximum capacity can be achieved by properly allocating bearer services in different networks. Assume a set of M networks, each has different capacities for a set of N bearer services. The problem can be formulated as properly allocating the bearer service type j in networks, so that the total capacity is maximized. That is

$$\begin{aligned} &\text{max.} \quad x_j \quad j = 1, 2, \ldots, N, \\ &\text{s.t.} \quad \rho_i = \rho_i(x_1^i, x_2^i, \ldots, x_N^i) \le 1, \quad i = 1, 2, \ldots, M. \end{aligned} \quad . (3)$$

This problem belongs to the General Assignment problems, which are *NP*-complete. However, when the allocation of each bearer service in each network, x_j^i, is small enough, the optimal solution can be found by comparing the relative efficiency of the networks. Intuitively, for each single bearer service x_j, it should be allocated to the network with the smallest marginal rate of load increase, so that the network load can be kept at a minimum level. With linear capacity region (2), the marginal load increase in network i when bearer service j is allocated can be calculated as

$$\frac{\partial \rho^i}{\partial x_j} = \partial \sum_{j=1}^{N} \frac{x_j^i}{C_j^i} / \partial x_j^i = \frac{1}{C_j^i} . \qquad (4)$$

However, when different services have to be allocated in one network, the influence of one service on other services has to be considered. We can use a scaling factor λ^i for each network, which makes the marginal load increase of each service in all networks comparable. The scaling factor is calculated as the marginal load increase of a network i for all services shown in (5).

$$\lambda^i = \frac{\partial \rho^i}{\partial x^i} = \partial \sum_{j=1}^{N} \frac{x_j^i}{C_j^i} / \partial \sum_{j=1}^{N} x_j^i = \sum_{j=1}^{N} \frac{1}{C_j^i} . \qquad (5)$$

Combing (4) and (5), we get the relative load increase for the service j in the network i:

$$\frac{1}{C_j^i} / \sum_{j=1}^{N} \frac{1}{C_j^i} \qquad i = 1, 2, \ldots, M . \qquad (6)$$

Thus, for each bearer service, the network with the smallest marginal value in (6) should be first allocated to. Take the example of Fig. 1, and combine (1) and (6), the results show that GSM has smaller marginal load increase for the voice service, and UMTS has smaller marginal load increase for the high rate data service. Thus voice users should be allocated as much as possible to GSM, and data users to UMTS. Fig. 1 shows the maximum combined capacity as well as the minimum capacity region when the opposite allocation is applied. It should be noted that, though the final allocation result is the same with the example in [3], but the method here is a more general.

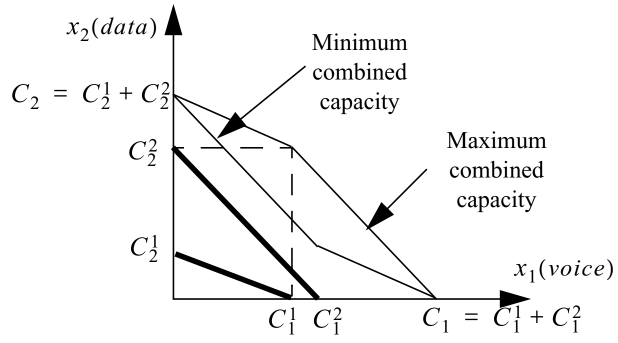

Fig. 1. Bearer service allocation in GSM and UMTS

C. Performance based bearer service allocation

In this section, different methods of improving the QoS of bearer services are studied, and a simple allocation policy is proposed. For simplicity and without losing generality, two real time services and one non-real time service are considered, i.e. the voice, high bandwidth streaming and elastic data service. Three possible approaches to increase QoS are considered. The first is bandwidth adaptation of real time services, where an adaptive service can degrade its bandwidth in case of congestion, which is effective in mitigating the varying level of resource availability and mobility [15][16]. The second is the integration of traffic from different kinds of bearer services. Since multiple bearer services will coexist in heterogeneous networks, thus the integration performance is also important. The third is overflowing traffic of different kinds of bearer services from one network to another, which is special feature in heterogeneous wireless network.

Due to the complexity of the problem, simulations have been carried out to evaluate the performance. Assume each wireless cell has the same coverage area, six neighbour cells, and the capacity of 32 units bandwidth. 19 adjacent cells have been simulated, and a wrap around model is used in order to eliminate the border effect. User movement is characterized by the dwell time in a cell, which is assumed to have a Lognormal distribution based on the reported measurement data [17], and a small dwell time mean of 15s is used in the simulation. The voice and streaming service are characterized by Poisson arrival, and exponentially distributed service time duration with mean 60s. Their maximum bandwidth requirements are 1 and 4 units respectively. Elastic data traffic is modelled based on a WWW traffic model introduced in [18], data requests have a Poisson session arrival process, each session contains a certain number of page requests separated by certain thinking time. Only flow level performance of the data service is studied, assuming data users share unused bandwidth equally in the range between its maximum and minimum bandwidth.

For real time services, a call can be lost either due to blocking at call initiation or dropping at handover. Simulation results show that the high bandwidth steaming service suffers from a high loss probability even at very low load; and mobility has little influence on the data service thanks to its elastic nature. The details are not shown for space reason.

A simple bandwidth degradation scenario for real time services is simulated, i.e. in case of congestion, real time services can reduce the maximum bandwidth to the half of it. Two performance metrics for degradation are studied: *bandwidth degradation* refers to the average reduced bandwidth normalized by its maximum bandwidth; *degradation probability* refers to the percentage of users, who have experienced degradation. Fig. 2 shows that degradation reduces the loss probability considerably, meanwhile, users have relative low bandwidth degradation but a high degradation probability.

Next, we present the integrated performance of bearer services. It is assumed that all bandwidth is completely shared by traffic from all bearer services, and real time traffic has priority over data traffic. Fig. 3 compares two integration scenarios. The degradation of the streaming traffic is reduced when it is integrated with the data traffic; and it has no performance improvement when integrated with voice. Actually, the average data rate of the data traffic mainly depends on the load of the network, and will not suffer much from the integration with the streaming traffic, details are not shown here.

To compare different overflow scenarios, we consider two layers of overlay networks, e.g. GSM and UMTS networks. These two types of networks can have the same and fully overlapped coverage area, which is technically feasible, and may also save installation cost. For simplicity, both systems are assumed to have the same capacities for all types of services. The preferred network for voice is GSM, for the streaming and data service is UMTS. Four different overflow scenarios are compared: only allow users of one of the three types of services select the least loaded network when they enter a cell, and other users overflow to the less preferred network only to avoid degradation or loss; allow all users select the least loaded network when they enter a cell. The performance comparison of the streaming service is illustrated in Fig. 4. All four types of overflow scenarios improve the performance as compared to the results in Fig. 3, revealing the benefit of overflow, and allow only data users to make overflow outperforms other scenarios.

Based on the results above, a simple bearer service allocation policy for GSM/UMTS networks can be derived, i.e. allocating voice users to GSM and high bandwidth streaming users to UMTS, while allowing data users to choose the least loaded network. Consequently, the number of overflow is reduced and network selection can be simplified, in that overflow is mainly limited to data users, and real time service users stay in the preferred networks whenever possible. Actually, this result can also apply to the interworking of other types of cellular networks.

III. PRICING BEARER SERVICES

A. *Pricing in a capacity limited network*

In wireless networks, we assume that the aggregate traffic demand of users for a certain service depends only on its price, and it can be described by a constant price elasticity model proposed in [19]:

$$x_j = A_j p_j^{-\varepsilon_j}. \qquad (7)$$

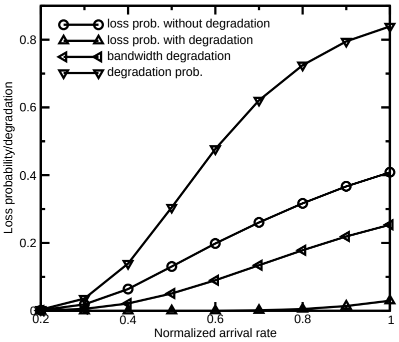

Fig. 2. Streaming service performance improvement by degradation

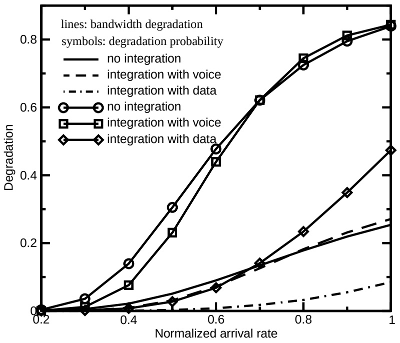

Fig. 3. Streaming service performance of different integration scenarios

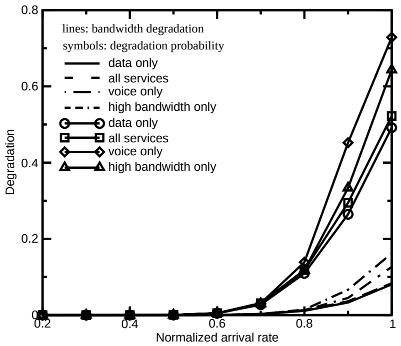

Fig. 4. Streaming service performance of different overflow scenarios

and ε_j are the demand and the price elasticity of the service type *j*, and A_j is the demand potential. Estimations show that the demand elasticity for data services is higher than that of the voice service, and both are approximately constant and greater than one. That is, a reduction in price leads to relative larger increase in demand and thus increases network revenue. Normally, the capacity of wireless networks remains stable within a certain period of time. The network revenue *R* with the capacity constraint shown in (8) is differentiable and concave, its maximum value can be calculated by defining a Lagrangian *L* [20], as shown in (9).

$$\text{max.} \qquad R = \sum_{j=1}^{N} x_j p_j = A_j^{1/\varepsilon_j} x_j^{1-1/\varepsilon_j} \tag{8}$$
$$\text{s.t.} \quad \rho = \rho(x_1, x_2, \ldots, x_N) \le 1$$

$$L = \sum_{j=1}^{N} x_j p_j - \lambda(1-\rho) \tag{9}$$

Assuming linear capacity region characterized by (2), the condition for maximum revenue is

$$\frac{\partial L}{x_j} = 0 \Rightarrow \frac{\partial R_j}{x_j} = p_j\left(1 - \frac{1}{\varepsilon_j}\right) = \frac{\lambda}{C_j} \Rightarrow p_j = \frac{\lambda}{C_j}\frac{\varepsilon_j}{\varepsilon_j - 1}\,. \tag{10}$$

The result in (10) indicates that network revenue is maximized when the marginal revenues of all services are the same. Services with low price elasticities are charged with high prices, which is called price discrimination [4], and the bearer services which are more efficient in the network are charged relatively less.

B. *Pricing bearer services in GSM and UMTS*

Different capacities of GSM and UMTS may lead to different prices for the same kind of bearer service. One possible policy is to charge different prices for the same bearer service in order to motivate users to use the more efficient network. This has the consequence that the cheaper network will attract more users and thus will get more congested. In fact, it is similar to PMP [7], which partitions the network into two parts differing only in price. Optimal PMP performance depends on optimal prices and optimal partition of the total capacity, which might not be feasible in an integrated GSM and UMTS networks. Critics on PMP argue that PMP is inefficient in an competitive environment [8], and a network may have lower revenue by implementing PMP [9]. In addition, charging the same price for the same service allows easy charging and network selection. So in the following, we only consider the same price for the same service. We assume a service provider can set the service price, and the competition between service providers will be a topic in the future.

As revealed in Section II, the streaming service has low trunking efficiency in wireless networks. Without degradation, in order to keep an acceptable low loss probability, the offered load of the streaming service has to be kept low. Degradation can increase its load, but high degradation also leads to reduced QoS, and thus reduces revenue. This is illustrated in the following example. Assume the price elasticity for the streaming service is similar as voice, and the elasticities for it and the data service are 1.1 and 1.5 respectively based on the reported estimation [19]. Two pricing scenarios are compared, the first case is that the streaming and data service have the same price, and the second case is that the streaming and data service are priced according to their price elasticities. In order to compensate the reduced QoS due to degradation, degraded users are given 20% discount off the normal price. Fig. 5. shows that the revenue increases with the demand potential, and the network has higher revenue for case 2. The reason is that for case 1, the price for the streaming service is relative low, thus leads to high demand and high degradation; while for case 2, the price for the streaming service is higher due to its low price elasticity, thus the demand of it is kept at a low level.

We propose that the pricing for bearer services should be based on the price elasticity, demand potential, as well as the network capacity. In addition, the QoS of bearer services has to be considered, because low QoS reduces user satisfaction and may lead to reduced revenue. The performance results of the streaming service indicate that there is a trade-off between the capacity and degradation, the higher the capacity, the higher the degradation. Thus, a good understanding of user behaviour with respect to degradation is important in finding the optimal price. However, only limited preliminary result is reported [16], and further research is still required.

C. *WLAN pricing*

In this section, we discuss the pricing for cellular networks and WLAN interworking. We assume both cellular networks and the WLAN are owned by the same service provider. Due to its high capacity and low cost, the service price of the WLAN can be lower. Suppose there are *y* WLAN cells deployed within a cellular cell, each WLAN cell has a relative coverage area compared with the cellular cell, denoted as *m*, normally we have $m \ll 1$. Suppose users are homogeneously distributed, and will use the WLAN when they are in its cov-

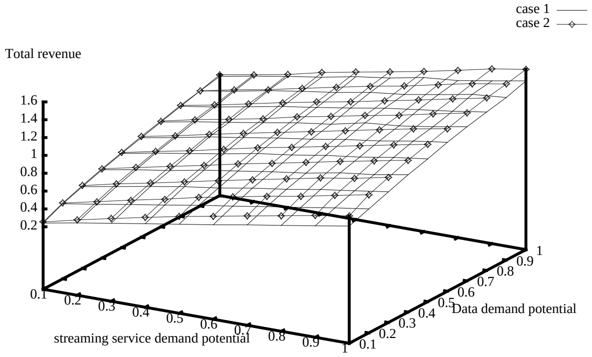

Fig. 5. Network revenue comparison

erage area. So the demand potential in the cellular cell and in the WLAN cell are $A(1-m)$ and Am respectively. The total revenue R_N is the sum of the revenue from the cellular network and the WLAN shown in (11), by using the result from (8), where C_C and C_W are the capacity of a cellular and WLAN cell respectively.

$$R_N = [(1-ym)A]^{1/\varepsilon}{C_C}^{1-1/\varepsilon} + (mA)^{1/\varepsilon}{C_W}^{1-1/\varepsilon} \tag{11}$$

When $m \ll 1$, using Maclaurin series, (11) can be approximated by

$$R_N \approx A^{1/\varepsilon}{C_C}^{1-1/\varepsilon}\left[1-\frac{ym}{\varepsilon}+ym^{1/\varepsilon}\frac{{C_W}^{1-1/\varepsilon}}{C_C}\right]. \tag{12}$$

The part in the parenthesis in (12) can be interpreted as the ratio of the total revenue to the revenue in the cellular system. (12) indicates that the revenue increases approximately linearly with the number of the WLAN cells, and more revenue can be obtained when the demand potential is high.

Suppose the average cost of unit bandwidth of cellular system is K_{CU}, and the cost of each WLAN access point is K_{WU}. When the user demand in cellular networks increases from C_C to C_C', The cost of extra cellular capacity is shown in (13), and the number of WLAN cells required to accommodate the extra demand in cellular systems can be calculated using (14). Combining (13) and (14), the relative cost of installing extra capacity in cellular systems and WLAN is derived in (15), which reveals that the higher the demand in cellular systems, it is more beneficial to use WLAN to increase the capacity of cellular systems.

$$K_C = K_{CU}(C_C' - C_C) \tag{13}$$

$$C_C'(1-ym) = C_C \Rightarrow y = \frac{C_C' - C_C}{mC_C'} \tag{14}$$

$$\frac{K_C}{K_W} = \frac{K_{CU}(C_C' - C_C)}{yK_{WU}} = \frac{K_{CU}C_C'm}{K_{WU}} \tag{15}$$

Both (12) and (15) show that a high relative coverage of WLAN *m* provides more revenue and saves installation cost. However, the physical coverage of WLAN cells is limited, an equivalent method to increasing its coverage is to increase its usage, e.g. charging a lower price in WLAN to encourage the usage of the WLAN, and postponing non-real time services in cellular networks till users can access the WLAN.

IV. CONCLUSIONS AND FUTURE WORKS

This paper describes bearer service allocation and pricing in next generation networks. Analysis and simulation results show that real time services should be allocated to the capacity efficient network and non-real time data service should be allowed to select the least loaded network. In cellular networks, the same bearer service should be charged the same price based on the its price elasticity and QoS. In addition, preliminary analysis shows that the WLAN can share the load and increase the revenue of cellular networks.

In order find the optimal price, a proper understanding of the QoS influence on user demand is necessary, and further research in this direction is still required. In addition, the pricing policy considering competition between service providers is also a topic in the future.

REFERENCES

[1] Gustafsson, E., Jonsson, A., “Always best connected, “IEEE Wireless Communications, Vol. 10, Issue 1, pp. 49 - 55, Feb. 2003.

[2] A. Tölli, P. Hakalin and H. Holma, “Performance of Common Radio Resource Management (CRRM),” Proceedings of IEEE International Conference on Communications 2002, pp. 3429-3433, Vol. 5, 2002.

[3] A. Furuskär, “Allocation of Multiple Services in Multi-Access Wireless Networks,” Proceedings of the IEEE Mobile and Wireless Communication Networks conference 2002, pp. 261-265, Sep. 2002.

[4] P.R.G. Layard, A.A.Walters, “Microeconomics Theory,“McGraw-Hill, 1978.

[5] F. P. Kelly, A. K. Maullo, and D. K. H. Tan, “Rate control in communication networks: Shadow prices, proportional fairness and stability,” J. Oper. Res. Soc., Vol. 49, pp. 237--252, 1998.

[6] Vasilios A. Siris, “Resource control for elastic traffic in CDMA networks,” Proceedings of the 8th annual international conference on Mobile computing and networking, Atlanta, USA, pp. 193 - 204, 2002.

[7] Odlyzko, A., “Paris Metro Pricing for the Internet,” Proceedings of the first ACM conference on Electronic commerce, pp.140-147, Nov. 1999.

[8] Gibbens, R., Mason, R., and Steinberg, R. “Internet service classes under competition,” IEEE Journal on Selected Areas in Communications, Vol. 18, No. 12, pp. 2490-2498, Dec. 2000.

[9] David Ros, and Bruno Tuffin, “A mathematical model of the Paris Metro Pricing scheme for charging packet networks,” Computer Networks, Volume 46, Issue 1, pp. 73-85, Sep. 2004.

[10] 3GPP TS 23.107 V5.10.0, “Quality of Service (QoS) Concept and Architecture,“ Sep. 2003.

[11] Doufexi, A., Tameh, E. Nix, A. Armour, S. and Molina, A., “Hotspot wireless LANs to enhance the performance of 3G and beyond cellular networks,” IEEE Communications Magazine, Vol. 41, Issue 7, pp. 58-65 Jul. 2003.

[12] 3GPP TR 101 112 V3.2.0, “Selection procedures for the choice of radio transmission technologies of the UMTS (UMTS 30.03 version 3.2.0),” Apr. 1998.

[13] Peter Rysavy, “Data Capabilities for GSM Evolution to UMTS,” 3G Americas white paper, Nov. 2002.

[14] Bernhard H. Walke “Mobile Radio Networks, Networking, Protocols and Traffic Performance,” second edition, John Wiley&Son Ltd, 2002.

[15] Naghshineh, M., and Willebeek-LeMair M., “End to end QoS provisioning multimedia wireless/mobile networks using an adaptive framework,” IEEE Comm. Mag., Vol.: 35, Issue: 11, pp. 72-81, Nov. 1997.

[16] Das, S.K., Sen, S.K., Basu, K., and Haitao Lin, “A framework for bandwidth degradation and call admission control schemes for multiclass traffic in next-generation wireless networks,” IEEE JSAC, Volume: 21, Issue: 10, pp:1790 - 1802 Dec. 2003

[17] K. Saitoh, H. Hidaka, N. Shinangawa, and T. Kobayashi, “Vehicle Motion in Large and Small Cities and Teletraffic Characterization in Cellular Communication Systems,” IEICE Transaction on Communications, Vol. E84-B No.4 pp.805-813, Apr. 2001.

[18] P. Tran-Gia, D. Staehle, and K. Leibnitz. “Source Traffic Modeling of Wireless Applications,” International Journal of Electronics and Communications (AEÜ), Vol. 55, issue 6, pp. 27-36, 2001.

[19] S. Lanning, D. Mitra, Q. Wang, and M. Wright, "Optimal planning for optical transport networks," Philosph. Trans. Royal Soc. London A, Vol. 358, No. 1773, pp. 2183-2196, Aug. 2000.

[20] M.S. Bazarra, H.D. Sherali, and C.M. Shetty. “Nonlinear Programming, Theory and Algorithms,“ John Wiley, second edition, 1993