Performance of Real-time and Data Traffic in Heterogeneous Overlay Wireless Networks

Wenhui Zhang

University of Stuttgart
Institute of Communication Networks and Computer Engineering
Pfaffenwaldring 47, D-70569 Stuttgart
e-mail: zhang@ikr.uni-stuttgart.de

Abstract: In next generation heterogeneous wireless networks, a user with a multi-mode terminal may have different bearer services using different access technologies. Various approaches can be applied to improve user QoS, such as bandwidth degradation for real-time services, integration of real-time and data traffic, and overflow traffic in overlay networks. In this paper, new techniques are derived to analyse real-time and data traffic performances in wireless networks, revealing the different influences of mobility on real-time and data traffic. In addition, simulations have been carried out to study the performance of the joint allocation and overflow control of real-time and data traffic. A simple bearer service allocation policy is proposed, which not only maximizes network capacity and increases user QoS, but also simplifies overflow management.

Keywords: heterogeneous networks, bandwidth degradation, resource allocation, integration, vertical handover.

1. INTRODUCTION

Wireless communications have undergone fast growth in the last two decades. The most widely used wireless communication system GSM and the third generation mobile communication system UMTS are now evolving towards integrated networks, which support multiple bearer services using different access technologies. Four QoS classes, the conversational, streaming, interactive, and background class, have been standardized by 3GPP for both systems, which allows easy integration of these two systems [1]. It is essential to improve the joint performance of both networks with respect to various types of bearer services, which covers a wide range of topics.

The performance of circuit-switched real-time traffic in wireless networks has been extensively studied in the literature. Real-time traffic typically requires certain bandwidth guarantee, and common performance metrics are the call blocking probability at call initiation and the call dropping probability at handover. To improve the QoS of real-time traffic, adaptive applications are shown to be effective, in that real-time services can degrade bandwidth in case of congestion [2], and the call blocking and dropping probability can be reduced at the cost of degraded bandwidth [3][4][5]. But degraded bandwidth also reduces the perceived QoS of users, and it may have negative influence on network revenue. It is shown that an optimal degradation policy may maximize network revenue [6]. Many analytical studies on real-time traffic in mobile networks assume Poisson arrival processes for both new calls and handover calls to allow tractable mathematical analysis, however, it is not realistic and may underestimate mobility influence on the performance [7].

Data communications are expected to play an important role in future wireless networks. Compared with real-time traffic, data traffic is elastic, i.e. it can adapt its rate to available bandwidth. Assume bandwidth is equally shared by all data flows with instantaneous adjustment as the number of flows changes, with the Poisson flow arrival process, the number of flows in progress behaves like that in the M/G/1 processor sharing queue [8], which is insensitive to the flow size distribution. Moreover, this model can accurately predict the packet level performance of TCP connections [9]. When each flow has the maximum or minimum rate limit, the Generalized Processor Sharing (GPS) model can be applied [10].

A complete analytical evaluation of integrating real-time and data traffic is shown to be very difficult even on a single link, and performance bounds have been investigated in [11], showing the integration gain in access networks. Analysis becomes more complicated when mobility is introduced. In most reported studies, Markovian processes are assumed for mathematic tractability [6][12], which are based on simple traffic and mobility models. In heterogeneous wireless networks, total carried traffic can be improved by properly allocating bearer services to more capacity-efficient networks [13][14]. The performance of bearer services can further be improved by overflowing a connection from a congested network to a less congested one by means of load balancing as reported in [15][16].

The contributions of this paper are twofold. Firstly, new techniques are applied to analyse real-time and data traffic performance in wireless networks, revealing the different influences of mobility on real-time and data traffic. Secondly, based on performance results of the joint allocation and overflow control of real-time and data traffic, a simple bearer service allocation policy is proposed, which not only maximizes network capacity and increases user QoS, but also simplifies overflow management. This paper is organised as follows: Section 2 shows the flow level performance of real-time and data traffic in wireless networks using both analytic and simulation approaches. Section 3 proposes a simple allocation policy by comparing different integration and overflow scenarios. Section 4 concludes this paper.

2. FLOW LEVEL PERFORMANCE

In this section, new techniques are derived to analyse real-time and data traffic performance using simple traffic and mobility models. In addition, simulations are used to validate the analysis and to study the performance using more general mobility and traffic models.

2.1 Real-time traffic

Assume circuit-switched real-time traffic is allocated with the maximum bandwidth for normal operation and may degrade bandwidth in case of congestion. For simplicity, call blocking and dropping are not distinguished, and both of them are denoted as call loss. In addition, two other performance metrics are evaluated: *bandwidth degradation* stands for the average reduced bandwidth of a real-time call normalized by its maximum bandwidth, and *degradation probability* stands for the percentage of users who experience bandwidth degradation. Normally, for real-time services, bandwidth degradation can be done in several discrete steps [2], and the smaller the step, the less influence users will perceive, but more frequently users will

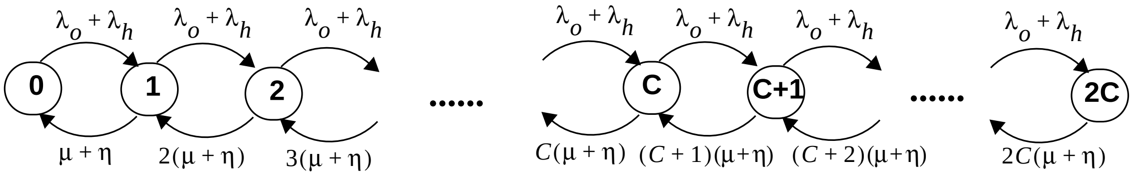

Fig. 1: State transition diagram

be involved [5]. We argue that frequent bandwidth degradation could deteriorate the perceived QoS and increase network signalling, therefore, the degradation frequency and the percentage of degraded users should be limited. In this paper, we consider a simple case that a user only degrades bandwidth from the maximum bandwidth b_{max} to the half of it, $b_{max}/2$, and upgrades bandwidth only when the user moves to a new cell with enough bandwidth.

Assume there is a maximum of C full rate channels or $2C$ half rate channels in a cell. New call arrivals in the cell follow the Poisson process at rate λ_o, and the call duration and cell dwell time are both exponentially distributed with mean $1/\mu$ and $1/\eta$ respectively. Calls in a cell are a mixture of both new and handover calls. Due to the memoryless property of the Exponential distribution, the average call duration and dwell time of handover calls are the same as those of new calls, i.e. each call either leaves the system with rate $1/\mu$, or makes handover to neighbouring cells with rate $1/\eta$. Therefore, the attained service time for each call is the minimum of the call duration and dwell time. Since both the call duration and cell dwell time are exponentially distributed, the attained service time for each call is again exponentially distributed with the termination rate $\mu + \eta$. When there are x active users in a cell, it follows that the handover departure rate is $x\eta$, and the total departure rate is $x(\mu + \eta)$. With the above assumptions, the handover departure process from a cell is a Poisson process [7]. The handover arrival in a cell from one of its neighbours is the decomposition of handover departures from that neighbouring cell. It is known that the decomposition of a Poisson process results in Poisson processes; moreover, the total handover arrivals in a cell are the composition of all Poisson arrivals from all neighbouring cells, which again form a Poisson process. If we assume a homogeneous cell deployment, and each cell has K neighbours, and statistically the same average number of active users $E[X]$, the probability that a handover call arrives in a neighbouring cell is $q = 1/K$, thus, the handover arrival rate in a cell λ_h is

$$\lambda_h = K \cdot \eta E[X] \cdot \frac{1}{K} = \eta E[x]. \qquad (1)$$

Actually, if cells are homogeneously deployed, each cell has statistically the same behaviour. When the system reaches a stationary state, the average handover departure rate in a cell must be equal to the average handover arrival rate. It holds also when the handover departure process is not a Poisson process. The total call arrival rate in each cell is the sum of both the new and handover arrival rate i.e. $\lambda = \lambda_o + \lambda_h$. The state transition diagram for the number of active users in a cell is shown in Fig. 1, and the well known Erlang-B formula can be applied to calculate the state probability:

$$\pi_x = \frac{(\lambda_o + \lambda_h)^x}{x!(\eta + \mu)^x} \Big/ \sum_{i=0}^{2C} \frac{(\lambda_o + \lambda_h)^i}{i!(\eta + \mu)^i} \qquad 0 \le x \le 2C \qquad (2)$$

The calculation of λ_h here uses a different method, but yields the same result as proposed by Lin et al. [3], and the loss probability due to blocking and dropping is calculated as follows using their result:

$$p_l = 1 - \frac{1 - \pi_{2C}}{1 - \eta\pi_{2C}/\mu}. \tag{3}$$

To calculate the state probability, they also proposed an iteration procedure, however, for certain values in our analysis, it does not converge. An improved procedure is as follows.

Step 0. Let $\lambda_h = 0, \delta = 1$.
Step 1. If $|\delta| < 0.00001\lambda_h$, finish, otherwise go to Step 2.
Step 2. Compute the state probability using (2).
Step 3. Compute λ_h' using (1), and let δ be the difference between λ_h' and the old λ_h, the new λ_h is set to be $1/n$ of the sum of λ_h' and the old λ_h, where $n > 1$. Go to Step 1.

Fig. 2: Iteration method to calculate state probability

2.2 Elastic data traffic

In this section, we derive new techniques to calculate the average data rate of elastic data traffic in wireless networks. Assuming an ideal fluid model for bandwidth sharing, where bandwidth is equally shared by all data flows with instantaneous adjustment as the number of flows changes. Generally, the maximum data rate of a user is limited, in addition, the minimum rate for data flows should also be ensured since extremely low data rate might be interpreted as broken of a connection and lead to manual disconnection by users. Bandwidth sharing on a link with rate limit can be modelled as the Generalized Processor Sharing (GPS). Assume a server has a service rate C, and users have the maximum rate limit r_{max} and the minimal rate limit r_{min}. Let $N = \lfloor C/r_{max} \rfloor$, $M = \lfloor C/r_{min} \rfloor$ ($\lfloor y \rfloor$ stands for the largest integer less than or equal to y), and x be the number of flows present at a server, the service rate for each of the flow is

$$f(x) = \begin{cases} r_{max}, & x \le N, \\ C/x, & N < x \le M, \end{cases} \tag{4}$$

and according to the GPS model, define the function $\phi(x)$ as follows:

$$\phi(x) = \begin{cases} \left\{ \prod_{i=1}^{x} f(i) \right\}^{-1}, & 1 \le x \le M, \\ 1, & x = 0. \end{cases} \tag{5}$$

Assume flow arrivals at a server follow the Poisson process at rate λ_o, each flow has an independent size distribution with mean σ, and let $\rho = \lambda_o\sigma$. Following Cohen's results [10], the probability that there are x flows in progress reads as

$$\pi_x = \frac{\rho^x\phi(x)}{x!} \Big/ \sum_{k=0}^{M} \frac{\rho^k\phi(k)}{k!}, \qquad 0 \le x \le M. \tag{6}$$

The average service time of all flows is

$$E[T] = \sigma \sum_{x=0}^{M-1} \frac{\rho^x \phi(x+1)}{x!} \Big/ \sum_{x=0}^{M} \frac{\rho^x \phi(x)}{x!}. \tag{7}$$

Combine (6) and (7), using simple mathematical manipulations, (7) can be rewritten as

$$E[T] = E[X]\frac{\sigma}{\rho} = E[X]/\lambda_o, \tag{8}$$

which is actually Little's Theorem, and the average data rate R can be derived:

$$R = \sigma/E[T] = \rho/E[X]. \tag{9}$$

In the following, we derive the average data rate of data flows in wireless networks. Assume that new flow arrivals in a cell follow the Poisson process at rate λ_o, the flow size and dwell time of each user have the Exponential distribution with mean σ and $1/\eta$ respectively. Each cell behaves like a GPS node in open and closed queueing networks, in that external flow arrivals follow the Poisson process, and internal flow arrivals are the result of handover from neighbouring cells. Flows in a cell are a mixture of new arrivals and handover flows. Each flow leaves a cell either because it makes handover or obtained its full required service size, thus, both the arrival rate and service rate for a flow in a cell are different from those on a single server. When there are x active flows, each flow has a service rate $f(x)$, without mobility, the mean termination rate for all flows is $\varepsilon_x = f(x)/\sigma$. When mobility is introduced, using the similar reasoning as in Section 2.1, the attained service time for each flow in each cell is the minimum of the residual life time and dwell time, i.e. the attained service time for each flow is exponentially distributed with the termination rate $\mu_x = \varepsilon_x + \eta$. Denote $g(x)$ as the service rate for each of the flow when there are x flows, and define a new function $\phi'(x)$, we have

$$g(x) = \mu_x\sigma = (\varepsilon_x + \eta)\sigma = f(x) + \eta\sigma, \text{ and} \tag{10}$$

$$\phi'(x) = \begin{cases} \left\{ \prod_{i=1}^{x} g(i) \right\}^{-1}, & 1 \le x \le M, \\ 1, & x = 0. \end{cases} \tag{11}$$

The total flow arrival rate is the sum of both the new and handover arrival rate, i.e. $\lambda = \lambda_o + \lambda_h$. Let $\rho' = \lambda\sigma$, the probability that there are x flows in progress is:

$$\pi_x' = \frac{\rho'^x \phi'(x)}{x!} \Big/ \sum_{k=0}^{M} \frac{\rho'^k \phi'(k)}{k!}, \qquad 0 \le x \le M. \tag{12}$$

If we assume homogeneous cell deployment, the average handover arrival rate in a cell is the same as the average handover departure rate out of a cell i.e. $\lambda_h = \eta E'[X]$. The state probability can be calculated recursively using the same iteration method as in Fig. 2. Since homogeneous cell deployment is assumed, the average numbers of users in all cells are the same. Consider a system with total number of y homogeneous cells, the external arrival rate to the system is $y\lambda_o$, and the average number of users in the system is $yE'[X]$, thus, the sojourn time of all flows can be calculated using Little's formula, and the average data rate R' is given by

$$R' = \sigma/E'[T] = y\lambda_o \cdot \sigma/yE'[X] = \rho/E'[X]. \tag{13}$$

2.3 Numerical results

The analysis above assumes the call duration and cell dwell time of real-time traffic are both exponentially distributed. When the cell dwell time is not exponentially distributed, handover arrivals do not form the Poisson process [7], thus the Erlang-B formula cannot be applied. In addition, it is also not realistic to assume exponentially distributed flow sizes for data traffic. In this section, simulations have been carried out to validate the analytic results and also to study the performance of real-time and data traffic using more realistic mobility and traffic models. We assume each wireless cell has the same coverage area, six neighbour cells, and 32 units bandwidth. 19 adjacent cells are simulated using a wrap-around technique to eliminate the border effect. User movement is characterized by the cell dwell time, and the distribution of the dwell time is assumed to have the Lognormal distribution based on reported measurement [17]. Extensive simulations have been carried out and due to space reason, only the results for cell dwell time distribution with mean 15s and Coefficient of Variance (CV) 10 are shown here. Two types of real-time traffic, voice and high bandwidth streaming traffic, are characterized by Poisson arrivals, the maximum bandwidth requirement of 1 and 4 units respectively, and exponentially distributed service time duration with mean 60s. In case of congestion, real-time services can reduce the maximum bandwidth to half of it. The data traffic model is based on the WWW traffic model in [18], it has the Poisson session arrival process, each session consists of a certain number of flows separated by certain viewing time. Only the flow level performance is studied, assuming each data flow shares bandwidth unused by real-time traffic equally between the maximum of 8 units and the minimum of 0.5 unit bandwidth.

Simulations show that the voice and streaming traffic have similar performance results, since the voice traffic requires low bandwidth, it has higher trunking efficiency, thus a low loss probability and degradation compared with the streaming traffic at the same traffic load. Fig. 3 shows the loss probabilities of the streaming traffic for different dwell time distributions. When the dwell time has the Exponential distribution, the analytic result matches the simulation result well. The streaming traffic suffers from a high loss probability even at very low load, and mobility leads to a higher loss probability especially when the dwell time has the Lognormal distribution. Fig. 4 demonstrates that the loss probability of the streaming traffic is reduced considerably by bandwidth degradation; meanwhile, users have relative low bandwidth degradation but a high degradation probability. Therefore, without degradation, in order to keep a low level of loss probability for real-time traffic, especially streaming traffic, its offered load has to be kept low. Though degradation can reduce the loss probability, a high percentage of users are involved, which can also reduce QoS experienced by users.

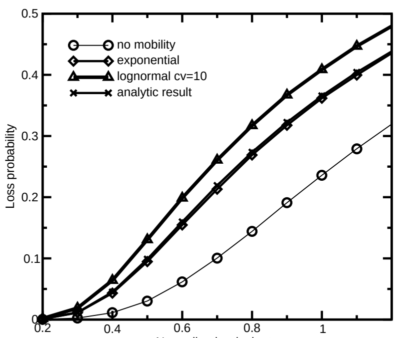

Fig. 3: Mobility influence on streaming traffic

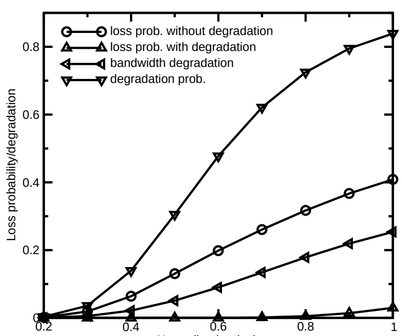

Fig. 4: Streaming traffic degradation performance

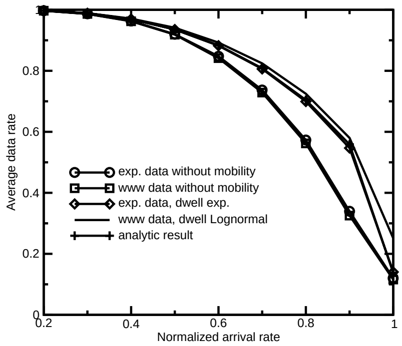

Fig. 5: Mobility influence on data traffic

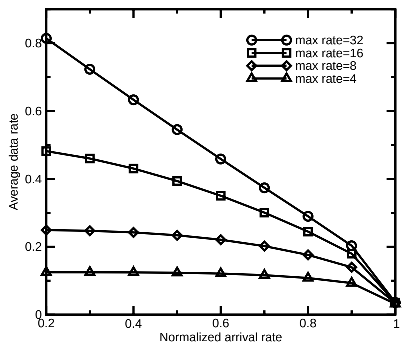

Fig. 6: Maximum rate limit influence on data traffic

Compared with the real-time traffic, mobility has little influence on the data traffic thanks to its elastic nature. We only show the average data rate here since the loss probability of the data traffic is very low in our simulation results. Fig. 5 presents the average data rate for different mobility and traffic models. Without mobility, the WWW traffic model and the model with exponentially distributed flow sizes have identical results, i.e. the average data rate depends mainly on the network load and is insensitive to the traffic type, which conforms with the insensitivity property [9]. When the cell dwell time is exponentially distributed, the analytic result matches the simulation result well, indicating that mobility even improves the average date rate, and the results do not differ much when the WWW traffic model is applied and the dwell time has the Lognormal distribution. Fig. 6 illustrates the influence of the maximum data rate limit on the average data rate. Clearly, with a high rate limit, the average data rate is mainly affected by the network load; and with a low rate limit, the maximum rate limit is the determinate factor of the average data rate. Moreover, as indicated in Section 2.2, the handover rate is determined by the average number of data users in the system, and (9) indicates that the average data rate is inversely proportional to the average number of users, thus, a lower average data rate also implies a higher handover rate. Therefore, the performance of data traffic is quite good as long as the network load is not too high, and it may not necessary to manually limit the maximum rate of a user for the benefit of both users and networks.

3. ALLOCATION PERFORMANCE IN OVERLAY NETWORKS

The above studies only focus on the individual traffic type, and reveal the different influences of mobility on real-time and data traffic. In heterogeneous networks, multiple bearer services will coexist, and two special features need to be highlighted. One feature is that different networks may have different efficiency in supporting traffic of different bearer services, and the maximum network capacity can be achieved by properly allocating different types of traffic to more efficient networks [13][14]. For example, voice traffic should be allocated to GSM, and high data rate traffic to UMTS. The other feature is overflowing a call from one network to another, which can be done either at the initiation of a call or by vertical handover during a call, and it is shown that overflow can improve the performance of real-time and data traffic [15][16]. There are many possibilities to allocate different types of traffic, and it is yet not clear if integrating different types of traffic based on capacity efficiency is beneficial when the stochastic nature of traffic and mobility are considered. In addition, as discussed in Section 2, handover has different influences on the real-time and data traffic, which suggests that vertical handover of different types of traffic may have different performance results. Therefore, it is necessary to evaluate the performance of allocating different types of traffic in heterogeneous networks. In this section, different allocation and overflow scenarios are compared with the focus on the performance of the streaming traffic, since it has a high loss probability and degradation. Due to the complexity of the problem, simulations are applied to study the performance, and analytical study is left for the future. Based on the results, a simple allocation policy is proposed, and the benefits are also discussed.

3.1 Performance of integration and overflow

For simplicity and without losing generality, three types of traffic, voice, streaming and elastic data traffic are simulated, assuming all bandwidth is completely shared by all traffic flows, and the real-time traffic has priority over the data. Three allocation scenarios are compared: allocating only the streaming traffic in a network; integrating it with either the voice or data traffic in a network, with each type of traffic amounting to 50% of the total traffic. The loss probability and degradation of the streaming traffic are compared as shown in Fig. 7 and Fig. 8. Obviously, the loss probability, bandwidth degradation and degradation probability of the streaming traffic are reduced considerably when it is integrated with the data traffic. Nonetheless, the average data rate of the data traffic is only slightly affected by the integration. By comparison, the streaming traffic can not benefit from the integration with the voice traffic.

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 [1]. For simplicity, we assume both systems have the same capacity for all types of traffic. Three types of traffic, voice, streaming and data traffic are simulated, and each accounts for 40%, 30% and 30% of the total traffic load. The preferred network for the voice traffic is GSM, for the streaming traffic and data traffic is UMTS. Four different overflow scenarios are compared. One scenario is to allocate all types of traffic to the least loaded network when they enter a cell, i.e. all types of traffic can be used for load balancing. Among the other three scenarios, each scenario only uses one type of traffic for load balancing, i.e. only one type of traffic is allocated to the least loaded network, and other two types of traffic are allocated to their preferred networks. The performance comparison of the streaming traffic is illustrated in Fig. 10 and Fig. 11. All four scenarios improve the performance as compared with the results in Fig. 7 and Fig. 8, revealing the benefit of overflow, and allocating only data users to the least loaded network outperforms other scenarios. The reason can be attributed to the elastic nature of the data traffic, when data flows enter a less loaded network, they can efficiently grab more bandwidth compared with real-time traffic flows. The results imply that it might not be necessary to allow all traffic to select the least loaded network; when there is certain amount of

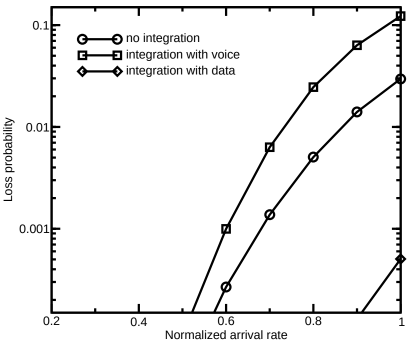

Fig. 7: Streaming traffic loss probability comparison of different integration scenarios

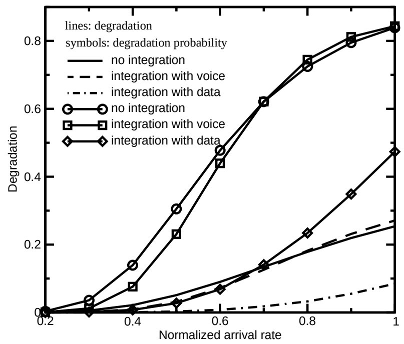

Fig. 8: Streaming traffic degradation comparison of different integration scenarios

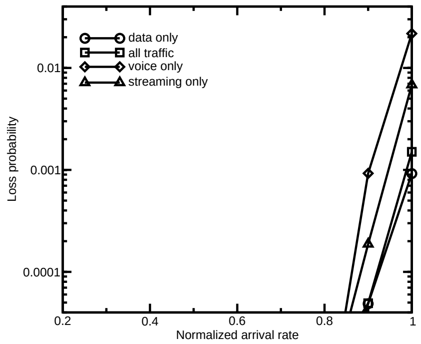

Fig. 10: Streaming traffic loss probability comparison of different overflow scenarios

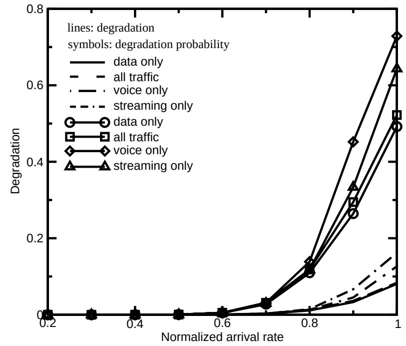

Fig. 11: Streaming traffic degradation comparison of different overflow scenarios

data traffic, allocate data traffic to the least loaded network can efficiently improve the performance of all types of traffic.

3.2 Allocation policy

Based on the performance comparison of different integration and overflow scenarios, a simple allocation policy is proposed for GSM/UMTS interworking, i.e. allocating voice users to the GSM network with priority, and streaming users to the UMTS network with priority, while allowing data users choose the least loaded network. This allocation policy improves network capacity by allocating bearer services to efficient networks, and also reduces network signalling since overflow is mainly limited to data users. When overflow is necessary during an active data flow, data vertical handover is required, and compared with real-time vertical handover, data vertical handover has less stringent delay requirement, thus, this allocation also reduces vertical handover complexity. Another important implication is that the decision for network selection can be simplified. Considerable amount of effort has been devoted to handover decision in heterogeneous networks, which requires signalling protocols to distribute and retrieve network capability and user preference information as well as comparing different access networks with respect to different criteria [19]. With the proposal in this paper, handover decision in cellular networks can be simplified and also be transparent to end users.

4. CONCLUSIONS AND FUTURE WORK

We have analysed mobility influence on real-time and data traffic in wireless networks, in particular, we have developed new techniques to calculate the state probability and handover arrival rate of real-time and data traffic. The results indicate that mobility reduces the QoS of the real-time traffic, and has little influence on the performance of data traffic. Bandwidth degradation is effective to reduce real-time traffic loss probability, but a high percentage of users are involved in bandwidth degradation. In GSM/UMTS networks, an efficient allocation strategy for traffic of different bearer services is to allocate voice traffic to GSM, and streaming users to UMTS, while allowing data users choose the least loaded network. Overflow is mainly limited to data users. This allocation policy not only improves network capacity by allocating bearer services to efficient networks, but also simplifies vertical handover management.

It is interesting to analyse the performance of integrating real-time and data traffic in wireless networks. However, a complete analytical evaluation of integrating real-time and data traffic is shown to be very difficult even on a single link, and it becomes more complex when mobility is introduced. As as a starting point, the analysis can be based on simple mobility and traffic models. This will be examined in the future work.

References

[1] Harri Holma, Antti Toskala, "WCDMA for UMTS," John Wiley&Sons, Ltd., 2nd Edition, 2001.

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

[3] Yi-Bing Lin, Noerpel A.R., and Harasty D.J., "The sub-rating channel assignment strategy for PCS handoffs," IEEE Transactions on Vehicular Technology, Vol. 45, Iss. 1, Feb. 1996, pp. 122 - 130.

[4] Chun-Ting Chou, and Kang G. Shin, "Analysis of combined adaptive bandwidth allocation and admission control in wireless networks," IEEE INFOCOM - The Conference on Computer Communications, Vol. 21, no. 1, 2002.

[5] Nikos Argiriou and Leonidas Georgiadis, "Channel sharing by rate-adaptive streaming applications," Performance Evaluation, Vol. 55, Iss. 3-4, pp. 211-229, Feb. 2004.

[6] 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 Journal on Selected Areas in Communications, Vol. 21, Iss. 10, pp. 1790 - 1802, Dec. 2003.

[7] Y. Fang, I. Chlamtac and Y. B. Lin, "Channel Occupancy Times and Handoff Rate for Mobile Computing and PCS Networks," IEEE Transactions on Computers, Vol. 47, No. 6, pp. 679-692, 1998.

[8] L. Kleinrock, "Queueing Systems Volume 2: Computer Applications," John Wiley and Sons, Inc., New York, 1976.

[9] S. Ben Fredj, T. Bonaild, A. Proutière, G. Regnié and J. Roberts, "Statistical bandwidth sharing: A study of congestion at flow level," Proc. of SIGCOMM 2001, pp. 111-122, Aug, 2001.

[10] J.W. Cohen, "The multiple phase service network with generalized processor sharing," Acta Informatica 12, pp. 245-284, 1979.

[11] F. Delcoigne, A. Proutière and G. Régnié, "Modelling integration of streaming and data traffic, Performance Evaluation," Vol. 55, Iss. 3-4, pp. 185-209, Feb. 2004.

[12] Bin Li, Lizhong Li, Bo Li, Sivalingam K.M, and Xi-Ren Cao, "Call admission control for voice/data integrated cellular networks: performance analysis and comparative study," IEEE Journal on Selected Areas in Communications, Vol. 22, Iss. 4, pp.706 - 718. May 2004.

[13] A. Furuskär, "Allocation of Multiple Services in Multi-Access Wireless Networks," Proceedings of IEEE Mobile and Wireless Communication Networks Conference, pp. 261-265, Sep. 2002.

[14] W. Zhang, "Bearer Service Allocation and Pricing in Heterogeneous Wireless Networks," Proceedings of the IEEE ICC, Seoul, Korea, May 2005.

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

[16] Susan Lincke-Salecker, and Cynthia S Hood, "The Efficiency of Engineering Traffic across Network Boundaries," International Journal of Wireless Information Networks, Vol. 10, no. 3, pp. 117-125. Jul. 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, Iss. 6, pp. 27-36, 2001.

[19] W. Zhang. "Handover Decision Using Fuzzy MADM in Heterogeneous Networks," Proceedings of the IEEE Wireless Communications and Networking Conference, Mar. 2004.