[Mog05] *Parag Mogre:* **A Dynamic Network Architecture for Cellular Access Networks;** Kommunikation in Verteilten Systemen (KiVS), Kaiserslautern, März 2005, S. 321-331.

A Dynamic Network Architecture for Cellular Access Networks

Parag S. Mogre

Multimedia Communications Lab (KOM), Technische Universität Darmstadt,
Merckstrasse 25, 64283 Darmstadt, Germany

Abstract. High user mobility coupled with high bandwidth demands and bursty nature of traffic is expected in beyond 3G cellular access networks. Such a scenario leads to the creation of highly congested areas or hot-spots in these cellular access networks. The location and duration of existence of these hot-spots is closely related to the mobility patterns of the users and varies over time. We observe heavy loss, high delay, and congestion in parts of the cellular access network as a result of the mobility induced load variation, although the network as a whole can support the user load. This paper proposes a novel dynamic network architecture (*DNA*), which enables better distribution of load and can adapt the network topology dynamically to provide relief to congested areas in the network. Our mechanism outlined here is online, distributed, and does not require advance knowledge of traffic demand.

1 Introduction

The evolution of networks beyond 3G leads to a heterogeneous set of access technologies. In such a scenario, there are several challenging issues, including more efficient utilization of the available bandwidth, low power consumption and provisioning of QoS guarantees. In particular, we see a need for flexible, distributed network architectures, which can achieve higher throughput and are capable of dynamically responding to high variation in traffic demand over time. QoS provisioning is crucial for future cellular access networks to enable voice and multimedia applications to be supported. The high bandwidth requirements of such applications coupled with the mobility of end-users leads to fluctuation of the required bandwidth in individual cells over a period of time. This leads in turn to the creation of highly congested areas in the cellular access networks, so called hot-spots, despite the fact that the overall capacity of the network may be greater than the total load offered to the network at a given point in time.

Hollick et al. [1] provide a detailed study of the changes in traffic demands as a function of user mobility. In general, we observe that congestion develops in parts of the network due to an uneven distribution of the load. In a cellular network, partial relief from this phenomenon can be provided by relaying some traffic from heavily loaded cells to some neighbouring cells which may be less loaded. This enables a more even distribution of the load. Our dynamic network architecture, *DNA*, builds on relaying concepts similar to the work of Wu et al. [2]. However,

we believe that a simple relaying mechanism alone is not enough. Therefore we introduce and propose a wireless relay router (WRR), which works in close co-operation with our QoS-routing mechanism presented in [3]. Our framework enables dynamic modification of the network topology in response to congestion in parts of the network. This alleviates the need for heavy overprovisioning of resources in all parts of the cellular access network. At the same time, this mechanism enables additional bandwidth to be made available wherever and whenever it is required.

The remainder of this paper is organized as follows. In Section 2 we give a short overview of the related work. Then, in Section 3 we outline our proposed architecture and also define the roles of the individual network components in detail. This is followed by a description of the functions of the individual components. We outline the mechanism by which the network architecture adapts itself in response to the likely build up of congestion in parts of the network. Finally we present a summary of this work, also giving an outlook for a real implementation of our proposed DNA.

2 Related Work

As has been highlighted in the introduction, the beyond 3G scenario requires us to design smart technology which has the ability to handle high data traffic, at the same time enabling provision of QoS. The work [1] models the user mobility for the case of a city. The results clearly highlight the creation of transient hot-spots in the network. The work also shows that the location of these hot-spots varies over time. Given the above observation, it is necessary to manage the very high variation in the offered load. However, it is expensive to overprovision network resources. Besides, despite a high overprovisioning factor, it is possible that transient bursts of traffic lead to congestion in parts of the cellular access networks. Therefore, we believe that the network should be flexible enough to allow an on-demand dynamic allocation of resources.

A promising approach to address the congestion problem arising due to unbalanced traffic in a cellular network is proposed by Wu et al. [2]. They propose the use of ad hoc relaying stations (ARS) to relay traffic from a heavily loaded cell to a less loaded cell dynamically. This increases the overall load that the system can handle. They identify three different relaying mechanisms, namely primary relaying, secondary relaying and cascaded relaying.

The primary relaying and secondary relaying mechanisms are shown in Fig. 1. For example, with respect to Fig. 1(a), a mobile subscriber X wants to set up a new call in a heavily loaded cell A. However, the call fails due to the unavailability of channels in the cell. In such a case the mobile subscriber X switches to an alternative radio interface and the call will be relayed through the (ARS) to a neighbouring cell B that has less load. The authors describe this mechanism as primary relaying. Fig. 1(b) shows the secondary relaying mechanism. Here, as in the case of primary relaying, the mobile subscriber X wants to establish a new call in a highly loaded cell A. The call initially cannot

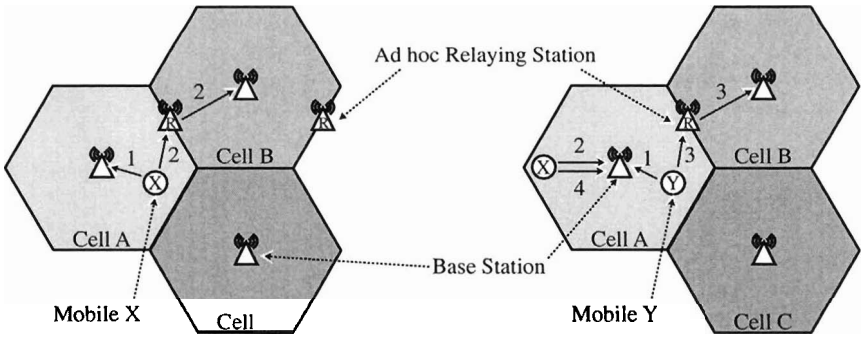

Fig. 1. Overview of relaying mechanisms

be setup due to unavailability of channels in the cell A. In this case, however, the mobile subscriber is not within the range of a suitable *ARS*. The base station for cell A then transfers an existing call, in this example that of mobile subscriber Y, via an *ARS* to a neighbouring cell. This frees up channels which can then be used to permit mobile subscriber X to set up the new call. The third relaying mechanism, cascaded relaying, described in [2] is a combination of both primary relaying and secondary relaying. In summary, the authors describe a mechanism, which enables an even distribution of load. They show that the overall load supported by the system is maximized if the load per cell is equal.

In our work [3], we develop a near-optimal multiclass, multipath, minimum-delay routing algorithm. The routing algorithm is distributed in nature and computes for a given destination a next-hop set and a corresponding load distribution along various paths to that destination. Fig. 2 gives an overview of the routing mechanism. Consider a router i, as shown in Fig. 2, and its neighbours. At time $n * T_l + p * T_s$ the routing table at router i has as next-hop, for some destination, the set of neighbours k_2, k_3 and k_4. The numbers besides the links show the fraction of traffic for the destination under consideration at router i, which is to be sent along that link to a neighbour in the next-hop set. As shown in figures 2(b)-2(f), the routing algorithm operates at two timescales. At every time interval T_l, the next-hop set for a destination is updated; whereas, at a smaller time interval T_s, the fraction of traffic sent to individual neighbours in the next-hop set is adjusted. This enables an efficient load-balancing mechanism which takes into account both, the long term delay along paths to the destination (by changing the next-hop set) as well as the short term transient changes in delay along individual paths to a destination (by adjusting the fraction of traffic sent along individual paths). The routing mechanism in [3] utilizes as the link cost, the incremental delay that is experienced by traffic along the link. Here, by link we refer to a directed link, i.e. the cost if link (i,k) may not be the same as the cost of link (k,i) at a given point of time. Incremental delay as link cost was introduced by Gallager [4]. He presents the optimal routing problem and also

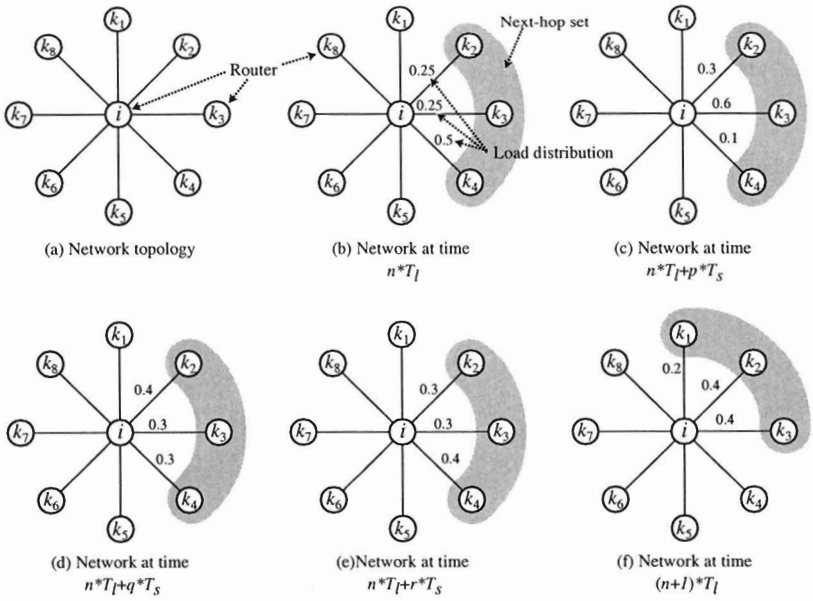

Fig. 2. Overview of routing mechanism used by Mogre [3]

proposes a solution to this problem. The work by Vutukury et al. [5] extends [4] so that it is applicable in real networks. We, in our work [3], extend and improve on [5], so that it is QoS aware and can cater to different classes of traffic and their varying tolerances of loss and delay. We show through an extensive simulation study that our routing mechanism gives excellent results with respect to both average delay and overall throughput in the network.

The work [6] by Lin et al., presents a new architecture (MCN) for wireless communications. The goal of this work is to either reduce the number of base stations needed to service a given area or to reduce the transmission power and range of individual base stations. To reach the above goals, the work makes use of multi-hop relaying mechanisms involving mobile user nodes. However, the usage of end-user nodes in the system makes it difficult to guarantee improved coverage or throughput. Hence, in our work here, we avoid the usage of the end-user nodes at the time of design of our proposed *DNA*.

The work [7] by Lott et al., discusses a possible scenario to handle communications beyond 3G. It discusses the possibility of a hierarchical overlay of multi-hop communication or WLAN, HiperLan/2 and UMTS, GSM/GPRS, and satellite systems. Such a vertical handover mechanism is also possible in our network architecture. In [7], the authors also discuss the use of multi-hop enabled nodes or extension points in order to extend coverage. Similar to the work [6], they assume that these extension points can be mobile end-systems.

The work by Yeh et al. [8], advocates that a completely novel approach is needed to deal with challenges thrown up by multi-hop cellular networks. They propose a complete overhaul of existing architecture and routing mechanisms. They propose a selective table-driven integrated routing with different roles for different mobile devices in the system, depending on the capacity and the ability of the node in question. An implementation of this mechanism makes radical changes necessary both at the end-user nodes and in the provider infrastructure. Although, such an overhaul may be inevitable, we in our work design our architecture to be more easily implementable given current infrastructure. At the same time the architecture we develop also serves the needs of future applications.

3 Dynamic network architecture for cellular networks

The focus of our architecture is to support load balancing and QoS routing in a beyond 3G network. We build up on the work of Hollick et al. [9] to derive the network architecture shown in Fig. 3.

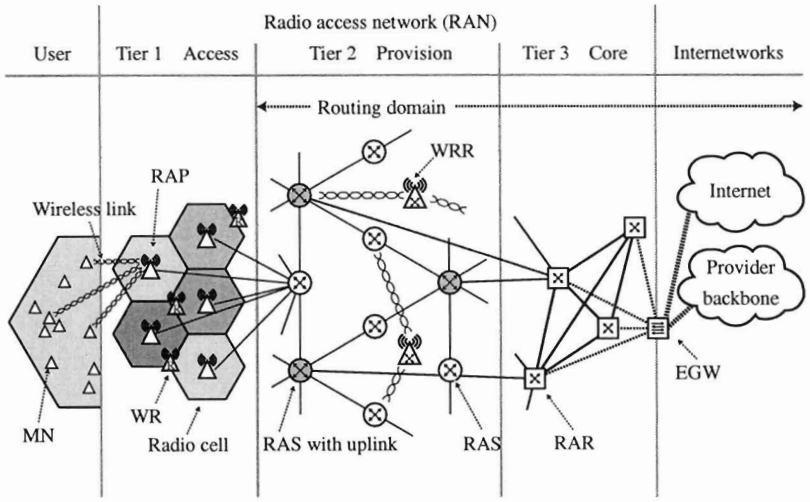

Fig. 3. Proposed dynamic network architecture (*DNA*)

Similar to the work of Hollick et al. [9], we assume a three tier architecture. The mobile nodes or terminals (*MNs*) are associated to the wireless base stations, the so called radio access points (*RAPs*), representing the last hop of the provision network. The function of the first tier can thus be described as radio access. The balancing of load between neighbouring radio cells is performed by the wireless relays (*WRs*). The *WRs* are supposed to have a functionality similar

to that outlined for *ARSs* in the work by Wu et al. [2]. The second tier of the radio access network comprises of radio access servers *RASs* and the wireless relay routers (*WRRs*). *RASs* are used to attach multiple *RAPs*. The *RASs* are partially meshed with additional connectivity being provided by the (*WRRs*). As in [9], we assume our network to be a routing network beginning at the *RAS* level. Some of the *RASs* have uplinks to the so called radio access routers (*RARs*), which together with the edge gateways (*EGWs*) form the third tier of the proposed *DNA*, and are assumed to be fully meshed. A crucial point is the distributed and decentralised nature of the topology instead of the traditional tree-structured ones in telecommunication networks like GSM or UMTS (see Fig. 4).

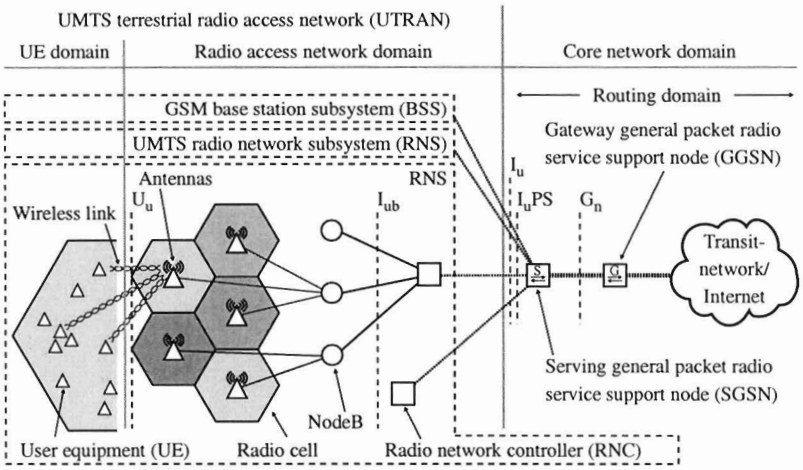

Fig. 4. Traditional cellular network architecture UMTS

Traditionally, radio access networks have not been routing networks because of the strict tree structure. We, in contrast, enable routing beginning at level two (ie. *RAS*/*WRR*) of the network. This enables us to deploy suitable routing mechanisms, which enable QoS and load balancing in the cellular network starting at a much lower level in the hierarchy as compared to traditional telecommunication networks.

3.1 Dynamic behaviour of the *DNA*

In the following we discuss the dynamic behaviour of our proposed architecture (*DNA*) in response to congestion in the network. Dynamic reaction to congestion in the network is enabled by the wireless relay (*WR*) and the wireless relay router (*WRR*), respectively. We next give an overview of the working of the *WR* and the *WRR*.

The *WR* operates at tier one of the proposed *DNA*. The goal of the *WR* is to distribute load from a heavily loaded cell to some neighbouring cells. In our scenario it enables load balancing between neighbouring *RAPs*. We assume that the *WR* is able to support all the relaying mechanisms which may be supported by the *ARS* proposed in [2]. Consequently, the *WR* is assumed to be able to support primary relaying, secondary relaying as well as a cascaded relaying mechanism. However, we assume that the *WR* is part of the provider network and not an ad hoc end-user node. Thus, the *WR* enables the *DNA* to achieve a higher call acceptance ratio as compared to a traditional cellular network, by enabling a more even amount of load per cell.

The other dynamic component of our proposed *DNA*, which enables dynamic action in face of congestion in the network, is the *WRR*. The *WRR* operates at tier two of our proposed *DNA*. The function of the *WRR* is to adapt the actual network topology in response to congestion in the network. We assume the *WRR* to be located such that it can reach a number of *RASs* by means of wireless links. The *RASs* are assumed to have a suitable wireless interface to interact with the *WRR*.

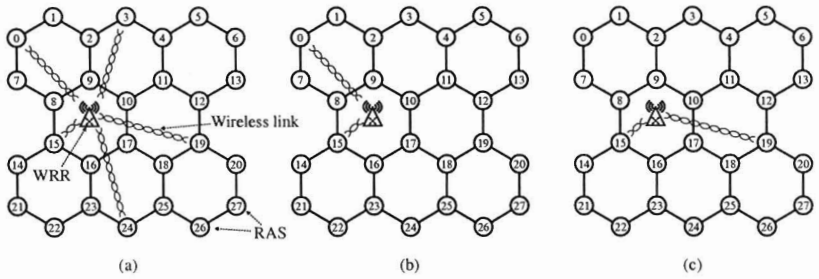

Fig. 5. WRR mechanism

Fig. 5(a) shows a sample instantiation of the second tier of the *DNA*. Here, the lines from the *WRR* to the individual *RASs* denote that the *WRR* is able to reach these *RASs*, which have a suitable wireless interface capable of interacting with the *WRR*. The solid lines connecting individual *RASs* indicate wired links between the respective *RASs*. Also, in order to enable a higher number of *WRRs* to be deployed in the network, we assume that the wireless links are established by means of directional antennas, thus reducing the resulting interference. These wireless links are assumed to be high bandwidth links, and may permit duplex data transfer. Fig. 5(a) shows that the *WRR* is able to communicate with a number of *RASs* in its range, but need not and may not be able to support high bandwidth simultaneous data transfer with all these *RASs*. For the purpose of this paper, without loss of generality, we assume that the *WRR* is able to simultaneously support high capacity, duplex links using directional antennas to a minimum of two other entities (either both *RASs* or both *WRRs* or a

combination of a *RAS* and a *WRR*). Given the limited number of wireless links with the *RASs* that the *WRR* can simultaneously support, the mechanism we develop works to provide this additional bandwidth where it is most needed in the network. This mechanism works with the aid of feedback from the routing mechanism developed in [3]. For example as shown in Fig. 5(b) at a given time t_1 the *WRR* may enable the wireless links to *RAS* 0 and *RAS* 15, thereby providing a virtual link between *RASs* 0 and 15. While at some other time t_2, based on the network traffic conditions and feedback from the routing algorithm, the *WRR* enables the wireless links to *RASs* 15 and 19, thus providing a virtual link between the *RASs* 15 and 19. Thus, it can be seen that the *WRR* should adapt the network topology (provide virtual links between pairs of *RASs*) in response to feedback from the routing mechanism. The interworking of the topology changes with the routing mechanism is crucial for the understanding of our architecture. As already introduced, the routing mechanism used in [3] computes multiple loop-free paths, possibly of unequal costs, to a destination at each router in the network. This route computation is done at time interval T_l. On a smaller timescale T_s, the amount of traffic sent along individual paths is adjusted so that more load is sent along low cost paths than along higher cost paths. The link cost used by the routing mechanism is the incremental delay over the link. Fig. 6 qualitatively shows the variation in the link cost with respect to the link utilization.

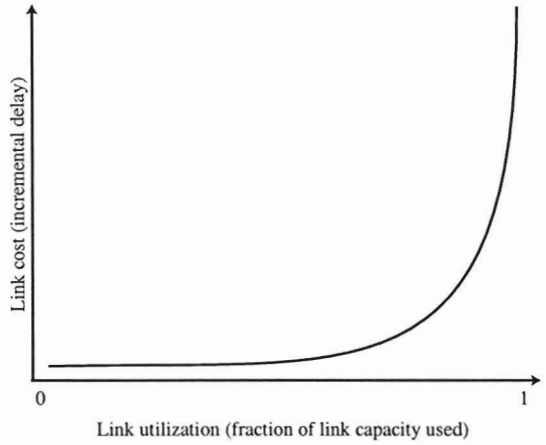

Fig. 6. Variation in link cost with the utilization of the link

As can be observed from Fig. 6, the link cost starts increasing exponentially as the traffic on the link approaches the link capacity. With the incremental delay as link cost more than an order of magnitude increase in the link cost is observed as the link utilization moves from 0.7 to 0.9. This region corresponds

to the build up of congestion at that link. For the wireless links through the *WRR* on the other hand we keep the link cost a constant value and set it to be more than the average wired link cost observed in the network at 0.7 link utilization. We call this constant cost for the wireless links through the *WRR* as *threshold-cost*. We return to Fig. 5 for the following discussion. Consider Fig. 5(a), the *WRR* at startup discovers the set of *RASs* in its range having a suitable interface for communication with the *WRR*. In this example it would be the set of *RASs* 0, 3, 15, 19 and 24. It then transmits this neighbourhood information to the above *RASs*. As mentioned, these *RASs* have a suitable radio interface for communication with the *WRR*. The data transmission and reception over this interface is handled by an agent called the virtual interface proxy (*VIP*).

On receipt of this connectivity information by the *RAS* 0 (for example) with the aid of the *VIP*, the *RAS* 0 updates its set of neighbours to 1, 7, 3, 15, 19 and 24. Here, it should be noted that *RAS* 0 has permanent wired links to *RASs* 1 and 7 but may only sometimes be able to communicate with *RASs* 3, 15, 19 and 24 via the *WRR*, depending on whether the required wireless links are enabled or not. Hence, we denote these wireless links as virtual links (in the sense that they are not always existent). Here, the links (0,3), (0,15), (0,19), and (0,24) are virtual links. As we mentioned previously, these virtual links are advertised as *threshold-cost* links. Due to their high cost as compared to the normal wired links in the network, these virtual links are normally not used for forwarding data by the routing tables computed at the individual *RASs*. However, with the increase of load in parts of the network, the cost of some of the wired links may increase exponentially and at a point in time, a path through the virtual link may become cheaper in cost and will then be used by the routing tables at the individual *RASs*. An indication as to which of the virtual links are to be activated is closely coupled to the feedback of the routing mechanism. Here, the *VIP* at the *RAS* plays a significant role. It is responsible for periodically requesting the *WRR* to activate a particular virtual link based on whether it has data to transfer over the virtual link, which is triggered by the routing algorithm at the *RAS*.

The *VIP* at a *RAS* periodically transmits requests for activation of a virtual link to the *WRR* along with a measure of how much traffic has to be sent along the virtual link. The *WRR* then analyzes all the currently pending requests (based on which virtual link has to carry more traffic), and decides on which virtual link to activate next.

In case the *VIP* receives data for some destination to be transfered over a virtual link that is not currently active, then the data is distributed over the other neighbours in the next-hop set for that destination. For example, consider Fig 5(b). Here, the virtual link (0,15) is active, i.e. data can be sent via the *WRR* from *RAS* 0 to *RAS* 15. Now consider that the next-hop set at *RAS* 0 for the destination 4 is {1, 3} (here *RAS* 3 is connected via a virtual link (0,3) to *RAS* 0 and is said to be a virtual neighbour of *RAS* 0). If the routing algorithm at this point sends some data intended for *RAS* 4 for transfer over the virtual link (0,3) it is handled first by the *VIP* at *RAS* 0. In this case the *VIP* detects

that the required virtual link is currently unavailable and distributes the data, where possible, over the other members in the next-hop set for the destination 4. In our example, this data will be transmitted over the interface to *RAS* 1. At this point the *VIP* also sends a request for activation of virtual link (0,3) to the *WRR* along with related information about what fraction of traffic at *RAS* 0 is to be sent along link (0,3). The information about the fraction of traffic to a particular neighbour can be computed from the routing tables as used in [3].

The *WRR* selects periodically the next virtual link to be activated based on which virtual link has to carry the maximum amount of traffic. For seamless interworking with the routing algorithm, we also have to determine an appropriate frequency for the switching between the various virtual links to be activated by the *WRR*. As already explained, the routing mechanism updates the routing tables at a time interval T_l and at a smaller time interval T_s adjusts the fraction of traffic to be sent along individual paths. Hence, we decide to choose the virtual link switching interval to be some multiple of T_s but less than T_l. This enables the *WRR* to respond to transient congestion as well as long term overload in parts of the network. Detailed protocols and heuristics to be run by the *VIP* and the *WRR* are given in [3].

4 Conclusion

In this paper we have given an overview of a part of our work in [3]. We have discussed our dynamic network architecture and have highlighted the roles of the individual components. We also described our mechanism to provide additional bandwidth in the network where it is required. In order to enable such a dynamic bandwidth allocation and topology adaptation in the network we rely on feedback from the routing mechanism. We have outlined a distributed mechanism by which this can be achieved. We believe that the concept of the *WRR* may be implemented using currently available standard technologies. For example to enable high capacity, long range wireless links for the *WRR* we envision the use of components based on IEEE 802.16-2004 standards [10].

The ultimate goal for such topology adaptation mechanisms, as also discussed by Kleinrock [11], is to achieve optimization goals such as minimization of delay or maximization of throughput.

5 Acknowledgements

This work was partially funded by Siemens AG, CT IC 2, and the DAAD. We would like to thank Dipl.-Ing. M. Hollick, Dipl.-Inf. T. Krop, Prof. Dr.-Ing. J. B. Schmitt, Prof. Dr.-Ing. R. Steinmetz, and Prof. G. Barua for their guidance and support.

6 References

1. Hollick M., Krop T., Schmitt J.B., et al.: Modeling mobility and workload for wireless metropolitan area networks. Computer Communications, vol. 27, 751–761, 2004.
2. Wu H., Qiao C., De S., et al.: Integrated cellular and ad hoc relaying systems:iCAR. IEEE Journal on Selected Areas in Communications, vol. 19, 2105–2115, 2001.
3. Mogre P.S.: Near-optimal multiclass minimum-delay routing for cellular networks with variable topology, M.Tech. thesis, Dept. of CSE, IIT Guwahati; Multimedia Communications Lab, Dept. of ETIT, TU-Darmstadt, 2004.
4. Gallager R.G.: A minimum delay routing algorithm using distributed computation, IEEE Trans. on Communication, vol 25, 73–84, 1977.
5. Vutukury S.: Multipath routing mechanisms for traffic engineering and quality of service in the Internet, PhD. thesis, Univ. of California Santa Cruz, 2001.
6. Lin Y.D., Hsu Y.C.: Multihop cellular: a new architecture for wireless communications, IEEE INFOCOM 2000,1273–1282, 2000.
7. Lott M., Weckerle M., Zirwas W., et al.: Hierarchical cellular multihop networks, EPMCC03, 2003.
8. Yeh C.H.: Acenet: architectures and protocols for high throughput, low power and qos provisioning in next-generation mobile communications, IEEE PIMRC 2002, 2002.
9. Hollick M., Krop T., Schmitt J.B., et al.: Comparative analysis of quality of service routing in wireless metropolitan area networks, IEEE LCN 2003, 470–479, 2003.
10. IEEE 802.16-2004: IEEE standard for local and metropolitan area networks – part 16: air interface for fixed broadband wireless access systems, 2004.
11. Kleinrock L.: Queueing Systems, vol 2: computer applications. Wiley Interscience, 1976.