Wireless Broadband Services using Smart Caching

Stephan Goebbels
sgs@comnets.rwth-aachen.de
Chair of Communication Networks,
Faculty 6, RWTH Aachen University

***Abstract*— Broadband access for all will be a key functionality in future mobile radio networks. Depending on the length and the characteristic of the radio wave propagation the performance of the network is substantially impacted. Especially in urban application scenarios a ubiquitous reachability is hard to achieve. In combination this results in high variations of link quality so that broadband services are hard to realize. This paper discusses a technique, named Smart Caching, which tries to overcome the problems of network performance variations. It focuses herewith on non real time and non interactive services.**

Smart Caching follows the approach of pre-fetching and buffering data at the edge of the core network and transmitting it whenever the terminal comes into service range of an Access Point. Especially on motorways this technology can provide broadband services even under the conditions of intermittent network coverage.

The mutual reaction of Smart Caching enabled and legacy services is shown. Additionally dimensioning aspects and the resulting performance gain is discussed.

The performance of the Smart Caching service is evaluated with a sophisticated queuing model based on the Markov Arrival Process. All network characteristics, like arrival and service process are modeled in very detail.

***Key Words*—Smart Caching, Heterogeneous Radio Networks, Traffic Theory.**

1. Introduction

The integration of various radio communication networks in future system design is a widely adapted approach. In order to benefit from the special advantages of each network their combined employment is necessary. However, the heterogeneity of the different networks causes the problem that some of the networks will provide ubiquity with basic Internet access while others allow broadband connections only in specific regions. The reason for that is that the distance between access node and terminal and the achievable data rate is directly related to each other. Hence, with a reasonable number of access nodes full coverage can only be achieved if simultaneously the data rate is reduced. This results in wide area networks which overlay the different regions of broadband access provided by short range networks.

Services which rely on continuous broadband coverage are hard to provide in such environments. Their requirements in means of bandwidth are usually not met by full coverage networks. For a mobile user who handover from one network to the other this implies that the service is always interrupted if the new system cannot provide the necessary data rates. The handover of the service to the wide area network might be successful but the continuation of the service would consume an unreasonable portion of the radio resources in the new system, if possible.

Nevertheless the problem can be circumvented as video streaming and other non interactive and non real time services have low delay constraints. This does not mean that in legacy protocols delays of more than several seconds can be tolerated. But whether the video scene is shown within one second after transferred from the server or a lot of seconds later is of no relevance for the user. As long as the play-back of the video is not interrupted a time shift is acceptable.

This property is exploited by Smart Caching. A massive reservoir of user data is pre-fetched at the edge of the wired backbone network. This data is kept until the terminal comes into service range of an Access Point or a timer expires. If the user terminal connects to a new Access Point the buffered data is transmitted to the terminal and buffered there by exploiting the full capacity of the wireless link. If this approach is compared with legacy video streaming the amount of buffered data in the terminal is substantially increased by Smart Caching. By doing so it is possible to hide even long intervals of low network performance or service interruptions of up to tens of seconds from the user. The service is continued by using up the pre-fetched data in the terminal.

But communication networks have to deal with multiple services. Each of them has special requirements. In contrast to video streaming bidirectional communication services like Voice over IP (VoIP) have very limited bandwidth constraints. But on the other hand their delay tolerance is very low. Packets delays of more than 200 ms are noted by the end user as quality degradation. Higher delay values are even intolerable in modern communication systems. For such services wide area networks, e.g. like UMTS, are perfectly suitable.

The paper shows which impact the different services have on each other and additionally how the employment of Smart Caching in integrated networks can improve the system performance.

2. The Concept of Smart Caching

Smart Caching allows the decoupling of end-to-end data flows in combined wired and wireless networks. At the edge of the core network a buffer, the Smart Cache, is located which introduces two subsections to the end-to-end connection. The first subsection consists between the application server in the Internet and the Smart Cache. Usually this part relies only on wired connections. Although each link in this subsection may provide a higher bandwidth than wireless links the combination of several links lead to the fact that each chain is as strong as the weakest link. This results to the fact that end-to-end only a limited bandwidth can be supported.

If the capacity of end-to-end connections in the (wired) Internet is compared to wireless links a discrepancy occurs. While the Internet usually provides data rates of around 10 Mbit/s the

wireless link might offer often smaller data rates. But at their best quality radio links can provide data rates of several tenth of Mbit/s.

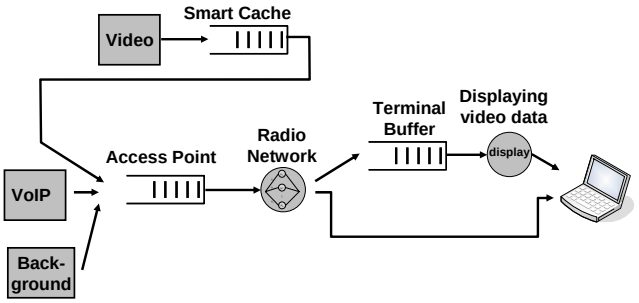

Figure 1: The of Smart Caching enabled Wireless Broadband Networks

The second subsection between Smart Cache and user terminal comprise both wired (between Smart Cache and Access Point) and wireless links. However, the dominating factor is the radio transmission and the wired hop has only minor impact on the network performance. The purpose of Smart Caching is to simulate a continuous radio connection even if the wireless link is broken. The data stream on the first subsection goes on and the data is buffered until the wireless link is re-established. This separation might require the break up of TCP end-to-end connections. Developed for a similar purpose the SNOOP protocol [1] can be used in such cases.

To reveal its best performance the radio link should always be used by its current capacity. By buffering data in the Smart Cache the data flow on the first subsection is not influenced whether the wireless link is broken or not. But the buffered data offers the advantage that it can be used to fill up otherwise unused resource areas. The data rate on the second hop can now always be adapted to the currently available network capacity.

Therefore the Smart Cache decouples the two subsections so that on both of them the best performance in means of throughput can be reached. Smart Caching allows exploiting the actual available radio link capacity to transmit as much information as possible to the end device so that its buffer is filled and to be able to bridge periods of no radio service.

Figure 1 illustrates schematically a Smart Caching enabled network. For real time services like VoIP the network behavior does not change. If the data has traversed the backbone network it has to wait until resources of the wireless link get available. Then they are sent to the terminal and directly consumed. If the wireless link fails the service is interrupted immediately. But for non real time services like video streaming the data is routed through a Smart Cache and from there on forwarded to the access node. It now depends whether the wireless link is available or not if the data is forwarded to the access node. If no radio connection to the user terminal consists the data is buffered and as soon as the connection is available this data is forwarded. Of course, all services are sharing the scare resource which is the radio capacity.

The incoming data of the video services is used in the user end device to fill up a second cache, the Terminal Buffer. The play-back of the video should not start until a certain fill level in the Terminal Buffers is reached. By doing so it should be guaranteed that this buffer is not drained even in case of longer network service interrupts. Initial fill size and length of the idle period are strongly related. The later analysis will provide a dimensioning instruction for that.

3. Traffic Model of the System

Applied queuing theory is a well understood tool for the analysis of communication networks. The combination out of queue and serving processor perfectly matches to the waiting of packets in a network node for getting forwarded via the wireless link. The limiting resource especially in wireless networks is the radio capacity. Packets of the different traffic streams compete for the access to the radio link. In order to apply queuing theory each traffic stream has to be modeled by a stochastic process. Also the service process must be described by a random distribution which approximates the duration a packet needs to be transmitted via the air interface of the Access Point.

To achieve significant results in the analysis of wireless networks it is important to model the radio transmission most accurately.

OFDM based wireless networks like the chosen WiMAX system take different PHY modes to match their data transmission to the characteristic of the radio connection. More robust PHY modes are chosen if the radio signal gets weaker. The robustness of a PHY mode mainly expresses in a decrease of the transmission rate. Switching points, in means of signal to noise values, between the different PHY modes are specified by the standardization bodies. Thus a certain PHY mode is applied between an upper and lower signal strength level. Together with the radio propagation conditions each location of a scenario can be assigned with a PHY mode by which a user is served if residing there. Together with the user's mobility for each PHY mode a certain probability of being employed can be calculated.

In order to model the data rates of the different PHY modes a hyperexponential service time distribution is used. According to path probabilities p_i packets are processed with different service rates μ_i. Both parameters can be derived out of the residence probabilities p'_i and corresponding data rates r_i of the WiMAX system.

The arrival process of the queuing model can be modeled by using the Markov Arrival Process (MAP). In [2] a method is introduced which allows the feeding of on queue with a mixture of streams represented by MAPs. Additionally the packets of each stream can be assigned with separate service time distributions G_i. It allows calculating the queuing delay of each arrival stream separately. The resulting queuing system has the notation MMAP(i)/G(i)/1. This allows a very detailed modeling and analysis of the investigated scenario setup. To describe a MMAP a set of matrices, D_0 and D_i, is necessary. While D_0 contains state transmissions without a packet arrival, the matrices D_i cover transitions with a simultaneous arrival of a type *i* packet.

In the following analysis three different types of Internet traffic are considered. These are VoIP and Video streaming traffic as well as background traffic, which combines all other types of network streams. While the latter can be modeled by a simple Poison process ($D_{0,i}=-\lambda$ and $D_{1,i}=\lambda$) the first two require more sophisticated approaches. Already in [3] the superposition of several VoIP sources to a Markov chain is introduced.

Furthermore this can be transformed in a Markov modulated Poison Process with two states (MMPP[2]) [4]. Hence, this allows the modeling of an arbitrary number of VoIP sources by a simple Markov process. In order to parameterize the model the length of talk spurts (352ms) and idle periods (650ms) in between are taken from [3] while the other parameters (Packet rate = 100 s^{-1}; Packet size = 80+40 Byte) are adopted from the G.711 recommendation.

In [5] it is shown that also video sources can be regarded as a superposition of a predefined number (M=20) of so called Mini-Sources. It is possible to model N video sources by M*N Mini-Sources. Since this approach is very similar to the one mentioned before the video sources can be represented by another MMPP[2]. However, the parameters of the video source modeling have to be updated to fit the new video standards. In the later presented analysis MPEG4 streams are considered with parameters taken from [6] (Cov(0)=0.509; Cov. param. = 0.09; Data rate = 3.37 Mbit/s; Packet size = 188+40 Byte).

MMPPs can be directly used for the input matrices of MMAP queuing systems. While the underlying Markov process is used in the transition matrix $D_{0,i}$ the different Poisson arrival rates in each state are included in the matrix $D_{1,i}$. Since several different traffic streams are considered and each of them has an own matrix $D_{0,i}$ the superposition of the different processes is necessary. It can be done by composing the Kronecker sum of the different matrices $D_{0,i}$ and $D_{1,i}$, which is described in [7]. At the end there is only one matrix D_0 and i different matrices D'_i according to the traffic streams. The dimension of the resulting matrices arises form the product of the size of the separate input matrices $D_{0,i}$.

Although the advantages of Smart Caching are obvious it is necessary to compare its performance with legacy network setups to numeralize the benefit. As mentioned before especially bandwidth consuming services like broadband video streaming reveal problems in heterogeneous network structures. For such services Smart Caching shows its full potential. By employing a Smart Cache user data can be stored at the edge of the wired backbone close to the Access Points of the wireless network if the forwarding of data is obstructed. The reason could be an insufficient radio link performance or a broken connection. In legacy networks packets which can not be forwarded via the wireless link are usually discarded almost immediately. To really benefit from broadband coverage it is necessary to fully exploit its capabilities which means transfer as much data as possible, so that the Terminal Buffer is filled to its maximum. Without a Smart Cache it is very likely that the wireless network offers more transport capacity then data can be delivered by the backbone. While the backbone is limited in an end-to-end connection to a data rate of several Mbit/s e.g. the WiMAX system allows under optimal conditions up to 55 Mbit/s. Therefore the available data rate perceived by a mobile user is cut to the limit of the backbone connection, as shown in Figure 2.

The grey shaded area reflects the capacity which is wasted as the data transport is limited by the backbone. The wireless network idles as the backbone cannot deliver the required data fast enough. Depending on the backbone limit the average user data rate can be reduced so far, that for example the video streaming service has to be interrupted since the Terminal Buffer is drained.

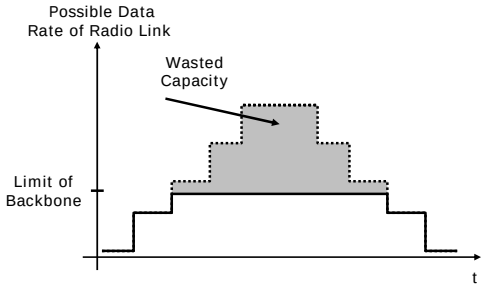

Figure 2: Impact of missing Smart Cache

If a customer traverses a coverage gap between two successive Access Points the incoming packets of e.g. the video stream are buffered in the Smart Cache. Due to this buffering the end-to-end delay of the packet is drastically increased. Which influence the gap has on the actual packet delay is depicted in Figure 3. The scene is separated in three periods. In the first, denoted by x, the customer leaves the coverage zone and the Smart Cache starts the buffering so that the fill level increases. After reentering a new coverage zone the incoming traffic is still buffered as there are still earlier arrived packets left in the Smart Cache. Since the arrival rate of new packets is smaller than the transmission rate of the air interface the buffer starts to be drained and the fill level decreases. The period until the whole buffer is emptied is called y. And finally the period z just denotes the normal operation of the wireless network. Packets which arrive have to wait a short period until the radio resources gets available and they are transmitted.

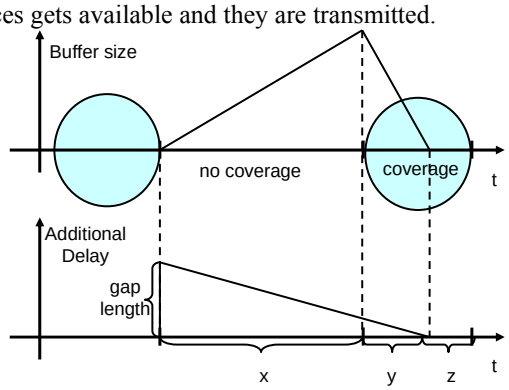

Figure 3: Influence of Coverage Gap on Buffer Size and Packet Delay

In the lower part of Figure 3 the additional delay depending on the time of arrival is shown. It is caused by the coverage gap. The maximum delay is suffered from the first packet which arrives after leaving the coverage zone of the first Access Point. It is delayed for the whole period of no coverage and instantly transmitted when the next coverage zone is reached. The packet delay decreases until the point is reached where the complete Smart Cache is drained and the operation migrates to the normal behavior. For later evaluations it is necessary to determine the ratio of packets which are affected by the gap and the portion which is transferred during the normal operation. It holds that:

$$\frac{x+y}{x+y+z} = \text{ratio of influenced packets.}$$

The portion of time which is covered by wireless access (cov) is given by the quotient out of x and the overall size x+y+z.

4. Network Setup

In the following a motorway setup is investigated to show the potential and capabilities of Smart Caching enhanced wireless networks.

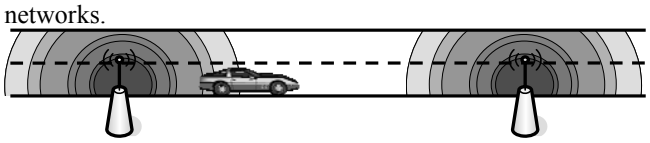

Figure 4: Motorway Scenario

In order to reduce deployment costs not the full motorway is covered with broadband access but only at certain distances WiMAX Access Points are mounted. They cover the street only partially so that there are also uncovered areas in between. The degree of coverage depends on the range of the Access Points and the distance between two consecutive ones. It is assumed that the Access Points are operating with 100mW transmission power. The switching points between the different PHY modes and the corresponding data rates are taken from [8].

In such a scenario setup it can be assumed that the attenuation between sender and receiver is reduced to a minimum. Therefore the path-loss [dB] is derived by the D1 Line-of-Sight model taken from the report D 5.4 [9] of the Winner project.

$$21.5 * \log(dist.) + 44.6$$

The duration of the gap and its distribution can be derived from the vehicles velocity and the gap size distribution. The gap size ($f_l(l)$) is assumed to be normally distributed while he standard deviation is one fifth of the mean ($\sigma=0.2*\mu$).

The velocity distribution ($f_v(v)$) is based on measurements made in [10]. It is as well normally distributed and the mean is given by 104 km/h and the deviation counts to 12.5 km/h.

From the distributions of the car velocity and the gap size the density of the duration of a gap period can be derived by applying the following formula.

$$f_t(t) = \int_0^\infty v f_l(vt) f_v(v) dv$$

5. Results

If Smart Caching is employed in future wireless networks it has to be assured that other services are not (too much) negatively influenced by it.

As Smart Caching buffers data at the edge of the wired network no data is lost during a communication session. In a legacy approach data which belonged to an interrupted communication session is discarded and the streaming is stopped. This additional amount of data heavily burdens the network. Although it allows the provisioning of continuous broadband services even in case of intermittent broadband coverage it may decrease the performance of other services.

In Figure 5 the number of VoIP sessions (One session includes up- and downlink) is shown which can be supported depending on the number of video sessions and the degree of coverage the scenario offers. For one or two simultaneous video session per Access Point the number of VoIP users is above 50 users even for coverage values of 20-30%. This corresponds two an Access Point distance of 12-15km. This is more than enough for the proposed scenario. Usually another factor would limit the number of VoIP users. As such a bidirectional real-time service cannot be enhanced by Smart Caching in times of no broadband coverage the VoIP calls have to rely on cellular or wide range networks like UMTS. Their capacity is much lower so that less VoIP calls can be supported. Nevertheless the temporarily shift of calls in the WiMAX network can disburden the UMTS network substantially.

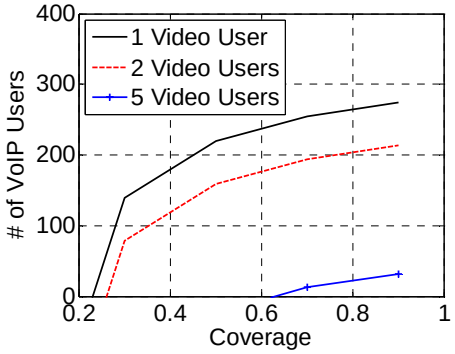

Figure 5: Capacity for VoIP Customers

For 5 video users the number of VoIP calls is reduced to values of less than 30. For a broadband network coverage of 70% (Access Point distance 5 km) only 12 users can be supported. But this is still enough for the given scenario. If the coverage would be further reduced the buffered data in the Smart Cache could not be transmitted during traversing the WiMAX cell so that the buffer would overflow after a while and the service would be interrupted.

The packet delay in a video session is massively determined by the size of the coverage gap and the resulting time the customer needs to traverse this gap. In Figure 6 the Cumulative Distribution Function (CDF) of the packet delay depending on the Access Point (AP) distance is shown. It is assumed that only one video user is traversing the motorway setup.

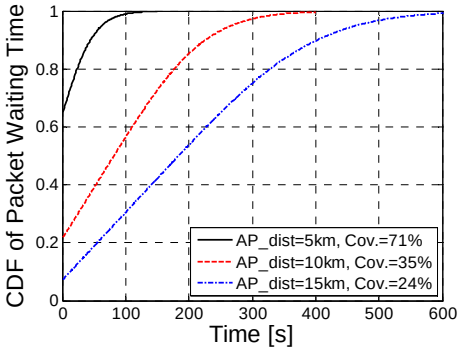

Figure 6: Delay for Video Streaming Packets

The seamless playback of a video stream always depends on the arrival of packets and usually even more important on the fill size of the Terminal Buffer. The more video information is buffered in the terminal the longer the playback can continue even if the radio connection quality decreases or the link breaks.

The maximum delay of a packet determines the amount of data the Terminal Buffer has to contain, in order to ensure a seamless playback. Due to the special setup of the Smart Caching scenario the packet delay usually decrease after a peak

value is met (Compare Figure 3). Hence if the maximum delay value can be tolerated the next packets will arrive even faster so that the Terminal Buffer can be filled up again. For a coverage degree of 71% which corresponds to an Access Point distance of 5 km the maximum packet delay is around 100 seconds. The Terminal Buffer has to store 100 seconds of video data before the play-back can start. Under this condition an interrupt of the play-back is very unlikely. If the distance is increased to 10 km the necessary reservoir of video data raises to 400 seconds. And for an average distance of 15 km is goes even up to 600 seconds. Although these values sound very high the question is how long it takes to download so much data. During traversing an Access Point cell it is possible to download almost 700 seconds of video data if as assumed only one user is present. So after traversing the first cell it is already possible to start the play-back even with a network coverage of 24%.

Already mentioned before it is a fact that - in means of bandwidth - a limited backbone connection can throttle the performance of the wireless network. In Figure 2 it is shown that the peak capacity of the wireless broadband network is unused if the backbone cannot deliver the data fast enough.

In Figure 7 a comparison is made between download rates of legacy and Smart Caching enabled network setups. With Smart Caching it is easily achievable to maximize the download rate as always enough data is cached close to the Access Point so that the capacity of the wireless link can be fully used. But if no Smart Cache is employed the backbone can limit the performance and therefore decrease the overall performance.

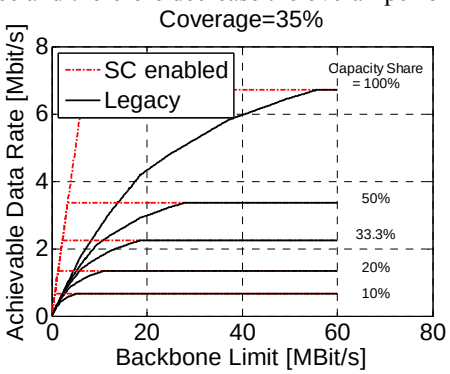

Figure 7: Capacity Estimation without Smart Cache

The calculation is based on the parameters x, y, and z out of Figure 3. The data which arrives during period x is transferred in period y. The resource or capacity share which is available for that is derived from the utilization of other services of the wireless network. The less free capacity is available the more data is stored in the cache and has to be forwarded during coverage. This extents the y period up to the maximum that it comprises the complete coverage region and period z vanishes.

This maximum is compared with the legacy case were the data is streamed from the application server somewhere in the backbone when coverage is re-established. Due to the backbone limit less data can be streamed compared to Smart Caching. While for Smart Caching the curve of the available data rate directly linearly raises to its maximum, which depends on the available capacity share, the curves of the legacy case show a different behavior. In such a situation they need high backbone limit in order to provide the same data rate.

6. Summary and Conclusion

For integrated networks of the next generation Smart Caching was introduced to provide broadband services even in regions of heterogeneous network performance. Especially for application areas like the discussed motorway scenario the employment of Smart Caching optimizes the network performance so that otherwise not supportable service can be provided. Mobile users of non interactive and non real time application can highly benefit from Smart Caching.New emerging applications like IPTV of Video on Demand can substantially benefit from it [11].

The analyses show that the impact of Smart Caching on other services is limited or not risky so that their quality is not decreased. Furthermore guidelines for the size of the Terminal Buffer are provided. The advantage of Smart Caching over legacy network setups is clearly outlined.

The introduced queuing system and the modeling of the different network parameters are very sophisticated. This allows further detailed analyses and can also be used for other investigations. It offers a wide range of application areas.

References

[1] H. Balakrishnan, S. Seshan, E. Amir, and R.H. Katz, Improving TCI/IP performance over wireless networks, In Proc. 1st ACM Int'l Conf. on Mobile Computing and Networking (Mobicom),pp. 2—11 , Nov. 1995.

[2] T. Takine, Queue Length Distribution in a FIFO Single-Server Queue with Multiple Arrival Streams Having Different Service Time Distributions, In Queueing Systems, Vol. 39, No. 4, pp. 349-375, Springer 2001.

[3] H. Heffes and D. Lucantoni, A Markov Modulated Characterization of Packetized Voice and Data Traffic and Related Statistical Multiplexer Performance Selected Areas in Communications, IEEE Journal on, 1986, 4, pp. 856-868.

[4] H. Hassan, J.M. Garcia, C. Bockstal, Aggregate Traffic Models for VoIP Applications, In International Conference on Digital Telecommunications (ICDT'06) , p. 70, 2006.

[5] B. Maglaris, D. Anastassiou, P. Sen, G. Karlsson, and J. Robbins, Performance models of statistical multiplexing in packet vide ocommunications, Communications, IEEE Transactions on, 1988, 36, pp. 834-844.

[6] P. Seeling, F.H. Fitzek, and M. Reisslein, Video Traces for Network Performance Evaluation - A Comprehensive Overview and Guide on Video Traces and Their Utilization in Networking Research Springer, 2007, p. 274.

[7] T. Takine, The Nonpreemptive Priority MAP/G/1 Queue, Operations Research, Vol. 47, No. 6, 1999, pp. 917-927.

[8] C. Hoymann, Analysis and performance evaluation of the OFDM-based metropolitan area network IEEE 802.16, Computer Networks, 2005.

[9] S. Baum et al., "Final report on link level and system level channel models," Winner Project, Tech. Rep. D 5.4, 2005.

[10] C.-H. Rokitanski, Ed., Validation of the Mobility Model of SIMCO2 with Dutch Motorway Measurements. PrometheusWorkshop on Simulation, 1992.

[11] S. Goebbels and K. Schmolders , Smart Caching for Supporting Video Streaming in Heterogeneous Wireless Networks, In Proceedings of European Wireless 2007, Paris, France.