SMART CACHING FOR SUPPORTING VIDEO STREAMING IN HETEROGENEOUS WIRELESS NETWORKS

Stephan Goebbels and Karsten Schmolders

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

Abstract - Broadband wireless networks will be a key feature of future communication technology. Mobilizing the services and applications used from desktop PCs is the next logical step in enhancing the Internet. But especially video streaming services burden wireless networks with resource consuming tasks. Due to this reason the realization of high quality streams is hard to achieve through wireless LANs.

Especially in dense populated urban areas WLAN systems suffer from shadowing and propagation effects which limit the coverage range and performance. A complete illumination in such environments with broadband access is not achievable. Moreover areas of good coverage will interchange with regions of only rudimentary or even no wireless access. This heterogeneity in network performance causes a lot of problems for streaming services. Usually they require a fixed certain level in bit rate. Although streaming protocols in the future will be adaptable to varying bandwidth the quality reduction if often not acceptable. Buffering techniques which are already employed on today's protocols may partly solve the problem but only more sophisticated approaches, namely Smart Caching, will fully succeed.

Smart Caching is a technique to overcome the problems caused by the diverseness of available radio bandwidth. Its purpose is to maximize resource utilization in regions of broadband access and to live on buffered data in periods of low network performance or broken connections. For video streaming services this implies that a sufficient amount of data is stored in the user terminal to bridge even longer areas without reception. The optimization of user data caching and transport is the main functionality of Smart Caching.

Keywords - Smart Caching, Video Streaming, Heterogeneous Wireless Networks.

I. INTRODUCTION

Video streaming in wireless networks has the problem that the required bandwidth of the stream and the offered data rated of the network seldom fit together. The major objective is to transfer as much data as possible to the end device so that the user can continue watching the stream even if no connection is currently available. Hence, it is necessary to optimize the download rate especially under heterogeneous network conditions [1]. However, for broadband wireless networks it exists the drawback that the capacity of the link is, under perfect conditions, much higher than the usual end-to-end connection in the Internet. In such situations the wireless transmission is thwarted by the backbone. On the other side the core network can deliver data much faster than the average throughput in the wireless network is. So in this case packets would pale up in front of the wireless link. Both aspects Smart Caching make use of.

For video streaming this means that either the user terminal is close to the access node and has a good reception so that the video traffic just constitute a fraction of the overall throughput. Or the end device is far away or in a shadowed area so that almost the whole resources of the access node are necessary to carry the required data volume.

Such a situation is depicted in Figure 1. A mobile user is traversing an urban environment. Although several access nodes cover the scenario the user often gets into shadowed regions. Since he is consuming a video stream it is important to buffer sufficient amount of data to bridge these gaps in radio coverage.

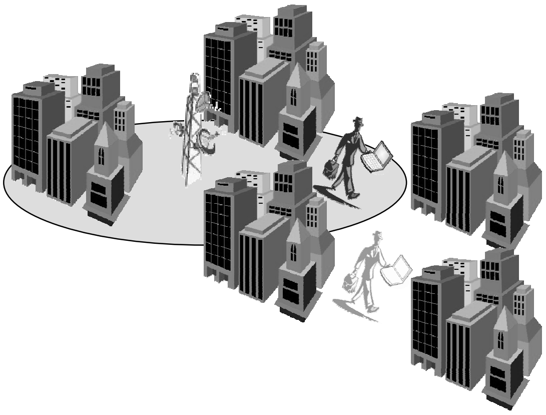

Fig. 1
Application Scenario for Smart Caching enhanced Video Streaming

The Smart Caching approach follows the aim of buffering user data at the edge of the core network. This data can be used to fully exploit the wireless link in periods of good channel conditions. Contrary to that at phases of worse condition the data can be buffered and used as supply for later transfer.

Thus, it is possible to fully exploit the wireless network as always enough data in the reservoir, namely the Smart Cache, is available which is dedicated for transmission. Furthermore the throughput of the network is increased as no idle periods or resources emerge anymore and the overall performance is maximized. The decoupling of cabled and wireless links by Smart Caching is the key idea of this approach.

Smart Caching is also suitable for other services like file transfer or advanced prefetching of data. Nevertheless due to the caching of data it suffers higher delays than legacy approaches so that it is only applicable to services with weak delay constraints. Bidirectional services or conversational applications are not in the focus of Smart Caching.

II. Smart Caching

The idea of Smart Caching rests upon the separation of the server - client connection into sections of cabled and wireless hops. The Smart Cache at the edge of the core network is the buffer which absorbs data currently cannot be forwarded and provides it to fully use up bandwidth at periods of perfect link condition.

Smart Caching is independent of the underlying transport protocol. The hops between server, Smart Cache, and client can employ all types of legacy IP and transport protocols. The interrelation of the different entities which are involved in a Smart Caching enabled scenario are depicted in Figure 2. A full overview is given in [2].

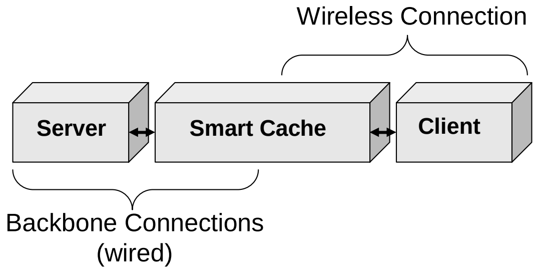

Fig. 2
Smart Caching Architecture

Client and server are the usual communication partners used from today's video streaming applications. Between them resides the Smart Cache which receives the data from the server, stores it for a while and forwards it finally to the client node. The caching of the data and the separation of the end-to-end connection makes it necessary to terminate the ingress data path in the cache. The data packets have to leave the transport layer and are transferred to a caching entity. The organization of the cache itself is not of further relevance for the analysis and therefore not discussed. It can be seen as a simple first come first serve buffer which perfectly fits in the later queueing analysis. Finally, depending on the available resources the packets are forwarded to its destination, the client node.

The Smart Cache itself is separated into three sub entities. The caching server is responsible for retrieving the data out of the storage and sending it to the actual communication client. For this an active request of the communication partner is necessary, as otherwise the server is not informed by which link and access node the data should be forwarded.

At the beginning of each communication session the client requests a video stream from a server somewhere in the Internet. Contrary to typical systems the data packets are not routed directly to the client but take a detour through the Smart Cache. At this node the packets are received be the caching client sub entity which hands them over to the actual storage. From this buffer the packets are retrieved and forwarded by the server sub entity when an actual client request is pending.

As long as the connection is available the data is forwarded to the client. If the wireless link is broken due to a handover or because the user has left the coverage area of the wireless network no more packets are requested and the delivery is held.

Not affected by this is the filling of the Smart Cache through the first intersection. As soon as the client terminal reaches a new coverage zone and has re-established a connection to the network a location update is send to the Smart Cache together with the request for new data. Now not only the actual incoming packets but also on top of it the cached data is forwarded to the client. Hence, as long as packets are available in the buffer the bandwidth exploitation of the wireless link can be optimized.

Obvious in this setup is the imbalance between management overhead on the first subsection between server and Smart Cache compared to the wireless hop. Most of the management traffic is restricted to the latter hop. Therefore the reaction on changing link conditions and broken connections is much easier to control and latency times are minimized. Furthermore protocol overhead will be kept away from the backbone network.

Nevertheless the data which is stored at the edge of the wired network has to be accessible from several network nodes otherwise a reuse of data is not possible. If the data is buffered directly at the AP the user could leave the corresponding cell and the data would be lost. Only if the user would return in the same cell he could reuse the already buffered data. Thus, it is necessary to combine all APs of a certain area by one Smart Cache. Than it is possible to reuse the data even if the current WLAN cell is left. As soon as another cell is reached the data can be directly downloaded form the Smart Cache instead of requesting it again from the server.

The grouping of a certain amount of APs to a cluster can even be extended to other wireless communication systems. The integration of UMTS like systems is easy to achieve. Especially if IP Multimedia System comes into play and the core network traffic is handled by the IP protocol. Thus, it is possible to install the Smart Cache on IP level so that a rerouting of data is very easy achievable.

III. System Model

For the further evaluation of the Smart Caching approach queueing theory is applied. The Smart Cache itself is seen as a first come first serve queue which stores the packets before they are forwarded to the final destination. The wireless link between WLAN Access Point (AP) and user terminal is modeled by the processor of the queue. The current size of the queue in this model reflects the number of cached packets and the service rate of the queue is the transmission rate of the wireless communication system. The available throughput of the system is the arrival rate of the queue as ingress and egress data rate of a queueing system are always the same. Otherwise the system would get overloaded.

A. User Scenario

For the investigation of Smart Caching the following user scenario is assumed. A mobile user is moving through a patchy covered urban environment. The scenario is partly covered by WiMAX [3] Access Points witch provide broadband wireless access. Some regions of the scenario are not covered with wireless access so that the fractions of coverage can be varied between 0 and 100%.

Depending on this coverage the distance between two successive APs is varied so that periods of coverage and phases of no coverage in between comply with the aforementioned fractions. Here a average user velocity of $1\frac{m}{s}$ is assumed. This means that in the case of 50% network coverage ($p_{cov} = 50\%$) and a coverage range per AP of for example 50m after a period of $100s$ network coverage the connection is interrupted for another $100s$ before the next AP is reached. During the idle gap in between all ingress packets are stored in a Smart Cache which is responsible for all APs in the scenario. After coming into range of the next access node these buffered packets as well as the current traffic is forwarded to the user terminal. Hereby the capacity of the wireless link can be fully used as enough data is already in the queue.

The arrival rate of the queue λ has to be calculated from the packet size and the ingress data rate. As the queue is based on the assumption that data is handled by chunks of packet size this conversion is necessary. This has to be performed as it is a well known fact that packet size has influence on the delay of information. The packet size is set to reasonable value of 100 byte.

Together with the utilization of the overall scenario ρ the parameters λ and coverage p_{cov} can be freely chosen. However, they are obviously dependencies between each other, so that the change of one of them affects the other parameters. For example it holds that $\rho \geq 1 - p_{cov}$. Even if no user traffic is transported the queue model is constructed in such a way that the utilization is always equal to the rate of no coverage. Therefore the real utilization may only vary between $1 - p_{cov}$ and one.

The WiMAX radio propagation model is based on free space propagation with adapted attenuation factor γ. For urban scenarios the range of coverage is very limited due to obstacles so that a γ of up to 5 is reasonable [4]. For the later analysis a less restrict value of 4 is chosen.

The transmission power of the APs is assumed to $100mW$. So the maximum range of a WiMAX cell with that sending power is around $50m$ where the annulus of the slowest PHY mode covers approximately 30% of the whole cell. All relevant data of the PHY modes is summarized in Table 1.

The fractions of coverage of the overall cell in Table 1 are used to calculate the path probabilities $p_1 - p_n$ in Figure 4.

B. Service Time Distribution

The modeling of Smart Caching with an analytic queue model requires special attention on the service process. The application of Smart Caching is based on heterogeneous network scenarios with varying link performances. Especially periods of lost coverage are of special interest. It is of vital relevance to include all these aspects in the later model. Due to changing PHY modes in systems like WiMAX the available transmission rate strongly depends on the link conditions between AP and user terminal. Assuming an ideal scenario where users are equally distributed within a WLAN cell. If the radio propagation conditions into all directions are almost the same than areas of equal C/I level constitute concentric circuits. Moreover regions which can use the same PHY mode forming an annulus. The size of such an annulus compared to the cell size is the residing probability of a user to be served by a transmission rate corresponding to the PHY mode. It is denoted by p_{PHY_i}. Furthermore in any partly covered scenario there is a certain likelihood p_{cov} to be within the range of coverage of any AP or outside of it ($p_{out} = 1 - p_{cov}$).

These facts perfectly fit to the Cox model [5] of a hyperexponential phase model for the service process. Each job which enters the server of the queue is distributed to one of several paths while each path has different average service rates μ_i and service time distributions. Service rate as well as path probability can be deduced from the residence probabilities in the PHY mode annulus and their corresponding transmission rates.

The likelihood of whether being covered by an access point or not has to be translated into path probabilities. Therefore it is important to understand how the absence of coverage can be modeled by a service time distribution. It can be solved by extending the service time of the packet which is in the processor right in the moment the coverage area is left. The service time is expended to a value which corresponds to the period of no coverage. This

Table 1
PHY Mode Parameters

Modulation & Coding Scheme	Throughput [6]	Switching Point [7]	Radius	Fraction
BPSK 1/2	6.1 Mbit/s	6.4 dB	48.32m	29.2%
QPSK 1/2	12.2 Mbit/s	9.4 dB	40.66m	13.25%
QPSK 3/4	18.6 Mbit/s	11.2 dB	36.66m	25.92%
16QAM 1/2	24.7 Mbit/s	16.4 dB	27.17m	5.92%
16QAM 3/4	37.2 Mbit/s	18.2 dB	24.5m	10.39%
64QAM 2/3	49.7 Mbit/s	22.7 dB	18.9m	2.72%
64QAM 3/4	55.8 Mbit/s	24.4 dB	17.15m	12.58%

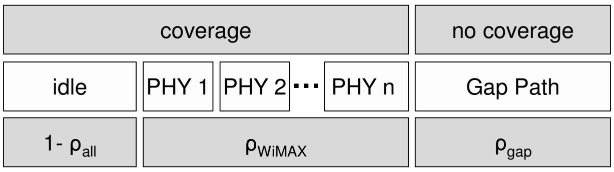

Fig. 3
Coverage and Utilization Interrelation

means that during the distance between two succeeding regions of coverage the whole time one and the same job, so called the ”gap packet”, is served and just finished if a new coverage region is reached. This means that virtually the processing of queued jobs is stopped. During the whole gap period the state of the queuing system does not change. The probability for a packet to use the ”gap path” is denoted by p_{gap} and the probability to use one of the other WiMAX paths is $p_{WiMAX} = 1 - p_{gap}$. An overview of the involved probabilities and there dependencies is depicted in Figure 3. The period of network coverage is divided in periods where packets are forwarded and phases when the queue is empty. Furthermore the server is busy if ”gap packets” are served, hence if a loss of coverage is modeled. During normal packet transmission depending on the current PHY mode different service paths are chosen. Therefore the utilization of the overall system consists of one part caused by the modeling of the idle phases (packets are using the ”gap path”) and a real utilization sourced by simple packet forwarding. Obviously p_{out} and p_{gap} do correspond in some way but do not have the same value. Otherwise a wireless network coverage of 50% would imply that half of the packets are served by a rate complies with the gap between APs. The interrelation between both factors is given by the utilization

$$\rho = \frac{\lambda}{\mu}$$

of the different paths of the phase model. The utilization of the lowest path of Figure 4 which reflects the periods of lost connection must match p_{out} so that the server is the same fraction of time busy with ”gap packets” as no coverage can be supplied. This can also be seen from Figure 3.

$$\rho_{gap} = \frac{\lambda p_{gap}}{\mu_{gap}} = p_{out} \tag{1}$$

The other parameters of Figure 4 are now related to the residence probability in each PHY mode area. However, the path probability of the phase model does only reflect the likelihood that a packet is processed by the corresponding service rate and is therefore **not** proportional to the residence probability. The residence probability denotes the fraction of time the system operates with one or the other transmission rate. In the system model this value is given by the quotient out of path probability and average service rate. So the fraction of time the server operates with transmission rate i is given by:

$$p_{PHY_i} = \frac{p_i/\mu_i}{\sum_{k=1}^{n} p_k/\mu_k} \quad \text{with} \quad \sum_{k=1}^{n} p_k = 1.$$

The corresponding linear equation system has to be solved to get the according path probabilities. For the service time distribution of each separate path negative exponential distributions with different mean values are chosen. Although other distributions are also possible analyses have shown that the distribution of the WLAN paths has only minor influence of the later results. Relevant for the choice of the distributions of the gap path is their second moment as this has some influence on the average waiting time. A deterministic distribution for the gap path would mean that the distance between two successive cells is always the same. As no further assumptions on the behavior of the inter cell distance is made the negative exponential distribution as it is of medium variance.

C. Arrival Process

For the arrival process of the incoming packets two different processes are taken. The first one is the well known Poisson Process which is widely distributed in queueing theory. It has special characteristics which allow an easy derivation of queue parameters like average queue length or packet delay. A detailed analysis is given in [8]. For further evaluations secondly a batch arrival process is taken. Here user data comes in groups of several packets as used from MPEG streams [9]. This model was basically developed by Lucantoni and the main results can be found in [10] and [11].

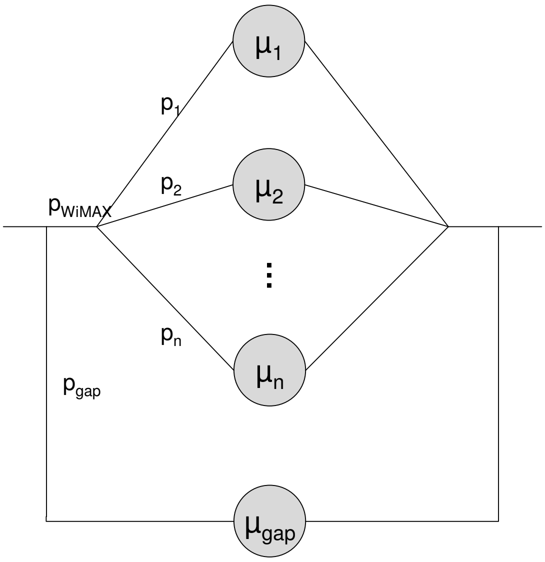

Fig. 4
Phase Model of Service Time Distribution

In both cases the incoming bit rate has to be packetized. As in each packet switched network data in broken down in chunks which have to be transmitted over the link. Each packet is one job for the queuing network.

IV. Results

The applicability and the benefit of Smart Caching depend on network parameters like coverage, required data rate and tolerable packet delay. In the following it is shown how these parameters depend on each other.

In Figure 5 the required data coverage depending on the average waiting time is shown. For smaller waiting times better network coverage is necessary. The less tolerate an application is to delay and latency the more network coverage is necessary. Delays less than a couple of seconds of course always require an unbroken network connection. But such kind of services is not in the focus of Smart Caching. It is also shown which influence the average data has on the curves. For higher average throughput rates the dependency between coverage and waiting time weakens. The closer the value gets to the maximum available data rate of a WiMAX cell (under the given assumptions) the smaller is the change in the coverage value. This can be explained by the fact that for higher data rates the most of the delays is caused by gaps in the Network connection so that less restricted delay constraints have only a minor impact on the parameters.

This is also second by Figure 6. Here the required coverage is shown as function of the average data rate. All curves with the different delay constraints meet in the same point. Rates of 21.8 Mbit/s which is the maximum value in the given scenario require 100% network coverage irrespectively of any waiting time requirements. But for smaller data rates it is of course of interest which average packet delay is assumed. As already seen in the previous figure small delays require high network coverage even for small bit rates. But on the other side for services which may tolerate longer waiting times the required coverage can be much reduced even for high bit rate services.

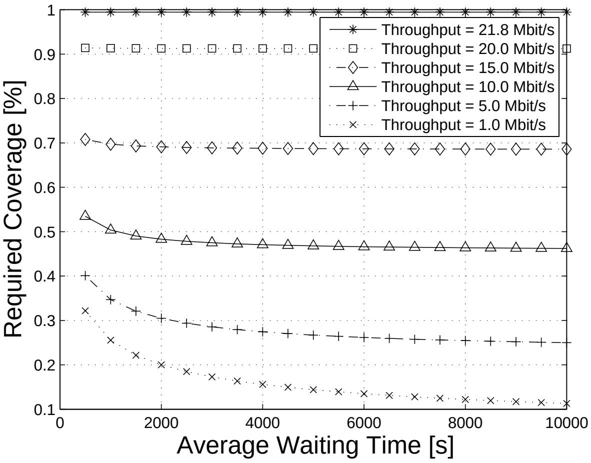

Fig. 5
Influence of average Waiting Time on required Coverage Portion

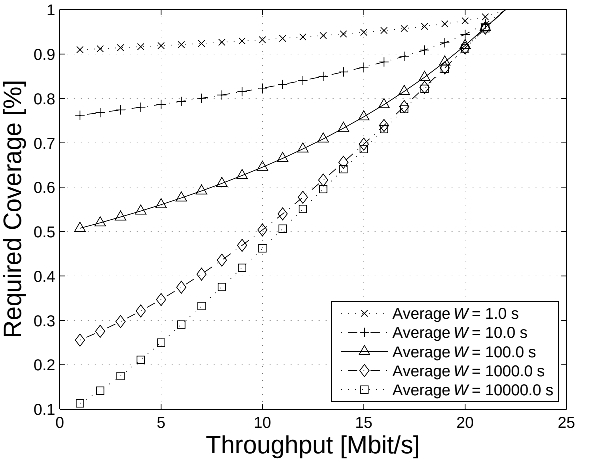

Fig. 6
Required Coverage depending on assumed Traffic Load

Another advantage of Smart Caching is illustrated in Figure 7. Here the relative impact of the variance in bit rate of the traffic source is shown. Usually the delay of packets is seriously influenced by the type of the arrival process. Pure Poisson arrivals usually have a much smaller delay than services with higher variances in the interarrival time. The more regular packets arrive the better the performance of the queueing system and therefore of the network. To investigate which impact this has on Smart Caching two scenarios are compared - firstly a Poisson stream which is often used for such purposes and secondly a batch arrival stream which has a much higher variance in the interarrival time. The relative difference between the waiting time of both models is shown in the Figure. The more Smart Caching is employed, which occurs for smaller coverage ratios, the less influence has a turbulent arrival process.

Due to the buffering of arriving packets any irregularity in interarrival periods is smoothed and in some cases completely removed. So if almost all packets are buffered before they are transmitted to the user terminal only the average data rate is of interest and no other parameter of the arrival process have to be considered.

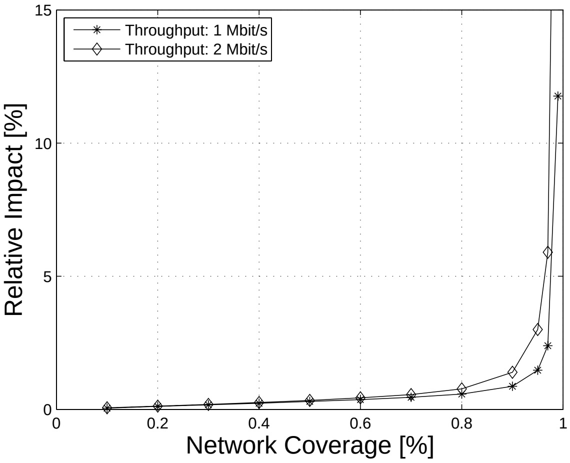

Fig. 7
Influence of average Waiting Time on required Coverage Portion

V. Summary and Conclusions

For the employment of video streaming services in future heterogeneous wireless communication systems applications like Smart Caching are necessary to increase performance and user convenience. By Smart Caching it is possible to overcome problems caused by patchy network coverage and variance in link performance. It is stated how Smart Caching fits into future wireless communication systems. Such enabled networks will provide much improved performance for unidirectional and non real-time services. Especially for video streaming applications like IPTV or Video on Demand [12] the new approach is extremely suitable.

The queueing model for the analysis of Smart Caching gives a good possibility to assess the capacity of the new approach. For finding the trade off between network coverage, available average data rate and packet delay the analytic model can be used. The dimensioning of smart Caching enabled wireless networks can be based on these results. It was shown that full wireless broadband network coverage is not required for video streaming services and that the employment of Smart Caching therefore can decrease deployment costs of new network setups. In the analysis it was shown that Smart Caching has specific boundaries inherited of the scenario preferences which cannot be crossed. It was shown that packet size and therefore variance of the arrival process have only minor impact on the overall performance

The applicability of Smart Caching as a possibility to improve the performance of future heterogeneous network structures could be verified. The analytic method will allow the easy and fast adaptation of the model towards new network architectures and communication protocols.

References

[1] R.H. Frenkiel et al., *The Infostations challenge: balancing cost and ubiquity in delivering wireless data*, Personal Communications, IEEE, Vol. 7, No. 2, p. 66-71, 2000.

[2] S. Goebbels et al., *Smart Caching in Service Specific Overlay Networks for Wireless Networks*, IST Mobile Summit 2006, Mykonos, Greece.

[3] IEEE 802.16 WirelessMAN, compare http://www.ieee802.org/16/.

[4] B. Walke, *Mobile Radio Networks: Networking, Protocols and Traffic Performance, 2nd Edition*, October 2001.

[5] D.R. Cox, *A use of complex probabilities in the theory of stochastic processes*, In Proceedings of the Cambridge Philosophical Society, 1955.

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

[7] C. Hoymann, *MAC Layer Concepts to Support Space Division Multiple Access in OFDM based IEEE 802.16*, In Wireless Personal Communications, p. 23, Springer Netherlands, May 2006.

[8] L. Kleinrock, *Queueing Systems*, Wiley, 1975.

[9] J.A. Zhao et al., *MPEG-4 Video Transmission over Wireless Networks: A Link Level Performance Study*, Wireless NEtworks, Vol. 10, No.2, p. 133-146, Springer, 2004.

[10] D.M. Lucantoni, *The BMAP/G/1 QUEUE: A Tutorial*,Lecture Notes In Computer Science, p. 330-358, Springer-Verlag London, UK, 193.

[11] D.M. Lucantoni, *New results on the single server queue with a batch Markovian arrival process*, Stochastic Models, Vol. 7, No. 1, 1991.

[12] T-Online Video on Demand, compare http://www.t-online-vision.de/.