Enhancements in Wireless Broadband Networks using Smart Caching

An Analytical Evaluation

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

Abstract— **Since ubiquitous mobile communication is already true in many countries in the future the provisioning of broadband access for mobile users is the key functionality. For mobile users the performance of the radio network vitally depends on the length and the characteristic of the transmission path between user terminal and access point. Due to the continuously changing conditions the link quality will be subject to high variations which make especially broadband services hard to realize. In urban environments a full broadband coverage is therefore hard to provide.**

In the following the Smart Caching technology is presented which allows the overcoming of these problems at least for non-real-time and non-interactive services. It uses the approach of pre-fetching and buffering data at the edge of the core network and transmitting it whenever the user gets connected to an access node with broadband connectivity.

The paper provides a detailed description of Smart Caching and an analytical evaluation of the advantages of Smart Caching enabled networks compared to legacy setups. Furthermore dimensioning issues are covered. For the analysis advanced queuing theory is used. User services are modeled in very detail as well as radio link, network and user behavior.

Key Words— Smart Caching, Queuing Theory, Mobile Broadband Networks.

I. INTRODUCTION

The integration of different wireless networks will form the communication system of the future. The capabilities of the different networks can so optimally used since always the best suitable system can be taken. Some of the networks can provide almost ubiquitous coverage so that users are always connected to the Internet. On the other side intermittent networks like WLAN or WiMAX provide much higher data rates at so called Hot Spot zones so that massive data transfer are only possible in such networks.

This heterogeneity is caused by the fact that the distance between access node and terminal and the achievable data rate is directly related to each other. Hence, 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.

Broadband services like video streaming cannot be supported by such an integrated network. The bandwidth requirements usually are not met by the overlaying network. For mobile users this implies that the service is frequently interrupted – as soon as they leave the coverage zone of a broadband system. 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.

The frequent disruption of the broadband connection nowadays implies a simultaneous interrupt of supported service. But non-interactive and non-real-time services like video streaming have low delay constraints. Between transmission of the actual video data and the play-back of the same a delay of several seconds is easily acceptable. More relevant in means of Quality of Service is the continuous and uninterrupted play-back of the video. Smart Caching makes use of this characteristic by exploiting intensive buffering techniques. At the edge of the fixed backbone network, close to the access point where a mobile terminal is roaming, reservoirs of data are build. Here data is pre-fetched and stored for services like video streaming. The buffered data is kept until the terminal comes into service range of an access point or a timer expires. As soon as a mobile terminal enters the service area of some access point the cached data is transferred to be again buffered in the user terminal at maximum possible speed. Compared to video streaming as supported by the Internet the amount of data buffered in the end device is much larger with Smart Caching, so that even long intervals of low network performance or service interruptions of up to tens of seconds are covered and become invisible to the user.

Of course, mobile networks have to handle more than just video streaming services. Furthermore each service has its special requirements so that e.g. Voice over IP (VoIP) is completely unsuitable in intermittent networks. But VoIP has only basic bandwidth requirements so that for such services wide area networks, e.g. like UMTS, can be employed. In spite of that, if other wireless networks with higher capacity are available the handover of the connection to them might be useful. The new customer represents only a minor load for the broadband network while the overlaying network is substantially disburdened.

II. SMART CACHING

To overcome the problem of the combination of wired and wireless networks Smart Caching decouples the end-to-end data path. By locating a buffer at the edge of the backhaul two subsections are introduced. The first section contains the links between application server and the new buffer, which is called Smart Cache. It completely relies on wired links. So usually one would assume a higher average bandwidth compared to wireless connections as each link of the backhaul can provide much higher data rates. Nevertheless the end-to-end connection consists out of several links and they are not exclusively reserved for that connection. This leads to the fact that each chain is as strong as the weakest link. So if only one link is overloaded or too small dimensioned the end-to-end data rate is diminished. The end-to-end connection performance in fixed Internet does not reach the capacity of wireless broadband links that achieve even today a data rate of several tenth of Mbit/s while the capacity of end-to-end connections in the Internet is usually smaller than 10 Mbit/s.

The subsection between the Smart Cache and the actual data sink, the user terminal, comprises 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 Smart Cache in the middle of the data path has now the function to pretend a continuous radio connection even when the link is broken in the meantime. The streaming of data on the first subsection simply continues but the data is not forwarded via the wireless link but instead buffered in the Smart Cache. If a radio connection is re-established the buffered data is transferred. This separation might require the break up of TCP end-to-end connections. Developed for a similar purpose the SNOOP protocol [4] can be used in such cases.

To achieve the highest average data rate which is necessary for an optimal streaming process it is necessary to always fully exploit the capacity of the wireless link. Due to the decoupling of the end-to-end connection by the Smart Cache the throughput on the first subsection is not affected by the current capacity of the radio link. But on the second hop now it is possible to always stream as much data as possible depending on the current quality of the wireless link. Since data is buffered in the Smart Cache there is kept a reservoir which can be used to fill up the otherwise unused resources of the wireless connection. The massive delivery of user data can in the following be used to bridge periods of no radio service.

But for real time services like VoIP such a proceeding is not applicable. Such services are operated like in nowadays networks - there is no change. For non real-time services like video streaming the data is routed through a Smart Cache and from there on forwarded to the access point. It now depends whether the wireless link is available or not if the data is forwarded to the user terminal or temporarily stored in the Smart Cache. As soon as the connection is available the 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. Only after reaching a certain fill level the play-back begins. This should guarantee that the play-back of the video is not interrupted even if the radio connection is lost for a while. Initial fill size and length of the idle period are strongly related.

III. NETWORK MODELING

For the analysis of communication networks applied queuing theory can be used. 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.

A. Modeling of Service Process

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.

B. Traffic Modeling

The arrival process of the queuing model can be modeled by using the Markov Arrival Process (MAP). In [1] a method is introduced which allows the feeding of one 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 transitions without a packet arrival, the matrices D_i cover transitions with a simultaneous arrival of a type *i* packet.

In the following analysis two different types of Internet traffic are considered. These are VoIP and Video streaming

traffic. Already in [6] the modeling of several superposed VoIP sources to a Markov chain is introduced. Furthermore this can be transformed in a Markov modulated Poison Process with two states (MMPP[2]) [7]. Hence, it 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 and idle periods in between are taken from [6] while the other parameters are adopted from the G.711 recommendation (Values are summarized in Table 1).

TABLE I. VOIP SOURCES' PARAMETERS

	Parameters			
	T_{on}[s]	T_{off}[s]	Packet Rate [1/s]	Packet Size [Byte]
VoIP (G.711)	0.352	0.650	100	80+40

In [8] 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 [5]. They are summarized in Table 2.

TABLE II. VIDEO SOURCES' PARAMETERS

	Parameters				
	Cov. (0)	Cov. parameter	Data Rate [10^6 bit/s]	Packet Size [Byte]	M
Video (MPEG4)	0.509	0.09	3.37	188+40	20

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 [3]. 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 from the product of the size of the separate input matrices $D_{0,i}$.

C. Comparison to Legacy Network Setups

In order to fully understand the benefit of Smart Caching it is important to compare its performance with the legacy case. Especially for bandwidth consuming services like video streaming heterogeneous broadband networks reveal a serious problem. Smart Caching allows the buffering of user data, if no sufficient wireless connection is available, at the edge of the wired network close to the access node and hence close to the consumer. In ordinary networks data is discarded if one of the links of the end-to-end connection is broken – most likely the wireless hop.

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 the upper part of Figure 1.

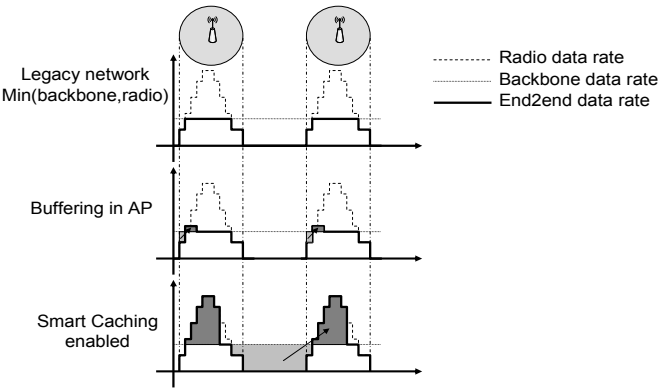

Figure 1. Smart Caching vs. Legacy Setup

The un-shaded area reflects the capacity which is lost for the customer if no Smart Cache is employed. 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. By simply buffering data in the access point the problem cannot be solved as can be seen in the middle part of Figure 1. Only a small amount of data can be stored and later on used to fill up the unused wireless network resources. Only by using a Smart Cache which combines several access nodes of a wireless network it is possible to use the buffered data even if the access point changes.

D. Impact of Radio Link Failure

Which impact the employing of Smart Caching has on the amount of stored packets in the Smart Cache and the additional packet delay it causes is depicted in Figure 2. If a user moves between two areas of network coverage without broadband connectivity all incoming packets are stored in the Smart Cache. Only after re-establishing the connection in the next access zone these packets are transmitted. Due to the length of the idle periods without connectivity the delay of the packets is drastically increased.

Which influence the gap has on the actual packet delay is depicted in Figure 2. 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 decrease. 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.

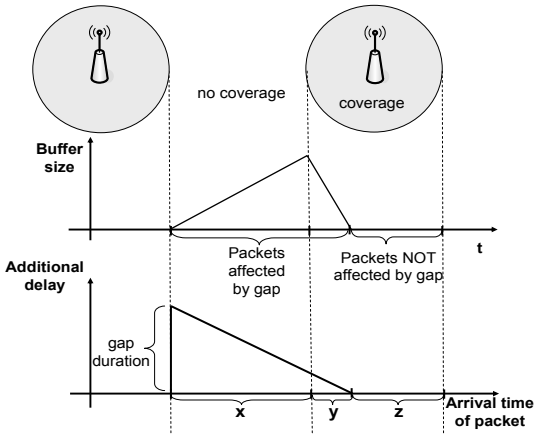

Figure 2. Impact of Smart Caching

In the lower part of Figure 2 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:

$$(x+y)/(x+y+z) = \text{ratio of influenced packets.} \quad (1)$$

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.

IV. User Scenario

For the evaluation of the Smart Caching approach an urban scenario is chosen. In a densely populated environment a mobile pedestrian user should traverse the streets while accessing a video stream. For the street grid the well known UMTS 30.03 Manhattan grid is used which contains 200m sized blocks and 30m wide streets. Access points of the WiMAX network are located at every second crossover. The crossovers between them are not covered which reflects the fact that in densely populated urban environments full coverage with broadband technologies is not possible and always areas exist which cannot be reached.

The WiMAX access points are operated with a transmission power of 100mW. The switching points between the different PHY modes and the corresponding data rates are taken form [2]. Since the signal strength of radio transmissions attenuates with the distance it is important to calculate the path-loss a radio signal perceives depending on the position of the terminal. The signal strength at a receiver is given by an adapted free space model with attenuation factor γ of 3 (2 is free space up to 5 for indoor).

In the previous section it was declared that the additional delay a packet perceives due to the coverage gap is mainly subject to the length of this period. It is determined by the distance between two WLAN cells and the average velocity of the user. The duration of the coverage gap is derived from mobility simulations where test users traverse the given scenario with a predefined average speed. From the measured values of the length of such idle periods the distribution of the gap duration is approximated.

V. Results and Evaluation

Smart Caching vitally depends on the ability of the wireless network to support much higher data rates than the average end-to-end Internet connection can provide. The otherwise unused resources which are shown in the upper part of Figure 1 are exploited by transferring the buffered data out of the Smart Cache to the user terminal. But such an approach is only possible if there are resources which are unused and not required by other services. The ratio by which the wireless network is available for the video streaming process is called capacity share in the following. Obviously the best performance of Smart Caching can be reached it the network can be exclusively used by the video streaming process (Capacity share = 100%). This is also shown in Figure 3. The red dashed line shows the available average data rate depending on the capacity share and the backbone limit. Up to a certain maximum the Smart Caching enabled network can transfer all data which arrives from the video server. The maximum is influenced by the capacity share which simply determines the average data rate which can be provided from the wireless network for the video session. The backbone limit denotes the data rate by which on average the data can be transferred between server and Smart Cache.

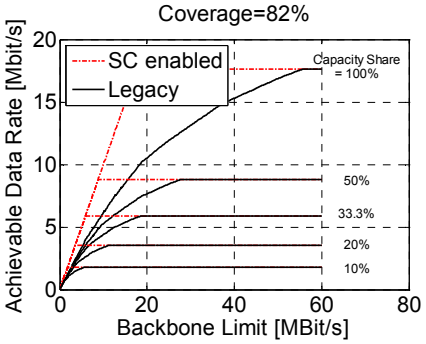

Figure 3. Achievable Data Rate

In legacy scenarios the backbone limit is of much higher relevance. As the end-to-end data flow is interrupted if the wireless connection fails there cannot be build a reservoir like with Smart Caching. If now the connection is re-established some of the wireless resources might be unused as the backbone cannot deliver the data fast enough. Therefore the backbone limit is of much interest in the legacy scenario.

In order to understand the benefit of Smart Caching enabled networks in Figure 3 also the achievable data rate in the legacy approach is presented (black solid lines). The area between them and the Smart Caching curves reflects the performance gain which can be achieved. Out of these curves it is also apparent that even for small capacity shares Smart Caching might improve the performance. The required backbone limit can be reduced substantially. But the question is if the backbone limit values are still realistic.

To cover dimensioning issues it is of interest how much video streams can be supported simultaneously. This depends on the capacity share of each single session and the available data rate – which might be limited by the backbone. In Figure 4 the number of concurrent streams for Smart Caching enabled and legacy systems is shown for the given scenario. It depends on the backbone limit. Obviously for high values of the backbone limit the results do not differ. If the limit gets close to the maximum wireless data rate there is no longer a difference between Smart Caching and legacy setups. Nevertheless it is shown that with Smart Caching right from the beginning the maximum number of streams can be supported while in the legacy setup the number slowly increases with raising backbone limit.

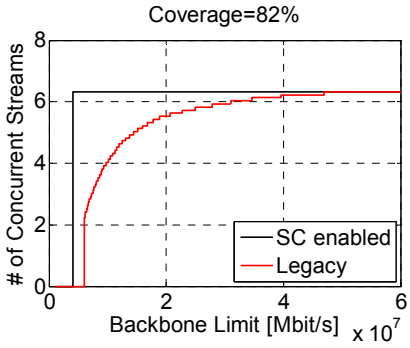

Figure 4. Number of Concurrent Video Streams

But in addition to the number of users for dimensioning issues it is important to know which packet delays have to be handled. The outcome of the queuing analysis is the Cumulative Distribution Function (CDF) of the packet waiting time. In Figure 5 different CDFs are shown depending on the user velocity and the number of VoIP users which are using the wireless network at the same time.

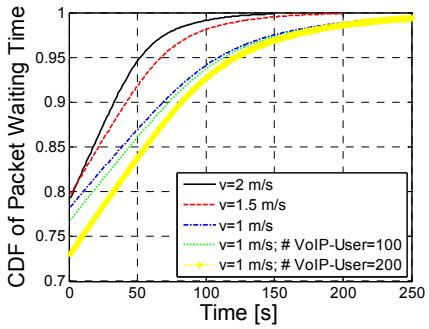

Figure 5. Packet Waiting Time

For fast moving pedestrians with a velocity of 2 m/s the 95 percentile of the waiting time is around 50 seconds. This means that if in the user terminal 50 seconds of video data are buffered before the play-back starts there is a good chance to watch the video without interrupts. If the velocity decreases this value grows to more than 100 seconds for v=1 m/s. By burdening the network with additional VoIP users the value is still increased but only slightly. So the major impact is provided by the duration of the gap which results out of the user's velocity.

VI. Summary and Conclusion

In this paper the approach of Smart Caching is clearly outlined and an analytical evaluation is provided. Smart Caching is demonstrated as one possible solution to overcome the problem of intermittent broadband coverage in heterogeneous networks at least for non-real-time services. Especially for new upcoming appications like IPTV or Video on Demand the benefit is essential [9].

It is shown that Smart Caching improves services and furthermore allows them in situations where otherwise no service provisioning would be possible. The impact on other service also carried by the same wireless network is marginal. Altogether a short waiting and pre-buffering period at the beginning of a video download allows mobile customers the seamless play-back of videos even under patchy network coverage. 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] 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.

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

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

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

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