Smart Caching in Heterogeneous Wireless Networks

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

***Abstract*—Future mobile radio networks will aim on broadband access for all 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. Furthermore in urban environments a full broadband coverage is hard to provide. Together, this will cause a high variation of link quality which makes especially broadband services hard to realize. This paper introduces a technique, called Smart Caching, which allows overcoming the problem of variation in network performance at least for 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.**

The paper shows the interaction of Smart Caching enabled services with legacy applications, it provides dimensioning aspects, and finally the resulting performance gain is discussed.

In the following for the analysis of system performance and the dimensioning of Smart Caching enabled radio networks advanced queuing theory is used. User services are modeled in very detail as well as radio link, network and user behavior.

***Key Words*—Applied Queuing Theory, Heterogeneous Wireless Broadband Networks , Smart Caching.**

I. INTRODUCTION

Future mobile radio communication will consist out of different integrated wireless networks. Some of them will provide ubiquity with basic Internet access while others allow broadband connections only in specific regions. 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.

However, video streaming and other non-interactive and non-real-time services have low delay constraints. Smart Caching is exploiting this property by pre-fetching and buffering a massive reservoir of data at the edge of the fixed backbone network, close to the access point where a mobile terminal is roaming. 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 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.

Nevertheless wireless networks have to handle a multitude of different services and user applications. This diversity comes along with several requirements. Contrary to video streaming services like Voice over IP (VoIP) have only basic bandwidth constraints but rather hard packet delay limits. For such services wide area networks, e.g. like UMTS, are perfectly suitable. 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.

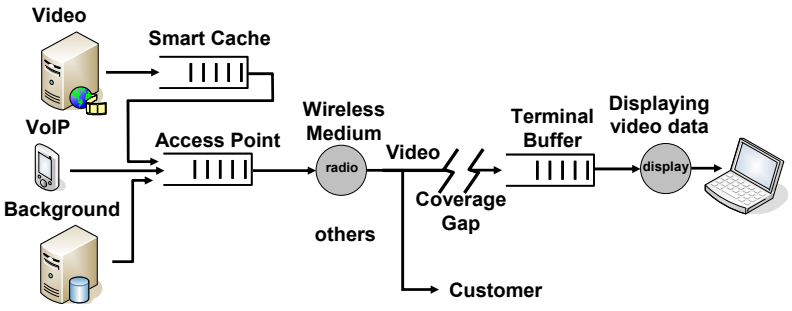

Figure 1: System Model for Smart Caching enhanced Wireless Broadband Networks

In the following it is shown which effects the different services have on each other and how Smart Caching improves the performance of integrated wireless networks.

II. SMART CACHING

Smart Caching allows the uncoupling of end-to-end data flows in combined wired and wireless networks. A buffer is positioned at the edge of the core network which introduces

two subsections. The first is between the application server in the Internet and the Smart Cache. It relies completely on wired connections. Although such connections 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. And hence, the end-to-end bandwidth is usually very limited compared to a separate link.

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 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.

For an optimal streaming performance over wireless the radio link must be exploited according to its current capacity. If data is pre-fetched in the Smart Cache the data throughput on the first subsection is not affected by the performance of the radio link. On the second subsection the data rate is adapted to the capacity of the radio link according to the resources available and the current link quality. The Smart Cache serves for decoupling of both sections. 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.

In Figure 1 a schematic illustration of a Smart Caching enabled network is given. Real-time services like VoIP are operated like in nowadays networks - there is no change. 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 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 playback begins. This should guarantee that the playback 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. The later analysis will provide a dimensioning instruction for it.

III. System Modeling

A natural approach to analyze communication networks is given by applied queuing theory. The combination out of queue and serving processor resembles the waiting of packets in a network node and the forwarding of them through the next link. The limiting resource especially in wireless networks is the radio capacity. Different traffic streams compete for radio access in order to transmit their packets to the customer. To use a queuing theory approach traffic streams have to be modeled by stochastic processes. Similarly, a random distribution for the service time has to be defined which approximates the behavior of the service process. This process describes the time which is necessary to transmit one packet through the air interface of an access point.

New evolutions in queuing theory allow a detailed modeling of such setups. In [2] methods are introduced which allow the modeling of the different incoming streams by a Markov Arrival Process (MAP). Furthermore each stream can be equipped with a separate service time distribution G. And the queuing delay of each arrival stream can be calculated separately. Then the arrival process is called marked MAP (MMAP) and 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.

A. Traffic Modeling

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 ($D_{0,i}$=-λ and $D_{1,i}$=λ) 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 and idle periods in between are taken from [3] while the other parameters are adopted form the G.711 recommendation (Values are summarized in Table 1).

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

Table 1: Parameter of VoIP Sources

In [5] it is shown that also video sources can be regarded as a superposition of a predefined number (M) 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 form [6]. They are summarized in Table 2.

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 $D_{0,i}$.

	Cov. (0)	Cov. parameter	Data Rate [10^6/s]	Packet Size [Byte]	M
Video (MPEG4)	0.881	0.075	1.317	188+40	20

Table 2: Parameter of Video Sources

B. Service Process

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

Radio networks like the chosen WiMAX system use different PHY modes to adapt the data transmission to the current quality of the radio link. The weaker a radio signal the more robust PHY modes are used. The robustness of a PHY mode mainly expresses in a decrease of the transmission rate. In the WiMAX standard there are certain specific quality levels, in means of signal to noise values, defined at which the PHY modes have to be changed. Therefore a certain PHY mode is applied between an upper and lower signal strength level. Together with the radio propagation conditions and the mobility of customers for each PHY mode a certain probability of being employed can be calculated.

To reflect the different data rates of each PHY mode in the queuing model a hyperexponential service time distribution is taken – compare Figure 2. 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.

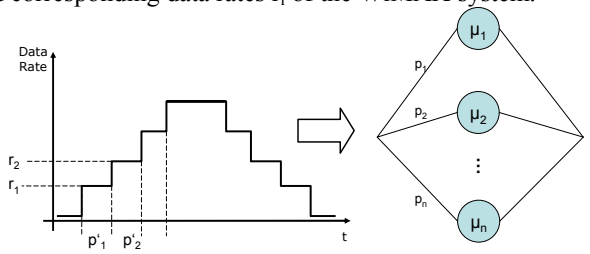

Figure 2: Modeling of Radio Transmission

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 Figure 3.

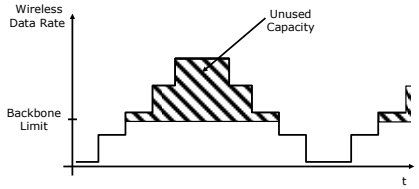

Figure 3: Impact of missing Smart Cache

The 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.

C. Influence of Coverage Gap

If a customer traverses a coverage gap between two successive access points the incoming packets of the e.g. 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 4. 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.

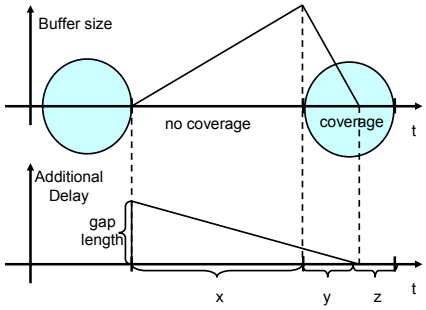

Figure 4: Influence of Coverage Gap on Buffer Size and Packet Delay

In the lower part of Figure 4 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.

IV. Network Scenario

The scenario to analyze the capability of Smart Caching enhanced wireless networks is derived from the frequently used Manhattan scenario out of the UMTS 30.03 technical report.

The street grid consists out of 200m sized blocks and 30m wide streets. WiMAX access points are placed every second crossover, so that areas around crossovers in between are NOT covered by them. This 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 [8].

Since the signal strength of radio transmissions attenuates with the distance it is important to calculate the pathloss 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. A lognormal distribution with small variance is chosen to model the time in the coverage gap. The small variance of the distribution is inherited from the fact that on longer distances the velocity of a user is subject to averaging effects so that outliers are rather unlikely. Additionally the distribution provides low probability for values much smaller than the average value. Since such small values would imply that user velocity is much higher than the average value it is a reasonable assumption. But furthermore much larger values are allowed with a certain probability. This reflects the behavior of user which took a short break in between or have to wait at traffic lights.

Despite of VoIP and video customers it is assumed that each WiMAX access point has to carry additional 10 Mbit/s of background traffic which comprises all other Internet applications currently used by the different customers.

V. Results

The applicability of Smart Caching depends on the impact it has on other services which have to coexist in the same wireless network and the performance of the enhanced services (e.g. video streaming). Both aspects have to be considered.

A. Impact on real-time Services

The buffering of customer data in the Smart Cache increases the amount of data which has to be carried by the wireless network. Before data which belongs to an interrupted connection was discarded at the access point and usually the corresponding communication session was finished. But in the enhanced network the data is cached and transferred later on. On the one hand this dramatically increases the perceived data rate for customers of non-real time services – on the other hand it burdens the network with additional load. In Figure 5 the packet delay of VoIP packets is shown depending on the number of sessions (One session includes up- and downlink) and the number of carried video streams. Since quality criteria of VoIP are not based on average delays but rather of percentiles here the 95% value of the distribution function is depicted. In the G.114 recommendation of the ITU a one-way delay of 150ms is given. This border is also included in the figure. For just one video customer the capacity of a WiMAX cell allows more than 200 simultaneous VoIP calls. With increasing number of video sessions this quantity naturally reduces. But even with seven video customers the network allows more than 30 VoIP calls. Due to the extensive bandwidth of WiMAX networks the employment of Smart Caching does not imply a overload of the radio resource.

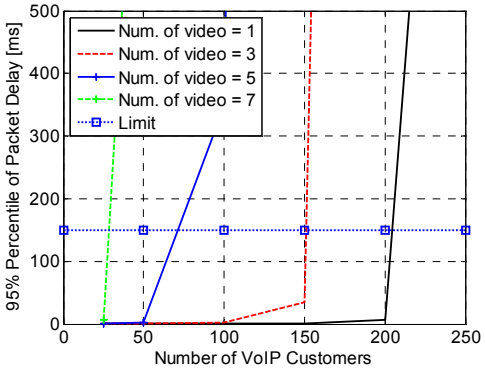

Figure 5: Capacity for VoIP Customers

B. Lead time

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 average customer velocity is shown.

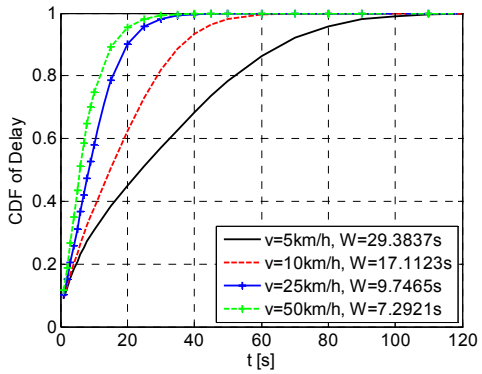

Figure 6: Delay for Video Streaming Packets

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 4). 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 the results of Figure 6 this means that a customer with an average velocity of 50 km/h has to have enough video data in its terminal buffer to overcome delays of up to 35 seconds. In such a case the playback of the video never interrupts under the given scenario conditions. For slower customer velocities this value increases to around 100 seconds. Although such value sounds extremely high for packet delay in case of video streaming the condition can be easily fulfilled by starting the download of data a certain period in advance before the playback starts. If the period exceeds the maximum delay the service continues without interrupts, since there is always a reservoir of video data in the terminal buffer available to keep the video running.

C. *Capacity Limit*

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 4 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 this situation is analyzed for an application with high data rate requirements like video streaming. If the connection to the Internet is frequently broken it is important to store enough data in the terminal buffer in order to bridge the idle gaps. This requires a massive data transfer in periods of good link quality. With Smart Caching this is easily achievable 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.

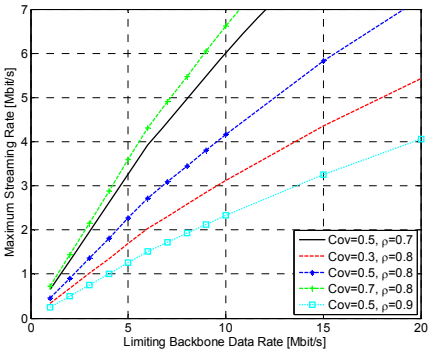

Figure 7: Capacity Estimation without Smart Cache

Figure 7 shows the achievable average data rate under the condition that the backbone is limited to a certain average data rate. The calculation is based on the parameters x, y, and z out of Figure 4. The data which arrives during period x is transferred in period y. The resource which is available for that is the free capacity (1-ρ) of the wireless network. The less free capacity is available or the smaller the coverage ratio (cov) 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 used for the evaluation displayed in Figure 7.

It is depicted that average data rates of around 5 Mbit/s can require backbone rates of up to 10 Mbit/s. But these values even increase if the network is more loaded (raising \rho) or the coverage ratio decreases. Under worse network conditions a backbone data rate of more than 20 Mbit/s is necessary to keep the video playback running. For end-to-end connections in the current Internet this usually is not achievable. The employment of a Smart Cache close to the access point can reduce the data rate required from the backbone to the average value of the video stream. Hence, if the service can be supported for immobile customers than it is also suitable for mobile users if the Smart Cache enhanced wireless network is used. Only a small period of pre-buffering is necessary to equip the terminal buffer with a sufficient reservoir of video data.

VI. Conclusion

Smart Caching has been introduced as a new approach to overcome the problems of heterogeneous wireless networks for broadband services. Their provisioning to mobile users even in areas of discontinuous network coverage is aimed. Smart Caching enabled networks will provide improved performance for non-interactive and non real-time services. New emerging applications like IPTV of Video on Demand can substantially benefit from it [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 playback 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] 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. Goebbels and K. Schmolders , Smart Caching for Supporting Video Streaming in Heterogeneous Wireless Networks, In Proceedings of European Wireless 2007, Paris, France.