SMART CACHING JOINS HIERARCHICAL MOBILE IP

Stephan Goebbels

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

***Abstract* – The provisioning of wireless broadband services will be a key functionality of next generation wireless networks. Hereby the intelligent caching of data at strategic positions within the network will improve its performance as well as consumer satisfaction.**

Furthermore future wireless communication systems will integrate several heterogeneous networks which form a combined overlay network [1]. For the integration of these different systems, which most likely all base on IP backbone connections, Mobile IP will be employed [2].

The combination of smart caching and Mobile IP is therefore an obvious solution. This paper presents the advantages and performance improvements of the new approach.

Keywords – Smart Caching, Hierarchical Mobile IP.

1. Introduction

The expansion of the Internet towards the wireless sector is a leading development for the near future. The availability of mobile end devices like PDAs or laptops comes along with new requirements for radio transfer technologies and data communication protocols. The difference between wireless and cabled communication properties has to be covered by future integrated networks.

A cabled connection usually offers better performance than a wireless link. However the concatenation of several links to an end-to-end connection cause a degradation of the overall performance. Each chain is limited by its weakest link. So it comes to the situation that the wireless link can outperform the rest of the connection although each cabled link, seen for itself, may offer better performance. Due to patchy network coverage and radio transmission shortcomings the availability in time of the wireless link is rather intermittent. To align these antagonistic properties the server-client path should be separated. Smart caching provides this by using a cache at the edge of the network which has on the one side a cabled connection through the backbone to the server and on the other side a section which is mostly dominated by the wireless hop.

The buffering of data at this position allows a continuous and steady downstream of data on the backbone links. Furthermore it is possible to fully utilize the actual bandwidth of the wireless link. In addition to the currently streamed ones the packets stored in the buffer can be transmitted. This combined rate can always be adapted to the load situation of the current base station / access point (AP). It results in a buffering close to the terminal if the connection is interrupted and a massive downstream as soon as a connection is (re-)established.

The integration of different communication systems as well as the cooperation between Mobile IP and smart caching is depicted in Figure 1.

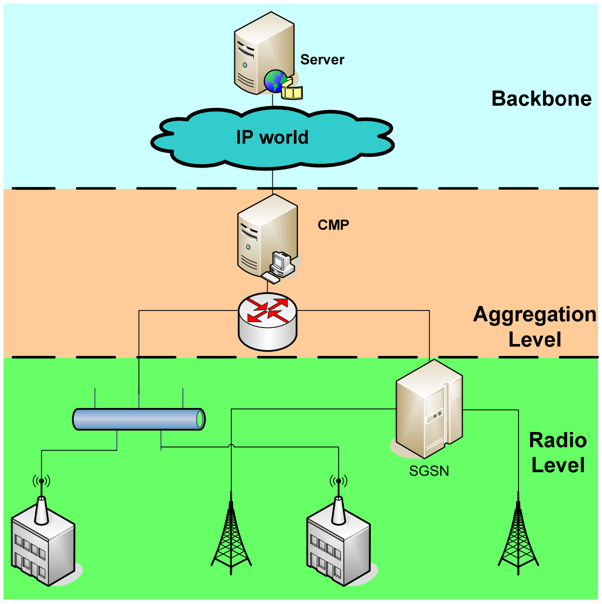

Figure 1 : All-IP System Architecture

The proposed technique is suitable for download of huge data amounts and unidirectional streaming services. An example for the later is Video on Demand [11], a currently upcoming broadband service. The most utilized protocol for streaming services is the Real Time Transport Protocol (RTP) [3]. As it uses UDP/IP packets for data transport the buffering of packets has no direct influence on the transport protocol.

Functionality like timing recovery or packet loss detection is already included in this protocol so that the unreliable UDP traffic can be used. The control of data streaming will of course continue on TCP/IP connections. But due to the limited bandwidth requirements of such control traffic connection could be provided by reliable cellular systems like GPRS.

The re-routing of multimedia streams is committed to the IP layer so that Mobile IP [4] will be used to assure, that the packets reach the mobile node. Due to this strong interrelation between data delivery and Mobile IP (MIP) the integration with smart caching is apparently.

2. Motivation

A strong tendency towards a hierarchical organization of IP networks is observable. First, it begins with the introduction of IPv6. Globally unique addresses are separated into 3 parts: a global routing prefix (GBR), a subnet ID, and the interface identifier. This allows routers in the backbone to forward packets between sites by only using the GBR. This first hierarchy layer is followed by the second step. Each link (e.g. IEEE 802.3) is identified by its subnet ID. And furthermore the remainder of the address is use to locate the appropriate terminal on the current link. These three layered structure allows a lightweight and fast routing mechanism.

Also the MobileIPv6 (MIPv6) [5] protocol may be improved by introducing hierarchy levels. A major drawback of MIP is the triangular routing between Correspondent Node (CN), Home Agent (HA), and Mobile Node (MN). Although MIPv6 allows route optimization between CN and MN the setup of connections and the reaction to link breakdowns is still burden with long latency.

The time consuming update of the HA can be partly overcome if another intermediate node in between is included. Such an intermediate node, called Mobility Anchor Point (MAP), is responsible for several Access Routers (AR), which connect the wireless link with the core network. All possible ARs are grouped to clusters which are controlled by one MAP. If the MN roams from one AR to another one which is controlled by the same MAP it is not necessary to update the HA but the MAP takes over the rerouting of packets. This incorporates a hierarchy level into the MIPv6 protocol which leads to the idea of Hierarchical Mobile IP (HIP) [6] [7]. On top resides the HA which forwards packets to the currently responsible MAP and from there on to the AR which serves the MN.

The inclusion of hierarchies by the principle “divide and conquer” is therefore a widely used approach to solve Internet protocol problems. The separation into different sub areas allows an easier handling of problems on each level. Routing in IPv6 can be accomplished by only considering parts of the overall address. For mobility updates it is enough to proceed to the MAP instead of proceeding to the home link with the HA. Therefore the next logical step is it to extend this approach also to data delivery.

3. Smart Caching joins Mobile IP

If one wants to include a hierarchy level into data delivery to mobile nodes a reasonable attempt would be to use this level to separate wireless and cabled sections. A lot of problems of the migration of the Internet towards the wireless world are sourced by the paradigm that protocols were for a long time developed either for cabled or fixed connections. One example for such a problem is the application of TCP/IP over wireless links. While in the past TCP could see transmission errors almost exclusively as congestion caused, this is no longer valid for WiFi connections. Therefore decreasing of congestion windows and re-initiation of slow start mechanisms are a completely false response to radio transmission errors.

The separation of an end-to-end path into a wireless dominated and a cabled section would overcome these shortcomings. An additional node in the center which serves as a client to the host application and as a server to the application on the Mobile Nodes could fulfill this requirement. Therefore MAP and smart caching node are combined to the Caching Media Point (CMP).

As mentioned before the RTP protocol covers packet losses on the end-to-end connection. But if the CMP is aware of the RTP protocol it is even possible to request discarded packets directly at the CMP so that it is assured that the content of the buffer is a full replica of the sent data. The reaction on retransmission requests of the mobile node can therefore be much faster as not the complete end-to-end path is involved and therefore latency is reduced.

The integration of smart caching and TCP is also possible. However due to the end-to-end oriented protocol design of TCP it is necessary to break down the communication path. Therefore the CMP has to react as a TCP client and server for incoming and outgoing data traffic.

Another aspect of Mobile IP is route optimization between CN and MN. With the introduction of another hierarchy level also a route optimization between CN and MAP is allowed. Here the effect of triangular routing, one of the biggest drawbacks of MIP can be significantly reduced. Therefore the integration of smart caching and HIP is an approach that improves the streaming and delivery of data without burdening the network with remarkable additional load.

4. Communication Protocol

The workflow of a communication session in a HIP and smart caching enabled network environment is displayed in Figure 2 in terms of a Message Sequence Chart (MSC). At the beginning it is assumed that the Mobile Node is already attached to an Access Router and all HIP announcement and binding protocol procedures are already carried out. Therefore the MN has a fully operative connection to the Internet which allows communication with every possible Correspondent Node as long as it does not lose the link to the current AR.

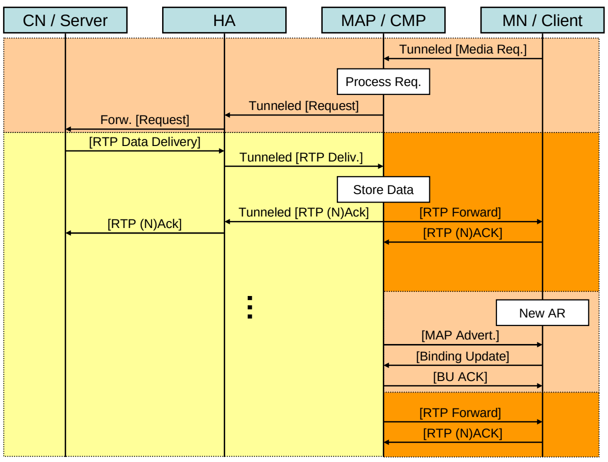

Figure 2 : MSC of Smart Protocol / Caching

A request for any type of media is first sent out to the MAP/CMP entity. An IP-in-IP tunnel is used like in legacy MIP implementations. However the Caching Media Point does not only unpack the packet and forward it to the Home Agent but adapts the source address of the request. As the MP acts as an end-to-end path separator the first section starting at the CN is terminated in the CMP which requires the change of the source address. Afterwards the request is forwarded on another tunnel to the HA which unpacks the packet and delivers it to the CN.

For the transport of the multimedia data RTP is used. It is widely adopted in the Internet and usually employed for real-time streaming services.

The RTP packets are first delivered to the home address of the mobile node. Hence, the HA will receive and intercept this data as it knows that the MN is currently not reachable through his home link. Therefore like in legacy MIP the packet is wrapped and tunneled. However the destination is in HIP the regional Home Agent, the MAP. At this node the routing logic detects that the MAP/CMP entity is responsible for packets with the home address of the MN. But also the MP indicates that the current session has to be cached. Therefore the packet is queued in a buffer. As soon as all earlier packets are transmitted, which may happen immediately as no other packet is queued; the current packet is transmitted towards the Mobile Node. Therefore the reception and the delivery of data at the Media Point are completely independent from each other. Both actions can be carried out in parallel. This fact is indicated by the two arrows, one for the (negative) acknowledgement and one for the data forward, which are at the same horizontal level in the MSC which denotes simultaneousness.

The (N)ACK also points out that no longer the MN is the direct communication partner of the server but the CMP. The CN is always informed by this node if RTP packets are missed out or irreparable corrupted. If the MN detects any problems with a packet this information is reported to the CMP. Either the packet is retransmitted by using a replica in the buffer or the packet is demanded from the CN by a NACK signal.

Afterwards the data delivery continues accordingly. Packets are first sent to the Media Point where they are stored until they can be delivered to the Mobile Node. If the wireless link breaks down the transmission between MP and MN is interrupted. However, this has no influence on the data traffic on the first section of the end-to-end path. The transport of RTP packets continues permanently whether the wireless hop is available or not. Packets which cannot be forwarded by the CMP are stored within the cache.

If the MN comes into range of a new radio access node, which can be a WiFI Access Point as well as an UMTS node or any other (wireless) access technology, the first task of the Mobile Node is to detect a MAP which is responsible for the current Access Router of the access node. By periodic advertisements of the MAP which are forwarded by all connected AR to all access nodes within the controlled region each mobile terminal gets informed about the ruling MAP.

Hereon the MN sends a binding update to the advertised MAP which contains the link address of the new Access Router. In case the request is accepted an acknowledgement is responded. If the old MAP and the new one are identical there is no need to update the Home Agent as the stored information is still up to date. This implies the major advantage that the stored data of the ongoing download session can be directly sent to the MN. All packets which were cached during the connectionless time and the (re-)association period are now available for the transmission over the wireless link. This allows a full utilization of the currently available bandwidth of this link regardless of any bottlenecks or congestions on the first section in the backbone between CN and MAP/CMP.

The separation and the independence of the two sections of the end-to-end path are also reflected in Figure 2. The upper-most rectangular and the field on the left reflect protocol messages due to MIPHIP functionality. While the bottom left area is reserved for the delivery of user data to the CMP the two dark areas on the right represent the forwarding of data. Both processes are very independent of each other and as long as data is available in the cache the two sections of the end-to-end path are operated autonomously.

5. Simulator and Simulation Scenario

The evaluation of the proposed smart caching protocol is carried out on a NS2 [8] based simulator which is enhanced by caching functionality. As in this version only MIPv4 is supported the approach was adapted to this. By collocating HA and CMP the assumptions of the HIP enabled approach are met. The AP are extended with the functionality of Foreign Agent (FA) which is obsolete in MIPv6. But despite of these changes the above described approach is met (Although in MIPv4 the user terminal is denoted by Mobile Host in the following the already introduced term Mobile Node (MN) is further used).

The fact that HIP has an additional hierarchy level with the actual HA on top would have no impact on the simulation as all backbone connections are represented by one link. Therefore it can be omitted in the simulator setup.

For the multimedia service requested by the user a multimedia stream with continuous bite rate (CBR) is assumed. Hence, packets of 500 byte are steadily streamed from the server/CN to the CMP. The streaming rate is set to 400 kbit/s which approximately matches nowadays Internet video streams. Instead of forwarding the packet towards the wireless link the packets are stored in a buffer of the CMP.

The decision when the packets are forwarded and transmitted over the wireless link is made by the CMP. For this three different approaches are implemented. In a first version the packets are automatically forwarded as soon as they arrive. This is identical to simple streaming over MIP without caching enhancements and can therefore be seen as a reference scenario (legacy scenario).

In a caching enabled version the packets are buffered in the CMP and the MN actively requests the stored packets. After receiving a packet the request for a new packet and the acknowledgement of the received one are piggy packed and responded to the CMP. Thus it is assured that packets are only transmitted as long as the end device is reachable.

In a further enhanced setup the control of the forwarding is given to the HA and its Mobile IP functionality. As soon as the new location is registered in the HA the CMP starts the forwarding of packets.

To detect the connection loss of the mobile terminal the HA takes the fact that two subsequent advertisement beacons are not positively responded by the MN. These beacons are send out by the HA to allow the mobile nodes the detection of available HAs and they should be acknowledged by the mobile if the connection allows this. The frequency of the advertisement has effects on the performance of the protocol. Frequent beaconing allows a fast reaction on link losses but at the cost of network resources.

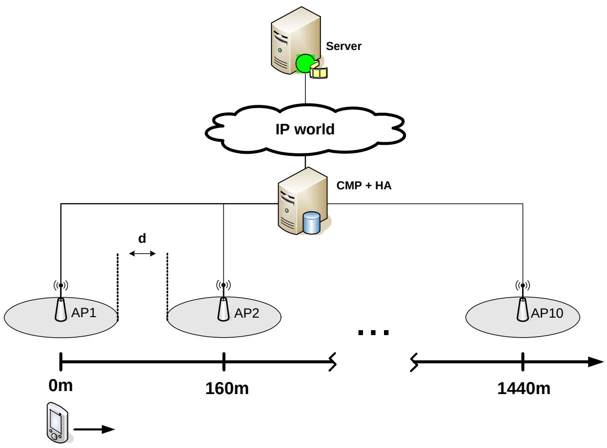

Figure 3 : Simulation Scenario

The simulation scenario consists out a server in the backbone and a mobile client. The CMP is connected to the server via a link which resembles the backbone connection. This link is limited to 1 Mbit/s in order to reflect the fact that the end-to-end connection is less capable than the wireless link. The CMP is connected via separated links to ten 802.11 APs [9] which are arranged behind each other in one row. The links between AP and CMP are set to 10 Mbit/s which almost matches the theoretical MAC bandwidth of 802.11b. The distance between two successive APs is set to 160m which together with a predefined maximum coverage range of the AP of 40 m corresponds to equal distribution of covered and not covered zones (Both are 80m wide). Within the simulation the mobile terminal starts at AP1 at the left-most side of the scenario depicted in Figure 3. From there on it traverse linearly the scenario until it reaches AP 10. The velocity of the mobile is chosen to 1 m/s which is similar to average pedestrian walking speed.

The frequency with which forward packets are requested by the MN respectively sent out by the CMP is called aspired transmission rate and varies in the different simulations.

6. Simulation Results

In the above introduced simulation scenario several different simulations were carried out. Within them the aspired transmission rate between HA and MH was varied. Starting from a transmission rate of 200 kbit/s it was increased up to a value of 4000kbit/s. Compared to the offered traffic from the server of 400 kbit/s this is for the lower rate an unchallenging of the wireless link so that even in the case of a continuous connection never all send data from the server could be forwarded to the mobile node. On the other side an aspired data rate of 4000 kbit/s is not permanently achievable as the provisioning by the server cannot fulfil the requested data rates. However, the discontinuous network coverage comes along with the buffering of data in the CMP which is close to the wireless link. Therefore after a longer connectionless period and the subsequent re-association to a new AP enough stored data should be available to fully use up the wireless bandwidth.

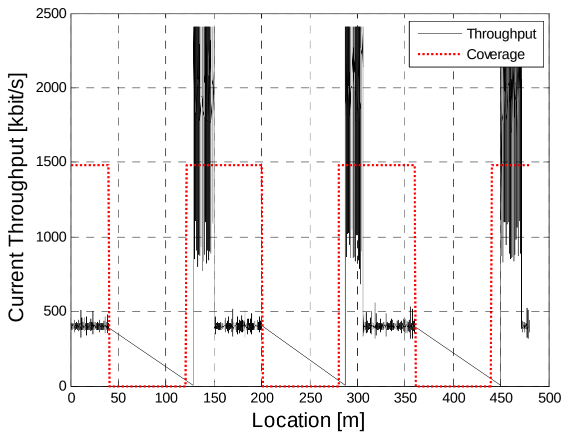

Figure 4 : Throughput of MIP-enabled System

In Figure 4 the current throughput perceived by the end terminal on IP level is depicted. Herewith the MIP enabled smart caching version is investigated. The doted rectangular curve depicts the coverage area of the different access points. However, the reached level on the Y-axis has no further meaning it just shows where a WLAN connection theoretically is possible with respect to the radio conditions and where not.

Since the terminal is close to an AP a wireless link is available at the beginning and the streaming of packets between server and client is possible without any need to buffer the packets in the CMP. Therefore the source rate of 400 kbit/s is matched in the end device.

After reaching the coverage boundary the wireless link breaks down. Obviously the current throughput goes down to zero. Since the evaluation of the current throughput takes place only at the arrival of a packet the curve in Figure 4 depicts a linear decrease to the point in time when the next data packet comes in.

After entering the new coverage zone the terminal can start to recover the ongoing streaming session. This means the execution of the ARP protocol, the establishment of a MAC connection and the redirection of the data flow by updating the HA within the MobileIP protocol. During this period no user data may be received although the network coverage is provided. It takes several seconds before all procedures are carried out and the first user data packet is received. The length of this period is highly dependable on the frequency of the advertisement beacons. The higher the frequency the longer it takes before the data transfer (re-)starts.

Since enough data is already buffered in the cache it is possible to use up the capacity of the wireless link and stream as many packets as possible. The CMP tries to forward the stored data with a data rate of 2000 kbit/s. This level is matched by the received download rate which oscillates around this level. The variance of the experienced data rate is sourced by the granularity of the measurement procedure. The smallest detectable event is the reception of one data packet of 500 Byte.

After exhausting the packet buffer the CMP can only directly forward packets which arrive from the server. Therefore the data rate drops down again to 400 kbit/s. Reaching the edge of the coverage area the procedure just described repeats and continues till the last AP is reached.

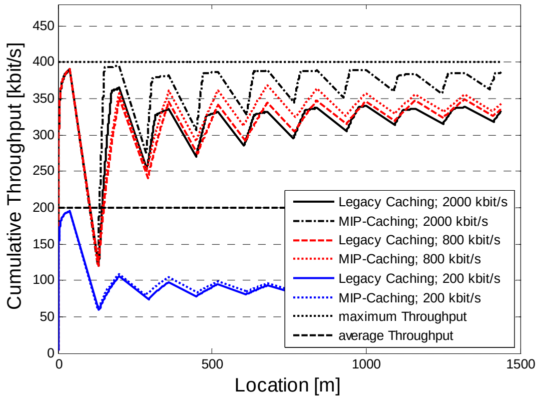

Figure 5 : Cumulative Throughput

To assess the performance of the combination of the smart caching approach with the HIP protocol it is necessary to compare the legacy version with the HIP enabled approach. In Figure 5 both attempts are compared while the aspired download rate is varied by one decade from 200 to 2000 kbit/s.

It is plotted the cumulated throughput, the average throughput of the whoel communication session, against location of the mobile terminal. For reference reasons the maximum available throughput of 400 kbit/s (corresponding to 100% network coverage) and the average throughput for a network coverage of 50% is depicted. The two lower curves correspond to a data forward transmission rate of 200 kbit/s. As this low rate is easily available in the 802.11 network both curves do not differ very much. The slight advantage of the HIP enabled approach is sourced by the fact that the HA is the first entity which is informed about the new IP connection and this node also triggers the retransmission of packets. If the CMP has to wait until it receives a request by the MH is takes in average one half of a requesting period after IP connectivity is supplied before the transmission starts.

For a value of 800 kbit/s, which corresponds to an average level of 400 kbit/s if 50% network coverage, as applied in the simulations, is assumed, the advantage of the smart caching approach allows an explicit improvement of the cumulated throughput. Again differences between the MobileIP-enabled and the legacy approach are visible.

To come as close as possible to the theoretic limit of 400 kbit/s it is necessary to completely empty the buffers in the CMP during connectivity periods. As re-association and protocol overhead shifts the proportion of radio and non-radio coverage to the disadvantage of the user it is necessary to stream more than the double of the average streaming rate when connected to an AP. The last two curves represent a rate of 2000 kbit/s. It can be recognized that even here the upper limit is not reached although the buffers get exhausted. This is reflected by the break in the gradient of the curves. This is due to the fact the packets on the MAC layer are discarded if interface queues are exhausted.

But even more important is the behaviour of the legacy smart caching graph. It is distinctly separated from the MIP-enabled curve. This manner is more illustrated in the next Figure.

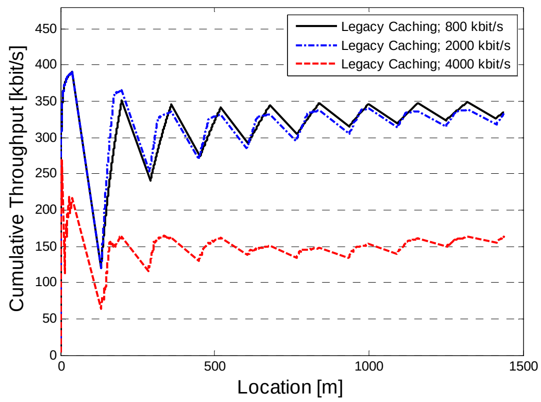

Figure 6 : Performance of Legacy Approach

Again the two graphs for 800 and 2000 kbit/s are displayed. It can be seen that the curves cross each other several times and at the end the 800 kbit/s curve even outperforms the other one. Moreover for an aspired download rate of 4 Mbit/s the performance completely collapses. This occurs due to 802.11 MAC protocol and its limitation in bandwidth. The active polling (even if piggy packed with acknowledgements) of packets requires small packets which dramatically decrease the overall available throughput of an 802.11 cell. Such a high polling rate consumes most of the bandwidth so that nothing remains for the actual transport of data.

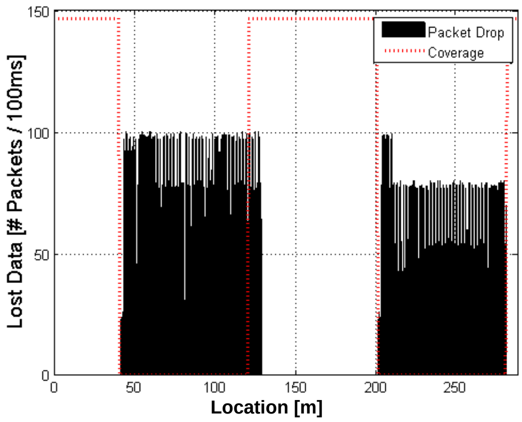

Figure 7 : Packet Drops in Reference Scenario

Smart caching relies on the principle to transfer the buffered data in coverage areas “as fast as possible”. Figure 6 has shown that with the legacy approach this working point is hard to find as a good proportion between polling and data transport has to be found. For HIP enabled systems an acknowledgement on IP level is not necessary as the Mobile IP protocol surveys the connectivity state of the terminal. Packet errors can be handled either on the link layer level or are recognized by the RTP protocol itself. Both protocols add significantly less overhead to the communication session.

Finally Figure 7 illustrates the behavior of a usual streaming service in an environment of patchy network coverage. It depicts the amount of dropped packets depending on the current location of the terminal. The same simulation as before carried out but with switched off caching functionality. The streaming rate is set to 2000 kbit/s. It can be seen (coverage zones are marked again by the doted line) that outside the AP cells all packets are dropped by APs.

Again it can be noticed that after entering the cell of AP2 still some packets are dropped until the IP connection is re-established. Contrary to that one AP later the re-connection works faster and almost no packets get lost. This behavior as well as variations in the dropping rate depend on the other involved communication protocols like MAC, MIP, or ARP.

The dropping rate of the smart caching simulations is missed out in the graph as it is negligible small.

In the simulation the case is missed out where old and new AP are not belonging to the same MAP/CMP. This would imply that the cached data can no longer be used. However this case can be seen as a simple restart of the transmission beginning just after the last received data chunk.

7. Summary and Conclusion

Optimal support of smart caching requires that in connection sessions the “as much as possible” principle is applied. Therefore the transmission has to start as early as possible and the rate should be adapted to the available bandwidth. The first fact can be provided by the MIP enabled approach so that other attempts are easily outperformed. Obviously the MAP, respectively the HA in the simulation scenario, is the first point which is informed about the new connection. Furthermore for heterogeneous systems this fact will gain more importance if overlay networks get involved and different network coverages are overlapping and end devices switch their radio access network. The connection point will then be the Mobile IP aggregation node which is either the HA or the MAP.

It was also shown that smart caching significantly reduce dropping rates. As each dropping of packets implies a waste of radio bandwidth is has a major impact on the overall network performance.

For the future it remains to include other wireless networks in order to reach the described heterogeneous network. Also an adaptation of the WLAN protocol towards newer standards 802.11g or WiMAX would make the scenario more realistic. However more important for the simulation results is the relation between wireless and wired bandwidth since the absolute raise of bandwidth only shifts the quantity of the results but has less impact on the quality.

Summarizing HIP enabled smart caching is an intelligent solution [10] for the provisioning of multimedia steaming and download services in next generation heterogeneous networks.

REFERENCES

[1] C.W. Lee, L.M. Chen, M.C. Chen and Y.S. Sun. "A Framework of Handoffs in Wireless Overlay Networks Based on Mobile IPv6". IEEE Journal on Selected Areas in Communications, Vol. 23, No. 11, Nov. 2005.

[2] J. Li and H.-H. Chen. "Mobility Support for IP-Based Networks". IEEE Communications Magazine, Vol. 43, No. 10, Oct. 2005.

[3] C. Perkins, "RTP Audio and Video for the Internet". Addison-Wesley, 2003.

[4] J. Solomon, "Mobile IP: The Internet Unplugged". Prentice Hall, Englewood Cliffs, 1998.

[5] The Internet Engineering Task Force. "RFC 3775 Mobility Support in IPv6". www.ietf.org .

[6] H. Soliman. "Mobile IPv6". Addison-Wesley, 2004.

[7] S. Aust, N.A. Fikouras, M. Sessinghaus and C. Görg. "Hierarchical Mobile IP NS-2 Extensions for Mobile Ad Hoc Networks". Proceedings of Wireless Networks and Emerging Technologies, 2004.

[8] NS-2, Network Simulator 2, Version 2.2.7, www.isi.edu/nsnam/ns/.

[9] IEEE 802.11 standards. Available at http://standards.ieee.org/.

[10] S. Goebbels and R. Probokoesoemo. "Intelligent Caching Strategy for Mobile Communication Networks". Proceedings of European Wireless 2005, Nicosia, Cyprus.

[11] T-Online Video on Demand. Compare http://www.t-online-vision.de/.