Smart Caching in Service Specific Overlay Networks for Wireless Networks

Stephan Goebbels[1], Ralf Jennen[1], Stephen Herborn[2], Bertrand Mathieu[3]

***Abstract*—The provisioning of multimedia services in heterogeneous environments will be a key issue in the evolution of upcoming communication networks. The construction of customised overlay networks has been established as a means of handling dynamic network conditions. Strategies for intelligent caching of user data, together with the overlay networking methodology, will provide support for the delivery of multimedia content in mobile and dynamic use scenarios.**

This promising and interesting combination of this so called smart caching and service specific overlay networks has been extensively investigated in the context of the ongoing IST Ambient Networks project. This paper presents the current results of the research.

***Index Terms*—Smart caching, Service Specific Overlay Networks, Ambient Networks.**

I. Introduction

Today's broadband networks are quite unbalanced in terms of up- and downlink utilization, and this is expected to remain the case in the future. While the channel from the terminal towards the core network is more likely to carry short data packets containing requests and acknowledgements the main portion of the data traffic is carried on the downstream channel. Therefore optimization of the downstream channel is of special interest for the evolution of new wireless broadband systems.

In the future, user devices will be able to choose and switch automatically between various wired or wireless access technologies to be "always best connected". Although the provided data rates of the core network are relatively stable the last-hop link is likely to offer somewhat less stable performance guarantees, especially if it is wireless. Compared to the section of the end-to-end connection that lies within the core network, narrowband links like GPRS are rather overextended whilst WLAN access is usually underutilised.

In order to mitigate problems caused by unstable and/or narrow last-hop channels, multimedia applications typically buffer data internally before rendering it to the user. For multimedia applications which usually require a high-speed downlink access the use of buffering techniques to overcome variances in end-to-end data rate may greatly enhance the perceived quality of service. This paper further extends the concept of caching and buffering by adopting it to the special needs of wireless networks and proposing the placement of buffers at the edge of the core network before the last-hop link. This special arrangement of caches and its specific application to wireless networks is called smart caching.

In [1] it was shown that the caching of data at the edge of the core network in close proximity to the user partly overcomes problems caused by wireless last-hop link instability. In periods during which the wireless link is narrow and/or overloaded the data may be buffered; and during subsequent periods of broadband access the buffered data may be transmitted in addition to the normal data traffic, at the expense of some time shift in rendering to the user.

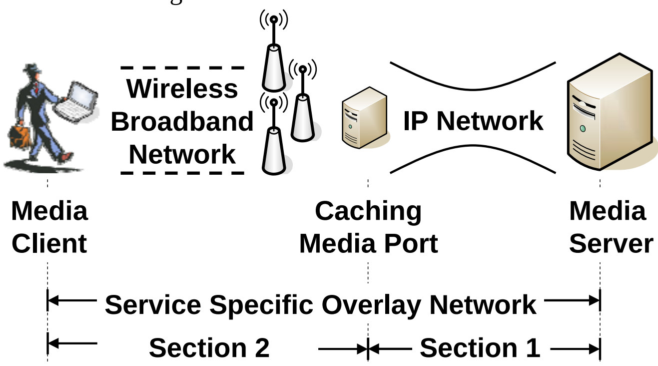

Figure 1: Separation of end-to-end connection

The idea to subdivide an end-to-end connection into several sections is an established overlay networking approach [2]. The European research project Ambient Networks [3] has developed the application of Service Specific Overlay Networks (SSON) to support new types of transport and multimedia services like smart caching [4][5]. The setup of SSONs is strongly bound to characteristics of the currently ongoing communication session (e.g. service profile, access network…). Depending on this overall context, a SSON is constructed, with some specific intermediate nodes (MediaPort, or MP) continuously adapted to the needs and requirements of the end-to-end connection between the client (MediaClient, or MC) and the server (MediaServer, or MS). Smart caching is an example of an intermediary service that may make use of this service specific overlay networking.

[1] Chair of Communication Networks, Faculty 6, RWTH Aachen University, Kopernikusstrasse 16, D-52074 Aachen, Germany; Email: {sgs|jen}@comnets.rwth-aachen.de.

[2] National ICT Australia, Bay 15, ATP, Eveleigh, NSW 1430, Australia; Email: stephen.herborn@nicta.com.au

[3] France Telecom, R&D Division, CORE/M2I, 2 Av. Pierre Marzin, 22300 Lannion, France; Email: bertrand2.mathieu@francetelecom.com

The remainder of the paper is structured as follows. In section 2 the motivation for the use of SSONs in conjunction with smart caching is given in more detail. Section 3 describes the required system architecture for the proposed system enhancements while the following section 4 explains the workflow of a combined SSON and smart caching architecture during a communication session. Section 5 provides a discussion of possible improvements, and the paper is concluded in section 6 together with a future outlook.

II. MOTIVATION

Future consumption of multimedia content will be dominated by wireless mobile clients who access information that resides on high speed links connected to the core network. This fusion of cabled and wireless links reveals new problems of IP based data exchange.

The higher variance in the throughput and reachability of clients connected via wireless links requires fast reactions from the backbone network. However, due to the potentially high number of hops between server and client, a fast reaction to link losses and session re-establishment is hard to achieve. In addition, the maximum throughput of wireless links might never be reached by the overall end-to-end connection. Although each wired link for itself is much more capable than a wireless connection the concatenation of several links and the occurrence of a high number of parallel communication sessions in the backbone leads to a much smaller end-to-end throughput per session and user. Therefore capacity might be wasted due to some bottleneck link on the backbone path for an ongoing session.

The caching of data close to the wireless hop at the edge of the core network would partly overcome these problems. The communication between a data source and the consumer is separated by a caching entity so that the first segment is exclusively confined to wired networks and therefore manageable by legacy technology. All issues regarding the wireless data transfers are then restricted to a very limited region and much easier to handle.

Figure 1 shows the separation between server and client. The overlay consists of two subsections. Section 1 lies between the MediaServer (the provider of the requested service) and the caching node, so called Caching MediaPort (CMP). The next connection (Section 2) is established between the CMP and the MediaClient (the requester and consumer of the service).

Although this idea is also included in [6] the applicability in this approach is closely tied to the underlying network infrastructure. The dependency on system functionality and architecture entities limits the concept to very specific mobile communication systems. Within this paper the concept is widened and by the integration of SSONs it is assured that smart caching is not bounded to a specific underlying network structure.

Content Distribution Networks (CDN) also tries to deal with the problem of allowing fast access to requested information. But they are usually limited by the fact that only most popular content is worth to get included in the CDN. Otherwise the overhead would easily overwhelm CDN and backbone network. For personalized data they are not applicable at all. Contrary to that the user may benefit in each communication session from the support and advantages of smart caching.

This paper investigates caching nodes strategically located in areas of predicted large-scale wireless communication like densely populated urban areas or highly frequented motorways. One characteristic of such wireless broadband systems is that they are highly dependent on line of sight wave propagation (especially in higher frequency bands). That is why there are hot spots of good coverage alternating with areas of limited coverage, if at all. MediaServer and Client together with at least one CMP in between form a SSON. While the endpoints of the overlay remain the same for the whole communication the CMPs might be switched.

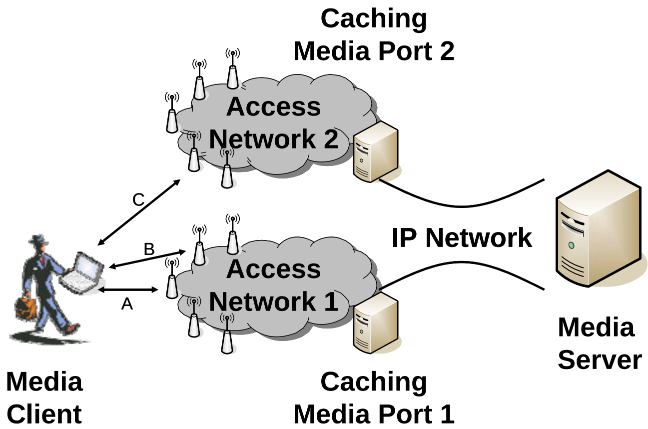

Figure 2: Deployment of Caching MediaPorts

By grouping any kind of wireless access network into clusters, a user moving in an area covered by such a cluster would be continuously served by a single CMP. This implies that after switching over to another access node of the same cluster, the same CMP can be employed and accessed. Figure 2 shows the case of MediaClient mobility in an area of two access networks that belong to different clusters. As long as the MediaClient is connected via the Access Network 1 (case A or B in the figure) the CMP 1 can be used as a part of the SSON. When the media client performs a handover to the Access Network 2 (case C in the figure) for which the CMP 1 is not responsible, the SSON must be adapted. CMP 1 will be substituted by the CMP 2. As long as a terminal is served by access nodes supplied by the same CMP the service can continue without any problem. But if the user changes in Figure 2 his access network the cached data in CMP1 would be lost. To overcome such problems all access networks of a specific area should be clustered and dominated by one CMP so that even an intersystem handover would not interrupt the end-to-end service provisioning. Also the SSON does not need to be adapted in the later case as the re-routing of packets take place on the lower layers and has impact on the overlay routing.

The idea of smart caching has been previously discussed in relation to the MediaPort concept [7]. This approach is refined in this paper to benefit from the promising features of SSON, defined in the Ambient Networks project.

The concept of overlay networks tailored to the needs of specific services allows both service provider as well as end users to take advantage of 3rd party network-side services which improve or even allow new types of multimedia delivery [8].

The SSON concept is designed in order to include MediaPorts arbitrary located within the network into the end-to-end path. Those MPs perform media processing tasks to enable the correct media delivery of the service, when needed: e.g. media transcoding, adaptation, synchronization, or caching. Different overlay networks may be deployed for every media delivery service. The MPs may provide any kind of service ranging from media transcoding and adaptation to caching functionality. Furthermore the data transport can benefit from more sophisticated overlay routing algorithms which outperform legacy IP packet forwarding.

The combination of both ideas makes it possible to take advantage of overlay networks, like dynamic adaptation of the end-to-end path, routing on overlay level and easier handling and maintenance of multimedia sessions.

III. Architecture and Node Structure

The applicability of smart caching is heavily dependent on the used overlay architecture and the employment of the SSONs. For the purpose of smart caching the overlay basically consists out of three main entities, the MediaServer, the MediaClient and the Caching MediaPort. All of these entities are Overlay Nodes (ONodes) by nature of the fact that they are enabled to participate in the overlay network itself. To join a service specific overlay the network stack of each participating ONode needs to be supplemented by a common Overlay Support Layer, or OSL. The OSL provides overlay level routing and addressing functionality on top of the legacy protocol stack.

While the MS and the MC are in each case the endpoints of the overlay which can contain an arbitrary number of MediaPorts in between, for a basic caching use case, for instance, one Caching MediaPort is sufficient.

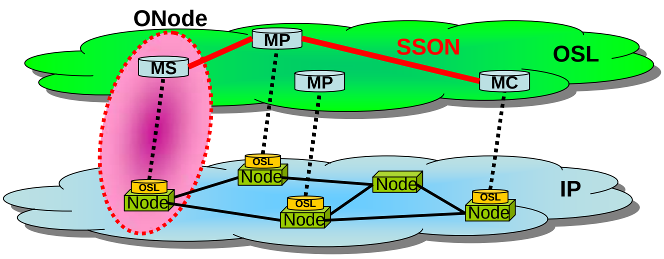

Figure 3: Service Specific Overlay Network

The SSONs in the given architecture are located on top of the IP world in the Overlay Support Layer. This layer, depicted as the upper cloud in Figure 3, is where SSONs take shape and packets are forwarded from one ONode to the next. The operating of routing via the OSL is similar to that of IP routing where packets are forwarded on the basis of routing labels. Each OSL entity contains an overlay routing table which provides instructions about the link that should be used to forward overlay packets prefixed by a specific overlay address label. In order to allow the IP layer to route the packet to the correct destination ONode the label is translated by the OSL into an IP address. Therefore a kind of tunnelling is used to bridge the hops between the OSL nodes. Thus design of decision logic used to govern the construction of OSL routing tables in ONodes is a crucial issue.

This administration and maintenance of SSONs is provided by a control entity referred to as the Overlay Control Space (OCS). The OCS contains logic for the setup and the restructuring of overlays during their life cycle, the Media Routing Logic (MRL). The MRL contains the Media Caching Logic (MCL), which deals with all issues regarding the caching of multimedia data.

The MRL determines which MPs should be included into an SSON and the order in which order they are traversed. This decision is taken by gathering context information from the underlying network as well as capability and location information of potential ONodes. The choice of suitable ONodes for a requested parameter set is supported by the MediaPort Information (MPI). It provides a list of feasible ONodes which is then used by the MRL and MCL to implement the SSON. The MPI may either be an active ONode discovery or simply make use of precompiled information stored in a database. For the purpose of smart caching it is of special interest to choose a CMP which is closely located to the end customer respectively the MC. Due to this requirement an active search for CMPs is more suitable. Since the architecture is focused on mobile environments the storage of potential MPs in a database is fast outdated. Furthermore the distance between MC and CMP can only be determined by direct measurements and usually cannot be compiled out of database information. This information is processed together with the requirements for the requested service and used to find the best suitable SSON setup to support the multimedia service.

As the caching of multimedia data requires not only information about the external properties of a multimedia flow but also information about the flow contents itself, in order to decide whether the multimedia data worthwhile to cache (i.e. whether the data is current) or if a replica of the contents is already available elsewhere, the MCL is an extension of the MRL intended to deal with such issues. Each multimedia request has to be intercepted by the MCL in order to determine whether the data can provided directly by the cache or if the incoming response data should be simultaneously forwarded to the end user as well as stored in the cache. The MCL makes use of the MRL functionality to setup and maintain the SSONs, as such it can be seen as a module on top of the MRL which just provides additional information for the routing and SSON setup decision.

The aforementioned entities all reside within the OCS.

However for the whole OCS it is not specified whether it is organized centralized or distributed. Although each ONode needs a small interface towards the overlying OCS the actual decision handling and database maintenance could be concentrated on more capable servers within the backbone.

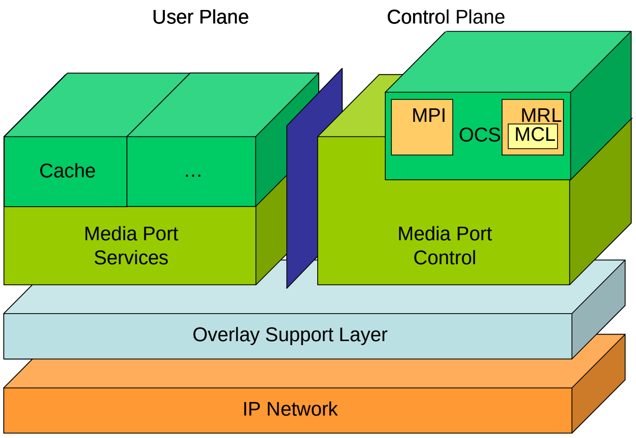

Figure 4: Protocol Stack of ONodes

The configuration of an ONode and the separation into user and control plane is depicted in Figure 4. Both are operating on top of the OSL, and each node has to be integrated in an IP network in order to take part in an overlay. The OSL layer, as a transport sub layer, allows the communication between the different ONodes within the network. If routable OSL identity labels are globally unique then the OSL could also conceivably enable communication between ONodes connected with different versions of IP.

The maintenance and administration of each ONode is integrated in the MP control. It takes part in the distributed OCS and can therefore provide information for e.g. routing decisions and can also be remotely controlled by centralized entities like the MRL. Therefore the OCS and its sub entities, the MPI, MRL and MCL, is shown in the figure as a self-contained component which extends into the MP control but also lies outside it.

In the user plane additional features are depicted as part of the MP Service entity. On the user plane side datagrams are forwarded by the OSL and potentially received and processed by the service entities which may include caching or media adaptation such as transcoding.

IV. Workflow of Smart Caching Session

The provisioning of smart caching in an SSON enabled network requires several actions in order to setup, maintain, and tear down the appropriate overlay as well as the caching entity. In Figure 5 the complete operating sequence of a caching SSON is illustrated. The workflow is based on an object oriented analysis which has been performed according to the Object Engineering Process of Oose.de GmbH [9] and the Rational Unified Process [10].

Each session is initiated by an external service request which may be sourced by customers or service providers, too. If one of the communication partners decides that the session could be improved or enriched by smart caching the appropriate actions has to be taken. A detailed request is given to the OCS. This includes a description of the requested contents as well as a set of QoS parameters the connection should fulfil. The OCS processes this request, decides if the requested content is already stored in the cache and if not it is determined whether the end-to-end connection is sufficient or needs support by smart caching.

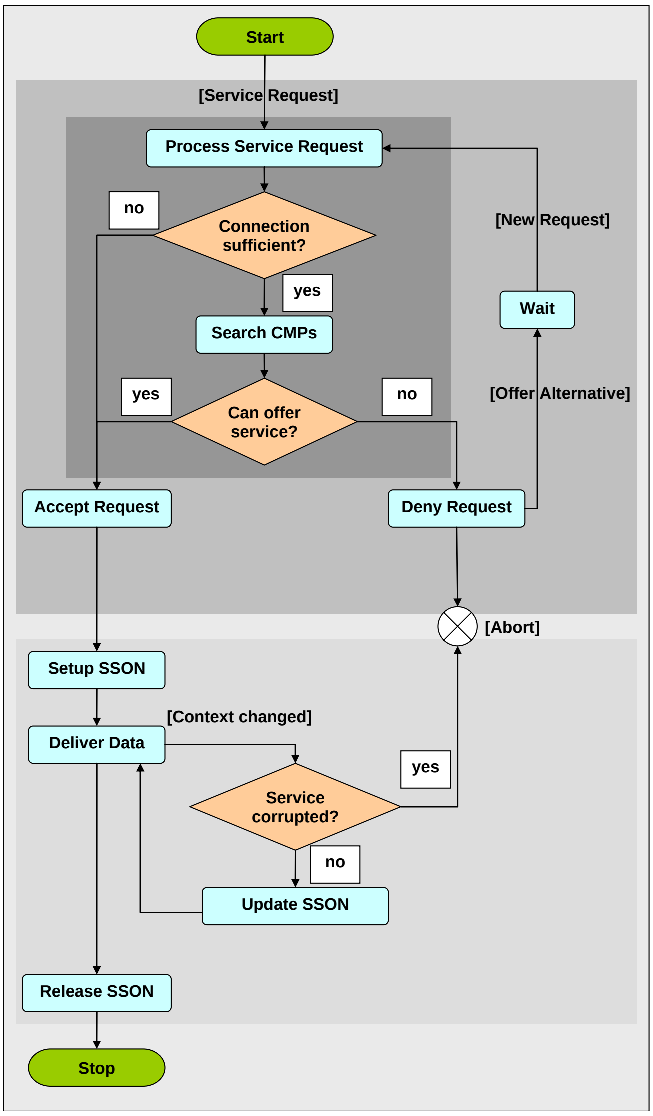

Figure 5: Workflow of smart caching

If smart caching is needed then a search for appropriate CMPs is performed. MPI is used to search for candidate CMPs that are close to the endpoint of the connection. After having found a set of nodes the OCS, and more precisely the MRL in combination with the MCL, decides whether it is possible to establish a SSON with the given CMP candidates and wether these candidates are able to provide the desired level of service quality. In the case that the desired level of QoS is not able to be guaranteed then an alternative offer can be negotiated with the entity that initiated the SSON creation request. By offering the maximum level of QoS that can be guaranteed the SSON initiator may decide to accept this opportunity or to deny the proposal.

After the request is accepted the SSON has to be established. The OSL level connections between MS, CMP, and MC have to be set up through configuration of OSL routing tables, and each MP on the end-to-end path needs to know which action they should perform on the incoming media data flow. After this has been completed the actual delivery of data takes place. Data packets are sent from the MS to the CMP and a replica is stored in the cache. The decision when packets are forwarded or retransmitted towards the MC is taken by the CMP.

Each time any of the relevant context parameters changes the OCS has to find out if the services provided by any ongoing SSON are influenced by this change. Context changes may refer to such events as the mobile client losing the radio connection to its current access point, or the client terminal switching to a different wireless network. Therefore in response to context change events the OCS needs to determine if the new access node may also be supported by the current Caching MediaPort. Other context events such as altered link or traffic conditions or even updated service requirements have to be taken into account. The OCS checks if the service can still be performed. If not, the service is terminated and the SSON is destroyed. Otherwise it is continued as long as the lower QoS threshold is not crossed.

After all data is delivered and the service is finished the SSON's life cycle concludes and the resources allocated to it are released.

V. Further Improvements

The proposed architecture is not limited to the support of smart caching. The concept of overlays is flexible enough to handle the short comings of today's IP infrastructure. One of the main aspects is the improvement of QoS provisioning in end-to-end connections. After approaches like IntServ and DiffServ have not yet proven their effectiveness, overlays are seen as a promising alternative. Although QoS is an aspect within SSONs it also includes other features. By making the overlay aware of the requested service it is possible to adapt the network connections to the transported content. Furthermore network sided processing and modification of the user data is possible which will allow completely new application and business cases. The network is extended from a pure packet transport medium to an active service provider. Services can include security enforcement by third party certification of content or codec adaptation of multimedia streams.

For the future work, there is still a wide area of application opportunities for SSON enabled networks. For instance the deployment of smart caches within wireless networks will help to handle overload situations by providing the same contents to several users [11].

The implementation of a joined smart caching and SSON concept is currently ongoing. The investigation of the performance and the validation of the concept is remaining for future work.

VI. Summary and Conclusion

Smart caching is a new concept to substantially enhance wireless network performance and user convenience. The intelligent provisioning of user data at strategic points in the network is the key innovation of this concept. It will guarantee decreased download and streaming times for multimedia content. It will help to overcome the effects of patchy network coverage and of bandwidth variation within heterogeneous wireless communication networks.

It is supported by the employment of Service Specific Overlay Networks. SSONs can be seen as an extension to basic overlay networks, such as those presented by Anderson et al. in [12].

The paper has demonstrated the applicability and importance of smart caching within future mobile communication systems. SSONs were proven a feasible extension of the smart caching approach. The integration of both technologies will come along with several synergy effects which ease the migration and deployment of the new architecture.

The wide range of SSON applications will guarantee the deployment and the spread of the new technology. Smart caching will also benefit from this distribution as it will automatically follow SSONs in each new application area.

Acknowledgment

The authors would like to thank their colleagues of Work Package 5 in Ambient Networks project (phase 1), who have shaped their work with active participation in discussions and cooperation regarding the topics discussed in this paper.

This document is a by-product of the Ambient Networks Project, partially funded by the European Commission under its Sixth Framework Program. It is provided "as is" and without any express or implied warranties, including, without limitation, the implied warranties of fitness for a particular purpose. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Ambient Networks Project or the European Commission.

References

[1] S. Goebbels and R. Probokoesoemo, *"Intelligent Caching Strategy for Mobile Communication Networks"*, In Proceedings of 11th European Wireless Conference 2005, Vol. 1, p.p. 294-300, Nicosia, Cyprus.

[2] Zhu Li, Prasant Mohapatra, *"QRON: QoS-Aware routing in Overlay Networks"*, IEEE JSAC Special issue on Advances on Service Overlay Networks, 2003.

[3] Ambient Networks Project, Part of EU's 6th Framework Program, www.ambient-networks.org.

[4] Ambient Networks, "*D5-3 SMART – Final Architectural Design* ", AN WP5, December 2005, www.ambient-networks.org.

[5] Ambient Networks, "*D5-4 Proof of Concept and Demo*", AN WP5, December 2005, www.ambient-networks.org.

[6] I. Herwono, J. Sachs, and R. Keller, "*Integration of Media Point System with the 3GPP IMS*", In Proceedings of the 11th European Wireless Conference 2005, Vol. 1, p.p. 301-307, Nicosia, Cyprus.

[7] I. Herwono, S. Goebbels, J. Sachs and R. Keller, *"Evaluation of Mobility-Aware Personalized Services in Wireless Broadband Hotspots"*, In Proceedings of Communication Networks and Distributed Systems Modeling and Simulation 2004, San Diego, CA, USA.

[8] J. Rose et al., *"Media Aware Overlay Routing in Ambient Networks"*, In Proceedings of PIMRC 2005, Berlin, Germany.

[9] oose.de GmbH. Available at www.oose.de/oep.

[10] Rational Unified Process, Best Practices for Software Development Teams, A Rational Software Corporation White Paper, 1998

[11] L. Yin and G. Cao, *"Supporting Cooperative Caching in Ad Hoc Networks"*, In Proc. of IEEE INFOCOM 2004.

[12] D. Anderson et al., *"Resilient Overlay Networks"*, In Proceedings of ACM SOSP, Banff, Canada, October 2001