System Architecture of a Mobile Message Transport System

Bernd Gloss
University of Stuttgart, Institute of Communication Networks and Computer Engineering, Pfaffenwaldring 47, D-70569 Stuttgart, Germany
email: gloss@ikr.uni-stuttgart.de

Abstract—Many scenarios of beyond 3G mobile communications describe the integration of various access technologies into one system. Being always best connected under certain optimization criteria will be a crucial point for network operators and mobile users and requires network changes of mobile devices at runtime, the so-called vertical handovers. For those, bandwidth fluctuations up to the order of one or two magnitudes, e.g., when changing from an IEEE 802.11 to a GPRS System, have to be expected and applications have to cope with them in an user friendly way. Multi-modality and flexible data representations exploiting weights and semantic of transmitted data as means for making applications resource adaptive are currently under investigation. On top of that, device and system-wide adaptation control instances are needed to solve cross-layer and inter-application issues. This requires a rethinking of the classical communication paradigm of OSI-like protocol layering. With this paper, an overview on adaptation in communications is presented and an experimental framework providing system support for application adaptation and adaptation control is introduced as part of that discussion.

I. INTRODUCTION

It is obvious that mobile user equipment and wireless technologies currently evolve with blazing speed. Mobile devices become a composition of a communication device and a personal digital assistant (PDA), and their performance increases comparable to the evolution of PCs and notebooks during the last decade. Along with this, new wireless technologies emerge or refine. They offer packet-oriented data service, a higher system capacity, and a higher per-user bandwidth. A further trend is the convergence of wireless technologies. Wireless LAN technologies merge into the world of cellular systems providing hot-spot services, and, driven by Internet mobility solutions like Mobile IP and SIP, wireless LAN islands and cellular systems can be combined to one communication environment. It can also be identified as a trend that mobile devices will incorporate more than one wireless network technology.

The aforementioned aspects of mobile devices and wireless networks are the basis for new applications and services. Some well known examples are the multimedia messaging service (MMS), mobile tourist guides, and navigation support. Even today there are many more applications and it is likely that the majority of the future ones are not yet foreseeable. There is a chance that from a certain point in device and communication system evolution the applications will become the driving force comparable to today’s PC software, and that network capacity will be the limiting factor. Given that this vision becomes true, highly loaded wireless networks will be seen, and the need for advanced flow control mechanisms and means for application and service adaptation can be derived. Both for optimizing the wireless network infrastructure usage and maximizing user’s fun.

With this paper, a middleware-based framework providing (1) mobility support, (2) means for communication adaptation in the dimensions transmission time and transmission volume, and (3) mechanisms to provide cross-layer and intra-service adaptation control are introduced. This framework serves as base platform for research in adaptation control mechanisms and system support for adaptive applications. The whole architecture is motivated and evaluated against the requirements defined by the Nexus Centre of Excellence [1] at the University of Stuttgart. The main objective of Nexus is the definition and realization of world models as basis for location- and context-aware applications and services. In this scope, issues concerning information management, methods for model representation and sensor data integration as well as communications are covered. Exemplary applications employing this world model are used to derive requirements to the presented communication subsystem and to evaluate it.

The reminder of this paper is structured as follows: in section II, background information on mobility support and adaptation is presented and requirements of a mobility and adaptation supporting system architecture are derived. The proposed system architecture is presented and discussed in section III and in section IV, the current state of the art and related work are outlined. In section V, conclusions are drawn and an outlook to further research efforts is given.

II. BASICS AND REQUIREMENTS OF THE ARCHITECTURE

The most important properties of the base scenario are user mobility and the ability of mobile devices to communicate via several wireless technologies. A major effect of such a scenario are resource fluctuations due to network availabilities. To illustrate those fluctuations seen by a mobile device, a Gantt diagram with potentially available wireless access systems over time is derived from an access network scenario and

1 The Nexus Centre of Excellence at the University of Stuttgart is funded by the Deutsche Forschungsgemeinschaft (German Research Foundation) as project SFB 627
a user’s path and speed in Fig. 1. This exemplary Gantt diagram represents the cause for resource fluctuations and spans one part of the decision space, i.e. the assignment of communication processes to access systems. The other part of decision space is the adaptation of offered communication processes to network availabilities.

The aim of the presented system architecture is to provide means for mobility support and communication adaptation as a system function. In the following subsections, the basics and requirements of communication support in beyond 3G scenarios are presented in four groups: mobility support, adaptation means, adaptation control, and resource sharing.

A. Mobility Support

Mobility support is a widely noticed field and thus it is touched only lightly within this paper. For classifying mobility support, there are two basic approaches: (1) mobility support by changing node addresses and (2) mobility support by changing routing procedures. The principle of mobility support by address changes can be seen in the architectures of Mobile IP [2], and mobility support by changes of the routing procedures are demonstrated by some micro mobility approaches like Cellular IP and HAWAII [3]. Mobility supporting architectures can be realized either by using one of them exclusively or by assembling one or both of them in hierarchies to more complex systems for achieving distributed mobility handling.

Depending on the required type of mobility support, e.g., device mobility, intra- or inter-technology mobility, personal mobility, etc., mobility support can be handled at different layers in the communication protocol stack. When looking at the environment and usage scenarios, it is obvious that at least device mobility support for inter- and intra-technology cases is required. In the general case of mobile devices with more than one network interface and the ability to have several of them active at the same time, it is necessary to have mobility support at least at transport layer, i.e., for connections and streams. In this case, certain connections can change the network while others can remain as it is desirable when data connections change to IEEE 802.x hot spots, while voice or video connections requiring special QoS-Mechanisms are kept on a UMTS system.

For providing device or connection mobility support, a system architecture needs the following mechanisms and components:

- network detection
- network selection
- routing support
- handover mechanisms
- paging (optional)

B. Adaptation Support

Systems that can adapt their traffic volume to available resources need operations on the transmitted data and building blocks to accomplish those tasks. As background for the presented system architecture, the aspects media adaptation, transmission adaptation, and adaptation instances are introduced and classified as follows.

1) Media Adaptation: Media adaptation means the adaptation of the content of an information flow to a transmission context. Depending on the type of media, several approaches can be taken. Some examples are

- the change of picture size or frame rate of video streams
- the change of colors and compression of bitmaps
- the variation of the level of detail of city maps

More approaches to adapt the content to a transmission context are discussed in the related work part in section IV. In Fig. 2, a classification scheme for content specific adaptation is depicted.
Content adaptation mechanisms with a knowledge of the media structure are highly application specific and thus are out of scope of the presented framework but might and should be supported by hooks and adaptation control functions.

2) Transmission Adaptation: Depending on the transmission delay requirements, the transmission time can also be subject of adaptation. Elastic traffic defines the amount of data to be transmitted, but not a deadline until when this has to be accomplished. In this case, the transmission time and the bandwidth requirement can be adapted to resources according to

\[ \int_{t_0}^{t_1} r(t) \, dt = V \]  

In this formula, \( t_0 \) and \( t_1 \) denote the start and stop time of the transmission and \( r(t) \) is the transmission rate as a function of time. \( V \) is the total amount of data to be transmitted. For transmissions with hard real-time requirements, this type of adaptation is not applicable, and direct bandwidth adaptation, e.g., when applying \( a(t) \) as a scaling function of time, goes along with information loss

\[ r(t) = a(t) \cdot r_{\text{source}}(t) \]  

In this case, we are again in the content adaptation domain. Adaptation of elastic and non-elastic traffic is also discussed in [4].

3) Adaptation Instances: Adaptation of communication streams can be done at several places as shown in the classification scheme in Fig. 3. The probably most obvious approach for adaptation is in the end nodes. Another approach is to adapt transmitted data along the transmission path. For this, the two categories (1) with and (2) without knowledge of the structure of transmissions can be distinguished. Adaptation along the communication path can be necessary if short reaction times for adaptation control are required and no long signaling loops are acceptable or if no signaling paths are available due to existing non-adaptation aware communication protocols.

Adaptation with knowledge of the structure of transmissions can be done in intermediary proxies, either explicitly or transparently configured. This is a well-known technique in HTTP scenarios and also fits other communication protocols. The transmitted data can be transcoded as shown in WAP and in several demonstrations for WWW access over wireless channels. In those cases, the adaptation preserves data structures but changes sizes, e.g., by reducing colors and sizes or any property else changing the transmission size.

The second approach of adaptation along the communication path is also well-known but maybe not always seen as adaptation mechanism. Simple packet loss in network buffers (i.e., random drops) can also be seen as adaptation mechanism since the transmission properties of a stream are changed, i.e., the bitrate of a stream is changed from \( r_1 \) at sender side to \( r_2 \) at receiver side. Unfortunately, this produces reactive load at the network path from the sender to the dropper, but means very little effort in signaling and processing for adaptation. Typically, this type of adaptation is acceptable for streaming media where hard time constraints prohibit long signaling loops and where robust codings are used, which allow to recover from transmission errors (Forward Error Correction). Without error correction and for smaller reduction ranges, packet loss just reduces the quality of the transmitted media in a user acceptable way\(^2\).

C. Adaptation Control

Adaptation control is the task of adjusting all adaptation aware instances in a system to co-operate on shared resources. It can be divided into the two categories explicit and implicit adaptation control. Explicit adaptation control collects state information, lets an optimization algorithm run on those collected state information, derives a system setting for adaptation, and propagates this information. This process can be carried out in cycles or as a continuous process where the adaptation control watches the system and only acts when something needs to be changed.

It is obvious that a completely centralized adaptation control is an extreme approach which will not be feasible for widely

\(^2\)Many times, adaptation by packet losses is coupled with an end-to-end adaptation mechanism to reduce reactive load.
deployed systems. Reasons for this are the concentration of signaling traffic, long delays, and the creation of a single point of failure. Additionally, such a solution could also not be feasible due to different administrative domains. Therefore adaptation control has to be distributed among several instances in the network, probably organized in hierarchies.

Adaptation instances and adaptation control instances in the case of explicit adaptation control are not the same. Thus, it is necessary to signal the adaptation decisions from the control instances to the adaptation instances. This distribution of adaptation control information will be an important research issue. Communication between adaptation control instances and adaptation instances can be done either explicitly or implicitly. Known methods are

- push-back
- explicit signaling
- implicit signaling (e.g., silent delay)

Another approach for adaptation control is an implicit and totally distributed solution. Adaptation entities sense for information about the transmission conditions and draw their conclusion. An example for this is the well-known TCP flow control, which adapts its transmission rate to an estimated available bitrate derived from probing the network bottleneck link by increasing the bandwidth usage and sensing for packet losses. TCP flow control is still an active research field. This kind of adaptation control is very easy to deploy, and under the presumption that every participating entity behaves TCP friendly, we will see a mostly fair resource sharing. A challenge of such a solution is that it is difficult to assure the fairness if unfair partners are in the game, and that other resource sharing models than fair sharing are hard to realize.

For the presented system architecture, an explicit adaptation control at the mobile node and within the cells is chosen. Further studies shall be carried out to evaluate alternatives.

D. Resource Sharing

Environments with multiple simultaneous consumers of a single resource typically need one or more resource sharing mechanisms. In LAN environments where several devices have to access one resource, i.e., the wire-bound or wireless transmission channel, such mechanisms are known as media access control (MAC) mechanisms. MAC mechanisms are realized as centralized or distributed approaches and can either be deterministic if explicit or implicit signaling for resource access is used, or non-deterministic, if contention-based access is applied. Another type of resource sharing is used if several incoming traffic streams have to be merged into one outgoing stream, as in routers or protocol stacks of communication devices. In this case, resource sharing is realized by a centralized mechanism.

Considering a whole communication system, resource sharing mechanisms are applied at several stages and between several instances. The most obvious and maybe best-known places are the sharing of resources between devices within one access network and the sharing of communication resources among applications of one end device.

A well-accepted approach for resource sharing is the fair share model. A better, but more complex solution is to differentiate among consumers to share the resources in a more adequate way according to their needs.

A resource-aware communication support has to deal with the problem of resource sharing and provide mechanisms to set and to control resource usage of its consumers. In the case of the present system architecture, resource management and adaptation control are tightly coupled aspects of the same problem. Topic of research to be done by means of this platform is the hierarchy of resource sharing mechanisms. This will be a trade-off between absolute control, administrative domains, and signaling overhead.

E. Summary

Before closing this requirements section, the major requirements should be subsumed. For the considered system support, the following requirements should be fulfilled:

- Mobility support at transport layer or above
- Network detection and network selection
- Adaptation means of traffic flows
- Adaptation control instances
- Controlled resource sharing

III. Architecture

Fig. 4 depicts the system’s node and cloud view. The mobile node, home proxy, and fixed network node depicted there each represent the entirety of all communication nodes of their type. The wireless access systems\(^3\) which are drawn as white clouds

\(^3\)Depending on the level of abstraction and the hierarchies of mobility handling, access systems are either single access cells or multiple access cells combined to a wireless access network
are permanently connected to the Internet as core network while bindings between mobile nodes and access systems are temporary due to user’s mobility. Different interfaces at a user device can bind to different access systems in the multihoming case. Sticking to the client and server paradigm, client and server applications can reside both, on a mobile node or on a fixed network node.

For supporting mobility, an approach with signaled address changes is chosen. This architecture minimizes the number of involved nodes since path reconfigurations can be handled by simple address updates between the communication endpoints or one communication endpoint and an intermediary node. Reachability of mobile nodes is achieved by a fixed point in the network topology, i.e., a home proxy, with a n : 1 relation between mobile nodes and home proxies. This basic architecture is comparable to that of Mobile IP except some aspects like route optimization and authentication issues which are currently not considered.

According to the requirements, the framework should be able to support mobility at transport layer or above. For this, it is based on a middleware component deployed between the TCP/IP protocol stack and the applications as depicted in Fig. 5. The middleware component is distributed between a mobile node and its home proxy. This approach also simplifies the implementation of a prototype since kernel domains are not affected.

The building blocks of the middleware component are shown in Fig. 6. The binding between the middleware component and applications and services is handled by so-called relay components. Those relay components can either be socket-based proxy functions like a modified HTTP proxy or a modified mail gateway, or an API interface providing more sophisticated access to the middleware functionalities for specialized applications. At the mobile node and the home proxy the same kind of relay components can be used and each instance of a relay component at a mobile node will have its counterpart instance at its home node and vice versa.

To reduce system complexity, currently only best-effort message-based communications are considered as seen by HTTP web access, SMTP mail transport, etc. A one-way transmission of an application layer protocol like mail delivery is handled by uncoupled local state machines at both sides of the middleware component and a simple message transmission of the middleware. Bidirectional application layer protocols like HTTP are handled by coupled local state machines at both sides of the middleware component and two or more message transmissions which are bound together by addressing mechanisms. It is assumed that the proposed system architecture is also capable to handle real-time traffic streams with minor changes as demonstrated in the TOMTEN [5] architecture.

Communication between a mobile node and a home proxy is handled by two types of transport modules. The one
which is called “control transport” is responsible for signaling information. It is responsible for transporting short notification messages between the middleware instances at the mobile node and the home node. The transmissions are protected by checksums and a simple sequence enumeration and are carried by UDP packets. Since the volume of the signaling traffic is considered negligible, a very simple network selection mechanism is assumed for this type of traffic and the mobility management is done the same way as Mobile IP does. A mobile node reports its current address(es) to the home node, where they are stored and used for control message transmissions in downlink direction.

For transmitting payload messages, which can be of any size, e.g., several 100 kB or even MB, a module called “message transport” is used which is built on top of TCP. For each message an instance of the message transport module is created which has the full control over exactly one message and can either transmit a message in uplink or downlink direction. In each case, the transport process of a message is initiated at the mobile node side. A message transport is continued until stopped by the message control unit, which is discussed next, or until a fatal error occurs, e.g., an interface reconfiguration due to user mobility. In this case, a message is postponed and maybe resumed later by another instance of the message transport module.

Messages received through one of the relay components, either at the mobile node or the home proxy, are handed over to the message control instance of that node. This stores the payload of the message in a local storage container for messages, i.e., a local disk, and the management information in a local database. This way, message queues in uplink and downlink direction are realized as depicted in Fig. 4. They allow transmission adaptation by deferring messages.

For messages in downlink direction, the management information is immediately transmitted to the mobile node using the control transport mechanism. At the mobile node, this management information is stored in a shadow database mirroring the current state of the queue at the home proxy. This allows to place all network selection and adaptation control functions at the mobile node. Motivation for this is that a mobile device will have the best knowledge about available communication resources – either through own scanning of the environment or by receiving reports from offering networks.

The message control modules are the central instances of the middleware components at home proxies and mobile nodes. They are the clue between the relay components, message stores, management information databases, and the transport modules. They handle the message forwarding between the relay components and local message stores, manage local control information for messages, possibly provide hooks for message transcoders, instantiate message transport modules, and coordinate the mobile node and home proxy part of the middleware component by exchanging control messages. Since the mobile node part of the middleware is assumed to have the best knowledge about available access systems, the real adaptation control functions, i.e., when to upload or download a message, are placed at the message control unit of the mobile node while the message control of the home proxy acts as its slave.
By now, the presented system is an experimental framework for (1) system side transmission adaptation, (2) system side content adaptation at intermediary nodes and (3) adaptation control instances. Due to complexity reasons, the system is currently restricted to message-based communications while stream-oriented services are left for future extensions. Transmission adaptation is done by queuing messages and adjusting their transmission bitrates. Content adaptation can be done at messages stored in the mobile node’s or home proxy’s message store by transcoding modules. With this framework the means for system support for adaptive communications are given and hooks for adaptation control and resource control algorithms are provided, which have to be filled next.

IV. STATE OF THE ART AND RELATED WORK

Currently, a lot of effort is spent in mobility supporting architectures, QoS awareness, cross-layer issues and adaptation. In this section, an overview of those efforts is presented and set in relation to the proposed mobility and adaptation supporting framework.

Leaving the established paradigm of anytime anywhere communication, many proposals for communication systems advance to heterogeneous environments following a many-time many-where approach. They indicate gains in system efficiency due to reduced infrastructure needs by exploiting user mobility and communication deadlines. Among others, examples for this are the infostation approaches of the Rutgers University [6] and proposals for opportunistic scheduling in cellular systems as proposed by Xin Liu [7]. Those approaches have set their focus on infrastructure aspects and feasibility studies of such a change in paradigms. The presented framework takes those heterogeneous environments as given and extends the communication processes to context-aware communications where adaptation and cross-layer issues are handled by management instances.

Many researchers identified weaknesses in the mobility support of Mobile IP. A major criticism is the disability of Mobile IP to differentiate traffic flows and to handle more than one network interface at the same time. To overcome this, many architectures and frameworks are proposed treating mobility in a different way. Most of them follow the approach to place mobility support at higher protocol layers. Examples are Session Layer Mobility (SLM) by Björn Landfeld et al. [8] and MSOCKS+ by Pravin Bhagwat [9]. Those architectures follow the same approach as chosen in the proposed framework of this paper, but they set their focus on more specialized mobility routing mechanisms for allowing better QoS support and split traffic routing over simultaneous connected interfaces. In principle, those approaches are adopted in the current framework.

QoS mechanisms are a first approach to treat resource fluctuations transparently for prioritized applications. A survey on those mechanisms in the area of wireless LANs (WLANs) can be found in [10]. Cellular networks like GPRS and UMTS inherently incorporate such mechanisms. Following the approach of combining different access technologies to one heterogeneous access system as discussed before, QoS mechanisms soon reach their limits in cases of vertical handovers so that they are complementary to adaptation approaches but cannot replace them.

In the field of content adaptation, much research work has been done and is currently in progress. Good proofs of the concept of content adaptation can be found in the work of Armand Fox [11] for WWW access and in that of Elan Amir et al. [12] for streaming applications. A survey of adaptation in communication environments is presented in the IST report of the project “ANWIRE: Adaptable Service Architectures” [13]. Among others, system architectures realizing adaptation capable communication systems are TOMTEN [5] by de Silva et al. and the work presented by Brian Nobel [14]. The latter architectures are similar to the framework presented here but differ in the way adaptation means are integrated and adaptation control is distributed.

Within the topic of adaptation, the management of technical context is considered as a critical issue. There are some discussions about management of context and the requirements to adaptation support, e.g., from C. Efstratiou [15] and Stavros A. Xynogalas [16]. A conclusion of those papers is that besides having adaptation means and technical context the automatism for doing adaptation decisions is a critical issue which is not understood comprehensively yet.

A conclusion of this related work section is, that a lot of effort has already been spent on communication adaptation, but it is still a wide and open field. Many solutions for incorporating adaptation means into communication systems have been proposed and investigated, but there is still a lack of intelligent and system wide adaptation control.

V. CONCLUSIONS AND FURTHER WORK

With this paper a middleware based framework providing mobility support, content adaptation, and adaptation control for mobile applications is introduced. This framework provides transport layer and session layer mobility support with a home proxy based system architecture. It is capable of doing transmission adaptation by queuing messages and adjusting transmission bitrates and provides access to communication flows for performing content adaptation. Application specific content adaptation mechanisms can be attached to it by hooks and adaptation-aware applications can be included into the adaptation control by special APIs.

With this framework, a system architecture is specified as a reference architecture for further research in the fields of adaptation control architectures and adaptation control algorithms. Open questions in the field of the system architecture are the placement of control instances, e.g., at the applications, in middleware components, at the access network
controller, or at a higher level. The most limiting conditions for this problem will be the given administrative domains of networks and nodes, and the signaling traffic needed to coordinate distributed adaptation control. Challenges in the area of adaptation control algorithms are the rules necessary to manage adaptation and to formulate optimization criteria. For this, it seems desirable to break with the paradigm of fair resource sharing and to find an opportunistic traffic management optimizing the user’s fun and the overall system capacity.

ACKNOWLEDGMENTS

I would like to thank my students contributing to this project, namely Oliver Fritz, Lars Rudolff, Stefanie-Heike Hinchliffe-Wörz, Daniel Maringer, Heitor Roriz-Filho, Mei-Huei Chiou, and Ju Wang.

REFERENCES