THE IMPLICATIONS OF IP ON SATELLITE UMTS: THE SATIN VIEW

M. Karaliopoulos, K. Narenthiran
Centre for Communication Systems Research, University of Surrey, UK

I. Andrikopoulos, E. Angelou
Space Hellas S.A., Athens, Greece

M. Mazzella, P. Henrio
Alcatel Space Industries, Nanterre, France

ABSTRACT

Efficient support of Internet-based applications to mobile/nomadic users is a key feature of the third-generation (3G) networks. In light of the shortage and the high cost of the T-UMTS spectrum, the operators are looking into the provision of integrated broadcast/multicast services through hybrid broadcast-UMTS systems. S-UMTS could play an important role in the efficient delivery of some UMTS services to which it is better suited. These services include broadcasting and multicasting applications such as audio/video, e-newspaper and live stock exchange data. The level of IP penetration into 3G networks is a decisive factor for the design of an efficient system, optimized for the delivery of these services. The paper identifies the respective requirements arising for the S-UMTS air interface, in the frame of the architecture scenarios envisaged within the EU IST project SATIN.

INTRODUCTION

Second generation mobile satellite systems failed to grab the mobile market, raising a lot of concerns about the future of commercial satellite systems in general. However the multimedia concept, strongly embedded within UMTS, introduced a new perspective for the mobile satellite systems as collaborative parts of terrestrial UMTS (T-UMTS) rather than stand-alone systems.

The satellite component of the UMTS (S-UMTS) system architecture has been extensively studied in a number of projects run over the last five years (such as EU ACTS projects SUMO and SINUS) and some basic principles have been established. The requirement for interoperability and integration with T-UMTS was one of the main drivers of these studies while the concept of discriminating between radio-independent and radio-dependent functions in the system design, as it was first coined in the RAINBOW project for T-UMTS, seems to have achieved wide acceptance[1]. However recent developments in the T-UMTS architecture generate some new challenges for the S-UMTS architecture design. The introduction of packet-mode into the system definition and the ever-increasing penetration of the Internet Protocol (IP) into the system functions constitute the basic reasons for a re-examination of the system architectures proposed so far and their modification/optimisation.

SATIN (Satellite-UMTS IP-based Network)[*] is an IST research and technology project partially funded by the EU. Its main objective is to introduce a new packet-based S-UMTS architecture as an integral part of UMTS. The derivation of specifications for the access scheme in packet mode, including functions and respective component interfaces, as well as the evaluation of key issues by means of simulation are tasks stemming from this objective.

The paper is organized as follows: The current penetration of IP into the T-UMTS is first reviewed and the future trends regarding the level of this penetration are presented. After a brief description of the architectures envisaged within SATIN, the corresponding requirements for the S-UMTS air interface are investigated. Different options are identified depending on the IP penetration into the Core Network (CN). The paper concludes by outlining the approach taken in SATIN.

T-UMTS AND IP

Internet is considered to be one of the big success stories in the telecommunications world. IP traffic is without doubt the dominant type of traffic in current data networks. More recently IP has been gaining acceptance as a platform for delivery of multimedia services (a role flirted in the early 90s by ATM). In fact IP is the protocol of today and seems -more than ever before- to be the protocol of the future as well.

[*] http://www.ist-satin.org

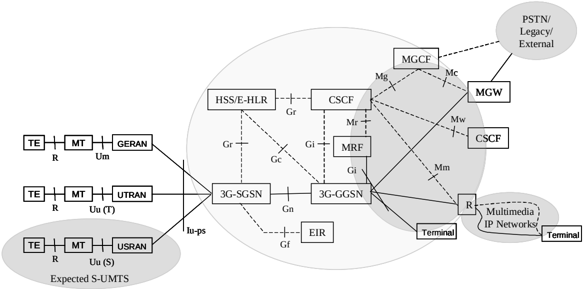

Figure 1: 3GPP UMTS architecture

UMTS (3G systems in general) have advertised quite early their vision to marry two success stories of 90s: cellular mobile networks and the Internet. IP is certainly present in the first release of T-UMTS (Release ’99). IP is deployed for transport of both data and signalling in the Core UMTS Network over the GTP tunnels, while IP addresses are allocated statically or dynamically to mobile hosts. However the level of this presence is limited so that we talk about interworking of the two technologies (Internet and cellular) rather than real integration.

Since mid-1999, UMTS specification work has been driven by a shift towards an all-IP UMTS network architecture. This shift formed the basis for the R00 specifications, which replaced the circuit-switched transport technologies used in UMTS R99 by packet switched and introduced multimedia support in the UMTS core network. Moreover, outside the official standardisation bodies (i.e. 3GPP, 3GPP2) a number of fora and partnerships, between manufacturers and operators (e.g. 3G.IP, MWIF), have greatly contributed to the success of the all-IP UMTS concept.

Mobile-IP, Session Initiation protocol (SIP), IP Integrated and Differentiated services are components of the IP ensemble envisaged for introduction in UMTS subsequent releases. The level of IP penetration into the UMTS architecture is constantly increasing, modifying the way certain system functions are implemented. In the following, the IP role in a number of UMTS functions is briefly reviewed, in an attempt to identify the terrestrial network that SATIN has to interface with.

Mobility

Regarding UMTS we can distinguish two different types (levels) of mobility:

- Macro-mobility, namely mobility between different RNC/SGSN nodes.
- Micro-mobility, namely mobility between different Node B elements within the same radio Access Network (RAN) - controlled by the same RNC.

In UMTS release ’99, macro-mobility is treated by the Core Network nodes/entities, SGSN, GGSN and HLR/VLR, assisted by the GTP protocol running over the Iu and Gp/Gn interfaces (Fig. 1). GTP tunnels are set-up and released when the mobile is moving, making sure that user and signalling data are routed via the appropriate 3G-SGSN. A more IP-oriented solution based on IPv6 was examined within the frame of the IST project Wine-Glass[2]. The basic idea is to replace the GPRS core network with an UTRAN-IP gateway that will terminate the Non Access Stratum signalling on the fixed network side and charge Mobile IP and its enhancements with the task of macro-mobility management.

Regarding micro-mobility, a number of IP-based schemes have been proposed for supporting mobility within the RAN (e.g. Hawaii, Cellular IP, etc). These solutions assume IP transport in the RAN and aim at replacing the standard UMTS mechanisms (soft handoff). However there are a lot of objections in relation to the applicability of these proposals in the UMTS case. The main obstacles are:

- The specific characteristics of the UMTS RAN radio network, namely the strict delay requirements for soft handoff, the absence of an end-to-end IP routing model and IP transport within the RAN, the radio-specific layers of the nodes. The same reasons also render inappropriate the use of application-level solutions to the micro-mobility problem[3].
- The lack of mature IP-based mechanisms that could, even in the case some of above constraints are overcome, provide a more efficient mobility mechanism than the specialized one currently deployed.

IP transport in the RAN

In release ’99 the transport of both data and signalling is provided by ATM/AAL2. There are quite strict requirements in terms of delay and transport efficiency that rendered IP suitability questionable. The lack of a mature QoS framework and the overhead imposed by the associated headers were the two main reasons for not considering IP as an option by that time.

The more recent progress in the definition and development of mechanisms that can provide service differentiation in combination with efficient multiplexing and compression schemes alleviating the headers’ overhead made IP a more attractive option. Furthermore there are expected benefits in terms of cost reduction, deployment flexibility and scalability.

Ten percent (10%) superiority in transport efficiency over ATM/AAL2 is claimed in Mobile Wireless Internet Forum (MWIF) studies. The introduction of IP as a transport option within RAN is ‘la raison d’ être’ of MWIF[4]. The outcome of their study is fed into 3GPP with the aim to be standardized there. The study addresses the conventional, and simplest at the

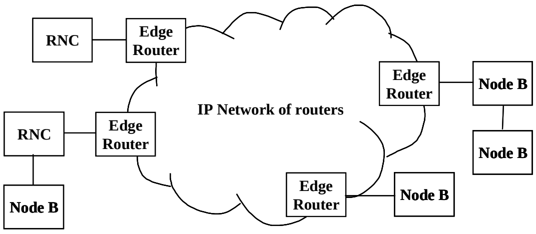

Figure 2: IP transport in the RAN[4]

same time, case of point-to-point connectivity between RNC and Node B. More complex topologies providing mesh IP connectivity between a number of RNCs and Node B elements are also envisaged (Fig. 2) but in that case the relevant efficiency is expected to be less due to the routing overheads.

Multicast

IP multicast defines an architecture that allows IP applications to send data to a set of recipients (a multicast group) specified by a single IP address. Audio/video conferencing, push-based data dissemination and remote education are examples of applications that make use of this architecture. Their efficiency lies in the fact that only a single datagram traverses a link rather than a number equal to the sum of receivers involved in such applications. The main functional elements of this architecture are the multicast routing and the multicast group management functions.

On the other hand, the standardisation process of UMTS point-to-multipoint services is undergoing within the Release 5 framework. The respective architecture is named Multimedia Multicast Broadcast Services (MBMS)[5].

Two main sets of multicast capabilities in the network have been identified in the MBMS:

- set 1: support of IP multicast over Gi (GGSN supports multicast), with conventional GPRS Tunneling Protocol (GTP).
- set 2: set 1 completed by providing multicast PDP context, multicast Radio Access Bearer (RAB).

It remains that both sets require multicast over the RAN.

The main option currently retained for the data path in the Core Network suggests sending multicast data from a multicast “source” (could be a multicast server or multicast capable node) to the selected GGSN which support multicast service (possibly identified using APNs) and their further distribution from the GGSN towards the RNCs via the SGSNs with registered multicast users.

Two main protocols could be envisaged in terms of Multicast data transport in the Core Network (i.e. two different functional architectures):

- GTP (i.e. GTP-U protocol in the User plane and GTP-C signalling).
- IP-multicast with IGMP, deemed more efficient over the transport network if it were to support multicast routers.

The respective selection remains a study item in 3GPP SA2.

End-to-end QoS support

As the Internet evolves towards the global multi-service network of the future, a key consideration is support for services with guaranteed Quality of Service (QoS). In recent years, the Internet Engineering Task Force (IETF) has proposed a number of QoS models and supporting technologies including the Integrated (IntServ) and Differentiated Service (DiffServ) frameworks.

When a Terminal Equipment (TE) receives some end-to-end service from another TE, the resulting traffic has to traverse the different bearer services of the underlying networks. In order for such a service to be realised a TE/MT Bearer Service, a UMTS Bearer Service and an External Bearer Service are used.

To provide IP QoS end-to-end, it is necessary to manage the QoS within each domain along the end-to-end path. Whenever resources not owned or controlled by the UMTS network are required to provide QoS, it is necessary to interwork with the external network that controls those resources. Interworking may be realised in a number of ways, depending on the QoS framework utilised in each network domain.

An IP Bearer Service (BS) Manager is required to manage and control the IP bearer services. Due to the fact that different standard IP mechanisms for QoS may be used within UMTS (e.g. IntServ RSVP, DiffServ packet marking/traffic conditioning), a Translation Function is necessary for the communication between the IP BS and the UMTS BS managers. Provision of the IP BS Manager – and hence the Translator Function – is optional in the UE and mandatory in the GGSN. In case there is an IP BS Manager in both the UE and the GGSN, then the two managers may communicate directly by using suitable signalling protocols.

Multimedia subsystem –call control

The IP Multimedia Subsystem was incorporated in Release 5 of UMTS and constitutes a significant step towards a closer integration with the Internet. The main components of this architecture are a number of functional entities and the Session Initiation Protocol (SIP).

SIP has been selected as the official IP multimedia call control protocol by 3GPP for the 3rd Generation mobile communications or UMTS. It is going to be used in the IP multimedia system in the UMTS all-IP network architecture as proposed in Release 5. Its main function is the provision of control functions for real-time, multimedia flows. The need to adapt the original SIP architecture to the needs (charging, accounting, authentication) of UMTS has motivated a close co-operation between IETF and 3GPP, aiming at accelerating the derivation of the respective specifications.

The main entities in the IM subsystem are the Call

State Control Function (CSCF), Media Resource function (MRF) and Home Subscriber Server (HSS)[†]. MRF performs multiparty call and multimedia conferencing functions (same function as an MCU in an H.323 network). HSS is the master database for a given user and hence, it contains the subscription related information to support network entities actually handling calls/sessions.

Neither SIP nor the aforementioned entities are examined within SATIN.

SATIN ARCHITECTURE

The architecture scenarios selected within SATIN are depicted in Figure 3.

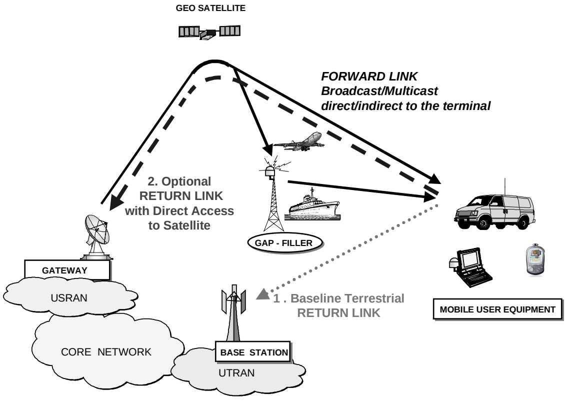

Figure 3: SATIN architecture

Two scenarios, a *baseline* and an *optional*, were selected to address the service requirements, as identified in earlier stage of the project[6]. In the baseline scenario, a handheld mobile terminal, receives data through the satellite and/or the intermediate module that features one-way repeater functionality. The satellite path would be the preferred communication link, but if the user's satellite path were blocked, the communication link would be sustained via the intermediate module repeater (IMR) stations. The introduction of IMR modules was deemed mandatory in order to overcome the inability of satellite-only systems to offer in-building and in-urban areas coverage (where the mass market resides) and support the moderate and high bit rate multicast/broadcast (MC/BC) services envisaged within SATIN. The return path is provided via the T-UMTS network (*baseline* case). A satellite receive-only terminal may well serve a given subset of services (pure broadcast, and non-highly interactive multicast). Alternatively, the terminal may also support direct transmission (in the return path) to the satellite (*optional* case), leading to the more conventional system configuration that allows a stand-alone system to be built at the expense of a more expensive and complex terminal[7].

IP IMPLICATIONS ON SATIN

Before proceeding with the IP implications, it is worth commenting on the relation of S/T-UMTS packet mode to the Internet Protocol. IP is a connectionless, network-layer protocol featuring packet (datagram) switching at the network nodes. In the traditional, best-effort (BE) manner, these nodes do not maintain any state about packets that come from a specific source and/or belong to the same information stream. Every packet is treated in an independent manner and there is no sense of connection at the network layer. However, this does not necessarily mean that IP is carried always in this 'pure' packet mode. Depending on the underlying transmission technology, which supports IP in a specific domain, the IP datagrams may be transferred in either packet- or circuit-mode. The IP transport over PPP (Point-to-Point Protocol) connections in the case of dial-up links is maybe the most straightforward example of circuit-mode transport of IP datagrams. In this case the connectionless service of IP layer is emulated over a strictly connection-oriented technology.

In UMTS the concept of packet mode encompasses specific transport methods that differentiate this mode from the traditional circuit-mode of the 2G networks. The major components of this method are the shared channels and the lack of explicit connection set-up procedures for the initiation of a so-called 'packet call'[8].

Therefore an IP-based packet-mode S-UMTS has to face two kinds of implications:

- The ones stemming from the adoption of packet-mode. These are not irrelevant to the IP suite, but are mainly a consequence of the transport protocols and the applications (e.g. TCP and HTTP induced traffic burstiness in case of the WWW) rather than a direct consequence of the Internet Protocol as such.
- The ones originating from the necessity or wish to achieve a closer integration with the Internet and the fulfilment of some functions in an end-to-end manner. To a great extent these implications are a consequence of the way these functions are performed in the terrestrial Internet (e.g. protocols, architectures).

[†] Note that only UMTS-UMTS calls are being considered here; other entities like MGCF, R-SGW, T-SGW and MGW will be needed if UMTS-PSTN/PLMN calls are considered.

Note that in some cases it may be difficult to discriminate between the two categories. While these issues also arise in T-UMTS, the specific characteristics of the satellite environment may magnify, alleviate or add new dimensions to them. Furthermore the two configurations chosen within SATIN introduce some extra considerations.

Multicast

The support of IP multicast in SATIN is mainly foreseen in:

- Taking benefit of the advantages of the UDP connectionless, datagram service for broadcast/multicast transport of applications and leaving acknowledgement processing at the application level (reliable multicast transport techniques).
- Targeting minimum acknowledgement of multicast transmission and retransmission needs.
- Optimising the content distribution by means of broadcast/multicast data servers and techniques such as web caching and mirroring, that are not necessarily located in the SATIN gateway and perform:
  - Routing to build multicast/broadcast IP streams of multimedia content (use of different multicast addresses, each corresponding to a service offer to the users in terms of content type and associated quality of service and security requirements) associated with content element segmentation, possibly QoS based routing (terrestrial versus satellite segment), scheduling as well as security features and reliable multicast transport techniques (FEC, retransmission).
  - Content serving to assign a service descriptor to each multimedia content; this descriptor being used all along the distribution chain to perform optimum routing, scheduling, and subsequently filtering, cache management as well as presentation to the user.

The implementation of both these functions will be based on open standards such as those devised within IETF or other fora.

The way multicast will be supported in SATIN (and more generally in any S-UMTS configuration) is heavily dependent on the level of IP penetration in the UMTS CN and its role in the macro-mobility support.

While it is agreed that it is not easy for IP-derived solutions/protocols to cope with the strict requirements of the UMTS micro-mobility functions, hence these functions are left to the native UMTS protocols, there are two approaches for the UMTS macro-mobility support.

The first is the solution currently implemented, up to Release 5, relying on the conventional SGSN, GGSN nodes and the GTP tunnels throughout the CN till the RAN edges. The second draws heavily from the IP-based solutions and promises better integration with the Internet. The standard GPRS network is replaced by (compressed into) a UTRAN/IP gateway, which is attached to a backbone of routers running pure IPv6, while Mobile IPv6 is charged with the macro-mobility task. The latter approach is considered to be a mandatory step towards the realization of fourth generation (4G) networks. In the following, the implications of each one of these approaches regarding multicast support are explored.

Multicast in an all-IP CN

The adoption of Mobile IPv6 in the CN makes the application of IP-derived solutions for multicast support more straightforward (or even mandatory):

- Multicast capable routers can be deployed at the CN for more efficient multicast transport
- IGMP can/must be used for group management purposes.

Support of IP multicast in this case has to address mainly the scaling problem; namely the standard IP multicast architecture implies a significant overhead of signalling/control messages, given the number of potential hosts per spot beam. These messages can generally be either multicast routing messages exchanged between the multicast-routing capable entities of the network or IGMP (Internet Group Management Protocol) messages. Within the SATIN context the problem is related to the IGMP messages. IGMP capable routers detect the presence of group members by sending IGMP queries, to which hosts answer with IGMP report messages. The messages are timer-driven and may constitute a significant portion of the network load, effectively reducing its available capacity for data traffic.

Nevertheless, there are two features of SATIN (and more generally satellite networks) that have to be noted and can be exploited for a more efficient support of multicast services.

The tree-like network topology and the 'IGMP proxying' principle

The aforementioned signalling load and the respective resource consumption can be avoided in certain topologies. This is the main reason why the 'IGMP proxying' (IGMP-based Multicast Forwarding) technique was conceived. The specification of this mechanism is still in a draft state but some of the ideas contained therein seem to fit well the considered multicast scenarios[9].

With respect to their position in the multicast spanning tree, the router interfaces can be divided into downstream interfaces (DI) and upstream interfaces (UI. There can only be one UI for an IGMP

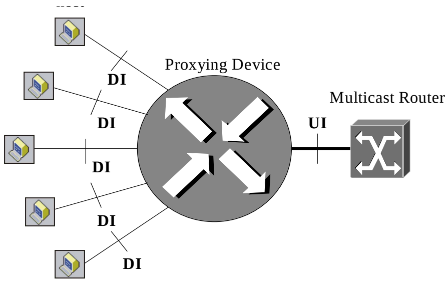

Figure 4: Standard IGMP proxying interface classification

proxying device (Figure 4). DIs are in the direction of hosts while UI is in the direction of another router. This differentiation is introduced since, depending on its type, each interface plays a different role in the protocol.

In the proxying technique, DIs run the so called router portion of the IGMP protocol; in other words, on each interface, the normal IGMP operations are performed, maintaining in a separate way a membership database. These databases are then merged to obtain a *global membership database* that takes into consideration the memberships on each interface.

UI runs the *host portion* of the IGMP protocol, so it has to send IGMP membership reports when it receives a query message, and has to send unsolicited reports or leaves when database changes occur.

As far as the forwarding technique is concerned, when a router (or proxying device) receives a multicast packet, it builds a record in a *forwarding database* consisting of a list of the interfaces (UI and DIs) where there is a subscription to the group except for the interface from which the packet arrives. Then it forwards the packet to those interfaces. This operation can be made simpler if the forwarding database is used as a cache, so that the creation of a record in the database is made once for all the packets belonging to the same group. This simplification comes however at the cost of updating the cache every time the situation in the membership changes.

In SATIN it is the S-RNC (Gateway) and potentially the UTRAN-IP gateway (physically they might be the same) that has to play the role of the proxying device(s).

The LAN-like nature of the network

Rather than implying a strict resemblance with a Local Area Network (LAN), the term “LAN-like” refers to the capability of all the hosts within a beam to receive all transmissions destined for this beam. This capability can be exploited to reduce the number of exchanged IGMP messages over the air interface. Rather than letting every mobile host (MH) send reports back to the gateway, which implements the IGMP querier functionality, one of the multicast group members is elected as the group representative for IGMP proxy[‡] functions for the whole group. The other hosts trigger a timer whenever they see a report from the designated group proxy and only send their own report when this timer expires. The underlying principle is that the gateway does not have to be aware of the exact number of MHs participating in a given group but rather whether there is one or more MH(s) in a specific beam so that a copy of the message is forwarded to this beam. The requirement for the *Max Response Time* field is to be higher than the roundtrip time but reasonably low so as to reduce the number of membership report messages sent after the receipt of the *General Membership Query*.

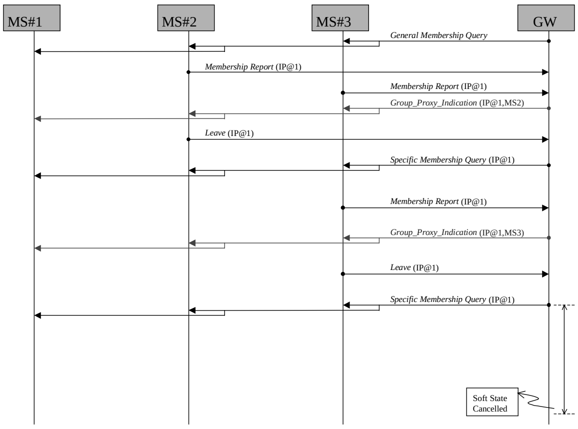

Figure 5: Example of IGMP proxying with overhead reduction

An instance of IGMP message exchanges, when both the aforementioned optimisations are adopted, is shown in Figure 5. The figure illustrates the case of three Mobile Stations (MS) associated with a given Gateway (GW) which acts as querier. The difference in comparison with the fixed broadband access system case lies in that MS terminals are not connected to a CPE (Customer Premises Equipment, a device with layer 3 functionality in this case), which would take on the role of the IGMP proxy for them.

A periodic *General Membership Query* (GMQ) message is broadcast to the cluster by the GW, which contains the selected *Max Response Time*. Upon receipt of the GMQ, the MS sets a delay timer for each group of which it is member. Timers are set to a random value selected from the range (0, *Max Response*]. In our example, MS#2 and MS#3 wish to receive traffic sent to the multicast address IP@1, while MS#1 has no active memberships. The first *Membership Report* message that is received comes from the MS#2, which, according to the overhead reduction protocol, becomes the elected group proxy. The GW broadcasts a *Group_Proxy_Indication* message (network signalling message) to inform all MSs in the cluster that MS#2 has been elected the

[‡] In the remainder of the paper the term ‘group proxy’ refers to the IGMP signalling overhead reduction, while the ‘IGMP proxy’ term refers to the IGMP-based multicast forwarding.

group proxy for the address IP@1. From now on, all members of the group IP@1 except MS#2 can suppress membership report and leave messages.

At the end of the session, MS#2 cancels its subscription to group IP@1 by sending a *Leave* message to the GW. As in standard IGMP, the latter sends a *Specific Membership Query* to make sure that no other member of the group is active in the cluster. In our example, MS#3 is the remaining member of the group IP@1, therefore it will send a *Membership Report* to the GW, and will become the new group proxy for address IP@1. When MS#3 - last and single member of the group IP@1 - finally leaves this group there is no reply to the *Specific Membership Query* sent by the GW. When the timer for the subscription to the group IP@1 expires, the GW cancels the relevant soft state.

The extra difficulty, when applying the second principle (IGMP signalling overhead reduction technique) in the case of mobile hosts, featuring no proxy device in front of them, is that modifications can no longer be transparent to the end hosts. Hence, it is necessary to modify the IGMP 'client' software at all hosts, while in the fixed satellite systems with end-hosts in a LAN behind a router, it would be enough to modify the latter.

For the *baseline scenario*, where the return link is provided by the T-UMTS, a solution for overcoming the unidirectional nature of the satellite link is provided by the Link Layer Tunneling Mechanism (LLTM), standardized in the IETF UDLR WG.

The UDLR LLTM

In the baseline scenario, it is necessary to come up with a solution to the problems posed to IGMP by the unidirectional nature of the satellite link. IGMP, much like the IP routing protocols, has been designed and engineered assuming a bi-directional link. Since this does not exist in the SATIN baseline scenario, it has to be emulated somehow over the T-UMTS link.

Such problems have mainly been addressed in the context of fixed satellite networks, where the unidirectional link is a satellite broadcast link (e.g. DVB-S) and there is a return terrestrial channel (e.g. dial-up line, PPP) that allows some form of interaction between the end-user and the provider/network operator. The IETF UDLR[§] WG concluded the first part of its activities with the specification of a link-layer tunnelling mechanism, which effectively allows the emulation of a bi-directional link over a unidirectional link. Within the SATIN context, UDLR feeder/hub functions are required in the GW and UDLR receiver/host functions are required in the terminals.

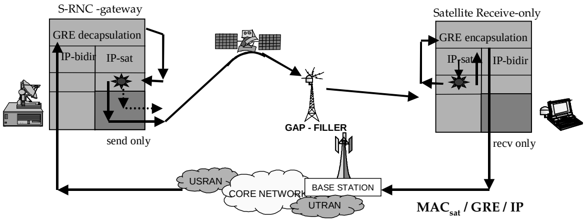

Figure 6: UDLR functionality within SATIN

Multicast in a GPRS-based CN

The implementation of multicast in this case seems to be a different case. The main reason for this is the different business paradigm of the two networks, namely the current, best-effort Internet and UMTS.

In the former, there is intensive, time-based, signalling at the edges of the network, between the hosts and the closest multicast-capable router because there is no detailed state at the router upon the exact number or addresses of the hosts that receive multicast content. In other words, the only thing required by the router is to know whether there are one or more hosts that want to participate in a multicast session. In order to facilitate this, the end hosts are subject to this frequent IGMP message exchange. In effect, the trade-off between signalling overhead and router complexity is determined in favour of the latter, the underlying assumption being that at the edge of the network the luxury of wasting some bandwidth on additional signalling is feasible.

On the contrary, in a mobile, wireless network like UMTS, there is generally much more information for the end user available at the network nodes. This information is available anyway in order to support the user mobility and AAA (Authentication, Authorization and Accounting) functions.

The additional capability in the GPRS networks is the capability to use this information when 'routing' each packet in pre-established tunnels that are created during the multicast PDP contexts. Therefore it is feasible for the SGSN to route traffic between the two access networks (T- and S- UMTS), since it is only necessary to set up the tunnels initially and make the respective bindings. This is not feasible, at least on the basis of the standard datagram routing paradigm, in a WineGlass-like all-IP CN.

Therefore the IGMP-related issues become of less (or even no) relevance in this case, since the information provided by IGMP is 'there' and, most significantly, can be used for routing packets to the users.

TCP and UDP/RTP transport

The TCP/IP suite provides applications mainly with two transport capabilities, expressed by Transport Control Protocol (TCP) and User Datagram Protocol (UDP). The former provides a reliable, byte-stream, connection-oriented service and is the workhorse protocol for the traditional (and most popular)

[§] http://www.ietf.org/html.charters/udlr-charter.html

Internet applications, while the latter a connectionless service that is often used by real-time services, with an intermediate session/application level protocol providing the respective control functions.

TCP flows

The provision of asymmetric, TCP-based services is envisaged within SATIN particularly in the optional scenario, given the interactivity these services require. In the baseline scenario the required interactivity will be provided via the terrestrial return link.

The problems TCP faces over satellite links have been a subject of research for quite some time, although it experienced a peak in the last 5-6 years. The propagation delay related to a GEO satellite and the wireless nature of the satellite link, are the main factors of TCP performance degradation. The former reduces the effectiveness of the window-based flow control of the protocol and its responsiveness to congestion incidents. The latter interacts badly with the protocol congestion control mechanisms. Although the requirement to be accommodated over satellite links was identified from the very early days of the Internet, the main TCP congestion control algorithms were devised subsequently with the underlying assumption of an error-free link: therefore any indication of packet loss is interpreted as congestion incidence at the terrestrial network. TCP flows are adaptive, in that when they sense congestion they cut down on their sending rate and adjust to the capacity that is available to them, at least their estimation of it. In a wireless network, especially one with a one-hop link without any buffer intervening (like in the case of the non on-board switching capable satellite envisaged for SATIN) losses are due to link errors. Given that TCP works end-to-end, it cannot differentiate between a corruption and congestion loss and reduces its sending rate even when there is no reason for that (e.g. congestion), leading to reduced throughput.

The SATIN architecture might feature another reason for TCP performance degradation: asymmetry, which is expressed as bandwidth asymmetry in the optional case and as both path and bandwidth asymmetry in the baseline case, where the T-UMTS network provides the return link. In fact this asymmetry is most of the times an engineering choice that fits the asymmetrical nature of some applications (traditional IP client-server applications). Asymmetric effects may be the result of asymmetry in the available capacity in the two directions of a TCP transfer, additional latency in one of the links due to medium sharing, usually in the return link, and can be exaggerated or smoothed depending on the traffic load and the respective queuing delays in each one of the two directions.

Such phenomena can have a negative impact on the TCP throughput, since the pacing of the ACK packets in a slower return link limits the sending rate of the TCP sender in the forward direction and prevents it from using potentially available capacity. When bandwidth asymmetry is combined with increased traffic load at the return link the effect can be even more dramatic due to the high queuing delays or losses of ACK packets, depending on the buffer sizes at the slow return link.

A number of ways to attack these problems have been proposed[10]. The majority of the solutions suggest transport (TCP)-level modifications, introducing different implementations of either the TCP sender or the receiver -or both. Timestamps, window scaling and larger initial window are considered a MUST in ‘long fat networks’, namely networks involving links with a large bandwidth-delay product. Furthermore TCP implementations with more elaborate flow control mechanisms (like TCP Vegas) or better response to congestion (Selective Acknowledgements, New Reno) than the standard Tahoe and Reno TCP have reported performance improvement, although they are not tailored for satellite links.

Proxying techniques report much more promising results[11]. TCP Spoofing, TCP Splitting and even TCP-aware link level schemes like Snoop outperform the standard end-to-end TCP connections. Although there are some arguments against their use, associated mainly with the extent to which they adhere to the ‘end-to-end’ principle and their incompatibility with the use of network-level security mechanisms like IPSEC, such techniques are known to the satellite community and have been widely used in fixed satellite networks. In fact a combination of split connections with link-level retransmissions yields superior performance and guarantees some resilience to the terrestrial network level of congestion[12].

The RLC retransmission protocol may be a further subject of differentiation with respect to T-UMTS. The ARQ protocol deployed in the case of optional scenario cannot have the persistence of the terrestrial analogues. It is also established that typical connection-oriented link layer protocols, attempting in order delivery of packets/frames interact badly with TCP.

Regarding asymmetry: for a start, the range of possible solutions that can smoothen the asymmetric phenomena may be divided into host-side and network-side ones: in the former case the improvement comes from modifications in the protocol stacks of the sender, the receiver or both, while in the latter, changes in the network elements – transparent to the end TCP host- are responsible for any performance enhancement. There are also proposals that necessitate combined action by network and users in order to yield some positive result.

A significant number of techniques are in an experimental stage[13]. Regarding the end-to-end mechanisms, an agreement appears to exist upon the benefit of the TCP connections from the use of the Path MTU Discovery mechanism that can save performance degradation due to potential network-level fragmentation. TCP Pacing is also promising in the sense that it does not necessitate major changes that could affect other standard connections in an adverse manner[14]. ACK Congestion Control, a technique that attempts to expand the TCP data congestion control algorithms to the ACK packets, is in an experimental stage. Cruder solutions like Modified Delayed ACKs, that reduce the number of ACK packets sent at the return direction, are not favored since they increase the TCP sender burstiness and may trigger unwanted effects in the forward direction.

Although some of the end-to-end mechanisms are promising, the SATIN context favors the network side, end-user transparent solutions for the mitigation of any asymmetry effects. The complexity argument that usually acts preventively in such cases is well outweighed by the promised performance improvement and the scalability of the approach. It is also the centralized, radio access architecture of SATIN that favors such solutions; the bottleneck link is placed between the mobile terminal and the gateway, allowing the latter to exercise a number of techniques to alleviate any undesirable asymmetry effects -note that in this case asymmetry is the result of engineering action rather than a physical medium limitation (as would be the case in terrestrial, dial-up connections for example).

The use of header compression techniques can limit the traffic load of ACK packets at the return link. Both the traditional Van Jacobson algorithm and more recent algorithms investigated mainly by the IETF ROHC[**] (Robust Header Compression) group fitting better to the non error-free wireless environments may be adopted in the system design. In fact the aforementioned schemes/frameworks have been adopted by 3GPP, while specifying the Packet Data Convergence Protocol (PDCP). Furthermore, differentiation in the treatment of ACK packets will be provided by the use of appropriate scheduling mechanisms at the S-RNC node. More innovative solutions like ACK Reconstruction and ACK Compaction/Companding that provide some form of the ACK packet stream regeneration, and are exercised immediately after the bottleneck router are not favoured within SATIN, given that the experience from the use of such techniques is rather limited and their use not recommended.

UDP flows

The efficient support of interactive real-time services (e.g. VoIP) is not possible within SATIN. The GEO satellite network introduces high latency, which becomes unacceptable in the case that a MS would like to communicate with another MS, implying a double-hop within the satellite network (MS-SRNC-MS).

On the other hand, UDP is also the common choice for the transport of audio and video services, which constitute core services of the SATIN portfolio. A whole family of protocols, namely the Real Time Protocol (RTP), the Real Time Control Protocol (RTCP) as well as the Real Time Streaming Protocol (RTSP) have been devised for the support of IP-based (streaming) multimedia services. Multicast transport channels are often used for the transport of these services.

There is recently a lot of interest in the efficient modelling of this type of traffic, given the growth of the respective services and despite the problems posed by the proprietary, in many cases, technology/protocols lying behind them.

IP QoS support

Support of IP quality of service in the SATIN scenario brings a number of issues that are worth investigating.

Making the assumption that the core network will be based on an IP DiffServ solution, given its proven scalability compared to IntServ, careful considerations must be made with respect to how DiffServ or IntServ will be used in the user and access domains taking also into account the mappings between the IP mechanisms and the underlying UMTS capabilities.

The SATIN baseline scenario assumes a bi-directional communication where each direction follows a different path, i.e. the forward link is via gateway/satellite/gap filler, whereas the return link is via T-UMTS. This has an impact on the way the IP QoS mechanisms will be used.

According to the RSVP specifications (within the context of IntServ), the RSVP PATH and RESV message objects must traverse the same route. This is done since the PATH message records the path along which the reservations will be made when the remote end send back the RESV messages. However, this is not the case with the SATIN architecture. The concept of the RSVP proxy may be used to overcome this problem. RSVP proxy functionality can also be used when RSVP client functionality is not implemented in the MS. More specifically, if the MS does not have the required IP QoS capabilities in order to provide end-to-end QoS, IP layer signaling may be performed in the IP and RAN gateway (e.g. RNC/GGSN) and be transferred transparently

[**] http://www.ietf.org/html.charters/rohc-charter.html

through the DiffServ core. The required QoS information can then be signaled between the MS and the RAN gateway using UMTS mechanisms. Moreover, admission control plays a vital role in the IntServ framework, as it is required in each RSVP-capable node. In contrary to wire-line networks, a number of additional factors need to be considered such as mobility and its interaction with other RRM functions. Indeed RSVP signalling should be avoided over the air. If SIP is being used then RSVP messages can be created by the GGSN by combining UMTS and SDP QoS information. The mapping of the IntServ service classes and associated QoS parameters to the UMTS classes is another aspect that needs investigation, taking particular care of the satellite link characteristics.

With DiffServ, the implications are mainly related to handover and mobility. When a mobile terminal with a specific SLS moves from one domain to another, there are no guaranties that the new domain will have enough resources to comply with the SLA because of the shared nature of the access network. In addition, the new domain may belong to another Service Provider that utilizes a different pricing scheme thus complicating things more. The application QoS is interpreted (PDP), the Serving RNC controls admission of new flow to the pre-established Iu-Packet Switched bandwidth pipe (DiffServ IP, IP bearer / QoS guaranteed pre-negotiated with ISP, traffic control is set based on SLS) and QoS mappings are performed inside and at the border of UMTS RAN (mapping between UMTS QoS Class and DiffServ done in the SGSN for Forward/Downlink and in the RNC for the Return/Uplink).

DISCUSSION - CONCLUSIONS

In SATIN the satellite component of UMTS is no more a standalone system that would only provide a coverage extension to the PLMN for a subset of services but rather a complementary means of service delivery in co-operation with the terrestrial access network, since Multi/Broad-casting over wide areas is best served by satellites.

Regarding the multicast support, which is the key issue within the SATIN context, the goal pursued in SATIN at first place is smooth integration into UMTS (3GPP Rel.5 is the reference although the MBMS architecture specifications are likely to be included in Rel.6). Therefore the architecture features a satellite RNC interfaced to GPRS backbone with some provision for evolving with the possible IP penetration in the 3GPP core.

For the SATIN baseline architecture the support of prime SATIN multicast services (i.e. no real-time interaction of the Return and Forward links required) favours the following main architecture features:

- Support of common features for the MBMS multicast and broadcast modes, e.g. both modes shall preferably use the same low-layer bearer for data transport over the radio interface.
- Support of external data sources in both modes (both IP multicast and unicast sources).
- MBMS shall be a point-to-multipoint service in the PS domain.
    - Home environment requirements of MBMS stage 1, inter alia Multicast Area
    - IP multicast capability in GGSN
    - Use of GTP (i.e. GTP-U protocol in the User plane and GTP-C signalling), in the terrestrial path and in the satellite path till the satellite RNC**.**

The transport service offered to multicast is an unreliable one leaving the responsibility for reliability provision to the upper layers.

REFERENCES

[1] J. De Vriendt, M. Schönborn, ‘Report on system concept studies in RAINBOW’, *3rd ACTS Mobile Communication Summit*, pp. 788-793, 8-11 June, Rhodes, Greece.

[2] IST Wine Glass Project, http://domobili.cselt.it/WineGlass.

[3] J. Kempf et al., ‘IP mobility and the CDMA radio access network: applicability statement for soft handoff’, Internet draft (work in progress), September 2001.

[4] Mobile Wireless Internet Forum, ‘IP in the RAN as a transport option in 3rd generation Mobile Systems’, Technical Report MTR-006v2, April 2001.

[5] Universal Mobile Telecommunications System (UMTS); MBMS; Architecture and functional description, 3GPP TR 23.846, Release 5, v0.0.1

[6] SATIN Project, ‘S-UMTS IP-specific service requirements’, Deliverable No. 2, October 2001.

[7] B. Evans et al., ‘Service scenarios and system architecture for Satellite UMTS IP-Based Network-SATIN’, *20th ICSSC and Exhibit*, Montreal, May 2002.

[8] ETSI Draft on S-UMTS packet mode, ‘Analysis and definition of the packet mode’, July 2000.

[9] B. Fenner, ‘IGMP-based multicast forwarding (`IGMP Proxying'), Internet draft (Work in progress), July 2001.

[10] M. Allman et al., ‘Ongoing TCP research related to satellites’, RFC 2760, February 2000.

[11] J. Border et al., ‘Performance enhancing proxies intended to mitigate link-related degradations’, RFC 3135, June 2001.

[12] M. Karaliopoulos et al., ‘TCP performance on split connection GEO satellite links’, *19th ICSSC Conference and Exhibit*, Toulouse, April 2001.

[13] H. Balakrishnan et al., ‘TCP performance implications of network asymmetry’, Internet draft (work-in-progress), November 2001.

[14] A. Aggarwal et al, ‘Understanding the Performance of TCP pacing’, *IEEE INFOCOM*, Tel-Aviv, Israel, March 2000, pp. 1157-1165.