Securing Network Access in post-3G Mobile Systems[1]

W. Buecker[1], G. Horn[1], K. Nyberg[2], H. Tschofenig[1]

[1] Siemens AG, Corporate Technology
81730 Munich, Germany
email: {wolfgang.buecker, guenther.horn, hannes.tschofenig}@mchp.siemens.de

[2] Nokia Group
…
email: kaisa.nyberg@nokia.com

ABSTRACT

This paper deals with the security in a "post-3G" mobile system. It focuses on the security features required to provide global IP connectivity and various forms of mobility.

I. INTRODUCTION

While the first release of third generation mobile systems is about to be deployed, and further releases are in the process of being standardised, much research is still needed relating to the mobile systems beyond the third generation. There is no common understanding of what a "post-3G" mobile system precisely is, there seems to be widespread agreement, however, that such systems will be characterised by an all-IP based core network to provide global connectivity, and a variety of heterogeneous access networks (e.g. WLAN, Hiperlan, Bluetooth, GSM, UTRAN and others) connected to this core network. A user in a post-3G mobile system should be able to use services from anywhere in the system (global roaming), and the use of a particular access network technology should be transparent to him when using these services. A distinction can be made between application layer services (e.g. web browsing) and network services (e.g. IP connectivity, mobility management, Quality of Service, session control). This paper focuses on the security features required to provide global IP connectivity and various forms of mobility to a roaming user in a "post-3G" mobile system. The paper is based on results from the IST project SHAMAN (Security for Heterogeneous Access in Mobile Applications and Networks).

II. REQUIREMENT AND SECURITY ARCHITECTURE

General technical approach
Protocols for IP-based networks are, in general, standardised by the IETF (Internet Engineering Task Force). The general approach taken for the work presented here is that any security architecture for post-3G mobile systems must be compliant with the IETF as far as applicable. However, the IETF (as opposed to e.g. 3GPP) does not standardise complete architectures. The challenge therefore remains to investigate how the protocols specified by the IETF can be used as building blocks for a post-3G security architecture in a consistent and efficient way.

Forms of payment and the need for authentication and authorisation
Network services may be provided free of charge in certain scenarios, in general, however, they will have to be paid for. For this reason, in general a user has to prove his ability to pay, and a network provider has to ensure that he will be paid for his services. Currently, this mostly involves a subscription with the network provider or some form of network-based prepayment. In the future there will be a larger variety of forms of payment based on e.g. local cash payments (Internet cafe model), credit cards, electronic purses or micropayments. The form of payment will have considerable influence on the security architecture. Depending on the form of payment used the identity of the user, or the validity of some payment token needs to be verified. This paper will explore the security architecture relating to the currently used forms of payment. A companion paper, also submitted to this conference, explores the impact on the security architecture caused by alternative means of payment.

[1] The work described in this paper has been supported by the European Commission through the IST Programme under Contract IST-2000-25350. The information in this document is provided as is, and no guarantee or warranty is given or implied that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability.

The currently used forms of payment require that a user be securely identified (authenticated), and his authorisation to use certain services be checked. In order to be able to authenticate and authorise a globally roaming user, an appropriate infrastructure is required. In this paper, in order to be compliant with the IETF approach, we will assume that a AAA infrastructure will be used. (The term “AAA” was coined by the IETF. It stands for “Authentication, Authorisation and Accounting”.) In general, it is also required that the network is authenticated by the user to avoid false base station-type attacks.

Types of traffic and the need for confidentiality and integrity

Various types of user and signalling traffic may require different degrees of protection. E.g. a public information services, provided free of charge which may be used at airports or in city centres, may not need any form of protection. Other traffic may require strong protection. In order to be as general as possible, we assume here that strong confidentiality and integrity protection for both, user and signalling traffic, is required. In order to be able to cope with the heterogeneity in the access part of the network, protection at the network layer, at least for user traffic, seems to be the straightforward approach. For compliance with the IETF, IPsec is then the obvious choice for the network layer security protocol. However, it requires careful investigation whether, in addition, link layer security is needed to protect access network specific signalling. This problem is discussed further below in this paper.

Forms of mobility

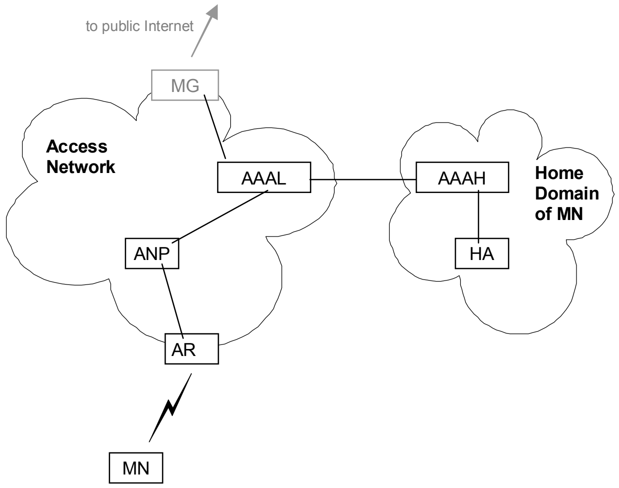
There are various degrees of mobility which may be satisfied by different protocols and may also affect the security architecture. Reachability under one global IP address is provided by Mobile IP. Nomadic mobility which only requires that the user can connect to the network anywhere does not require Mobile IP, but may still require a AAA infrastructure. Another mobility requirement is that of seamless handover. For real-time services, seamless handover may imply that there is no time to restart a full AAA exchange with every movement and to establish a new security association with the new point of attachment. Therefore, there is a need for a transfer of the security context from the old to the new point of attachment. Another useful distinction is that between macro- and micromobility. Whereas macromobility enables global roaming through Mobile IP and/or a AAA infrastructure, micromobility, in contrast, requires only mobility within a predefined local area. Micromobility may be used for optimal support of fast handover and minimisation of global signalling traffic.

Reference architecture

The work described here uses the simplified security reference architecture shown in figure 1. This architecture is compatible with the architecture for a post-3G mobile system developed in the IST project BRAIN which is used as a basis in the work of the SHAMAN project reported here. For a security architecture it is useful to distinguish between network domain security and user-related security.

Network domain security

Network domain security is based on static security associations between network entities. The security associations are not specific for individual users. Typically, network domain security is provided using IPsec tunnels between network entities. As the involved entities are powerful devices, and bandwidth restrictions are not a problem, key management can be performed using the standard Internet Key Exchange (IKE) protocol. In order to provide efficient key management for the protection of traffic between different networks, network operators have to agree on a public key infrastructure to provide the certificates required for IKE. For an example of a specification for network domain security see [1]



MN:	Mobile Node
AR:	Access Router
ANP:	Anchor Point, intercepting packets for MNs and tunneling them to the corresponding ARs
AAAL, AAAH:	AAA-server in the access network and the MN's home domain, respectively
HA:	Mobile IPv6 Home Agent
MG:	Mobility Gateway (Gateway from the access network to the public Internet)

Figure 1: security reference architecture

User-Related Security

User-related security requires the dynamic establishment and release of security associations based on user-specific keys. A roaming user has to perform a few fundamental steps in order to securely gain network access:

1) the MN needs to acquire an IP address
2) the MN and the visiting network mutually authenticate each other, and session keys have to be generated and distributed
3) the MN needs to register its care-of address at the home agent as part of the Mobile IP protocol (if applicable)
4) a security association between MN and AR needs to be established

5) Furthermore, when handover takes place a security context transfer needs to be carried out.

In the following, we will address these issues in turn:

Secure Address Configuration

Currently there exist two methods for an IPv6 node to automatically configure an IP-address while attaching to a network: *stateless* and *stateful* address autoconfiguration. In stateless address autoconfiguration the IPv6 node forms an IP-address by combining a prefix with topological significance and an interface identifier chosen by the node [2]. In stateful address autoconfiguration the IPv6 node obtains the IP address from a DHCP server along with other configuration information [3].

The mandatory duplicate address detection used with stateless autoconfiguration may be exploited for denial-of-service attacks, and the common use of the link layer address to derive the interface identifier may betray the user's identity. The latter threat can be avoided when choosing the interface identifier at random. Furthermore, duplicate address detection may cause latency problems, in particular during a handover because of the timeout values. But omitting duplicate address detection may be acceptable if the interface identifier is truly random, given its length of 64 bits which makes collisions extremely unlikely (see [4]).

Stateful autoconfiguration may also be susceptible to, among other things, denial-of-service attacks when an attacker tries to deplete the pool of IP addresses available at the DHCP server when no additional security measures are used. The security mechanism described in [3] is not applicable in a mobility scenario as it assumes the pre-existence of a security association between the DHCP server and the mobile node. Stateful autoconfiguration seems to be generally less efficient as it adds another roundtrip between mobile node and DHCP server to the overall access procedure. Attempts to integrate the stateful address configuration with other parts of the access procedure (e.g. with the Mobile IP registration procedure) seem to go against the principle of modularity of the security architecture, and no such integrated solutions currently exist. For this reason, a stateless address autoconfiguration with random selection of the interface identifier is the preferred solution in SHAMAN.

Authentication and Session Key Establishment

When a roaming user wants to attach to a visited network he never was in contact with before then the visited network needs to be assisted by some infrastructure to be able to authenticate the user, and, conversely, the user needs to be assisted when authenticating the visited network. The procedures by which a user can be authenticated vary a lot depending on the assumptions made on the system and the security mechanisms used. A fundamental distinction is made between symmetric (or secret key based) and asymmetric (or public key based) security mechanisms. Symmetric mechanisms typically require that the user has a permanent security association with an entity in his home network (i.e. with his AAA server) which is used to authenticate the user.

The required infrastructure in this case is provided by the AAA infrastructure which is used to transport authentication information among the AAA client (the access router AR or the anchor point ANP), the local AAA server (AAAL) and the home AAA server (AAAH) possibly through some AAA broker. The two main AAA protocols used for this purpose are RADIUS and DIAMETER whereby DIAMETER is preferred because of its modern design and offered extensions.

The AAA client exchanges authentication information with the mobile node. Currently no standardised protocol exists to carry authentication information on this part of a complete authentication exchange. However, the need for such a protocol has been recognised by the IETF, and the problem is being addressed by a newly founded working group named PANA (Protocol for Carrying Authentication for Network Access) [5]. The original idea underlying PANA seemed to be related to defining an EAP over IP (EAP = Extensible Authentication Protocol, cf. [6]), but the current work has a broader scope. It is important to notice that both, the PANA protocol and the DIAMETER base protocol, only specify how the authentication information is carried between the entities. But they do not define any specific cryptographic mechanisms; these have to be specified separately. One such mechanism is integrated in the proposal contained in [7]. Another mechanism is built into Mobile IPv4, cf. [8].

When using public key based security mechanisms then an on-line involvement of the home network need not be required as authentication and authorisation can be based on certificates. However, in this case a public key infrastructure is required, and additional entities such as certificate repositories or revocation servers may be required which may have to be accessed on-line.

Establishment of a Security Association between Mobile Node and Access Network

All these authentication mechanisms mentioned in the previous paragraph have to be suitable for the establishment of session keys together with authentication. (This rules out mechanisms like PAP and CHAP.) These session keys are later required to protect the confidentiality and integrity of any traffic between the mobile node and the access router (AR). However, in general it is not enough to have session keys available. What is required is the establishment of a security association. For IPsec, one option is to use existing protocols for the establishment of IPsec security associations, for example IKE or KINK [9], and, in the future, a son-of-IKE [10], [11], [12]. The use of these may, however, be a bit heavy-weight.

Security Context Transfer in Handover

The time constraints in handover may be such that a full AAA roundtrip for re-authentication and security association establishment between a mobile node and the new point of attachment may not be possible. In such a situation, only the transfer of the security context from the previous AR to the new AR may be appropriate. Protocols for such a context transfer are being defined in the IETF working group SEAMOBY [13]. We assume that an IPsec security association (for example ESP in tunnel mode) exists between the mobile node and an old AR, and that the MN is now handed over to a new AR. Then the old AR has to transfer all the attributes characterising the security association (SA) to the new AR. These include *static* attributes such as the IPSec Mode, authentication and encryption algorithm, key length, keys as well as selector fields that specify which traffic must be secured and the *dynamic* attributes like accumulated kilobytes relevant for the SA Lifetime, highest Sequence Number stored and flags which packets within the replay window have already been received. It is obvious that some SA attributes must be modified to address the fact that for example the IP address of the mobile node has changed through the handover. Other problems which may occur are: If the SPI value of the transmitted IPSec SA already exists at the new AR (SPI collision) then various alternative solution strategies are possible ranging from rejecting the transfer to re-negotiation of a new SPI value. In Section 7.3 of [14] an additional error case is mentioned that could occur during a context transfer. If the new AR does not support the algorithms of the transferred SA then the authors speak about a so-called SA Conflict. Our perception is that in a properly engineered network such error cases can be neglected and therefore no additional algorithm negotiations need to take place.

V. SECURITY FOR THE FIRST HOP

The first hop is the link between the mobile node and the first router packets encounter on the way into the network. The first hop may comprise both wireless and wired segments, e.g., in case wireless access points are linked by an Ethernet backbone to which the first hop router is attached. Security issues here are:

- at which layers to provide integrity and encryption?
- use or disable existing layer 2 mechanisms?
- what security context to use?

In a wireless access network the first hop contains link layer access to the underlying wireless service, e.g., GSM, UTRAN, HIPERLAN/2, Wireless LAN 802.11b or Bluetooth. Within this model the use of existing radio technologies does not imply that their authentication schemas apply. The purpose of this section is to develop a security architecture for the first hop within the SHAMAN context. The main problem is whether it is possible to handle the heterogeneous access procedures and security mechanisms of various wireless access systems within a unified framework.

A. First Hop Reference Model

The model for the first hop of an access network to be described is a further development of the wireless link in the BRAIN architecture, and serves basis for the development of the first hop security solutions.

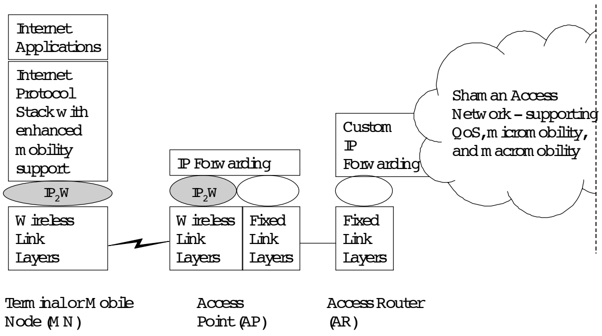


Figure V-1 – First hop reference model

The wireless link extends from the MN to the AP and the data packets encounter the AR only after passing through a fixed connection from AP to AR. The AP is the endpoint of the wireless link. It contains the wireless base station functionality, and the required link layer interface to communicate data and control information to the AR. This model allows the separation of network routing functionality from the radio access technology. In principle, the same general-purpose AR can serve IP packets originating from terminals using different radio technologies.

B. First Hop Interfaces

The BRAIN architecture [15] contains a model for IP to Wireless Convergence (IP_2W) and a specification of IP_2W interface. This interface is divided into two parts

- IP_2W Control Interface
- IP_2W Data Interface

For a wireless link layer technology to be IP_2W compatible, it is required in BRAIN that the functions performed at link layer can be divided into two categories according to these interfaces:

- Wireless Link Control Functions
- Basic Data Transport Functions

The first category contains all signalling functions, including security management, while second category contains functions related to the transport of user data.

The IP_2W provides a unified framework for controlling various capabilities of the wireless link. Also through it, useful information from link layer can be made available to the higher layers. This interface also allows for link layer specific optimisations of various IP specific features in a manner that is transparent to upper layers. In the BRAIN architecture, IP_2W interface shall be implemented both in the BAR and in the MN.

The first hop IP_2W interfaces are depicted in grey in Figure V-1. In case AP and AR are functionally in one unit, the fixed link can be omitted. The division of the IP_2W interface into control interface and data interface induces an analogous division to the interface in AP and to the interface in AR. For the purposes of defining first hop security functions, it is required that in AP link layer the control functions include functions for security management. IP$_2$W interface can be used to transport key information from higher layers to link layers to be used in protection the communication over the wireless link. On the network side the key material is first transferred to the AR, which then conveys it further to the appropriate AP to be used to protect the wireless communication between the AP and the MN. In order perform this security task the connection between the AP and AR needs to be protected.

The BRAIN architecture allows the wireless link to be protected using security mechanisms either on the link layer or the network layer or both. The security association used for deriving the appropriate keys resides on the network layer or higher.

In BRAIN, the IP$_2$W control interface offers means and appropriate commands for establishing keys and other necessary security parameters between the BAR and the MN on the link layer using the security association on the network layer.

C. How to Use Link Layer Security?

Since the goal is to provide access in heterogeneous access networks, it is necessary to make the architecture as modular as possible. The modules are connected using interfaces. For the security of first hop, as modelled in Figure V-1, basically two alternative approaches exist, the usefulness of which needs to be considered from the point of view of heterogeneous access.

1. The terminating points of the security mechanisms are identical to the endpoints of the first hop, that is MN and AR.
2. The terminating points of the security mechanisms are identical to the endpoints of the wireless link.

If the functionalities of AP and AR are geographically separated into two different network nodes, then these two approaches are different and mutually exclusive. Then the native wireless link encryption is not usable for approach 1. Instead, the security of the first hop must rely on Layer 3 security mechanisms.

In approach 2, native link layer mechanism can be used between MN and AP. The communication over the fixed link between AP and AR can be protected at either layer. The alternatives are summarized in Figure V-2.

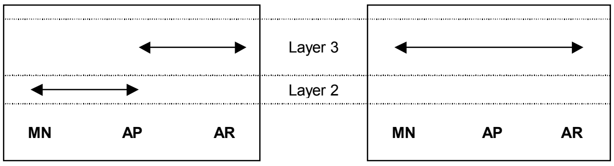


Figure V-2 Two approaches to the First Hop security

When considering the protection mechanisms distinction shall be made between at least the following types of communication:

- initial communication for authorized access,
- subsequent network layer signalling messages, such as IP package headers, QoS and resource allocation messages, etc., and
- communication payload (user data).

It is clear that the lower in the protocol stack the protection takes place the more communication can be protected. In particular, network layer signalling can be protected only by means of security mechanisms implemented at lower layers.

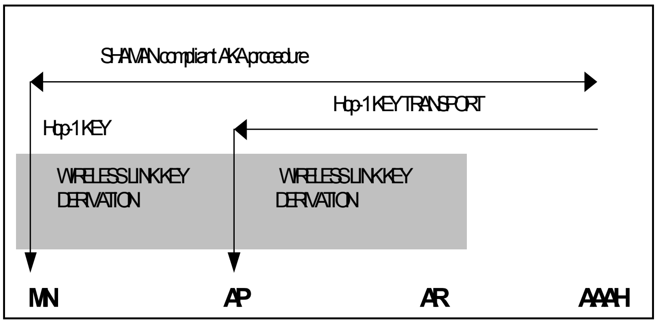
An important aspect to be considered is protection of the access network resources in wireless access. If the security context to protect the integrity of the wireless link is derived from the AAA procedures then the hijacking of the wireless link is protected. This principle is used in GSM and UMTS systems.

When discussing the *pro*s and *con*s of enabling wireless link layer security mechanisms, an undeniable *con* is the fact that the existing link layer security systems vary a lot with respect to what kind of protection is provided, but also with respect to the strength of security. For some wireless systems, the security mechanisms have been provided just for the sake of having some security as hardware cable replacement, while for true implementations, strong security solutions at network or application layer are recommended.

D. First Hop Security Associations

The security association for the first hop is derived in the authentication and key agreement procedure which is performed between the MN and AAAH of the home network of the mobile node.

The security association between the MN and the AAAH, the SHAMAN SA, makes it possible for the parties to establish shared secret key material, what we call SHAMAN First Hop (Hop-1) Key, for securing the first hop link between the MN and the AR. The SHAMAN Hop-1 Key is independent of the wireless link security technology. The keys specific to the used wireless security mechanisms are derived from the Hop-1 key. This functionality needs to be added on top of the wireless link layer stack in AP and MN. This procedure is depicted in FigureV- 3.


SHAMAN compliant AKA procedure
Hop-1 KEY TRANSPORT
Hop-1 KEY
WIRELESS LINK KEY DERIVATION
WIRELESS LINK KEY DERIVATION
MN
AP
AR
AAAH

FigureV- 3 Derivation of wireless link keys from Hop-1 key

For example, any SIM or USIM based authentication procedure can be used to derive the SHAMAN Hop-1 key. At the IP_2W convergence layer, the SHAMAN Hop-1 key can then be transformed to suitable keying material to be used to protect any wireless link using its native security mechanisms. In such a manner, keys can be derived for Bluetooth authentication (and subsequent encryption) as well as for air interface encryption, integrity and anonymity in UMTS.

E. Access Privacy

Privacy of the MN in initial access can be protected based on the security association between the access network and the home network of the MN. The identity of the MN is sent encrypted to AAAH. Such a protection method can be used when no previous security association exists between the MN and the access network. After the AAA procedures have been performed and the MN is authorised, a scheme for temporary identities is recommended for use between the MN and the AAAL.

Some link layer encryption protocols do not hide the MAC addresses in the link layer security protocols. A typical example is Bluetooth where the authentication is based on the true Bluetooth device address of the claimant. For the purposes of the anonymity protection of the MN a possible solution is that only unilateral link layer authentication between the MN and AP is performed, wherethe MN verifies the authenticity of the AP. If implemented in this way, the link layer security protocol does not reveal the identity of the MN. However, access for an unauthorised MN must be kept in restricted to prevent it from performing a denial of service attack by sending garbage data in place of encrypted identities.

REFERENCES

[1] 3GPP Technical Specification 33.210 V1.0.0, *IP network layer security (Release 5)*, December 2001, http://www.3gpp.org.

[2] S. Thomson, T. Narten, *IPv6 Stateless Address Autoconfiguration*, RFC 2462, December 1998.

[3] R. Droms (ed.), *Dynamic Host Configuration Protocol for IPv6 (DHCPv6)*, IETF draft, draft-ietf-dhc-dhcpv6-23.txt (work in progress), February 2002.

[4] M. Bagnulo et al., *Random generation of interface identifiers,* IEFT draft, draft-soto-mobileip-random-iids-00.txt (work in progress), January 2002.

[5] http://www.ietf.org/html.charters/pana-charter.html

[6] B. Aboba et. al, *Extensible Authentication Protocol (EAP)*, IETF draft, draft-ietf-pppext-rfc2284bis-01.txt (work in progress), Nov. 2001.

[7] P. Flykt, C. Perkins, T. Eklund, *AAA for IPv6 Network Access*, IETF draft, draft-perkins-aaav6-04.txt (work in progress), July 2001.

[8] Pat R. Calhoun, Charles E. Perkins, *Diameter Mobile IPv4 Application*, IETF draft, draft-ietf-aaa-diameter-mobileip-08.txt (work in progress), November 2001.

[9] M. Thomas (ed.), *Kerberized Internet Negotiation of Keys (KINK)*, IETF draft, draft-ietf-kink-kink-02.txt (work in progress), October 2001.

[10] D. Harkins, C. Kaufman, R. Perlman, The Internet Key Exchange (IKE) Protocol, IETF draft, draft-ietf-ipsec-ikev2-00.txt (work in progress), November 2001.

[11] W. Aiello et. al, *Just Fast Keying (JFK),* IETF draft, draft-ietf-ipsec-jfk-00.txt (work in progress), 2001.

[12] H. Krawczyk, *The IKE-SIGMA Protocol,* IETF draft, draft-krawczyk-ipsec-ike-sigma-00.txt (work in progress), November 2001.

[13] http://www.ietf.org/html.charters/seamoby-charter.html.

[14] Hamer, L-N.: *IPSec Context Transfer*, IETF draft, <draft-hk-seamoby-ct-ipsec-00.txt> (work in progress), May, 2001.

[15] *BRAIN architecture specifications and models, BRAIN functionality and protocol specification* - IST-BRAIN deliverable D2.2; March 2001.