Alternative Schemes for Dynamic Secure VPN Deployment in UMTS

Christos Xenakis and Lazaros Merakos

Communication Networks Laboratory

Department of Informatics & Telecommunications

University of Athens, 15784 Athens, Greece.

e-mail: {xenakis,merakos}@di.uoa.gr

ABSTRACT

Three alternative schemes for secure Virtual Private Network (VPN) deployment over the Universal Mobile Telecommunication System (UMTS) are proposed and analyzed. The proposed schemes enable a mobile node to voluntarily establish an IPsec-based secure channel to a private network. The alternative schemes differ in the location where the IPsec functionality is placed within the UMTS network architecture (mobile node, access network, and UMTS network border), depending on the employed security model, and whether data in transit are ever in clear-text, or available to be tapped by outsiders. The provided levels of privacy in the deployed VPN schemes, as well as the employed authentication models are examined. An analysis in terms of cost, complexity, and performance overhead that each method imposes to the underlying network architecture, as well as to the mobile devices is presented. The level of system reliability and scalability in granting security services is presented. The VPN management, usability, and trusted relations, as well as their behavior when a mobile user moves are analyzed. The use of special applications that require access to encapsulated data traffic is explored. Finally, an overall comparison of the proposed schemes from the security and operation point of view summarizes their relative performance.

KEYWORDS: Mobile Internet, UMTS, security, privacy, VPN, IPsec, IKE, NAT

INTRODUCTION

The mobile Internet, motivated by the continuous development of mobile technologies, the expansion of Internet services, the materialization of compact terminals, and the popularity of mobile data communications, creates new service paradigms. Wireless applications, such as e-business, e-government, e-finance, and e-health are emerging realizing the opportunities presented by the ubiquity of Internet and mobile devices. Moreover, seamless access to private networks by a mobile workforce is expected to drive the demand for anywhere - anytime access to corporate intranets, databases, and e-mail servers.

The Universal Mobile Telecommunication System (UMTS) [1] comprises a realization of mobile Internet and provides personal communication services. It intends to establish a single integrated system that supports a wide spectrum of operating environments. Users have seamless access to a wide range of new telecommunication services, such as high-speed Internet/Intranet applications, independently of their location.

Privacy and security are essential to the success of the new emerging applications over mobile systems. Mobile Internet users require flexible security mechanisms, which provide customized security services to data traffic, and are available anywhere – anytime. Confidentiality, integrity and authentication can be ensured by the deployment of Virtual Private Network (VPN) technology [2]. VPN authenticates and authorizes user access to corporate resources, establishes a secure tunnel, and encapsulates and protects data conveyance over a network. It extends dedicated connections between remote branches, or remote access to mobile users over a shared infrastructure. The advantages of using the transport facilities of a public network, combined with advances in the field of network security, make VPN services attractive compared to traditional private line services.

The most prominent technique for deploying VPN across IP networks, which guarantees interworking with any type of carried services, is the IPsec standard [3]. IPsec facilitates the authentication of the communicating entities, as well as the transparent encryption and integrity protection of the transmitted packets. It is especially useful for implementing VPNs, and for remote accessing private networks. However, mechanisms such as VPN and IPsec were originally conceived to address network security issues for fixed-point networks. Wired environment solutions can often be extended for applications to wireless environments, but they might need some changes or a complete rebuild. This is because of the limited bandwidth of the radio interface, as well as the limited processing, memory, and power resources of the majority of mobile devices. Moreover, mobility and private addressing might influence the tunnel

deployment and maintenance procedures. Therefore, it is critical to ensure that security services provided in wireline network are available in wireless environment too.

In this article, three alternative schemes for dynamic, client-initiated, secure VPN deployment over the UMTS network are proposed and analyzed. The mobile devices comprise the IP protocol stack including the TCP/UDP protocol, which enables the activation of any type of Internet service. The UMTS infrastructure provides to the mobile users access to the public Internet, and allows them to employ IPsec tunnel technique to traverse firewalls, access private networks, and convey sensitive data securely. This type of access is referred to as voluntary tunneling, since it enables a mobile node to establish a secure communication channel to a private network. The proposed schemes differ in the location where the IPsec functionality is placed within the UMTS network architecture (mobile node, access network, and UMTS network border), and whether data in transit are ever in clear-text, or available to be tapped by outsiders. The different security models are named as: a) the end-to-end, b) the network-wide, and c) the border-based.

The end-to-end security model [10] integrates VPN functionality into the communicating peers, which negotiate and apply security. Sensitive data traffic remains encrypted for the entire route between the sender and the receiver providing the best security services. For VPN establishment the Internet Key Exchange (IKE) [4] protocol is employed, which has to operate in a mobile UMTS environment where Network Address Translation (NAT) [5] is used. To overcome the incompatibilities occur from the coexistence of TCP/IP, IPsec, and NAT, the complementary UDP encapsulation is applied.

An alternative to the end-to-end approach pertains to a network-assisted security model [11, 12], which integrates VPN functionality into the network infrastructure. The network operator offers responsive, reliable, and flexible VPN services, thus, minimizing the administrative and the computational overheads for the end-user. By placing security functionality in the UMTS access network or in the UMTS border, the network-wide or the border-based VPN scheme are deployed respectively. In the network-wide scheme the deployed VPN is extended over the UMTS backbone and the pubic Internet, while in the border-based scheme it expands only on the public Internet segment. For VPN initialization and key agreement procedures an IKE protocol proxy scheme [12] is employed, which enables the mobile user to initiate a VPN, while outsourcing complex key negotiation to the network infrastructure.

Based on the security models analysis and their deployment attributes, the provided levels of privacy, as well as the employed authentication models are examined. An analysis in terms of cost, complexity, and performance overhead that each method imposes to the underlying network architecture, as well as to the

mobile devices is presented. The level of system reliability and scalability in granting security services is presented. The VPN management, usability and trusted relations, as well as their behavior when a mobile user moves are analyzed. The use of special applications that require access to encapsulated data traffic is explored. Finally, an overall comparison of the proposed schemes from the security and operation point of view summarizes their relative performance.

The rest of this paper is organized as follows. Section 2 introduces the security framework focusing on the UMTS network architecture, the current security solutions and the IPsec-based VPN technology. Section 3 presents the end-to-end security model. Section 4 describes the network-wide and the border-based VPN models, which are both based on the network-assisted deployment approach. Section 5 elaborates on critical features used for comparing the proposed alternative schemes. Finally, section 6 contains the conclusions.

SECURITY FRAMEWORK

UMTS Network Architecture

UMTS has been standardized in several releases, starting from Release 1999 (R99), and moving forward to Release 4 (Rel-4), Release 5 (Rel-5), Release 6 (Rel-6), supporting compatibility with the evolved Global System for Mobile communications (GSM) / General Packet Radio Services (GPRS) network [6]. The fundamental difference between the GSM/GPRS and the UMTS R99 is that the latter grants higher bit rates (up to 2Mbps) providing a wider variety of services. This is achieved through a new WCDMA (Wideband Code Division Multiple Access) radio technology for the land-based communications, named UMTS Terrestrial Radio Access Network (UTRAN). UTRAN consists of two distinct elements, Node B, and the Radio Network Controller (RNC). Fig. 1 depicts the UMTS R99 network architecture.

Consider a mobile subscriber using a mobile station (MS) and attempting to establish a secure remote connection to a corporate Local Area Network (LAN), and access a remote server through the UMTS infrastructure, as shown in Fig. 2. The security gateway (SG) that resides between the LAN and the public Internet functions as a proxy device providing security services to the private network nodes. It is assumed that the Internet and the UMTS backbone are based on IPv4. Both the Gateway GPRS Support Node (GGSN) and the SG use NAT.

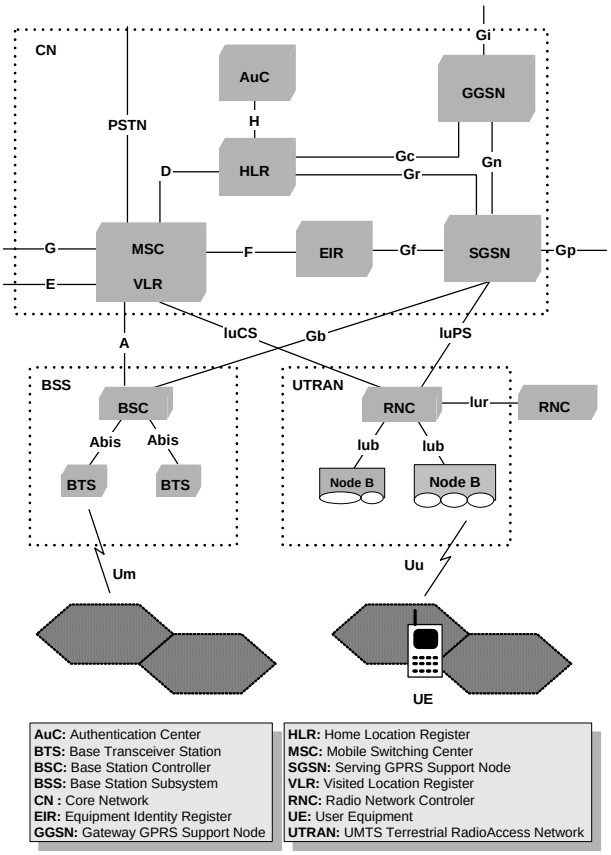

Figure 1: UMTS Release '99 system architecture

After power-on, the MS searches for a suitable cell in the UTRAN to provide services, and tunes to its control channel. Then, it performs the packet International Mobile Subscriber Identity (IMSI) attach procedure, which creates valid routing information for the packet switched (PS) connection in every node involved, and transferring the subscriber profile from the Home Location Register (HLR). When the IMSI has been attached, the MS initiates a Packet Data Protocol (PDP) context activation procedure, which negotiates the desired packet connection characteristics between the MS and the network [9]. The employed protocol for PS data transport in the UMTS R99 backbone network is the GPRS Tunneling Protocol (GTP) [7]. To be able to convey data packets from and to the MS, the Serving GPRS Support Node (SGSN) starts a radio access bearer (RAB) allocation procedure over the UTRAN, and a core network (CN) bearer is established between itself and the GGSN [9, 13].

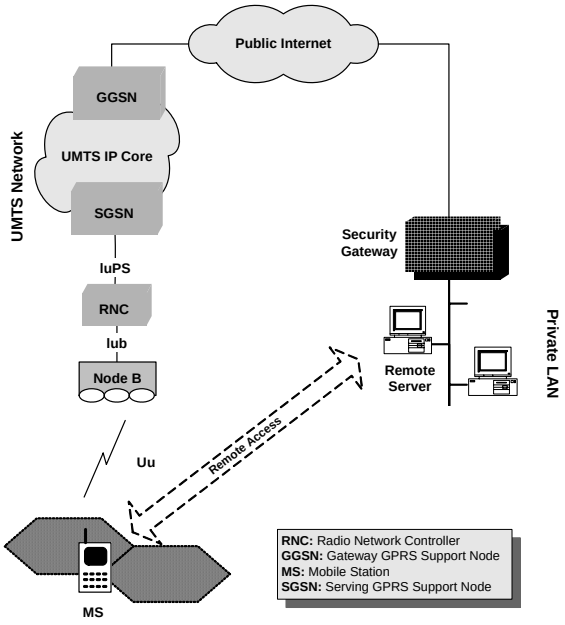

Figure 2: Network architecture

The result of these procedures is that, two types of bi-directional tunnels are set up: a) one tunnel between the MS and the RNC employing the Medium Access Control (MAC) [14] protocol over the WCDMA radio access interface, which also supports security protection; and b) one tunnel between the RNC and the GGSN employing the GTP without any security precaution. The latter consist of two parts: the Iu bearer over the Iu interface, and the PS domain backbone bearer between the SGSN and GGSN (see Fig. 3).

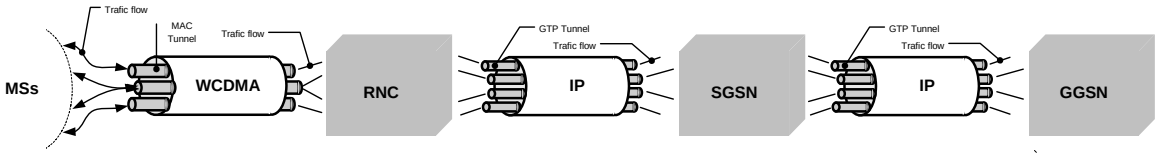

Figure 3: Schematic presentation of the UMTS tunnels

Despite the ciphering over the air interface, the IP traffic goes unencrypted all the way from the RNC to the corporate LAN SG, and vice-versa. Given that the GTP protocol operates over IP, and the UMTS is connected to the public Internet, the UMTS backbone may be considered as a vulnerable and easily accessible network segment. Firewall technology is inadequate against attacks that originate from malicious mobile subscribers, as well as from network operator personnel, or from any other third-party who gets

access to the UMTS core network [23]. Moreover, the current static VPN scheme supported by the UMTS involves the predefined establishment of security associations between the UMTS border and remote sites, failing to provide the necessary flexibility required by typical mobile users and ad hoc services [23].

Wireless security

In this section, a brief overview of the current solutions dealing with the security of wireless networks and applications are presented.

The Secure Sockets Layer protocol (SSL) is the default Internet security protocol [21]. It provides point-to-point security by establishing a secure channel on top of TCP where it supports server authentication using certificates, confidentiality, and message integrity. "KiloByte" SSL (KSSL) is an SSL client for the Mobile Information Device Profile of Java 2 Micro Edition platform (J2ME) [21]. This SSL implementation on J2ME devices (KSSL) provides an advantage by enabling these devices to communicate directly and securely with the huge number of Internet web servers supporting SSL. The main concept behind KSSL is represented in reusing previous session results such as certificate parsing results and master secrets, so as to avoid repeated SSL handshakes. This helps in avoiding complex, resource-intensive operations on the client device.

Wireless Application Protocol (WAP) is a suite of standards for delivery and presentation of Internet services on wireless terminals, taking into account the limited bandwidth of mobile networks, as well as the limited processing capabilities of mobile devices. To connect the wireless domain to the Internet, a WAP gateway is needed to translate the protocols used in WAP segment to the protocols used in the public Internet. The WAP architecture has been standardized in two releases (ver. 1.2.1 and ver. 2.0) [25].

To secure data transmission in the WAP architecture (ver. 1.2.1), the Wireless Transport Layer Security (WTLS) protocol [21, 25], which is based upon the Transport Layer Security (TLS) protocol, is employed. WTLS has been optimized for use over narrow-band communication channels providing also datagram support. It ensures data integrity, privacy, authentication, and denial-of-service protection.

WAP 2.0 proceeds to the re-design of the WAP architecture by introducing the existing Internet protocol stack, including the TCP, into the WAP environment. The new architecture allows a range of different gateways, which enables conversion between the two protocol stacks anywhere from the top to the bottom of the stack. A TCP-level gateway allows for two versions of TCP, one for the wired and another for the wireless network, on top of which a secure TLS channel can be established all the way from the mobile

device to the server. The availability of a wireless profile of the TLS protocol, which includes cipher suites, certificate formats, signing algorithms, and the use of session resume, enables end-to-end security support at the transport level allowing interoperability for secure transactions.

SPECSA [27] is a security architecture for wireless enterprise applications, which provides authentication, data confidentiality and integrity security services. It is based on a configurable security policy that controls security-related attributes such as the encryption algorithm, the hashing algorithm, the authentication mode, and the lifetime of the session keys and the user password. SPECSA was designed in a platform-neutral manner and can be implemented on a wide range of wireless clients.

Tiny SESAME [28] is a lightweight implementation based on the Secure European System for Applications in a Multi-vendor Environment (SESAME) architecture. SESAME is designed for operation in distributed systems where it provides access control, authentication, and data confidentiality and integrity. It supports the Kerberos authentication mechanism and extends it with additional services such as asymmetric cryptography based on public key technology, and access control and authorization certificates.

However, Tiny SESAME lightweightness is achieved through the employment of a dynamically reconfigurable component based architecture where resources can be loaded dynamically at runtime. This dynamic resource loading, although it helps in reducing the memory requirements of the application, increases network traffic, and raises significant security risks on low-end wireless platforms that lack the standard security verification and access control mechanisms for controlling the operation of dynamically loaded resources.

Application layer solutions, such as SSL and WAP security can be used to secure the communication of any application, but they must be integrated into the application, and, thus, to a large extent they are used for web-based applications. Moreover, for every new session between the communicating peers, a new security association needs to be established [21, 23].

SPECSA, Tiny SESAME and other [29] security architectures provide standard Application Programming Interfaces to allow application developers to utilize their security services. Hence, these security architectures cannot be applied to any type of application and ad hoc use. Furthermore, the introduction of specialized security modules required in mobile devices and remote servers minimizes the interoperability with the existing fixed network infrastructure.

Moving the encryption function from the application layer to the network layer removes the dependency on end applications. Network security protects traffic on a connection basis between specific source and

destination nodes or subnetworks. Encryption at the network layer has the advantage of operating transparently from the end user's perspective. This allows flexibility in the implementation of security policies within an organization, and enables subnetworks to be logically and securely separated via security devices. Additionally, facilitates mobile users to access securely remote corporate resources [30].

IPsec-based VPN Technology

It is commonly admitted that IPsec is the best security protocol available today. It aims at securing the network layer, and guarantees security for any application that uses it. It facilitates authentication of the communicating peers, and transparent encryption and integrity protection.

The IPsec works in two modes, transport and tunnel mode. Transport mode is typically used in peer-to-peer communications as only the payload of the packet is encrypted, not the IP header. Tunnel mode is used for site-to-site security given that the entire packet (header and payload) is encrypted. IPsec also grants two choices of security service, Authentication Header (AH), and Encapsulation Security Payload (ESP). AH provides support for connectionless integrity, data origin authentication, and protection against replays, but does not provide secrecy. On the other hand, ESP supports confidentiality, connectionless integrity, anti-replay protection, and optional data origin authentication [3].

A key concept that appears in both security services is the Security Association (SA) [3]. An SA is a one-way relationship between a sender and a receiver that affords security services. In order to establish an SA between two hosts, they must first agree to apply compatible policy and cryptographic algorithms. They must also share a secure mechanism for determining keying material over an insecure channel. The default IPsec method for secure key negotiation is the IKE [4] protocol. IKE consists of two sequential phases. Phase 1 creates an Internet Security Association and Key Management Protocol (ISAKMP) SA (or IKE SA) that establishes a bi-directional secure channel between the security endpoints. Phase 2 negotiates an IPsec SA using the pre-established secure channel. Multiple IPsec SAs can be established from a single ISAKMP SA, which may be considered as a "control channel" where IKE is the control protocol.

IPsec is especially useful for implementing VPNs and for remote access to private networks. Concerning VPN deployment there are two general approaches. The first is based on Customer Premises Equipment (CPE) approach, where the VPN capabilities are integrated into CPE devices. The second scheme pertains to network-assisted, where the security functionality and the VPN operation are outsourced to the

network operator, or a service provider. There is significant interest in such solutions both by customers seeking to reduce support costs, and by network operators seeking new revenue sources.

A principal issue that has to be considered in the IPsec-based VPN is the use of NAT. NAT maps an isolated address realm with private unregistered addresses to an external realm with globally unique registered addresses. The conjunction of NAT with IPsec arises many incompatibilities, listed in [5, 12], since the latter either hides private addresses through encryption and thus let them escape translation, or it experiences integrity violations as a consequence of NAT manipulation of protected IP addresses. A promising solution to the IPsec/NAT traversal problem based on the encapsulation of the IPsec-protected packets into UDP or TCP packets.

The NAT-Traversal (NAT-T) [8] specification, which is supported by the IETF and is in the final stage to become a standard, defines methods to encapsulate and decapsulate IPsec packets inside UDP packets. The UDP port numbers used for this functionality are the same as those used by the IKE traffic, so new holes do not need to be opened in the existing corporate security policy. Wrapping IPsec-secured packets into UDP packets allows modification of both the IP address and the port number, without affecting the secure functionality of IPsec. It is worth noting that the UDP checksum in the UDP-encapsulated ESP header, the floated IKE header, and the NAT-keepalive header should be transmitted as a zero value.

IPsec over TCP is a proprietary solution followed by specific vendors. It enables a VPN client to operate in an environment in which ESP or IKE cannot function, or can function only with modification of existing firewall rules. It encapsulates both the IKE and IPsec protocols within a TCP packet, and enables secure tunneling through NAT devices and firewalls. A functional advantage of using TCP encapsulation instead of NAT-T is that the port for the IPsec connection can be defined by the client, while the VPN terminating device has been configured to listen on that TCP port. Moreover, many network firewalls are configured to block all UDP traffic.

Although both NAT traversal solutions can be applied in mobile scenarios, in the proposed security models the UDP encapsulation is being selected, since it is in the final stage of standardization, and all vendors adopt it.

In the sequel, based on the VPN deployment approaches, three different security models for dynamic, on demand, IPsec-based VPN deployment over the UMTS network, are proposed and analyzed. These schemes, which place the security endpoints at different levels within the mobile network infrastructure, make feasible the realization of secure mobile Internet.

END-TO-END DEPLOYMENT SCHEME

Based on the principles of the CPE approach, the end-to-end [10] security model is implemented. The communicating endpoints (MS and SG) establish a pair IPsec SAs between them, which are extended over the entire multi-nature communication path, as shown in Fig. 4. Sensitive data are secured as they leave the originator site (MS or SG), and remain protected while they are conveyed over the radio interface, the UMTS backbone network, and the public Internet, eliminating the possibilities of being intercepted, or to be altered by anyone.

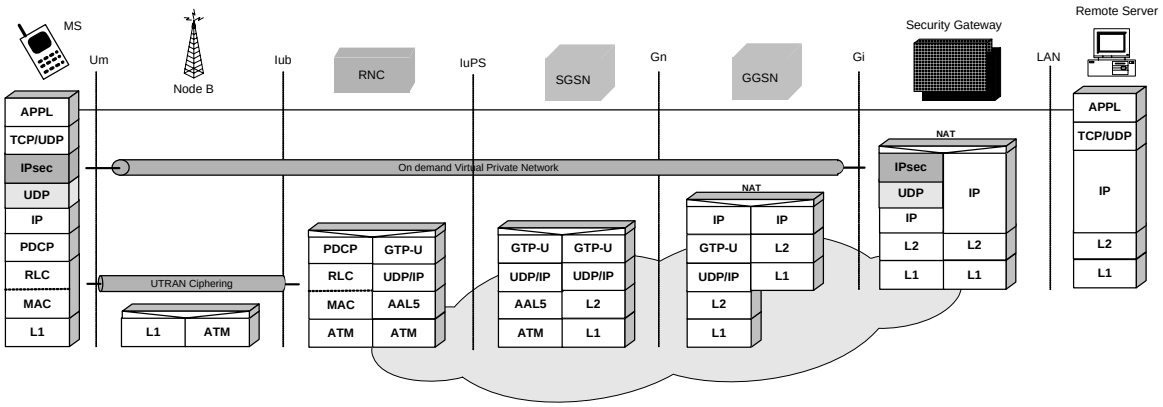

Figure 4: End-to-end VPN deployment scheme over UMTS

VPN establishment

For the end-to-end VPN establishment the IKE [4] protocol is employed. However, its standard version must be enhanced to resolve the problems arising from the NAT presence, and configured to operate in a mobile environment. IKE provides secure key determination via Diffie-Hellman (DH) exchanges [20] with authentication of participants, protection against reply, hijacking, flooding attacks, and negotiation of encryption and/or authentication transforms. The security endpoints exchange DH half-keys (X and Y) to arrive at a mutual session key, k. The key is at least as strong as the strongest half-key, and, thus, neither of the security endpoints can sabotage it. For the reader's convenience, Table 1 gives the notations and definitions used in the analysis that follows.

During IKE phase 1, an ISAKMP SA negotiation in aggressive mode (AM) and a NAT presence detection along path take place. The AM of the IKE key negotiation is an option defined to speed up the IKE transaction at a cost of slightly less security. Moreover, the authentication method used in AM doesn't involve the IP address of the initiator. Thus, it facilitates the IKE deployment in the UMTS network where dynamic (not static) IP addresses may be used. The authentication of endpoints is based on digital signatures,

such as those provided by RSA [26], which use the public key/private key pair technique. In order to prevent "man in the middle" attacks, both MS and SG must authenticate themselves to one another. This is performed by adding an exchange of digitally signed authentication information. Hence, even if an intermediate is able to intercept or read the messages exchanged, it will not be able to forge the signatures.

Symbol	Description
C_{MS} , C_{SG}	Cookies
$HASH_{MS}$, $HASH_{SG}$	Authentication information
ID_{MS} , ID_{SG}	Identification data
ISA_{MS} , ISA_{SG}	ISAKMP security association request - proposal
K	Mutual session key
M_{ID}	Message identifier
N_{MS} , N_{SG}	Nonce : a large random number between 64 – 2048 bits that adds randomness
$NAT\text{-}OA_{MS}$, $NAT\text{-}OA_{SG}$	NAT original address
$NAT\text{-}D_{MS}$, $NAT\text{-}D_{SG}$	NAT discovery payload
$[p, g]$	Diffie-Hellman group
$PRVKEY_{MS}$, $PRVKEY_{SG}$	Private key
SA_{MS} , SA_{SG}	Security association request - proposal
SIG_{MS} , SIG_{SG}	Digital signature of the authentication information
$SKEYID$	Authentication key
X, Y	Diffie-Hellman half-Keys

Table 1: Notations definition

The NAT presence detection between the security endpoints reveals whether the IP address, or the related IP port of the transmitted packets is changed along the path. It is performed by sending the hashed values of the IP address and the IP port of each end to the other end. When the hosts calculate those values and get the same result, they know there is no NAT between them. Otherwise, NAT occurs between the security endpoint, and, therefore, a NAT-traversal technique is required to get the IPsec-protected packets [12, 15].

To initiate the IPsec SA negotiation (see Fig. 5), the MS first generates a cookie (C_{MS}) (64-bit random number which facilitates prevention of flooding attacks). Then, the MS chooses a prime number, p, and an integer, g, (referred as DH group), it generates a large random integer, x, and it computes, $X = g^x \bmod p$. In message (1) the MS forwards the C_{MS}, the DH half-key (X) including the DH group ([g,p]), a nonce (N_{MS}) (a large random number between 64 - 2048 bits that adds randomness), the ISAKMP SA data (ISA_{MS}), and the Identification Data (ID_{MS}) to the SG. The ID_{MS} field contains a certificate of the mobile user, which

uniquely identifies him. The ISA_{MS} field includes a series of protection mechanisms and algorithms (e.g., encryption, hash function, etc.) proposed for the ISAKMP SA.

Upon receipt of message (1), the SG validates it. Then, the SG generates a cookie pair (C_{SG}) and a large random integer, y, and it computes, $Y = g^y \bmod p$, as well as the session key resulting from the DH exchange, $k = X^y \bmod p$. The SG replies with message (2), which contains the cookies, its ISAKMP SA response (ISA_{SG}), the DH half-key (Y), a nonce (N_{SG}), its certificate (ID_{SG}), the NAT discovery ($NAT\text{-}D_{SG}$) payload, its authentication information ($HASH_{SG}$), and the digital signature of the authentication information (SIG_{SG}). The ISA_{SG} payload contains the SG response to the security proposal made by the MS in message (1). The $HASH_{SG}$ field used for authentication is computed using the $SKEYID_a$ and the negotiated hash algorithm.

$$HASH_{SG} = hashfunc(SKEYID_a , Y |^1 X | C_{SG} | C_{MS} | ISA_{MS} |ID_{SG})$$

SKEYIDa is a key derived from SKEYID and is used as an authentication key. SKEYID is derived differently for each authentication method. Using the digital signature authentication method the SKEYID is computed as follows:

$$SKEYID=hashfunc(N_{MS} |N_{SG} , k),$$

$$SKEYID_a=hashfunc(SKEYID, SKEYID_d |k |C_{MS} |C_{SG} |1)$$

$SKEYID_d$, which is used to derive more keying material, is computed as follows:

$$SKEYID_d=hashfunc(SKEYID, k |C_{MS} |C_{SG} |0),$$

The $NAT\text{-}D_{SG}$ payload includes the hashed values of the IP address and the IP port of both security peers. The first field contains the remote end hash, and the rest contains the local end hash. The hash is calculated as follows:

$$HASH = hashfunc (C_{MS} | C_{SG} | IP | Port)$$

The C_{MS} and C_{SG} are included in the hash to make pre-computation attacks for the IP address and IP port impossible [15]. The SG digitally signs its authentication information using its private key ($PRVKEY_{SG}$) in order to defeat the possibility of man in the middle attack.

$$SIG_{SG}= PRVKEY_{SG} (HASH_{SG}),$$

[1] String concatenation

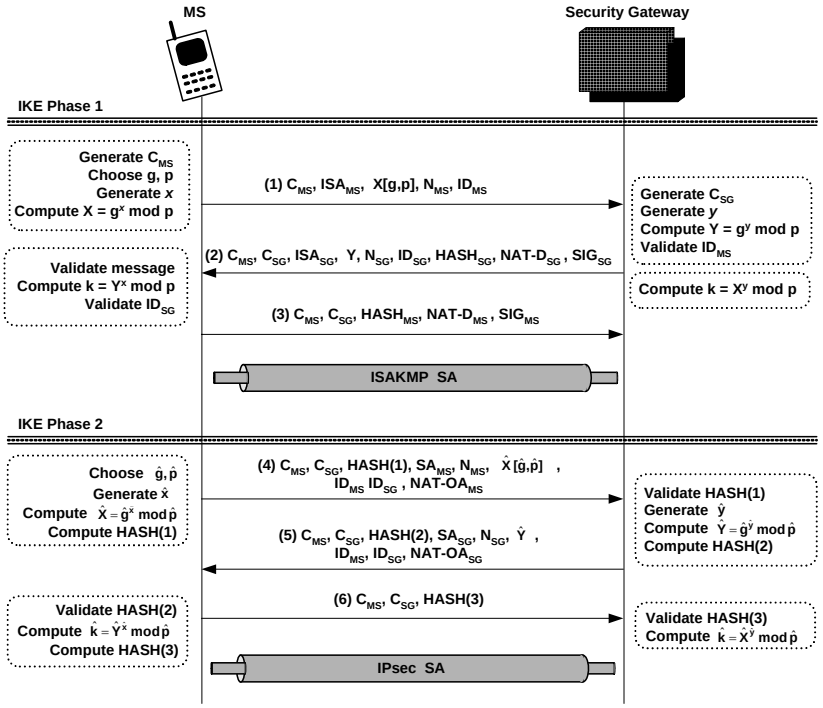

Figure 5: ISAKMP and IPsec SA negotiation

The MS use the SG public key to validate and verify the message. The SG public key is integrated in the certificate, which is included in message (2). Then, the MS computes the DH key, $k = Y^x \bmod p$. Finally, with message (3), the MS transmits its authentication information ($HASH_{MS}$), the digital signature (SIG_{MS}), and the NAT-D_{MS} payload to the SG along with the cookie pair. The $HASH_{MS}$ and the SIG_{MS} are computed as follows:

$$HASH_{MS} = hashfunc(SKEYID_a\ ,\ X \mid Y \mid C_{MS} \mid C_{SG} \mid ISA_{MS} \mid ID_{MS}\ )$$

$$SIG_{MS} = PRVKEY_{MS}\ (HASH_{MS}),$$

where $PRVKEY_{MS}$ is the mobile user private key.

Having established an ISAKMP SA, the communicating parties know whether a NAT device mediates between them, and have agreed on the following security attributes [14]: i) the encryption algorithm, ii) the hash algorithm for signing, iii) the authentication method for signing, and iv) the Diffie-Hellman exchange.

Following the successful completion of phase 1, the IKE phase 2 is performed to establish an IPsec SA, and negotiate a NAT-traversal technique. The latter requires the incorporation of the NAT Original Address (NAT-OA) field in the first two packets exchanged by the security peers [8]. Specifically, the MS includes the $NAT\text{-}OA_{MS}$ payload when proposing UDP encapsulation, and the remote SG replies with its $NAT\text{-}OA_{SG}$ payload if it agrees. The format of the NAT-OA field is presented in [15]. Since the IKE phase 2 is used to

derive new keying material, a new DH exchange occurs. All packets pertaining to phase 2 are encrypted using the pre-established ISAKMP SA.

First the MS chooses a new DH group ($\hat{g}$, $\hat{p}$), generates $\hat{x}$, and computes, $\hat{X} = \hat{g}^{\hat{x}} \bmod \hat{p}$ (see Fig. 5). Then, it transmits the cookies (C_{MS} , C_{SG}), the IPsec SA request (SA_{MS}), its nonce (N_{MS}), the DH half-key ($\hat{X}[\hat{g}, \hat{p}]$), the identities of the security endpoints (ID_{MS} , ID_{SG}), and the NAT-OA_{MS} to the SG (message 4). Since all negotiations in phase 2 use the cookie pair that was established during phase 1, each negotiation must be assigned a unique identifier, so that it can be distinguished. This is accomplished through the use of a message identifier, M_{ID}, which is part of the generic ISAKMP header that is included in all IKE packets. The SA_{MS} payload might contain one or more security proposals for negotiation. Moreover, the MS authenticates the message with HASH(1), which is computed as follows:

$$\text{HASH(1)}=hashfunc(SKEYID_a , M_{ID} | SA_{MS} | N_{MS} | \hat{X} | ID_{MS} | ID_{SG})$$

Upon receipt of message (4), the SG validates it, generates $\hat{y}$, and computes, $\hat{Y} = \hat{g}^{\hat{y}} \bmod \hat{p}$. Then, the SG forwards message (5) to the MS, which contains the cookies, its IPsec SA response (SA_{SG}), its nonce, the DH half-key ($\hat{Y}$), the (MS & SG) identities, and the NAT-OA_{SG}. The SA_{SG} payload includes the SG response to the security proposal made by the MS in message (4). The SG also authenticates the message with HASH(2), which is computed as follows:

$$\text{HASH(2)}=hashfunc(SKEYID_a , M_{ID} | SA_{SG} | N_{SG} | \hat{Y} | ID_{MS} | ID_{SG})$$

Finalizing this dialog the MS (message 6) replies with the cookie pair, and authenticates the transaction with HASH(3), which is computed as:

$$\text{HASH(3)}=hashfunc(SKEYID_a , 0 | M_{ID} | N_{MS} | N_{SG})$$

Both security endpoints are able to compute the DH session key, $\hat{k}$, ($\hat{k} = \hat{X}^{\hat{y}} \bmod \hat{p}$, $\hat{k} = \hat{Y}^{\hat{x}} \bmod \hat{p}$). Since an IPsec SA is used only in one direction, for bi-directional communications between the MS and the SG, two SAs are required.

NAT Traversal

Although the coexistence of NAT and IPsec is quite troublesome, both mechanisms can be configured to co-operate in the particular scenario for end-to-end VPN deployment. Specifically, there are two points (GGSN and SG) where NAT is applied. In the SG at the private network, both IPsec and NAT functionality are

combined in the same device entity. By placing the IPsec endpoint in the public address space, the incompatibility problems arising from their coexistence can be avoided. On the other hand, the NAT at the GGSN takes place between the VPN termination points (MS and SG), and, therefore, the incompatibilities presented in [5, 12] should be resolved.

ESP protocol is proposed for security services, given that it provides confidentiality and integrity protection as well. Unlike AH protocol, the ESP creates a message digest for packet authentication excluding the IP header, and, thus, allows NAT to modify the protected IP packets header without experiencing an IPsec integrity failure. However, the most prominent incompatibility issue that has to be considered in this scenario derives from the coexistence of TCP with NAT. As mentioned previously, a promising solution to this inconsistency lies on the use of UDP encapsulation. Wrapping the IPsec-protected packets inside a UDP/IP header leaves NAT modifications without acting on the encapsulated packet. The receiver is allowed to discard the UDP header, disregarding also the NAT changes [8]. The only requisite is that both IPsec peers have to support UDP encapsulation/decapsulation functionality.

Finally, concerning the incompatibility between the IKE address identifiers and NAT, the proposed VPN scheme employs the IKE in aggressive mode, which uses identification data instead of IP addresses for end-node authentication. The same authentication method should also be used during the IPsec SA negotiation.

Mobility Implications

Having established a pair of IPsec SA between the MS and the SG, a bi-directional private channel that allows for the secure data exchange between these two nodes is being set up. The MS may send and receive IP packets securely to and from a remote server, connected to the private LAN, through the UMTS network and the public Internet. The mobile subscriber may also freely move within the UMTS coverage area maintaining network connectivity and VPN service provision. The UMTS mobility management procedures keep track of the user location, and, therefore, the incoming packets are routed to the MS. Since the deployed end-to-end VPN model operates above the network layer (see Fig. 4), the security parameters, which are contained within the IPsec SA, are not affected by the MS movement.

NETWORK-ASSISTED APPROACH

Pursuing the network-assisted approach, the VPN functionality is integrated into the UMTS network infrastructure. The network operators provide the security aggregation facilities, which are shared amongst

the network subscribers, as a complementary service granting added value. They have solid network management expertise, and more resources to effectively create, deploy, and manage VPN services originating from mobile subscribers. However, the proposed security model involves two separate security domains (i.e., mobile subscriber – mobile operator, and mobile operator – remote site), and, thus, requires an explicit level of trusted relationships between them, which is analyzed below.

Security Infrastructure

For the deployment of a network-assisted security scheme, the MS must be enhanced with a security client (SecC), which is used to request for VPN services and express the end-user preferences. Moreover, a fixed UMTS node should incorporate a security server (SecS) that establishes, controls, and manages VPNs between itself and remote SGs at corporate LANs on behalf of the mobile users (see Fig. 6).

SecS comprises an IPsec implementation modified to adapt to the client-initiated VPN scheme, and the security service provision in a mobile UMTS environment. The main functional component of the SecS is the security manager, which manages the SecS submodules, and facilitates the VPNs configuration. The security manager maintains the security policy databases, handles the user requests, and reports on errors.

IKE authenticates IPsec peers, negotiates security services, and generates shared keys dynamically. It provides secure key determination via DH exchanges.

The policy manager contains the network security policy that specifies the set of users that are allowed to have security services, as well as the type of the offered services. It communicates with the HLR in order to acquire the users profile. The policy manager contents are used to configure the security policy database (SPD) and the security association database (SADB).

SPD is the primary policy database used by the SecS to decide on network traffic handling, such as encryption, decryption, authentication, discarding, passing through and modification. SPD contains an ordered list of policy entries, each of which defines the set of IP traffic encompassed by this policy entry, and is keyed by one or more selectors [3].

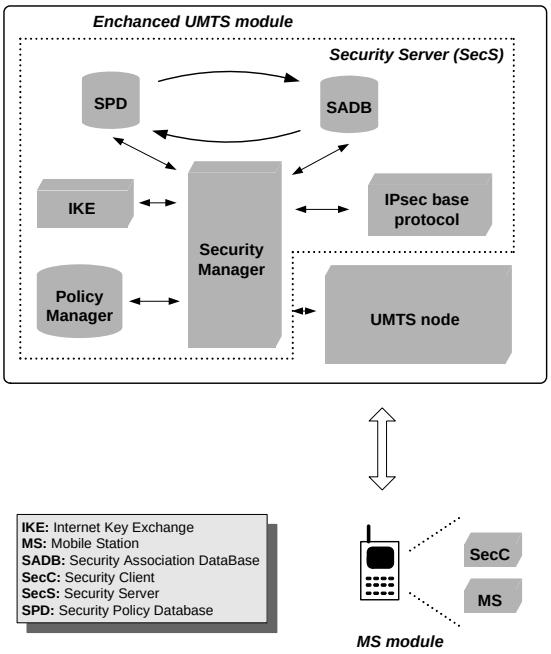

Figure 6: Security Client (SecC) and Security Server (SecS) modules

SADB maintains the contents of all active SAs used by the SecS for IPsec formatting. An SA is a management feature used to enforce a security policy. It represents all the necessary parameters (including protocols, modes, algorithms, etc.) that have been agreed between the IPsec peers. The security manager is responsible for filling out the contents of each entry in SADB.

Finally, the IPsec base protocol processes the authentication and encryption transformation defined in the IPsec framework. It handles all the network layer functions, such as fragmentation and path maximum transfer unit, and ensures that all traffic passing through the UMTS node is secure and authorized, providing firewall capabilities.

Network-wide Deployment Scheme

The network-wide scheme pertains to the network-assisted security model where the SecS is integrated into the RNC node of the UMTS infrastructure. It provides maximal security services to the communicating peers by employing the existing UTRAN ciphering over the radio interface, and extending a VPN over the Iu interface, the UMTS backbone, and the public Internet. Thus, sensitive user data remains encrypted in the entire network route between the originator and the recipient (see Fig. 7).

VPN Establishment - Deployment

VPN initialization and key agreement procedures are based on an IKE-proxy scheme [12], which enables the MS to initiate VPN establishment, while outsourcing key negotiation to the network infrastructure. When a mobile user wants to establish a secure remote connection towards a SG, it uses the SecC to request for an IPsec SA from the corporate SecS.

To initiate an IPsec SA negotiation, the SecC forwards a message destined for the SecS that includes the IP address (IP_{SG}) of the remote SG, the IPsec SA request (SA_{MS}), and the Identification Data (ID_{MS}) of the mobile subscriber. Upon receiving the request, the SecS verifies the mobile subscriber privileges and the mobile network capabilities in providing VPN services by asking the policy manager. Additionally, looks for an already active ISAKMP SA between the SecS and the SG on behalf of the particular user. If such an SA exists, then, the SecS proceeds to phase 2. If not, the SecS negotiates an IPsec SA using the IKE protocol, similarly to the end-to-end scenario. The only difference is that the SecS adds its certificate in the mobile user identification data sent to the SG. The SecS and the remote SG authenticate each other using digital signatures. Moreover, the SecS exchanges an additional message with the SGSN to inform it regarding the established VPN parameters.

After the VPN establishment between the SecS and the corporate SG (see Fig. 8), the MS may communicate with the remote server securely. Before transmitted over the radio interface, the IP packets, which are destined for the remote server, are processed by the Packet Data Convergence Protocol (PDCP) [16], the Radio Link Control (RLC) [17] protocol, and the MAC sublayer. The PDCP compresses the headers of the payload protocol in order to be carried over the WCDMA radio interface. The RLC performs link layer functions, such as segmentation and reassembly, flow control, sequence number check, ciphering, etc. The UTRAN ciphering is performed either in the RLC, if the radio bearer uses the non-transparent RLC mode, or in the MAC, if the radio bearer uses the transparent RLC mode.

The MAC layer multiplexes protocol data units from higher layer into transport block sets carried over common transport channels. Protected data packets (by UMTS ciphering) are tunneled, and forwarded to the serving RNC using the established RAB. The identification of MSs on the common transport channel is performed by a temporary identity, which can be either the Cell Radio Network Temporary Identities (C-RNTI), or the UTRAN Radio Network Temporary Identity (U-RNTI). The RAB ID identifies uniquely the RAB for the particular MS and the specific CN domain. It includes a binary representation of the Network

Service Access Point Identifier (NSAPI), which binds data streams from the access stratum and the non-access stratum [7, 18].

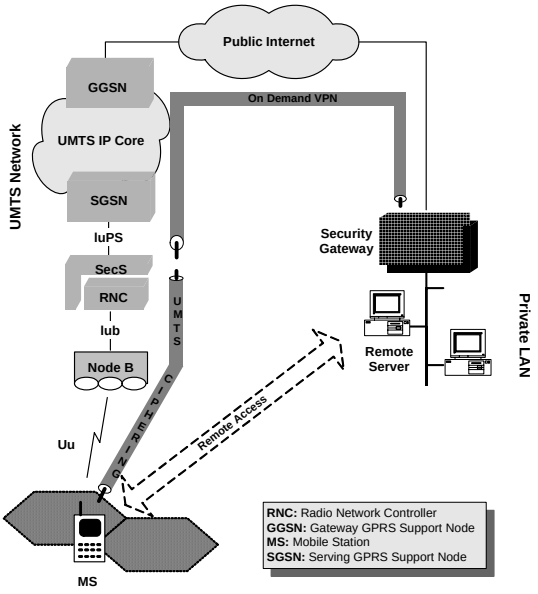

Figure 7: Network-wide VPN deployment scheme

In the RNC, the corresponding protocols terminate the MAC tunneling, decipher the packets, and acquire the active PDP context based on the RAB ID, the RNTI and the NSAPI values. A GTP tunnel, which is identified by the Tunnel Endpoint Identifier (TEID) and the IP address of the involved SGSN, corresponds to each PDP context [18]. Because of the presence of the SecS, every packet that is going through the RNC is subject to processing by the IPsec base protocol, which determines whether it will apply IPsec protection or not. For the specific VPN scheme, the default set of selectors [3], that facilitates the interaction with the SPD, should be enhanced comprising also UMTS routing parameters, such as the NSAPI, the TEID and the involved SGSN IP address.

In case that IPsec processing is to be applied, the original IP packets are encrypted and authenticated. The IPsec is configured in transport avoiding multiple encapsulations within the UMTS backbone. After the IPsec application, the protected data packets are wrapped within UDP/IP headers for NAT traversal. The wrapped packets are then tunneled between the RNC and the GGSN using the GTP protocol. The TEID, which is presented in the GTP header, indicates which tunnel a particular protocol data unit belongs to.

The GTP tunnel is ended in the GGSN, which removes the GTP header, and applies NAT on the encrypted IP packets. The packets are forwarded to the public Internet, and the latter delivers them to the SG

located at the private LAN. Upon receiving protected data packets, the SG discards the UDP header, terminates the IPsec tunnel, decrypts the packets, and forwards them to the inner LAN destination. Because of NAT application, the SG changes the destination address in the IP header. However, the NAT employment within the SG does not have any impact on the IPsec operation, since the IPsec tunnel has been terminated. In Fig. 9, the protected data flow from the MS to the corporate remote server over the UMTS network and the public Internet is presented.

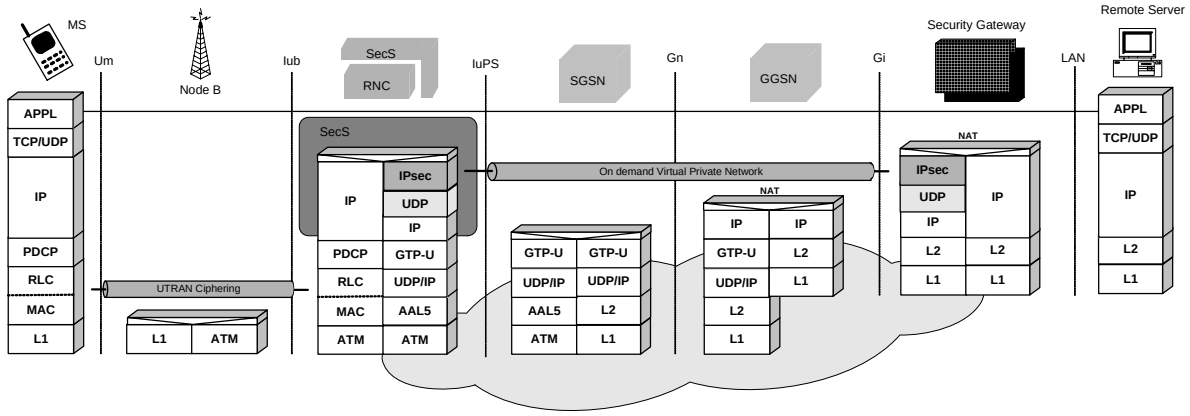

Figure 8: Protocol stack for network-wide VPN scenario

Whenever the remote server sends IP datagrams to the MS, the SG receives these packets, changes their source IP address (NAT), maps them to the appropriate SA, and, then, wraps the encrypted packets with UDP/IP header for NAT traversal. The encrypted packets are forwarded through the Internet, and are routed to the GGSN. The latter inquires the HLR to obtain the MS current location, and tunnels the packet through the UMTS backbone to the appropriate SGSN and RNC.

Within the RNC, prior to performing any IPsec processing, the UDP header, and the GTP encapsulation are removed. Each IPsec-protected datagram is identified by the appearance of the ESP value in the IP next protocol field. In order to determine the IPsec SA that is to be applied, a look up in the SADB is performed. If the SA lookup fails, then, the packet is dropped, and an error is reported. Otherwise, based on the SA found, the IPsec base protocol performs the IPsec processing (authenticates and decrypts the packet). In the sequel, it matches the packet selectors to these encompassed within the SA, and finds the incoming policy within the SPD. After that, it checks whether the required IPsec processing has been applied, and forwards the original IP packet to its destination.

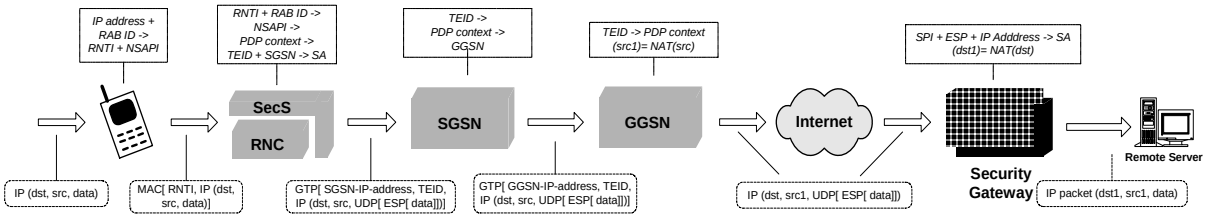

Figure 9: Protected data flow from the MS to the remote server in the network-wide VPN scheme

NAT Traversal

Similarly to the end-to-end scenario, the NAT employment in the GGSN requires special consideration for the network-wide VPN deployment. For that reason, the ESP is proposed for security services, given that it allows NAT to modify the protected IP packets header without experiencing integrity failure. Moreover, the complementary UDP encapsulation is used to overcome the incompatibilities arising from the TCP and NAT coexistence. Finally, the IKE protocol is configured in aggressive mode, since it uses identification data instead of IP addresses for end-node authentication. The same authentication method is also used for the IPsec SA negotiation.

Mobility Implications

The RNC partially handles the MS movement within the radio access network using the radio resource control (RRC) procedures [13]. The location information in PS transactions is distributed in such a way that the CN domain is aware of the location of the MS within the accuracy of routing area, and the RNC knows the location of the MS within the accuracy of cell. When the MS moves to an adjacent cell, which is served by the same RNC, it performs an RRC cell update procedure to inform the RNC about the current cell. The C-RNTI may be reallocated when the MS accesses the new cell, but since none of the UMTS routing parameters that are also involved in the VPN operation (NSAPI and TEID) is changed, the VPN between the RNC and the SG remains the same.

Every time the MS enters a new UTRAN registration area, and owns an RRC connection, it has also to perform an update procedure. In case this area is controlled by a different RNC but the same SGSN, the procedure is referred to as intra-SGSN routing area update, which corresponds to the relocation of the Iu interface. On the other hand, when the mobile enters an area, which is controlled by a different RNC and a different SGSN, then, it is referred to as inter-SGSN routing area update procedure [7, 9].

Fig. 10 illustrates the message sequence diagram of the intra-SGSN relocation procedure, which attempts to preserve the established network-wide VPN for the moving MS. Relocation procedure begins when the source RNC sends a *Relocation Required* message to the SGSN. The SGSN receives this, and tries to allocate the appropriate Iu resources towards the target RNC. It sends the *Relocation Request* message to the target RNC, which contains the information required to build the same RAB configuration as the one existing for the MS before the relocation. When the appropriate Iu resources are allocated, the target RNC responds with a *Relocation Request Acknowledge* message. The SGSN informs the source RNC that the preparation of the relocation is over, and, therefore, the appropriate actions may take place to perform the actual relocation of the serving RNC. For this purpose, the SGSN sends the *Relocation Command* to the source RNC. The source RNC requests from the target RNC to proceed with the relocation by sending *Relocation Commit* message. The purpose of this message is to transfer the serving contexts from the source RNC to the target RNC. These contexts are sent for each RAB, and mainly contain the appropriate sequence numbers of the user-plane messages to be subsequently transmitted in the uplink, and downlink directions. The *Relocation Commit* message should be enhanced to incorporate the status of the active SAs that the moving MS has established. It should contain the VPN context, which includes the SPD entries and the SADB entries referring to the involved MS SAs (incoming and outgoing). The VPN context transfer facilitates the target RNC to construct a copy of the security relations that exist between the source RNC and remote SGs for the particular MS, and, thus, it guarantees network-wide VPN service continuity as the user moves.

It is important to note that before sending the *Relocation Commit*, uplink and downlink data transfer at the source RNC is suspended. After having sent the *Relocation Commit* message, the source RNC begins the forwarding of data for each concerned RAB to the target RNC, and the target RNC sends *Relocation Detect* message to indicate to the SGSN that the execution of the serving RNS relocation has been detected. Then, the target RNC allocates a new RNTI to the MS, and forwards this value to the mobile with *RNTI Reallocation* message. When the MS acknowledges the correct reception of the RNTI, the target RNC considers that the relocation procedure has been completed, and responds to the CN with a *Relocation Complete* message.

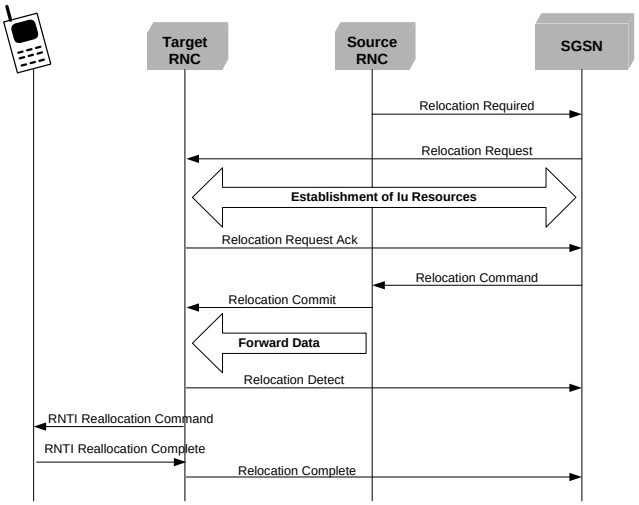

Figure 10: Intra-SGSN relocation procedure

In case of inter-SGSN routing area update (see Fig. 11), the MS first sends a *Routing Area Update Request* message, which contains the old and the new routing area identifier values to the new SGSN. Based on this information the new SGSN is able to determine the old SGSN, and it requests information about the subscriber contexts by sending an *SGSN Context Request* message. The old SGSN validates the presence of the MS, and responds with an enhanced *SGSN Context Response* message, which includes the active PDP context, the mobility management (MM) context, as well as the involved VPN context. Therefore, a copy of the current VPN context, which resides at the RNC and describes the active SAs between the RNC and remote SGs for the particular user, must also exist in the SGSN. This facilitates the VPN reconstruction and the VPN service continuation, in case that a moving MS handoffs to a different RNC or SGSN.

The new SGSN requests the subscriber Authentication Center (AuC) to provide authentication vectors, which returns them within a *Send Parameters* message. Now, the new SGSN is able to start authentication and security control for the moving MS. Furthermore, it acknowledges the contexts transfer, and informs the old SGSN that is ready to receive data packets belonging to the conveyed PDP contexts. When authentication activities have been successfully completed, the new SGSN informs the GGSN that the SGSN has now been changed, and the PDP context information has been changed accordingly. The new SGSN updates location information in the HLR, which in turns cancels the old location of the MS from the old SGSN. Then, the HLR starts to transfer the subscriber profile to the new SGSN, which sends a *Routing Area Update Accepted* message to the target RNC. The RNC receives and stores the VPN context values in order to reconstruct the security relations that exist between the source RNC and remote SGs for the moving MS. The target RNC verifies its availability in providing VPN services, updates its SPD and SADB with the relative IPsec SAs

contents, and, finally, sends a *Routing Area Update Accepted* message to the MS. The latter stores the new temporary identity on its USIM, and acknowledges the receipt by sending a *Routing Area Update Complete*.

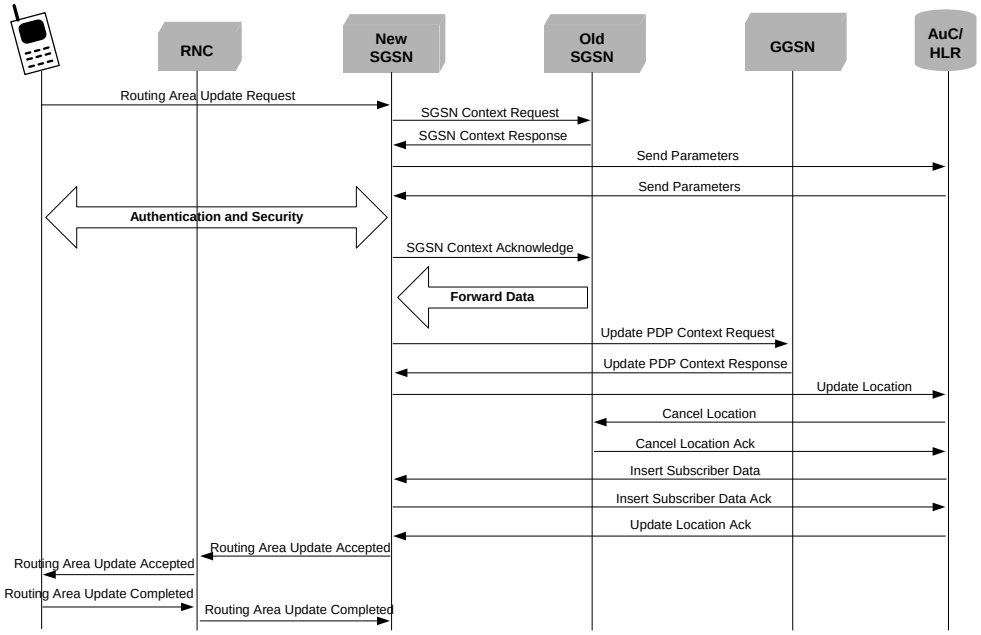

Figure 11. Inter-SGSN routing area update procedure

Border-based Scheme

By placing the SecS in the GGSN, the border-based VPN deployment scheme is realized. This scheme follows the network-assisted security approach, where the deployed VPN is extended between the UMTS border (GGSN) and a remote SG at a private LAN (see Fig. 12). Thus, this model protects data conveyance over the public Internet, which is a vulnerable network segment in the considered communication chain.

For VPN establishment and key agreement the IKE-proxy [12] scheme is employed, similarly to the network-wide scenario. The NAT presence in the GGSN does not have any effect on the deployed VPN, given that the latter is extended on the public address space. The IPsec is configured in tunnel mode, since both security peers are gateways. Moreover, the MS movement within the UMTS coverage area does not have any impact on the deployed border-based VPN, as long as the MS is under the same GGSN. Otherwise, the current SA is dropped, and a new SA should be established. Finally, the ESP employment is considered more advantageous, given that it can provide confidentiality and integrity protection as well.

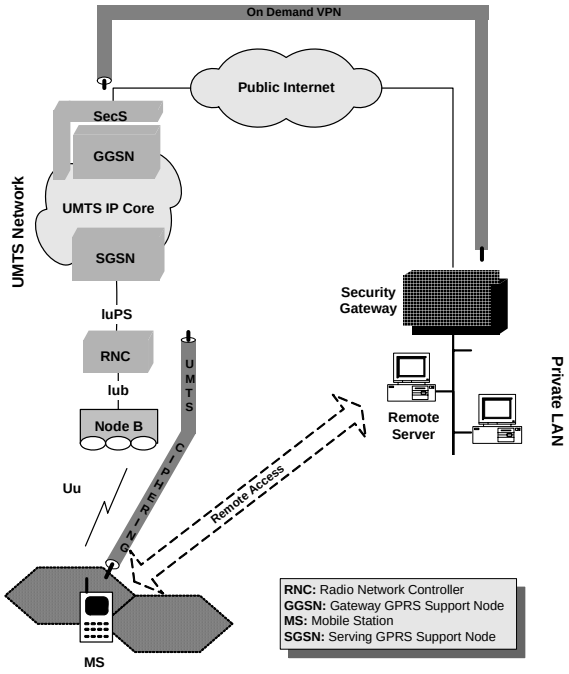

Figure 12: Border-based VPN deployment scheme

FEATURE EVALUATION AND COMPARISON

Dynamic Networking, Privacy, and Authentication

A common characteristic of the proposed schemes is that they support dynamic, client-initiated, secure VPN deployment providing any-to-any connectivity in an ad hoc fashion. A mobile user decides when, and where to establish a VPN across a public network managing the initialization process directly. This is well suited to mobile users, particularly those who roam globally and require access to their home networks, as well as off-the-street users needing a VPN for a short time, perhaps for a distributed game. Enterprises find this highly attractive, since their users may roam internationally avoiding the expense of international phone calls to connect securely to their home networks.

The required level of privacy depends on a risk assessment analysis. If the requirement for privacy is low, then, a simple abstraction of discretion and network obscurity may serve the purpose. However, if the requirement for privacy is high, then, strong access security and security protection on data passed through the network are applied. In any case, security must be implemented correctly; otherwise, either an increased communication cost, or ineffective security occurs.

The proposed schemes differ in the depth to which the secure VPN extends within the UMTS network architecture, and whether data in transit are ever in clear-text, or available to be tapped by outsiders. The

border-based scheme provides security over the public Internet, and, thus, it results in clear-text data transmission over the UMTS backbone network exposing them to potential IP spoofing. On the other side, both the end-to-end and the network-wide models provide maximal security services for accessing remote private networks. However, the first security channel, which is extended between the sender and the receiver, provides end-to-end security, and, thus, intermediate nodes are unable to eavesdrop on, or to modify the traffic. The network-wide scheme implies an uninterrupted sequence of two secured communication hops (UTRAN ciphering and VPN), and, consequently, protected data are presented in clear-text within the RNC node, where they change security tunnel.

The current standards for encryption are Data Encryption Standard (DES) and triple DES, and in the future will migrate towards the recently announced Advanced Encryption System (AES) [19]. VPNs carrying any form of sensitive traffic should use triple DES encryption to protect the data. However, when using triple DES, a VPN solution often suffers performance degradation, especially if it is applied on a mobile device. Thus, strong and resource consuming encryption algorithms can only be applied in the network-assisted security models, where there are adequate processing and memory resources to ensure security service feasibility.

A VPN solution should also address the identification and authorization process associated with the controlled access to it. Based on the deployment scheme, authentication can be performed either by the end-users (CPE approach), or it can be outsourced to the network (network-assisted approach). The latter option, which is already used in wired terrestrial networks, assumes that the communicating parties trust the involved network operator.

Cost and Complexity

The introduction of security services in the mobile network landscape does not come for free. It requires additional encryption/decryption algorithms and security management software, which increase the cost and complexity of the mobile devices and the underlying network infrastructure.

In the end-to-end model, the necessary enhancements for security service provision have minimal impact on the existing network infrastructure. Specifically, the UMTS network nodes, and the intermediate IP routers require no further enhancements or modifications to support the particular VPN scheme. The changes are limited to the security endpoints (MS and SG), which incorporate the IPsec functionality, including the IKE protocol to negotiate, establish, and apply security associations.

In the network-assisted schemes, the MS necessitates the introduction of a lightweight module (SecC), which does not entail considerable processing and memory capabilities, and, thus, it can be easily integrated in any type of mobile device causing minor performance overhead. Using this, a mobile subscriber initiates dynamically a VPN between itself and a corporate LAN SG, while outsourcing complex key negotiation and encryption/decryption functionality to the mobile network infrastructure. This minimizes the configuration and computation overhead associated with mobile users and their devices, as well as reduces the relevant cost compared to the end-to-end solution. Considering the constraints imposed by the nature of mobile devices (low processing power, limited battery power, and limited memory capabilities), it can be perceived that mobile subscribers can obtain significant advantages from outsourcing the management and operation of their VPNs to the network operator.

At the network side, the existing UMTS infrastructure requires enhancements. The RNC (in the network-wide scheme), and the GGSN (in the border-based scheme) need to incorporate the SecS, which is responsible for VPN establishment and operation, and the encryption/decryption process. The SecS module can be readily integrated in existing network infrastructure, and, thus, the proposed VPN schemes can be employed as add-on features of the UMTS. However, this requires additional investment from the mobile network operator, and the security functionality is expected to increase the signaling and the workload burden on the network infrastructure.

	End-to-end	Network-wide	Border-based
Number of messages the MS exchanges	6	2	2
Number of messages the network exchanges	0	7	6
Total number of messages	6	9	8

Table 2: The number of messages required for VPN establishment in the proposed security schemes

For VPN establishment in the end-to-end scenario, the involved MS exchanges six messages directly with the remote SG. On the other hand, in the network-assisted models, the MS exchanges only two messages with the corporate SecS. After the initial triggering, the SecS exchanges seven messages for network-wide VPN, and six messages for border-based VPN establishment. The former scheme requires an additional message, which transfers a copy of the VPN context from the RNC to the SGSN. Table 2 presents the number of messages that are exchanged in each of the three schemes for VPN establishment.

Performance

Security features may have an adverse impact on aspects of quality of service offered to end-users. The security procedures may cause longer delays during connection establishment and handover, and have a negative impact on system capacity, especially on the scarce radio interface. They occupy signaling channels, and the protection of data increases the required bandwidth. It is obvious that the VPN establishment and the consequent security transformation reduce network performance, and delay data processing and transmission.

The IPsec functionality imposes computational cost on the hosts that implement the required protocols. This cost is associated with the memory needed for IPsec code and data structures, the number of messages that are exchanged, and the computation of encryption and decryption, which is added in a per-packet fashion. The magnitude of this cost varies, depending on the mobile host capabilities, and the employed security algorithms. In addition to the IPsec protection, the required UDP encapsulation for NAT traversal imposes extra cost to system operation (e.g., longer session setup/release, encapsulation process, etc.).

The IPsec protection and the UDP encapsulation also increase the bandwidth utilization cost, due to the increase in packet size. The encapsulation overhead is related to the IPsec mode of operation. In tunnel mode, the entire IP packet including the IP header is protected, and an additional IP encapsulation is carried out. On the other hand, the transport mode protects only the IP packet payload, minimizing the operating cost. In general, the multiple encapsulation of the original IP packet induces a waste of valuable resources, and may cause network performance degradation.

In the end-to-end VPN model, the implementation of security functionality in mobile devices, which are characterized by limited power and processing capabilities, increases significantly the processing latency, and may result in service inadequacy. UMTS also employs an optimized ciphering for packet data transmission over the radio interface. This security scheme duplicates encryption (packet encapsulation) over the scarce radio interface, which increases the overall communication cost, and decreases the access network capacity. It is anticipated that the increased bandwidth demand will not noticeably affect the fixed IP infrastructure, but it will have significant impact on the scarce radio interface. For that reason, the IPsec is configured in transport mode in order to avoid an additional IP encapsulation of the original IP packets, in addition to that imposed by the IPsec and the UDP.

The network-assisted security models take advantage of the security measures provided by the UMTS, avoiding duplicate encryption (packet encapsulation) over the scarce radio interface. Hence, VPN deployment has no impact on the radio access network efficiency, as happens in the end-to-end security

model. However, a bottleneck might occur at the SecS level, since it processes the entire traffic from/to a large number of mobile users. To overcome this barrier, the use of specific hardware accelerator for faster and more efficient IPsec deployment, as well as for SecS scaling is discussed below. The network-wide model increases transmission overhead over the UMTS backbone network, and, thus, the IPsec protocol is configured in transport mode. Security implementation in the border-based scheme has minimal impact on the UMTS backbone network, since no extra encapsulation or process takes place over it.

	End-to-end	Network-wide	Border-based
Radio interface	512 bytes	480 bytes	480 bytes
UMTS backbone	512 bytes	512 bytes	480 bytes
Public Internet	512 bytes	512 bytes	524 bytes

Table 3: Security overhead in each network segment for the proposed security model

In Table 3, the protected packet size in each network segment for the proposed VPN models is presented considering a clear-text packet size of 480 bytes. It is worth noting that both encryption and authentication services are considered to be applied to the transmitted traffic using the DES, and the Message Digest (MD5) algorithms [3]. The mean packet size for mobile Internet applications is estimated to be 480 bytes [24].

Apart from the processing and encapsulation overhead in the data plane, the deployed VPN models may also cause control plane operating cost to the underlying mobile network infrastructure. In the network-wide scheme, the formation of security context and its transfer together with the PDP and MM contexts, in case of intra or inter routing area update procedures between the source and target RNCs and SGSNs, increase the necessary amount of information being exchanged during the mobility management procedures. However, the required bandwidth overhead for the control plane is expected to be negligible, since no additional messages for VPN maintenance are required. It might only affect the SecS capacity in case of frequent occurrence of routing area update procedures. In contrast to the network-wide scheme, the VPN establishment and maintenance procedures in the end-to-end and the border-based VPN models have no effect on the network control plane.

Reliability and Scalability

For a VPN deployment model to be actually applicable, it must be reliable. Reliability is perceived as the ability to use VPN services in terms of authorized usage at all times, at a level similar to that of the public telephone network. The VPN model reliability is highly related to the network connectivity, and the capacity of the underlying technology to provide security services.

In the end-to-end security model, VPN service reliability depends on the mobile network reliability. When the MS ensures network connectivity and guarantees the required resources, it is evident that it can establish an end-to-end VPN, although security features may reduce the system availability. VPN provision is restricted by the mobile network throughput, the access network capacity, and the terminal processing power.

In the network-assisted models, VPN reliability depends on the ability of the security infrastructure to provide the appropriate signaling for VPN establishment, and its capacity to apply the required security transformation. The capacity of the security infrastructure relates to the number of simultaneous VPNs it can support, as well as to the amount of information to be exchanged per VPN. From the network-assisted security models, the border-based scheme is more reliable, given that there is no interrelation between the network management and the VPN management.

System scalability in terms of processing and management also influences VPN model reliability. Scalability represents the ability of a particular approach to grow when it reaches the upper limit of its throughput. It should include ways to stack boxes together to work in tandem for increasing the overall capacity, and eliminating the total upgrades. The ability to cost-effectively adapt a VPN to meet changing bandwidth and connectivity needs is crucial in order to accommodate unplanned growth, and changes driven by customers demand.

In the end-to-end security model, it is clear that VPN scalability corresponds to mobile network scalability. On the other side, when the proposed network-assisted architectures have reached the upper limits of their throughput, then, system scalability can be applied to increase their capacity. Scalability can be achieved by changing the configuration of the SecS module by using more that one SPD, SADB, and IPsec base protocol submodules (e.g., one for incoming, and one for outgoing traffic). Another possibility is the incorporation of more than one SecS in the corresponding RNC or GGSN, for handling the incoming and the outgoing processing, separately.

Management, Usability, and Trust

The end-users demand for scalable and flexible VPNs shifts their deployment model from being "homegrown" towards being outsourced. Network operators offer security services at a lower cost enabling also the employment of specific hardware accelerator modules for faster and more efficient IPsec deployment. Since there is a lack of resources and expertise in deploying VPN, end-users are increasingly accepting that outsourcing offers them considerable cost savings. Moreover, the network-assisted implementations can continuously evolve, in order to respond to new end-user requirements, and allow for generic security strategies by network operators.

A VPN provides point-to-point connectivity, and, thus, both ends have to be configured to cooperate properly. The VPN configuration deals with complicated procedures, such as the endpoints authentication, key exchange, and selection of the appropriate encryption/decryption algorithm. Nevertheless, these procedures should be easy to be used and understood by end-users that are not security experts. In advanced VPN solutions, tunnel establishment and release, as well as encryption and decryption should be transparent to end-users.

Apart from the security association negotiation, a service level agreement (SLA) contract may be signed between the security provider and the involved communicating peers, which involves some challenging technical issues both for the provider and end-users. In general, network-assisted security solutions require from the communicating end-points to trust the security provider [2]. The SLA is the only binding tool at the subscriber's disposal to ensure that the VPN provider delivers the security services at the level agreed.

In the end-to-end security model, end-users are involved in the management and configuration of their VPNs. Thus, they must be aware of when encryption is required, since, the end-station software may require from the user to take decisions on the appropriate security policy. Considering also that the mobile devices are designed to be portable with small screen size and limited input capabilities, it can be perceived that VPN usability will be limited to experienced users. Generally, in this scheme, the SA negotiation is not transparent to the mobile subscriber and his device. However, the mobile network operator does not even realize the existence of an end-to-end VPN, and for that reason neither SLA nor trusted relations between the security endpoints and the network operator are required. In this model, the trusted relations are limited between the security endpoints and the certificate authority, which issues digital certificates and facilitates authentication process.

In the network-assisted models, VPN establishment and maintenance are performed by the network infrastructure, eliminating the technical skills and the specialized knowledge required by mobile users. An

end-station simply initiates a VPN establishment, while outsourcing authentication, key exchange, and encryption/decryption process to the network operator. Thus, the security association is transparent to the mobile subscriber. The mobile subscriber and the remote server administrator have to trust the mobile network operator, since no end-to-end authentication is possible. The transaction contents are encrypted/decrypted in the SecS, which resides in the UMTS, relying on the mobile network operator security policy. Moreover, all the involved parties (mobile users, mobile operator, and corporate LAN) have to trust the authority, which issues digital certificates.

Being able to provide security services, the network operator should sign SLAs with the mobile subscribers and the corporate LAN administrators. A trusted association between the mobile user and the mobile network operator already exists, since the latter controls the security issues for standard UMTS services. However, the trusted relationships required between the mobile network operators and the corporate LAN administrators might influence the network-assisted VPN models proliferation and scalability.

Mobility and Compliance

Security models, which extend over mobile networks, should consider end-user mobility. A deployed end-to-end VPN has no interrelation to the underlying network operation and the provided network connectivity. It supports user mobility and roaming, and operates transparently to the MS movement, since the security parameters, which are contained in the IPsec SA, are not affected by the mobility management procedures.

The network-wide VPN solution is not well suited for systems with high degree of mobility. The preparation steps for a handover include preparation for transfer of the corresponding security context to the new access point. This would enable the reconstruction of the security association to the target access point when the mobile user connects, providing continuous VPN services from the end-user perspective. However, this feature requires that some messages in the UMTS mobility management procedures incorporate the VPN context attributes in order to support security service continuity, as the user moves, and roams.

In the border-based security model, the subscriber mobility is transparent to the VPN operation, as long as the subscriber remains under the same network operator coverage, and is served by the same GGSN. However, whenever the mobile user roams to another GGSN, the existing security association cannot be used, and a new VPN should be established.

VPN deployment may cause problems to applications that need to inspect or modify encapsulated data packets. This means that traffic-shaping mechanisms, monitoring equipment, legal interception, firewalls, and NAT devices are unable to get access to protected data traffic and fail to perform their functions.

The end-to-end security model is not compatible with the legal interception option, or any other application that requires access to the traversing data within the mobile network. The enforcement of network security policy, traditionally performed by border firewalls, is devolved to end hosts, which establish VPN overlays. Despite this, the border firewalls remain to perform packet filtering, and counteract against denial of service attacks. Furthermore, this scheme tends to cause problems when NAT is used, and in order to overcome this incompatibility the UDP encapsulation is applied.

The network-wide scheme is compatible with legal interception; however, UDP encapsulation is also applied for NAT traversal. In this scenario, the network security policy is enforced by the RNC, which incorporates the SecS. The border firewall may also apply security policy, provided that it supports dynamic configuration and can share the security policy database included in the SecS.

Finally, the border-based VPN is compatible with legal interception option and NAT presence. In this scheme, the border firewall protects the trusted mobile network domain by assessing incoming and outgoing packets, and dropping the unwanted traffic. This is possible since both the border firewall and the SecS reside in the GGSN.

Comparison

The proposed security schemes have been designed to provide dynamic, secure, remote access of mobile users to corporate LAN resources over UMTS network based on IPsec. They support various levels of security services, and follow different models of deployment. Therefore, the particular scheme that will be selected in a potential scenario depends on the risk analysis performed, the required security services, the operator security policy, as well as the network topology. In Table 4, a comparison of the proposed security models, based on the deployment and operation features outlined above, is presented in a tabular form.

Table 4: Comparison table of the proposed security models

Evaluation Features	End-to-End	Network-wide	Border-based
Dynamic networking	√	√	√
End to end security	√		

Evaluation Features	End-to-End	Network-wide	Border-based
Security in one hop	√		
No clear-text data transfer	√	√	
No encryption/decryption in MS		√	√
Strong encryption algorithm		√	√
Authentication in the end-users hands	√		
Minimum enhancements in MS		√	√
Less MS cost		√	√
Applicable to any MS type		√	√
Minimum enhancements in the mobile network infrastructure	√		
Become an add-on feature of UMTS		√	√
Less computational overhead in MS		√	√
No encapsulation (IPsec, UDP) over UTRAN.		√	√
No encapsulation (IPsec, UDP) over the UMTS backbone			√
No point of bottleneck in the UMTS network architecture	√		
No interrelation between the VPN management and the network management	√		√
Less control plane overhead	√		√
System reliability corresponds to network reliability	√		
Scalable model in terms of performance		√	√
Scalable model in terms of proliferation	√		
VPN outsourcing model		√	√
Trusted third party	√	√	√
Less skilled end-user required		√	√
SA transparency to end-user		√	√
Solid VPN management		√	√
No SLA required	√		
No enhancements required to support mobility	√		√
MS movement transparency in VPN operation	√		
VPN continuity when user roams	√	√	
Compatible with traditional firewalls			√
Compatible with legal interception option		√	√
Compatible with NAT application			√

CONCLUSIONS

Mobile Internet offers global connectivity, and, thereby, facilitates the materialization of new services as well as the accessing of corporate resources for remote mobile users. Privacy and security are essential to the success of the new emerging applications in mobile systems. Three alternative schemes for dynamic, client-initiated, secure VPN deployment over the UMTS network have been proposed and analyzed. The proposed schemes differ in where the IPsec functionality is placed within the UMTS network architecture (mobile node, access network, and UMTS network border), depending on the employed security models, and whether data in transit are ever in clear-text or available to be tapped by outsiders.

The end-to-end security model integrates VPN functionality into the communicating peers, which negotiate and apply security. Sensitive data traffic remains encrypted for the entire route between the sender and the receiver providing the best security services. Authentication is performed by end-hosts using digital certificates, issued by a trusted certificate authority. This model requires minimum enhancements in the underlying mobile network infrastructure, and the deployed VPN operates transparent to the MS movement and roaming. There is no interrelation between the VPN management and the network management, and, therefore, no extra control overhead is caused by the VPN deployment. However, this scheme imposes extra computational costs on the lightweight end-user mobile devices, which have to incorporate the IPsec functionality. This increases the cost of mobile devices that should have advanced processing capabilities to perform security transformations. Moreover, it duplicates encryption (packet encapsulation) over the scarce radio interface, and it is incompatible with traditional border firewall presence and legal interception option.

An alternative to the end-to-end approach pertains to the network-assisted security model, which incorporates VPN functionality into the network infrastructure. The network operators have solid network management expertise, and more resources to effectively create, deploy, and manage VPN services that originate from mobile subscribers. These schemes (network-wide and border-based) eliminate the computational cost on the handheld mobile devices, which have limited processing capabilities, and the space overhead in the scarce radio interface resources. They are compatible with the legal interception option, which requires that public authorities are able to gain access on the traversing data within the mobile network for legal purposes. Since authentication is outsourced to the network operator, the network-assisted security solutions require that the communicating end-points trust the security provider. All the involved parties have to trust the certificate authority, which issue digital certificates. Since the security transformation is performed at the mobile network site, strong and resource consuming protections mechanism can be applied.

Moreover, the security server (SecS), integrated in the UMTS network infrastructure for VPN service provision, might become a bottleneck, since it processes the entire traffic from/to a large number of mobile users.

The network-wide scheme provides maximal security services by taking advantages of the existing UTRAN ciphering over the radio interface, and extending a VPN over the UMTS backbone, and the public Internet. However, it is not well suited for systems with high degree of mobility, and it is incompatible with firewall mediation. The border-based VPN is extended over the public Internet providing security services between the UMTS border and a remote corporate private LAN. Thus, it results in clear-text data transmission over the UMTS backbone, exposing data them to a potential IP spoofing. The subscriber mobility is transparent to the VPN operation, while the subscriber is served by the same GGSN. Moreover, it is compatible with firewall mediation at the UMTS border.

The proposed security schemes have been evaluated and compared from the security and operation point of view. The evaluation results summarize the relative performance of the considered security models, and provide useful insights for network security policy design and configuration.

ACKNOWLEDGMENTS

The authors would like to thank Nikos Loukas and the anonymous reviewers for their comments and suggestions.

REFERENCES

[1] 3GPP TS 23.002 (v3.6.0) “Network Architecture”, release ’99, Sept 2002.

[2] B. Gleeson, A. Lin, J. Heinanen, G. Armitage, and A. Malis, “A Framework for IP Based Virtual Private Networks,” RFC 2764, Feb. 2000.

[3] S. Kent and R. Atkinson, “Security Architecture for the Internet Protocol,” RFC 2401, Nov. 1998.

[4] D. Harkins and D. Carrel, “The Internet Key Exchange (IKE),” RFC 2409, Nov. 1998.

[5] B. Aboda, W. Dixon, “IPsec-Network Address Translation (NAT) Compatibility Requirements,” RFC 3715, March 2004.

[6] GSM 03.60, “GPRS Service Description,” Stage 2, 1998.

[7] 3GPP TS 23.060 (v3.16.0) “GPRS Tunneling Protocol (GTP) across the Gn and Gp interface”, release ’99, March 2003.

[8] A. Huttunen et al., "UDP Encapsulation of IPsec Packets," draft-ietf-ipsec-udp-encaps-09.txt, Internet Draft, May 2004.

[9] 3GPP TS 24.008 (v3.15.0) "Mobile Radio Interface Layer 3 specification; Core Network Protocols – Stage 3", release '99, March 2003.

[10] C. Xenakis, E. Gazis and L. Merakos, "Secure VPN Deployment in GPRS Mobile Network," Proc. European Wireless 2002, Florence Italy, Feb. 2002, pp. 293-300.

[11] C. Xenakis and L. Merakos, "Dynamic Network-based Secure VPN Deployment in GPRS," Proc. PIMRC 2002, Lisboa, Portugal, Sept. 2002, pp. 1260-1266.

[12] C. Xenakis and L. Merakos, "On Demand Network-wide VPN Deployment in GPRS," IEEE Network, Vol. 16, No. 6, Nov/Dec. 2002, pp. 28-37.

[13] 3GPP TS 25.331 (v3.14.0) "Radio Resource Control (RRC) protocol specification", release '99, March 2003.

[14] 3GPP TS 25.321 (v3.15.0) "Medium Access Control (MAC) protocol specification", release '99, March 2003.

[15] T. Kivinen et al., "Negotiation of NAT-Traversal in the IKE," draft-ietf-ipsec-t-ike-08.txt, Internet Draft, Feb 2004.

[16] 3GPP TS 25.323 (v3.10.0) "Packet Data Convergence Protocol (PDCP) Specification", release '99, Sept 2002.

[17] 3GPP TS 25.322 (v3.15.0) "Radio Link Control (RLC) protocol specification", release '99, Dec 2002.

[18] 3GPP TS 25.301 (v3.11.0) "Radio Interface Protocol Architecture", release '99, Sept 2002.

[19] E. Danielyan, "Goodbye DES, Welcome AES," Cisco The Internet Protocol Journal, vol. 4, no. 2, June. 2001, pp 15-21.

[20] W. Diffie and M. E. Hellman, "New Directions in Cryptography," IEEE Trans. Info. Theory, vol. 22, Nov 1976, pp. 644-654.

[21] V. Gupta and S. Gupta, "Securing the Wireless Internet," IEEE Communications Magazine, Vol. 39, No. 12, Dec. 2001, pp. 68-74.

[22] K. Lam et al., "Lightweight security for mobile commerce transactions," Computer Communications, Vol. 26, No. 18, Dec. 2003, pp. 2052-2060.

[23] C. Xenakis and L. Merakos, "Security in third Generation Mobile Networks," Computer Communications, Vol. 27, No. 7, May 2004, pp. 638-650.

[24] ETSI,Universal Mobile Telecommunication System (UMTS);Selection Procedures for the Choice of Radio Transmission Technologies of the UMTS, Technical Report TR 101 112 v3.2.0,1998.

[25] Wireless Application Forum (WAP), WAP specifications,
URL: http://www.wapforum.org/what/technical.htm.

[26] R. Rivest, A. Shamir, and L. M. Adleman, "A Method for Obtaining Digital Signatures and Public-Key Cryptosystems," Commun. ACM, vol. 21, Feb. 1978, pp. 120–26.

[27] W. Itani and A. Kayssi, "SPECSA: a scalable, policy-driven, extensible, and customizable security architecture for wireless enterprise applications", Computer Communications, Vol. 27, No. 18, Dec. 2004, pp. 1825-1839.

[28] J. Al-Muhtadi, D. Mickunas, R. Campbell, "A Lightweight Reconfigurable Security Mechanism for 3G/4G Mobile Devices", IEEE Wireless Communications Vol. 9, No. 2, April 2002, pp. 60-65.

[29] W. Itani, A. Kayssi, "J2ME end-to-end security for m-commerce", Proc. IEEE Wireless Communications and Networking Conference 2003.

[30] P. M. Feder, N. Y. Lee, S. Martin-Leon "A Seamless Mobile VPN Data Solution For UMTS and WLAN Users", Proc. 4th International Conference on 3G Mobile Communication Technologies, June 2003, pp. 217 – 221.

BIOGRAPHIES

Christos Xenakis (xenakis@di.uoa.gr) received his B.Sc degree in computer science in 1993 and his M.Sc degree in telecommunication and computer networks in 1996, both from the Department of Informatics and Telecommunications, University of Athens, Greece. In 2004 he received his Ph.D. from the University of Athens (Department of Informatics and Telecommunications). From 1998 – 2000 was with the Greek telecoms system development firm Teletel S.A., where was involved in the design and development of advanced telecommunications subsystems for ISDN, ATM, GSM, and GPRS. Since 1996 he has been a member of the Communication Networks Laboratory of the University of Athens. He has participated in numerous projects realized in the context of EU Programs (ACTS, ESPRIT, IST). His research interests are in the field of mobile/ wireless networks, security and distributed network management. He is the author of over 15 papers in the above areas.

Lazaros Merakos (merakos@di.uoa.gr) received the Diploma in electrical and mechanical engineering from the National Technical University of Athens, Greece, in 1978, and the M.S. and Ph.D. degrees in electrical engineering from the State University of New York, Buffalo, in 1981 and 1984, respectively. From 1983 to 1986, he was on the faculty of Electrical Engineering and Computer Science at the University of Connecticut, Storrs. From 1986 to 1994 he was on the faculty of the Electrical and Computer Engineering Department at Northeastern University, Boston, MA. During the period 1993-1994 he served as Director of the Communications and Digital Processing Research Center at Northeastern University. During the summers of 1990 and 1991, he was a Visiting Scientist at the IBM T. J. Watson Research Center, Yorktown Heights, NY. In 1994, he joined the faculty of the University of Athens, Athens, Greece, where he is presently a Professor in the Department of Informatics and Telecommunications, and Director of the Communication Networks Laboratory (UoA-CNL) and the Networks Operations and Management Center. His research interests are in the design and performance analysis of broadband networks, and wireless/mobile communication systems and services. He has authored more than 150 papers in the above areas. Since 1995, he is leading the research activities of UoA-CNL in the area of mobile communications, in the framework of the Advanced Communication Technologies & Services (ACTS) and Information Society Technologies (IST) programmes funded by the European Union (projects RAINBOW, Magic WAND, WINE, MOBIVAS, POLOS, ANWIRE). He is chairman of the board of the Greek Universities Network , the Greek Schools Network, and member of the board of the Greek Research Network. In 1994, he received the Guanella Award for the Best Paper presented at the International Zurich Seminar on Mobile Communications.