Security in Public Access Wireless LAN Networks

M.Sc. Thesis by

Fabio Moioli
(info@futucom.com)

Department of Teleinformatics, Royal Institute of Technology, Stockholm
Dipartimento di Elettronica e Informazione, Politecnico di Milano, Milan
Wireless LAN Systems, Ericsson Business Networks, Sundbyberg

Advisors:

Professor Gerald Maguire JR. – Royal Institute of Technology
Professor Riccardo Melen – Politecnico di Milano
Martin Johnsson – Product Manager, Ericsson Business Networks
Martin Rinman – NFI Manager, Ericsson Business Networks

12 June 2000

Abstract

Wireless LAN technology offers high-speed, wireless connectivity that enables mobile computing in many different environments. Many new services will be provided by usage of wireless LANs in public, home and corporate scenarios. Example of remote services might be downloading a video stream from a remote server or accessing news from an e-magazine. Local services could consist in travel information at airports, hotel-facilities at hotels, and conference-update at conference centers, as well as buying a fresh drink in a store.

Present wireless networks are based on IEEE 802.11 standard, operating in the unlicensed 2.4 GHz ISM band and providing a bitrate of 2 Mbps. A new version of the 802.11 standard (802.11b) already now allows a bitrate of up to 11 Mbps. Work is now in progress for a new high performance wireless LAN standard with initial data rates up to 54 Mbps. It will operate in the license free 5 GHz band, which is globally available.

Support for IP mobility is also on the verge of hitting the market on a broad scale. The requirement for this feature stems from the fact that terminals such as laptops, notebooks, and even palmtops, are on their way to regularly being connected to LANs by means of previous wireless interface. In a longer term, IP mobility is also seen as a requirement for the next generation of cellular networks and (of course) also for next generation of wireless LANs, e.g. those based on the HiperLAN/2 technology. While Mobile IP should be part of an overall mobility suit solution, it is best used selectively and in pop-up mode (i.e. using DHCP to obtain addresses) instead of using a Foreign Agent. Using Foreign Agents may in fact add additional complexity to the network, which may already be using DHCP.

Both Wireless LAN systems deployed today and the Mobile IP solutions specified by the IETF mobility working group implement and define different levels and parts of an overall security architecture. Current solutions lack an analysis of the security framework regarding requirements that would apply for the private, public, domestic, and VPN networking cases, which are quite different in their nature.

The main goal of this thesis has consisted in analysing and defining several security proposals for future wireless LAN network scenarios. Advanced services, such as IP mobility, accounting support for roaming, VPN services, and a secure interface for dynamic

assignment of IP addresses to mobile terminals (e.g. through DHCP or DRCP), have also been integrated in the overall framework. The most important results of this work consisted in:

- A security Functional Description (FD) for usage of wireless LAN networks in present and envisaged application scenarios.
- A detailed analysis of IEEE-802.11 and HiperLAN/2 standards, with particular attention to procedures for handover and security.
- An evaluation of Ericsson's present solutions for wireless LAN security and an extension of the latter to the case of public access networks.
- Several proposals for integration of (AAA) accounting schemes in wireless LAN systems.
- A complete and scalable architecture able to provide authentication, data confidentiality, and integrity, to usage of Mobile IP (an optimized interaction scheme between IPSec and Mobile IP has been produced through the definition of a new ISAKMP payload).
- Three possible proposals for secure dynamic assignment of IP addresses to mobile terminals (authenticated DHCP, DHCP with IPSec, and DRCP), with public key distribution support for roaming.
- An analysis of different Virtual Private Network solutions for wireless networks.
- Several appendixes with literature information regarding security policies, network and datacom security, several protocols for data confidentiality and authentication, key exchange and distribution, and IP mobility.

Acknowledgments…

I worked on this Master of Science Thesis as System Manager at Ericsson Wireless LAN Systems, Ericsson Business Networks, in Stockholm, during the period from September 1999 to March 2000.

I would like to express my sincere thanks to all my advisors, for their valuable help during this work:

- *Martin Johnsson, my supervisor at Ericsson, for his great support, his patience, and his theoretical guidance during the time that I have spent at the Wireless LAN Systems Department;*
- *Prof. Gerald Maguire JR, my advisor at the Royal Institute of Technology, for his directions, his precious ideas, and his help in reviewing my thesis;*
- *Prof. Riccardo Melen, my advisor at the Politecnico di Milano, for his directions, his kindness, and for the opportunity he gave me to join Ericsson for my degree project;*
- *Martin Rinman, for the administrative help he gave me at Ericsson.*

I would like also to thank Yi Cheng, Ph.D. student at Ericsson Wireless LAN Systems, for her significant help in understanding many security issues, as well as all people at Ericsson Business Networks, Wireless LAN Department, who helped me in many practical matters. Many precious suggestions were also provided from Pete McCann, from the Bell-Research Labs, Vipul Gupta, from SUN Microsystems, Moshe Litvin, from Check-Point Software, Basavaraj Patil, from Nortel Networks, and Charles Perkins, from Nokia Research Lab.

A special thanks to my family, who has been always supporting my choices.

Table of Contents

1 Introduction

Both Wireless LAN systems deployed today and the Mobile IP solutions specified by IETF implement and define different levels and parts of security architecture. Current solutions lack an analysis of the overall framework regarding security requirements that would apply for the private, public, domestic, and VPN networking cases, which are quite different in their nature.

Regarding the fact that roaming should be made possible between these environments, and that user and network security requirements must be fulfilled, together with the usability of the system from a user's perspective, several security proposals have been in this thesis considered for all these scenarios. When analyzing appropriate solutions, it has been taken into consideration existing as well as planned and envisaged security proposals. Advantages and disadvantages of these solutions have been evaluated. The criteria used for evaluation are very general and include both aspects of security as well as implementation feasibility, scalability, and performance. Unicast traffic and multicast traffic have been both considered.

The following themes have been used as input for this thesis project:

- Security solution in current WLAN products from Ericsson
- IPSec, architecture and protocols
- Mobile IP
- Simple Mobile IP (applying the requirement for dynamic IP address assignment)
- Applicable AAA requirements specified by the AAA WG in IETF
- Security solution for HiperLAN/2

1.1 Thesis Results

The main goal of this thesis has consisted in analyzing and defining a security framework for future wireless LAN networks. Advanced services have also been integrated in this framework, such as IP mobility, accounting support for roaming, VPN services, and a secure interface for dynamic assignment of IP addresses to mobile terminals (e.g. through DHCP or DRCP). The thesis work has been proceeded in three main phases:

- Analysis of the (wireless LAN) network architecture in different usage scenarios
- Specification of security requirements for each application area
- Study of present and envisaged security proposals for future application in WLAN systems

This analysis has finally lead to several author's proposals, with special consideration of what needs to be modified or added to actual solutions in order to produce a system able to scale and be used in public areas.

The most important results of this thesis consisted in:

- A security Functional Description (FD) for usage of wireless LAN networks in present and envisaged application scenarios.
- A detailed analysis of IEEE-802.11 and HiperLAN/2 standards, with particular attention to procedures for handover and security.

- An evaluation of Ericsson's present solutions for wireless LAN security and an extension of them to the case of public access networks.
- Several proposals for integration of (AAA) accounting schemes in wireless LAN systems.
- A complete and scalable architecture able to provide authentication, data confidentiality, and integrity, to usage of Mobile IP (an optimized interaction scheme between IPSec and Mobile IP has been produced through the definition of a new ISAKMP payload).
- Three possible proposals for secure dynamic assignment of IP addresses to mobile terminals (authenticated DHCP, DHCP with IPSec, and DRCP), with distribution support for public keys in case of roaming.
- An analysis of possible VPN solutions for wireless networks.
- Several appendixes with literature information regarding network security and IP mobility.

1.2 Report Structure

This thesis report is structured in three parts, where each of these parts covers a different and partially self-contained issue. Part 1 defines network and security requirements for several wireless LANs scenarios, also providing an overview of possible future application fields for this technology. A security functional description (FD) is provided for each of the major usage areas. Possible security threats are considered and related to the different networking cases. Following sections constitute the first part of the thesis:

- Chapter 2 introduces Wireless LAN networks
- Chapter 3 describes actual and envisaged usage scenarios for wireless (LAN) systems
- Chapter 4 gives an overview of security threats in different wireless application areas
- Chapter 5 defines a Functional Description (FD) for security in different environments

Part 2 introduces IEEE 802.11 and HiperLAN/2 standards for wireless LAN networks. It also describes security techniques adopted in present Ericsson's products for 802.11 wireless networks (wireless Guards). Planned solutions for HiperLAN/2 systems are then presented. Special attention is given to handover procedures between different access points belonging to the same network and to the key distribution infrastructure. Extensions for public and hot spot scenarios, based on present corporate solutions, are proposed. Accounting and IP roaming functionality are defined for the public usage case. Following sections forms part 2:

- Chapter 6 describes the IEEE 802.11 standard
- Chapter 7 introduces security mechanisms defined in 802.11 LANs and proposes some possible extensions
- Chapter 8 presents HiperLAN/2 standard
- Chapter 9 defines several extensions to wireless LANs security for usage in public environments

- Chapter 10 summarizes the overall security framework introduced in Part 2

The third part of the thesis focuses on several services offered at ISO layer-3, combining IP mobility, network security, key management distribution, secure dynamic assignment of IP addresses, and access to Virtual Private Networks (VPNs). All these services are analyzed and structured together. Security means are provided for each of these services, even if none of them is necessary to provide basic security for wireless LAN networks, defined in part 2. Following sections belong to Part 3:

- Chapter 11 provides a suit of security proposals for IP mobility
- Chapter 12 proposes several alternatives for secure assignment of IP addresses
- Chapter 13 presents Virtual Private Network services and applies them to wireless LAN networks
- Chapter 14 integrates previous services and highlights possible future works

Several Appendixes at the end of this thesis provides a general overview of network security. Most important security techniques and related themes are presented for all readers who are new in the area. Main attention is given to those specific security themes that are needed to understand the different parts of this thesis, as authentication and cryptographic algorithms, the IPSec framework, different key management protocols, and the public key infrastructure. Several different flavors of Mobile IP are also presented and compared, creating a mobility solution optimized for each network and usage case. Following appendixes are included at the end of the thesis:

- Chapter 15 (Appendix A) describes security as a multidisciplinary issue
- Chapter 16 (Appendix B) explains basic methods available to provide network security
- Chapter 17 (Appendix C) gives an overview of authentication schemes and algorithms
- Chapter 18 (Appendix D) introduces the IP Security (IPSec) suit
- Chapter 19 (Appendix E) presents different approaches to key distribution and management, with special focus on SKIP, ISAKMP, and IKE
- Chapter 20 (Appendix F) defines a hierarchical model for IP mobility

- PART 1 -

Network Architecture
(Network & Security Requirements)

2 Wireless LANs

A wireless local area network (WLAN) is a flexible data communications system implemented as an extension to or as an alternative for a wired LAN. Using radio frequency (RF) technology, wireless LANs transmit and receive data over the air, minimizing the need for wired connections. Thus, wireless LANs combine data connectivity with user mobility. With wireless LANs, users can access shared information without looking for a place to plug in, and network managers can set up or augment networks without installing or moving wires. Wireless LANs offer increased productivity, convenience, and cost advantages over traditional wired networks.

Wireless LANs have gained strong popularity in a number of vertical markets, including the health-care, retail, manufacturing, warehousing, and academia. These industries have profited from the productivity gains of using hand-held terminals and notebook computers to transmit real-time information to centralized hosts for processing. Today wireless LANs are becoming more widely recognized as a general-purpose connectivity alternative for a broad range of business customers.

Wireless LAN (WLAN) products based on the different flavors of 802.11 are available from many different vendors. Depending on transmission scheme, products may offer bandwidths ranging from about 1 Mbit/s up to 11 Mbit/s. Prices are expected to fall, making WLAN more and more a serious alternative to fixed Ethernet access.

To meet the networking requirements of tomorrow, a new generation of both WLAN and cellular network technologies are under development. These requirements include support for QoS (to build multiservice networks), security, handover when moving between local area and wide areas as well as between corporate and public environments, increased throughput for the ever-demanding need for better performance from both bandwidth-demanding datacom as well as for instance video-streaming applications.

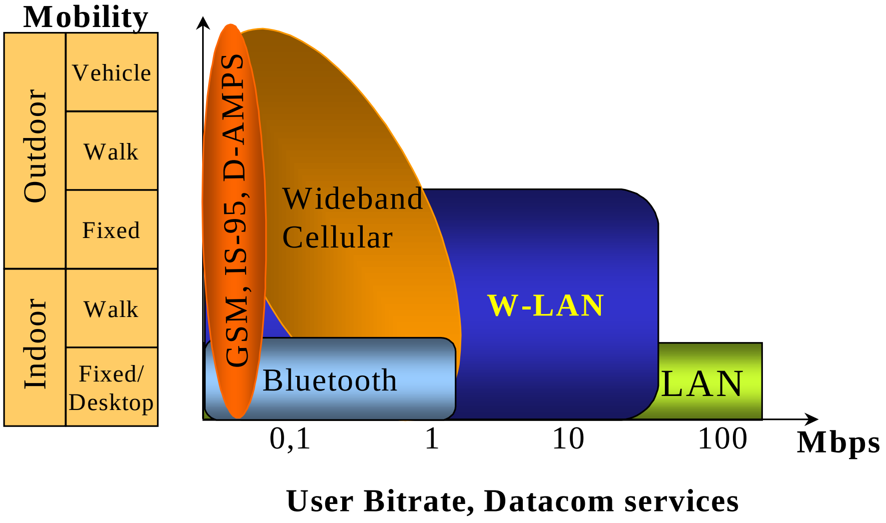
Part 1 of this thesis defines the network and security requirements applicable to the most important wireless LANs scenarios, also providing an overview of possible future application fields for wireless LANs technology. Possible security threats are considered and related to the different network usage cases. A functional requirement specification is provided for each of the major use cases.

3 Networks Scenarios

Wireless LANs may be used both in the case of private networks, e.g. corporate networks, and in the case of public networks, e.g. access to ISPs, as well as to create so called Virtual Private Networks (VPN). The home environment may be another profitable application scenario for future wireless LAN networks.

It is planned that wireless LAN will be finally used also as alternative access networks to 3rd generation cellular devices. Interoperability between future wireless LANs (HiperLAN/2) and Wide Band Cellular protocols (W-CDMA) will support this feature.

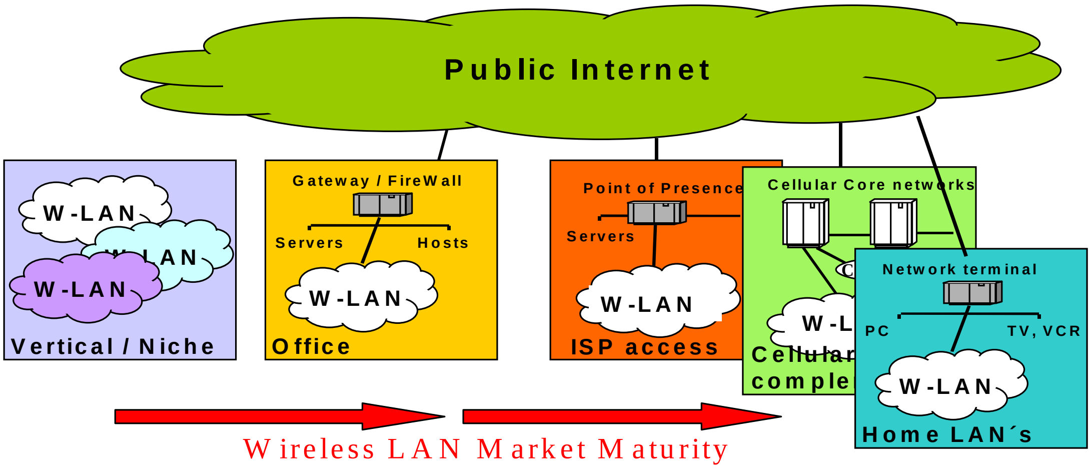
Wireless computing solutions - Positioning



Wireless LANs have been so far mainly used only for corporate networks. The lack of strong security in existing WLAN products has resulted in the development of for example Ericsson’s WLAN IEEE-802.11 security solution (Guards). This solution is studied for an environment where users are well defined and in a number that allows maintenance of a centralized users’ credentials database, i.e. only for securing corporate LANs.

When it is desired to have Wireless LAN networks deployed in public or home environments, it will be necessary to define new specific security policies. This is true also for the cases of remote access to corporate network at public hot spot areas, for ISP public access to the Internet, and for access to 3rd generation cellular networks. It must be remembered that scalability is critical for a communication network. Any technical solutions that should limit the system scalability must to be avoided.

Market development

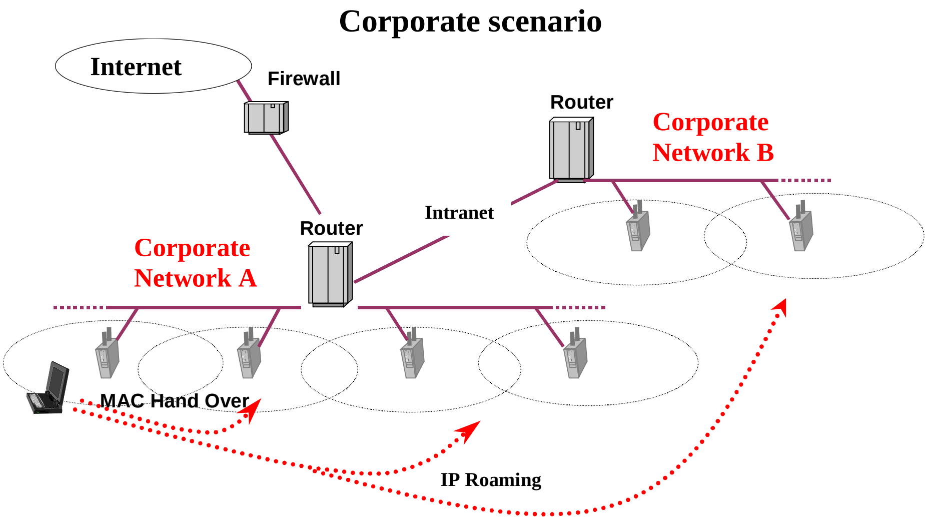


3.1 Corporate networks

There are many cases where it may be convenient to use wireless LAN in corporate environments, e.g. to avoid problems with wire installation in ancient prestigious building or to allow flexible re-planning of working groups. Small Office / Home Office (SOHO) applications may be another interesting area of corporate usage for wireless LANs.

A special case of corporate access need is when a user works between different corporate campuses in different geographic areas. In this case wireless access might be integrated with an appropriate mobility management protocol, so that the customer can move from office to office and even from country to country without the need of manually changing the network setting (IP roaming). Although wireless access is not mandated out from user mobility, it would certainly facilitate the mobility aspect.

Wireless LANs are usually used in corporate networks as last link segment between the MTs and the network wired LAN. The main security goal in corporate networks is to only allow authorized users to access the corporate wired LAN, at the same time providing confidentiality and integrity for on-the-air traffic. The wireless access network must support mobility within the same LAN/subnet (MAC Hand Over). IP mobility, i.e. movement and handover between different IP subnets, may also be provided.



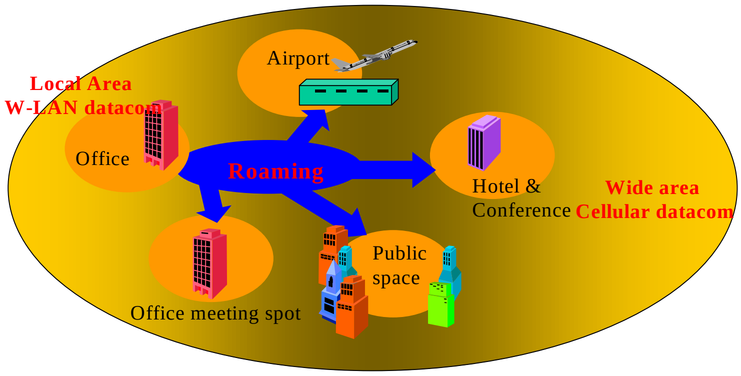
It is possible to make the following assumptions about a corporate environment, which ease the deployment of a secure network:

- Users' credentials can be easily located by means of a centralized database.
- Access to the corporate LAN is controlled. Only authorized users are granted access.
- Trusted system administrators take care of user registration, computer installation, network installation, long-term key generation and storage.
- Out-of-band key distribution, e.g. manual key distribution, may be feasible.
- Protection against unauthorized manipulation of computer hardware and network equipment is provided.

3.2 Public hot spots

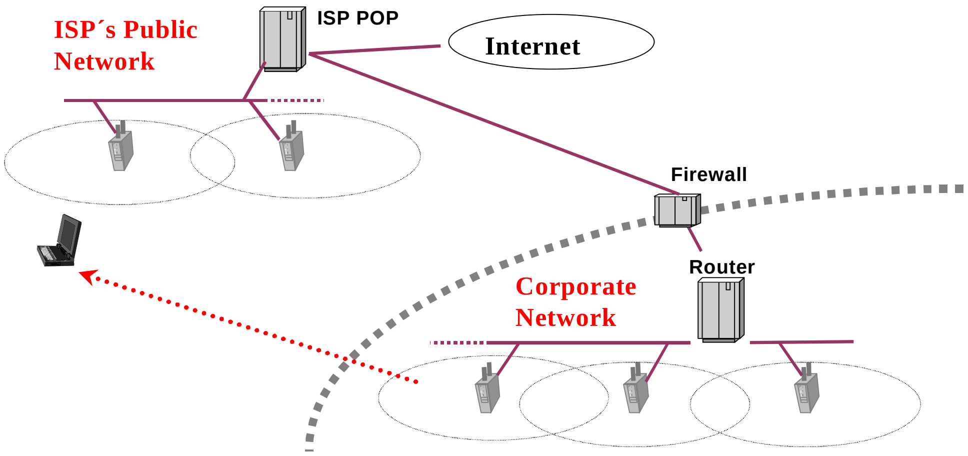
Wireless LAN networks could be deployed at hot spot public areas, e.g. airports, hotels, conference centers, etc., where people may desire to have access to datacom services. This would enable an easy way of offering remote access to the corporate network (VPN) and Internet services to business people. Access to Internet and Intranet has become as vital as voice telephony in business and many corporations may desire to offer datacom services to their employees in these hot-spot areas.

In a near future it is moreover forecasted that public users may as well be strongly interested in wireless datacom access at these hot-spot areas. In this case it is foreseen that Internet will be the main driving service. Specific local services may also be offered to wireless visiting users, e.g. travel information at airports, hotel-facilities at hotels and conference-update information at conference centers.



Wireless LAN usage in hot-spot areas

An access server to which the wireless network is connected may route a connection request either to the corporate network (possibly via a preferred ISP) or perhaps to an ISP for Internet access.



Remote Access to the Corporate Network and to the Internet

It is possible to make the following assumptions about a public hot-spot environment, which strongly influence the deployment of a secure network:

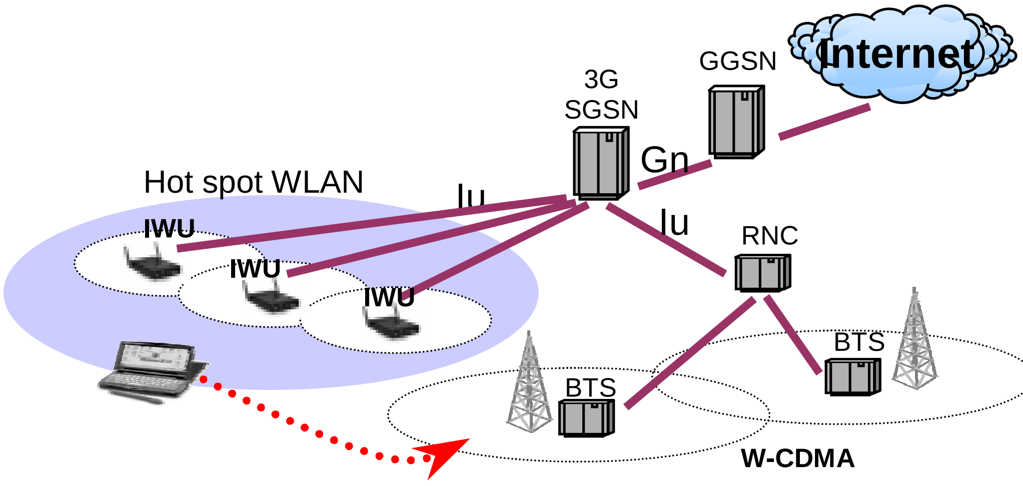
- Users' credentials can not be easily located by means of a centralised database. Some distributed system able to scale is required.
- Access to public hot spots is free and non-authorized users might grant access.
- Out-of-band key distribution, e.g. manual key distribution, is not feasible. Keys must be distributed through a public key infrastructure or some equivalent mean.
- Protection against unauthorized manipulation of computer hardware and network equipment is not always provided.

3.3 Access to 3rd generation cellular network

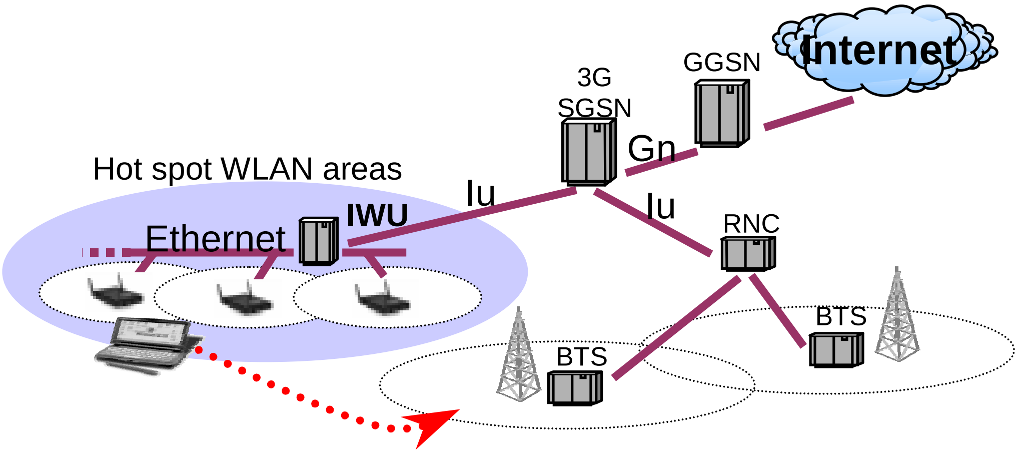
Wireless LAN should be used in the future as an alternative access technology to the 3rd generation cellular network. One may think of the possibility to cover hot spots and city areas with wireless LAN and the wide area with W-CDMA technology.

Combining wireless LANs and W-CDMA networks, a user can benefit from a high-performance network wherever it is feasible to deploy wireless LANs and use W-CDMA elsewhere. The core network will provide to the user automatic and seamless hand-over between the two types of access networks. Future Wireless LANs (e.g. HiperLAN/2) are planned to provide this within the UMTS scope. The virtual private network case, described in section 3.5, might also be part of the 3rd generation solution.

Wireless LAN networks may be either directly connected to the 3rd generation cellular system (through a SGSN) or connected through a wired Ethernet network. The latter would collect the traffic from several WLANs and provide inter-operability with 3rd generation backbone networks. However, these two different inter-connection schemes do not imply any significant change in the overall architecture.



Wireless LAN direct radio access to a 3G Backbone Network



Wireless LAN access to a 3G Backbone Network using a wired Ethernet LAN

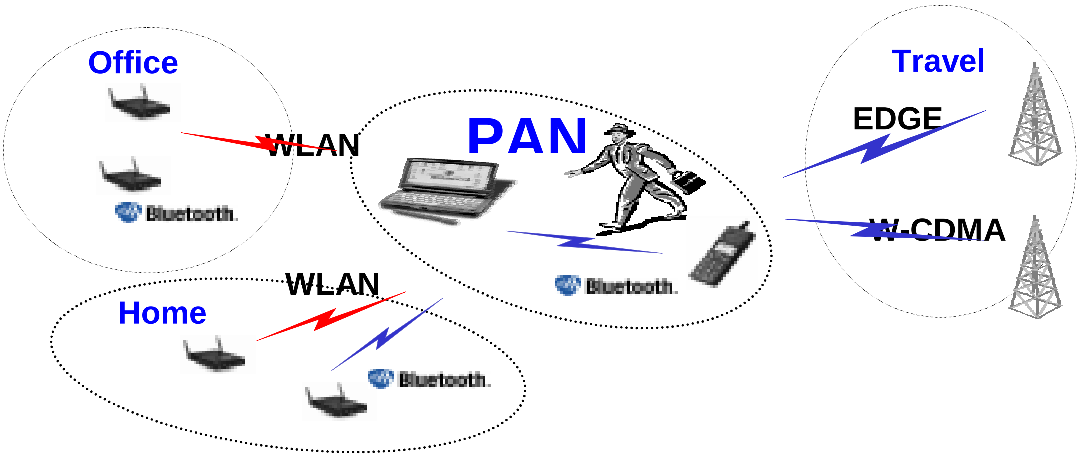
It is possible to make the following assumptions regarding the use of wireless LANs to access 3rd generation networks, which makes this scenario very similar to the use of wireless LANs in any other public environment:

- Users' credentials can not be easily located by means of a centralised database.
- Access to the physical medium is free and non-authorized users might grant access.
- Out-of-band or centralized key distribution, e.g. manual key distribution, is not feasible.
- Protection against unauthorized manipulation of computer hardware and network equipment is usually provided.

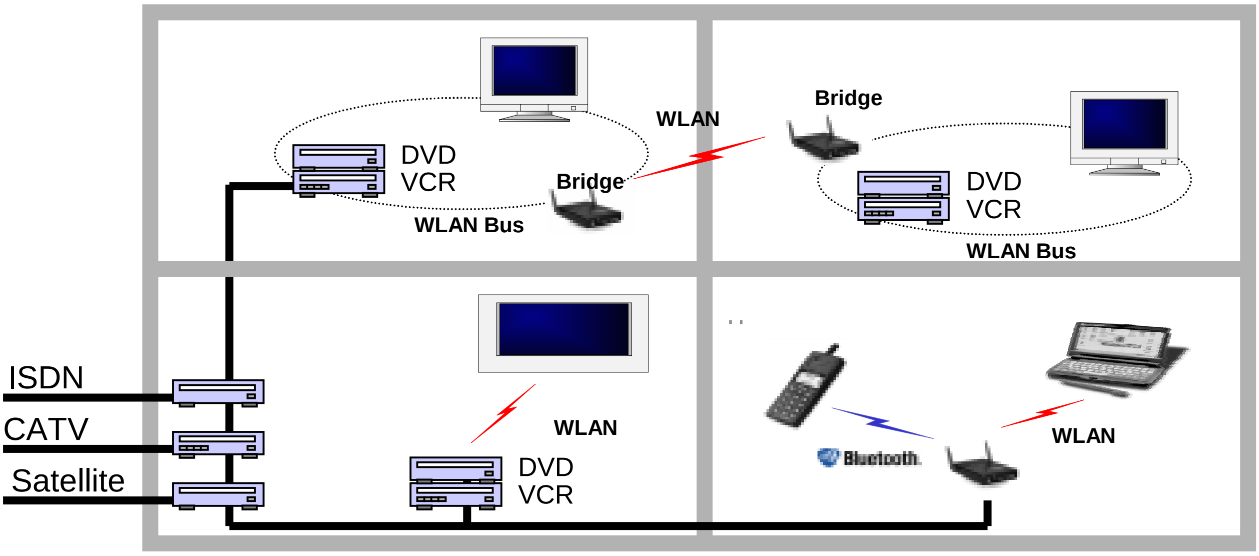
3.4 Home networks

The home environment is another possible example of wireless LANs usage. High-speed access to Internet and multimedia applications (e.g. video and music entertainment distribution) would be in this case the killer applications. Remote home access to the corporate network (i.e. a Virtual Private Network (VPN) solution) and Voice-over-IP (VoIP) may be other driving services.

Wireless LANs are also foreseen to have a major role in interacting with future W-CDMA services and so-called Personal Area Network (PAN), based on Bluetooth technology. Wireless LANs will provide connection in local areas, i.e. hot spot and home areas, while Bluetooth will work as a link medium between different devices and W-CDMA as backbone wide area network.



It may be even possible to create a wireless infrastructure for home devices (e.g. PC, VCRs, cameras, printers, etc). This infrastructure has been named Wireless Firewire. The high throughput and QoS features of future WLANs (i.e. HiperLAN/2) will support the transmission of video streams in conjunction with the datacom applications. The Access Point may in this case include an "uplink" to the public network, e.g. an ADSL or cable modem.



Wireless LAN as extension to an IEEE 1394 (Firewire) network

It is possible to make the following assumptions about the home environment, which simplify the deployment of a secure network:

- Users' credentials can be stored into the access points or easily located by means of a centralized database.
- Access to the home environment is usually controlled but there may be possibility of interference between different home environments, due to the usually short distance between different houses or flats.
- There is not a system administrators, who might take care of user registration, computer installation, network installation, long-term key generation and storage.
- Out-of-band key distribution, e.g. manual key distribution, is feasible.
- Protection against unauthorized manipulation of computer hardware and network equipment is not always provided.

3.5 Virtual Private Networks (VPNs)

It is often required that a mobile user may access its corporate resources from a remote access network. The latter may be either a public, corporate or home network. This kind of remote access is usually named Virtual Private Network (VPN).

In Virtual Private Network (VPN), security must be provided in the same trusted way as in a private network, even if access may be provided through a public access network and data may be transmitted over public links.

The Virtual Private Networks case is not an autonomous network scenario. It is instead a special and indeed critical service, which may be provided over previously described network scenarios. VPNs will be further analyzed in section 5.4.

3.6 Specific Application fields

All scenarios that have been so far described are very general and not relative to any specific application case. However, there are many usage cases where wireless LAN networks may be profitably used. The following is a partial list of some of these specific use cases.

Application oriented scenarios for Wireless LANs:

Places where it is difficult or even impossible to deploy wired connections:
- Historic buildings
- Remote buildings in public area (both building to building bridging and remote access to datacom services)
- Geographic areas inaccessible by wire (e.g. airport & harbor light beacon system, over-rail links in railway stations)

Temporary networks:
- Emergency networks (Emergency service providers' network at disaster site)
- Disaster recovery (Reactivation of affected user's network)
- Exhibitions
- Offices with frequent changes in topology
- Temporary offices
- Point of sale
- Point of entry (e.g. airports/ harbors - ship to shore)

Permanent networks:
- Corporate, Public and Home generic networks
- Conference rooms (Immediate access to wired LAN resources, participant-to-participant transfer of data, presentations accessed from network)
- University auditoriums (Student–students, lecturer–student, lecturer/students-resources)

Networks for mobile users:
- Manufacturing (Fast deployment of carts with on board computers)
- Schools (Mobile carts with PC being moved from classroom to classroom when needed)
- Warehouses (e.g. Barcode readers, mobile retailers)
- Multimedia audio and video distribution
- University campus
- Factory plants
- Hospitals

4 Secure wireless networks

The very traits and characteristics of wireless access as well as mobility aspects call for a security framework. The requirements that have been specified in this document, based on analysis of these characteristics, include both authentication and authorization to protect network resources, encryption for data, and user integrity. Privacy, both regarding content and user location, have been also considered.

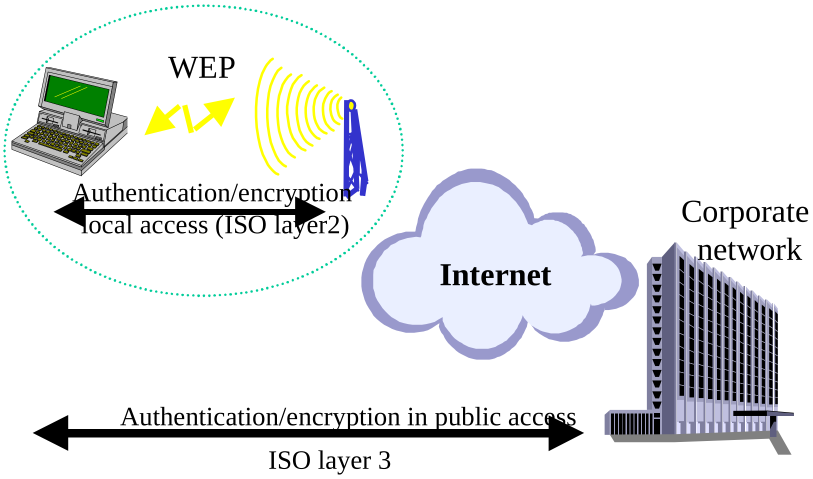
Both WLAN systems deployed today and the Mobile IP solutions specified by IETF implement and define different levels and parts of a security framework. Current solutions lack an analysis of an overall framework regarding the requirements that would apply for private, public, and VPN networking cases, which are quite different in their nature.

Regarding the fact that roaming should be made possible between these environments, and that user and network security requirements must be fulfilled, together with the usability of the system from a user's perspective, the end-to-end solution for all these cases has been provided. When analysing an appropriate solution, it has been taken into consideration existing as well as planned and envisaged security solutions.

4.1 System overview

Security should be transparent both to applications and end-users. It should also be transparent to the transport layer (ISO layer 4), except when it is required it to be visible for a certain service. Therefore, this thesis project will deal with security services either at the network IP layer (ISO layer 3) or at the link layer (ISO layer 2).

Security mechanisms for wireless systems that are provided at the link layer (ISO layer 2) aim to provide so-called Wire Equivalent Privacy (WEP). This means that the wireless physical medium should be as secure as an equivalent wire medium appears to be.


WEP
Authentication/encryption
local access (ISO layer2)
Internet
Corporate network
Authentication/encryption in public access
ISO layer 3

Local and Remote Access to Corporate Network

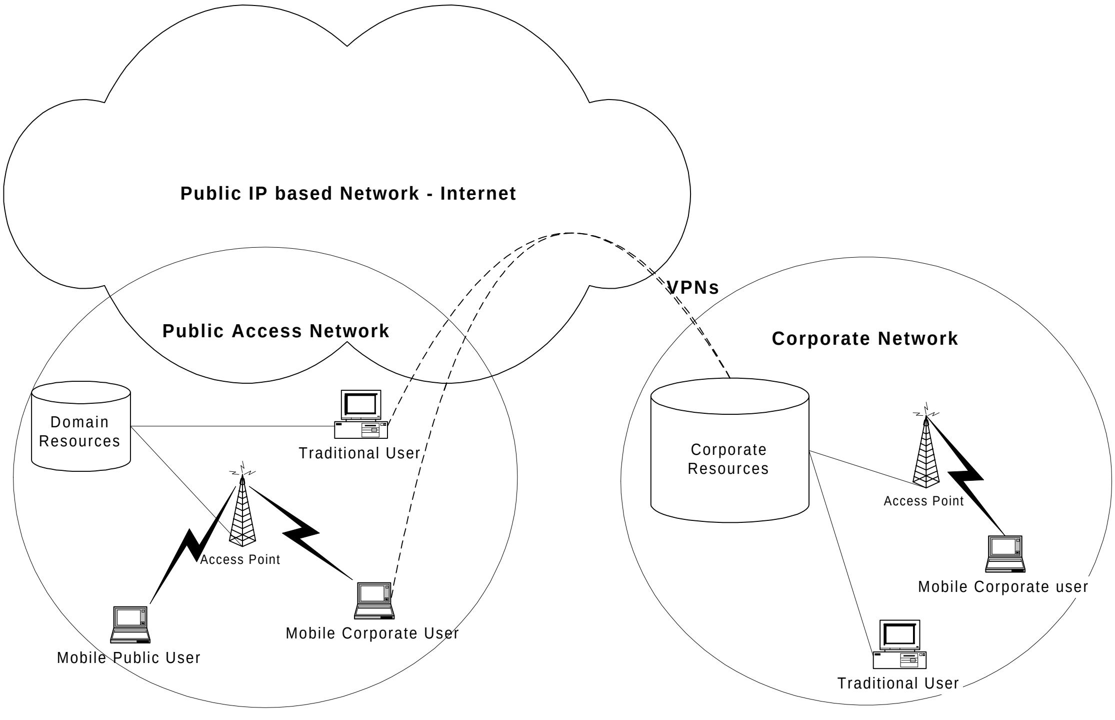
Wire Equivalent Privacy (WEP) is not so easily defined as it would appear, at least in a general environment. Wire security differs very much among different networks and WEP should be network dependent in the same way. This means that WEP for a corporate wireless network (where the corporate wired LAN is usually trusted) may sensibly differ from WEP for the public or home cases (where the wire is not so secure either). Section 5 will define which security services are necessary to be provided for the different network scenarios and especially for the corporate and public environments. The second part of the thesis will present some technical proposals to provide these services for the cases of network access based on IEEE 802.11 and HiperLAN/2 wireless systems.

Security services at the network IP layer (ISO layer 3) might include support for secure IP mobility, roaming between different network domains, the (AAA) accounting infrastructure, end-to-end data confidentiality with authentication, and remote access to Virtual Private Networks (VPNs). The third part of this thesis project will present a suit of protocol proposals, which will be able to provide all previous security services.

Security at ISO layer 2 may be based on authentication of the mobile physical interface. This would be consistent with the fact that the link layer is very related to the specifics of hardware devices, e.g. MAC addresses for DHCP requests. User-based authentication may anyway scale better and be more flexible when inter-domain roaming is provided.

Authentication at ISO layer 3 should be based on users' credentials, given that accounting and authorization data, services provided, and corporate resources are directly related to end-users and often independent from the devices utilized. This is moreover consistent with the fact that ISPs aims to provide connections to users, regardless of the device they use. It would also make it easier to divide mobile devices developers from access service providers, as many companies may be only interested in one of these areas.

Overall Network Architecture



(Visited Domain and Corporate Resources may include also Internet access)

4.2 Security threats

Appendix A at the end of this thesis will describe threats in security and provide a literature overview of present solutions to cope with them. All of these threats refer also to the specific case of WLAN environments. The severity of each threat varies among different networks, i.e. between public, home and corporate networks. Application and user scenarios have also a major role in determining the severity of these. In the following tables threats will be classified, referring to their potential severity, as high, moderate or low risk. When threat severity is highly dependent on specific uses cases, a range value has been provided instead of a fixed value.

Threat severity can be expressed in terms of parameters such as the attacker's capability and motivation, and how often they will try. Only effort in time, money and resources that potential attackers could be available to spend has been valued in defining the severity of each threat. It has been considered neither the vulnerability of the systems nor the ease of security mechanisms. These will be lately addressed and integrated in the risk management analyses. Policy decisions will be taken according to these tables but also to many other tradeoffs. Implementation cost and performance effects will be especially considered when it will be decided which technical solutions will be adopted.

- Threat severity - Wireless corporate networks (risk assessment):

Error and Omissions: low risk.

Users can be trained to behave in a secure way and tends to operate carefully. However, if specific instruments are not considered during the Network Security Policy (NSP) planning, this threat can compromise the security of the complete network.

Insider crimes: moderate-high risk.

Insiders commit 40% of the security violations in a corporation. These are mainly done to gain personal advantages.

Abuse of resources: moderate risk.

The frequency of abuses of resources can be high but the severity of each of these violations is usually small.

Sabotage: moderate risk.

Former or disgruntled employees, as an act of revenge, commit 50% of the network violations in a company. However, these violations are usually less severe than crimes committed for a specific personal purpose.

Hackers: moderate-high risk.

Hackers are an actual danger to any wireless network. They can often damage or sabotage a network but usually do not attack the corporation strategy.

Industrial espionage: high risk.

This can be considered as the most dangerous threat for a corporation, at least in the high technology field. It is often related with the insider threat.

Threats to Personal Privacy: low-moderate risk.

Personal private data should not be exchanged in the corporate network.

- Threat severity - Wireless public networks (risk assessment):

Error and Omissions: moderate-high risk.

Typical private users are well known to be technically incompetent. Any public system should not be confident in its users.

Insider crimes: low risk.

A public network should not have high confidential data that users could possibly reveal to the outside world.

Abuse of resources: high risk.

The situation is quite similar to the corporate case. What makes the difference is the number of users. Moreover, corporate networks are usually structured to tolerate these abuses. This is not the case for public networks.

Sabotage: low-moderate risk.

Sabotage has always to be considered as a possible threat. The lack of former or disgruntled employees decreases the frequency of these violations. However, disgruntled users could desire to sabot the network.

Hackers: moderate risk.

There is not sensible difference with the corporate case.

Industrial espionage: low risk.

Industrial espionage is unlikely to find interesting data on the public network. These data will in case be object of specific end-to-end security policies.

Threats to Personal Privacy: high risk.

Many people concern very much about threats to their personal privacy. This is of utmost importance when an access network is offered to private end users.

As it would have been easy to predict, security threats in corporate networks are usually more severe than in public environments, if no security protection at all is provided. It deserves anyway to be noted that deployment of security mechanisms is usually eased in a corporate environment, due to considerations already presented in section 3.1.

4.3 Security requirements

Security mechanisms are required to avoid the threats presented in section 4.2, at least as long as this can be done in a cost-effective way. When security services are not deployed in a convenient way, they represent system vulnerabilities that an enemy can exploit. The following tables present the most important security services that will be considered for WLAN networks. They will be later coupled with the severity for security threats both in corporate and public networks.

- **Confidentiality**: only the authorized entities can read the message.
- **Authentication of the User**: the network wants to be sure about user’s identity.
- **Authentication of the Access Point**: users want to be sure about network’s identity.
- **Data Integrity**: the receiver wants to be sure that the received message has not been modified.
- **Non-repudiation of origin**: the sender can not deny that he sent a message.
- **Non-repudiation of delivery**: the receiver can not deny that he received the message.
- **Auditing and logging**: both security violations and normal network activities are sensed and stored for security risk planning and incident handling procedures.
- **Denial of Service prevention**: enemies can neither prevent the use of any resources to legitimate users nor decrease system performances.
- **Traffic flow analysis prevention**: it is not possible to infer confidential data by passive attacks, e.g. analysis of traffic volume or destination.

Host security, data-driven attacks prevention, and organizational security policies, e.g. incident handling planes, are out of the scope of the link and network layers security (ISO layer 2 and ISO layer 3) and will not be further addressed in this thesis.

Non-repudiation services are also considered out of the scope of this thesis project. They are application dependent (needed for example in electronic commerce and home banking) and are better implemented on a transport or application security layer. For internal corporate use, for instance, non-repudiation services are usually not important. For a commercial or inter-corporate use they could instead be necessary.

Other technical requirements are defined by specific implementation necessity. However, it is possible to identify some basic characteristics required in all future wireless LAN networks. This refers both to companies with different security concerns but also to public or home networks. Wireless LANs should conform to the following statements:

- Be wired equivalent (WEP), providing link-layer encryption
- Sufficient both in business, residential and public access environment
- Be scalable
- Be flexible, allowing the negotiation of different security levels
- Have a low/moderate processing overhead, e.g. through the use of token SA
- Protect the identity of the mobile user
- Support pre-shared keys but also provide a Public Key Infrastructure (PKI)
- Be exportable
- Not differ or be a superset of security mechanisms in place for wired users

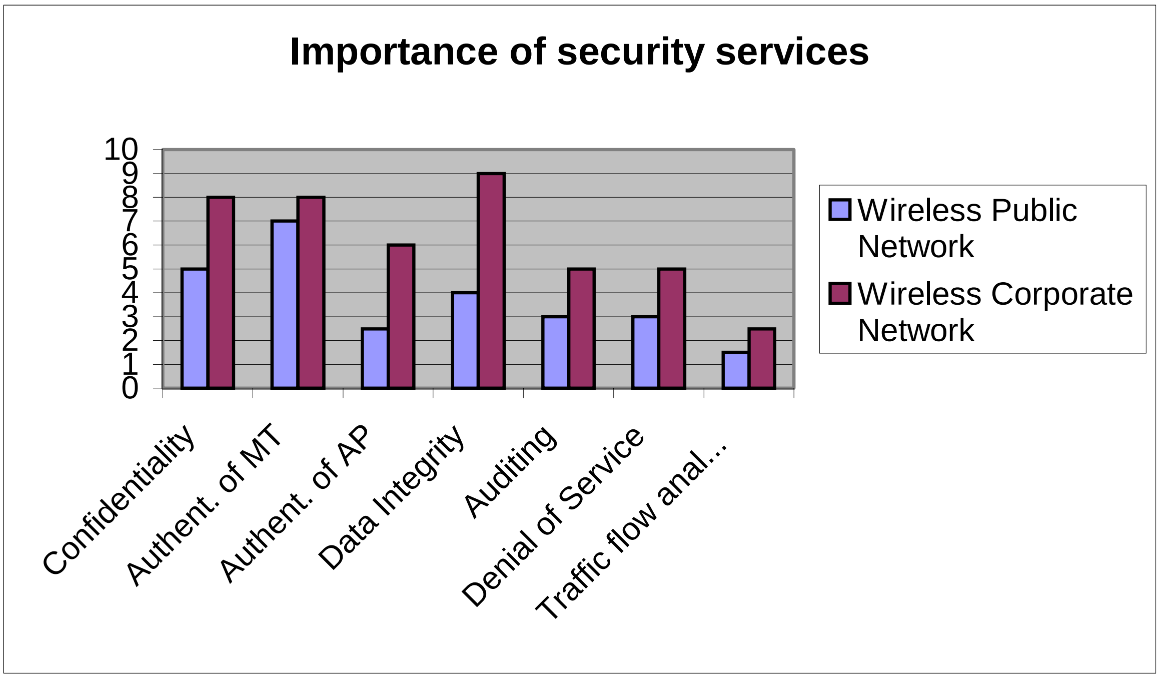
In the next table, security requirements are related with their main possible threats. Confidentiality of the exchanged data and the use of these data determine the type of security services to be adopted. Most of the threats affect all the presented services. Nevertheless, some main correlation can be individuated. Only these most severe threats have been indicated in the table.

The average importance of each security service has been assigned as given by the severity of related threats in that specific environment.

Average Importance for security requirements

Security Services	Most severe Related Threats	Wireless Public Network	Wireless Corporate Network
Confidentiality (Low security level)	Error and Omissions Insider crimes Hackers Industrial espionage Threats to Privacy	7	9
Confidentiality (Mean security level)	Error and Omissions Insider crimes Hackers Industrial espionage	4	8
Confidentiality (High security level)	Insider crimes Industrial espionage	1	4
Authentication Authorized Mobile Terminal (Low security level)	Abuse of resources Error and Omissions Hackers Industrial espionage Theft-of-service	9	9
Authentication Authorized Mobile Terminal (Mean security level)	Abuse of resources Hackers Industrial espionage Theft-of-service	7	8

Authentication Authorized Mobile Terminal (High security level)	Industrial espionage Theft-of-service	1	3
Authentication Authorized Access Point (Low security level)	Hackers Industrial espionage	3	7
Authentication Authorized Access Point (Mean security level)	Hackers Industrial espionage	2	5
Authentication Authorized Access Point (High security level)	Industrial espionage	1	2
Data Integrity (Low security level)	Hackers Sabotage Error and Omissions Industrial espionage	4	9
Data Integrity (Mean security level)	Hackers Sabotage Industrial espionage	3	8
Data Integrity (High security level)	Sabotage Industrial espionage	2	5
Auditing and logging	Insider crimes Sabotage Hackers Industrial espionage	3	6
Denial of Service prevention (Low security level)	Error and Omissions Abuse of resources Sabotage Hackers	4	6
Denial of Service prevention (High security level)	Sabotage Hackers	2	4
Traffic flow analysis prevention (Low security level)	Industrial espionage Threats to Privacy	4	5
Traffic flow analysis prevention (High security level)	Industrial espionage	1	2



5 Functional Requirements in Wireless LANs

From analysis done in previous sections it is now possible to draw some conclusions about the functional security requirements desired in wireless LAN networks. It is important to remember that security services are strongly dependent from the specific environment.

General Network and Security Requirements:

- User mobility must be possible between different IP networks (IP roaming).
- Session mobility must be supported at least between IP subnetworks and possibly also on a wider scope (the latter should be provided at least when 3rd generation network will be deployed).
- A remote mobile user, accessing the network through a visited domain, should be able to access corporate resources with the same security confidence as of any traditional remote user, i.e. with Virtual Private Network (VPN) end-to-end security. Other end-to-end security needs, e.g. banking services, should be supported in a similar way.
- Mobility may be implemented with an optimized solution, combining security means with mobility ones (e.g. Mobile IP tunnels with IPSec tunnels).
- A mobile user accessing the network through a wireless LAN system must have the same security protection as of traditional non-mobile users, i.e. must have Wire Equivalent Privacy (WEP). This, as already described, means that WEP requirements may be different, as long as wired networks security is also different, i.e., between corporate and public networks.
- An Authentication, Authorization and Accounting (AAA) verification infrastructure has to be supported to allow effective roaming between different domains.
- Protection from theft-of-service must be provided to visited networks.
- Any security mechanism has to be supported by a scalable and flexible Public Key Infrastructure (PKI) system for delivery of needed secrets.

5.1 Wireless Corporate Access network

Most of the data transmitted within a corporate network may be confidential. This calls for a secure mutual authentication of both the Mobile User and the Access Point. Data confidentiality should also be provided at the link layer.

The corporate environment allows, to some extent, the use of pre-shared keys distributed with out-of-band mechanisms, e.g. manually. The following lists the most important security and network requirements for a general corporate network.

A Wireless Corporate Access network,

MUST provide the following security services:

- Strong Confidentiality.
- Authentication of the user.

SHOULD provide the following security services:

- Protection from Denial of Service attacks.
- Authentication of the AP.
- Data Integrity.

MAY provide the following security services:

- Auditing and logging services.
- Traffic flow analysis prevention.

Additional security requirements:

- A Public Key Infrastructure (PKI) MAY co-exist with Out-of-band key distribution.
- Some AAA verification infrastructure (e.g. RADIUS and DIAMETER) MAY be integrated with the overall system.
- Security MUST be flexible, allowing the negotiations of different security levels.
- It SHOULD offer keys backup and recovery procedures.
- It MAY be possible for the access network to disable the authentication of the MT.
- It SHOULD NOT be possible for a user to disable the authentication of the AP.
- Cost and usability SHOULD be valued as important as network link layer security.
- Confidentiality and authentication of the mobile user SHOULD be considered as the main security issues.

Network requirements:

- Both link-layer and IP mobility MUST be provided.
- Network throughput and QoS SHOULD NOT be sensibly degraded from the introduction of the security mechanisms.
    - It MUST be wire compatible, i.e. ease inter-operability with wire networks.
    - There SHOULD be low/moderate security processing overhead.
    - Hand-over latency SHOULD be minimized.
- It SHOULD provide Automatic radio network planning, i.e. be Plug & Play

5.2 Wireless Public Access network

Most of data transmitted within a public network may be not highly confidential. This calls for an authentication scheme that is mainly needed out from subscription and accounting requirements. Strong data confidentiality has to be provided only for specific usage cases, e.g. VPN or remote banking. Some limited confidentiality may anyway be required as a basic service for all the data and especially for users' privacy. Theft-of-service should be considered as one of the most severe threats.

The public environment does not allow the use of pre-shared keys, distributed with out-of-band mechanisms, e.g. manually. Some scalable Public Key Infrastructure must be provided. The following lists the most important security and network requirements for a general corporate network.

A Wireless Public Access network,

MUST provide the following security services:

- Authentication of the user.

SHOULD provide the following security services:

- Confidentiality.
- Protection from Denial of Service attacks.

- Data Integrity.

MAY provide the following security services:

- Auditing and logging services.
- Authentication of the AP.
- Traffic flow analysis prevention.

Additional security requirements:

- A Public Key Infrastructure (PKI) MUST be used for key distribution.
- Some AAA verification infrastructure (e.g. RADIUS and DIAMETER) SHOULD be integrated with the overall system.
- It MAY be flexible, allowing the negotiations of different security levels.
- It SHOULD offer keys backup and recovery procedures.
- It SHOULD be possible for the access network to disable the authentication of the MT and for a user to disable the authentication of the AP.
- Cost and usability SHOULD be valued as more important requirements than network and link layer security. Any sensible data SHOULD be protected through an end-to-end security policy (e.g. between the mobile user and the corporate network gateway).
- Authentication of the mobile user SHOULD be considered as the main security issue. This to allow for correct billing procedures, prevent theft-of-service violations and to give some light protection from denial of service attacks.

Network requirements:

- Both link-layer and IP mobility MUST be provided.
- Inter-domain roaming SHOULD be provided.
- Network throughput and QoS SHOULD NOT be sensibly degraded from the introduction of the security mechanisms.
  - It MUST be wire compatible, i.e. ease inter-operability with wire networks.
  - There MUST be low/moderate security processing overhead.
  - Hand-over latency SHOULD be minimized.
- It MUST provide Automatic radio network planing, i.e. be Plug & Play

5.3 Wireless Home Access network

Most of the data transmitted within a home network are not confidential. This calls for a light authentication procedure that should be always provided (to protect for example from unauthorized access from neighbors) and a strong authentication scheme that should be only utilized for specific transaction, e.g. VPN or banking. Strong data confidentiality has also to be provided only for specific cases. Some limited confidentiality may anyway be required as a basic service for all the data and especially for users' privacy.

The home environment allows the use of pre-shared keys, distributed with out-of-band mechanisms, e.g. manually when the wireless network is installed. The following lists the most important security and network requirements for a general corporate network.

A Wireless Home Access network,

MUST provide the following security services:

- Authentication of the MT.

SHOULD provide the following security services:

- Basic confidentiality.

MAY provide the following security services:

- Protection from Denial of Service attacks.
- Data Integrity.
- Auditing and logging services.
- Authentication of the AP.
- Traffic flow analysis prevention.

Additional security requirements:

- Out-of-band key distribution (pre-shared keys) SHOULD be preferred.
- It MAY be flexible, allowing the negotiations of different security levels.
- It SHOULD offer keys backup and recovery procedures.
- It SHOULD be possible for the access network to disable the authentication of the MT and for a user to disable the authentication of the AP.
- Cost and usability SHOULD be valued as more important requirements than network link layer security. Any sensible data SHOULD be protected through an end-to-end security policy (e.g. between the mobile user and the corporate network gateway).
- Authentication of the mobile user SHOULD be considered as the main security issue. This to allow for correct billing procedures and to give some light protection from denial of service attacks.

Network requirements:

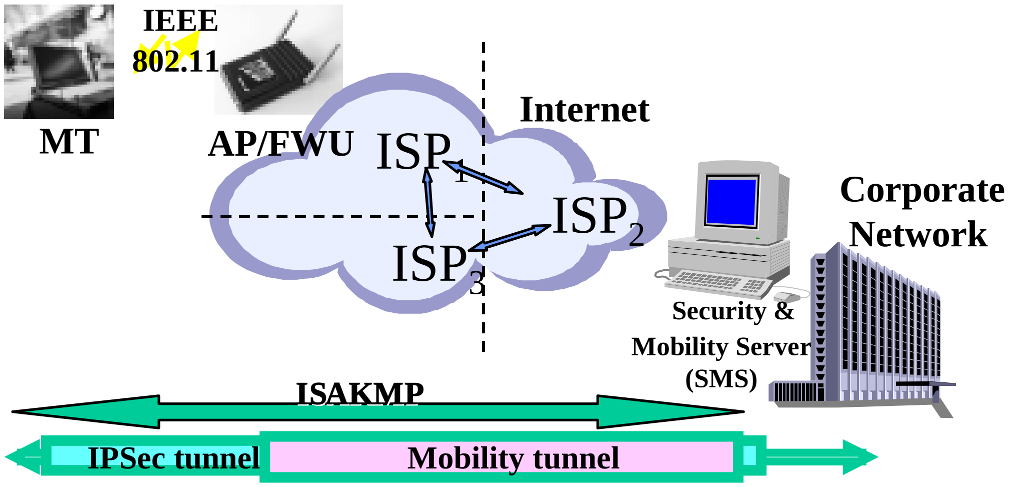
- It MUST be low cost.
- Link-layer and IP mobility SHOULD not be provided.
- High throughput and QoS SHOULD be considered as mandatory (e.g. video).
  - It MUST be wire compatible, i.e. ease inter-operability with wire networks.
- It MUST provide Automatic radio network planing, i.e. be Plug & Play
- It must provide inter-operability with audio & video equipment.

5.4 Virtual Private Networks (VPNs)

Wireless LAN networks may be used to provide remote access to corporate network. Most of data transmitted between a remote user and the corporate network may be highly confidential. Strong data confidentiality and authentication have to be provided, both for the case when the user is accessing the network from an home environment and when it is connected through a public hot-spot.

A set of security services is required for the case of Virtual Private Networks:

- Data confidentiality (e.g. using an IPSec tunnel) between the user and the corporate gateway (typically a Firewall Unit, FWU)
- Public-key based authentication between the user and the FWU.
- Confidentiality and integrity for the wireless traffic.
- Security association (SA) & session key management mechanisms.
- Secure SA transfer during hand-over.



Mobility and Security in VPN

Remote access to the corporate environment (i.e. Virtual Private Network services) is usually granted to the same users who have access to the corporate environment. This allows, to some extent, the use of pre-shared keys distributed with out-of-band mechanisms, e.g. manually and then used for remote connection. The following lists the most important security and network requirements for VPNs.

A Virtual Private Network,

MUST provide the following security services:

- Strong Confidentiality.
- Authentication of the user.
- Authentication of the corporate gateway.
- Data Integrity.

SHOULD provide the following security services:

- Protection for the corporation from Denial of Service attacks.
- Auditing and logging services.

MAY provide the following security services:

- Traffic flow analysis prevention.

Additional security requirements:

- A Public Key Infrastructure (PKI) MAY co-exist with Out-of-band key distribution.
- Security MUST be flexible, allowing the negotiations of different security levels.
- It SHOULD offer keys backup and recovery procedures.
- Cost and usability SHOULD be valued as important as network link layer security.

Network requirements:

- Both link-layer and IP mobility SHOULD be provided.
- Network throughput and QoS SHOULD NOT be sensibly degraded from the introduction of the security mechanisms.
- There SHOULD be low/moderate security processing overhead.
- Hand-over latency SHOULD be minimized.

- PART 2 -

Security for IEEE 802.11 & HiperLAN/2 Wireless LAN Networks

6 The IEEE 802.11 standard

The IEEE 802.11 standard [50] defines a Physical Layer (PHY) and a Medium Access Control (MAC) protocol for wireless LAN networks, which are local area networks (LANs) implemented as an extension to or as an alternative to wired LANs. Using radio frequency (RF) technology, wireless LANs transmit and receive data over the air, minimizing the need for wired connections.

Any 802.11-based wireless LAN appears as an ordinary 802 LAN to the logical link control (LLC) layer. The IEEE 802.11 standard also enables mobility within the MAC layer. Beyond the standard functionality usually performed by MAC Layers, the MAC protocol performs other functions that are usually related to upper layer protocols, such as fragmentation, packet retransmissions, and acknowledges. This is done in order to meet the assumptions that LLC makes about its lower layers. WLAN 802.11 systems have a connection-less structure.

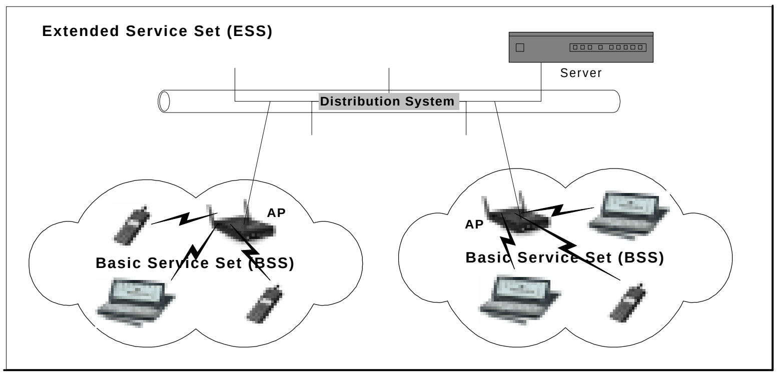
Some of the most important characteristics for IEEE 802.11 networks are hereby listed. They will be furtherly described in the rest of chapter one.

- 802.11 networks can work at the 2,4 GHz ISM band, which is unlicensed.
- 1-2 Mbps throughput for standard (3 Mbps proprietary)
- Coverage Scope: 150m indoors & 600m outdoors
- It is an IEEE standard since 1997
- Specifies 3 physical interfaces (DSSS, FHSS, IR)
- 11 Mbps throughput in DSSS for 802.11a extension

6.1 Network architectures

IEEE 802.11 wireless LANs are based on a cellular architecture where the system is subdivided into cells. Each cell (called Basic Service Set, or BSS) is controlled by a Base Station (called Access Point or, in short, AP). Although a wireless LAN may be formed by a single cell with a single Access Point (and it could also work without an Access Point), most installations are formed by several cells, where the Access Points are connected through some kind of backbone network (called Distribution System or DS). This backbone is typically Ethernet and, in some cases, is wireless itself.

The whole interconnected Wireless LAN, including the different cells, their respective Access Points and the Distribution System, is seen as a single 802 network to the upper layers of the OSI model and is known in the Standard as Extended Service Set (ESS).

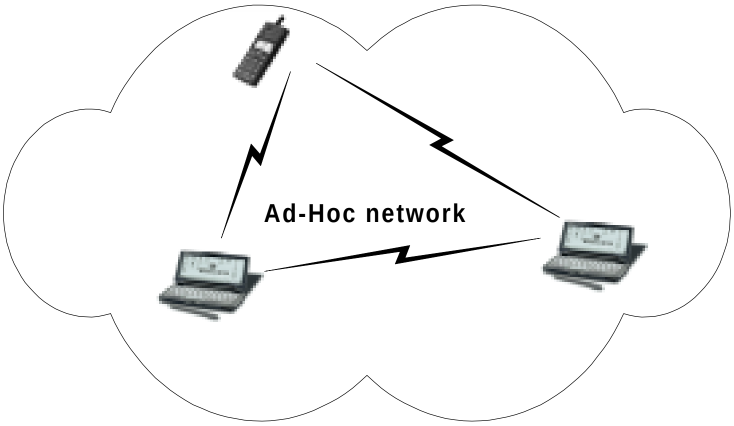


802.11 standard also defines the concept of a Portal. A portal is a device that interconnects an 802.11 with another 802 LAN. This concept is an abstract description of part of the functionality of a “translation bridge”. Even though the standard does not necessarily request it, typical installations have the AP and the Portal on a single physical entity. This is also the case with Ericsson’s Access Point (AP) product, which provides both functions.

Two network architectures are defined in the IEEE 802.11 standard: the Infrastructure Network and the Ad Hoc Network.

6.1.1 Ad-hoc networks

An Ad Hoc network is an architecture that is used to support mutual communication among wireless clients. Typically created spontaneously, an ad hoc network does not support access to wired networks, and does not need an AP to be part of the network. In the ad-hoc network, computers are in this way brought together to form a network “on the fly”. There is no structure into the network, no fixed points, and usually every node is able to communicate with every other node.



Although it seems that order would be difficult to maintain in this type of network, algorithms such as the spokesman election algorithm (SEA) have been designed to "elect" one machine as the base station (master) of the network with the others being slaves. Another algorithm in ad-hoc network architectures uses a broadcast and flooding method to all other nodes to establish who's who.

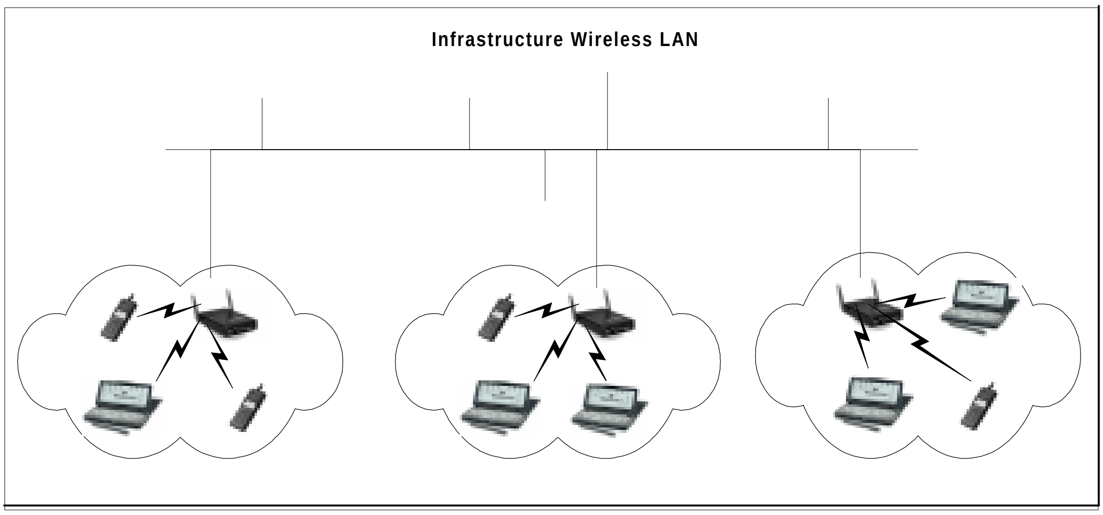
A good example of ad-hoc networks usage scenario might be a business meeting where employees bring laptop computers together to communicate and share design or financial information.

6.1.2 Infrastructure networks

The second type of network structure used in wireless LANs is the Infrastructure. An Infrastructure Network provides communication between wireless clients and wired network resources. The transition of data from the wireless to the wired medium is via an Access

Point (AP). The coverage area is defined by an Access Point and its associated wireless clients, and together all the devices form a Basic Service Set.

This architecture uses fixed network access points with which mobile nodes can communicate. These network access points are sometime connected to landlines to widen the LAN's capability by bridging wireless nodes to other wired nodes. If service areas overlap, handoffs can occur. This structure is very similar to the present day cellular networks around the world.



6.2 IEEE 802.11 Layers

The IEEE 802.11 standard defines the parameters for both the physical (PHY) and medium access control (MAC) layers of the network (as any 802.x protocol). The standard currently defines a single MAC protocol, which interacts with three PHYs (all of them running at 1 and 2 Mbit/s) as follows:

- Frequency Hopping Spread Spectrum (FHSS) in the 2.4 GHz Band
- Direct Sequence Spread Spectrum (DSSS) in the 2.4 GHz Band, and
- Infrared (IR)

Ericsson uses Frequency Hopping (FHSS) in its current products. The FHSS and DSSS layers are used to a fairly equal ratio by other vendors. Interoperability is possible only among 802.11 products using the same type of physical layer.

Beyond the standard functionality usually performed by MAC Layers, the 802.11 MAC performs other functions that are typically related to upper layer protocols, such as Fragmentation, Packet Retransmissions, and Acknowledges.

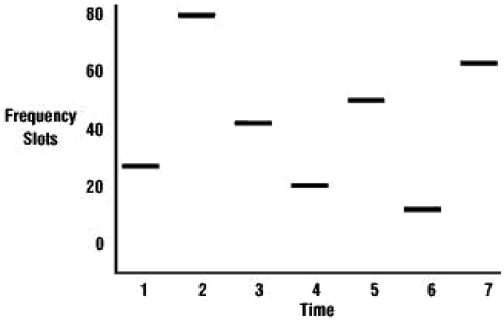
6.2.1 Physical layer

The physical layer (PHY), which handles the transmission of data between nodes, can use either direct sequence spread spectrum (DSSS), frequency-hopping spread spectrum (FHSS), or infrared pulse position modulation. The first two techniques are examples of spread-spectrum technology implementation.

Spread-spectrum technology is a wideband Radio Frequency (RF) technique developed by the military for use in reliable, secure, mission-critical communications systems. It is designed to trade off bandwidth efficiency for reliability, integrity, and security. In other words, more bandwidth is consumed than in the case of narrowband transmission, but the tradeoff produces a signal that is, in effect, louder and thus easier to detect, provided that the receiver knows the parameters of the spread-spectrum signal being broadcast. If a receiver is not tuned to the right frequency, a spread-spectrum signal looks like background noise.

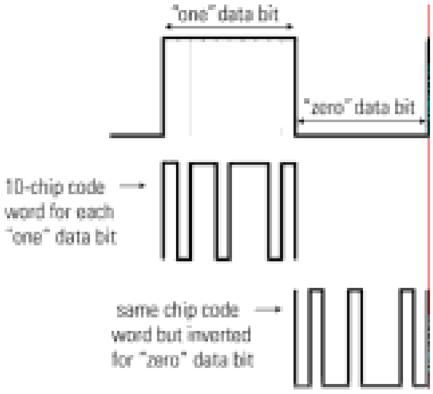
Opposed to the spread-spectrum technology, a narrowband radio system transmits and receives user information on a specific radio frequency. Narrowband radio keeps the radio signal frequency as narrow as possible just to pass the information. Undesirable crosstalk between communications channels is avoided by carefully coordinating different users on different channel frequencies. A private telephone line is much like a radio frequency. When each home in a neighborhood has its own private telephone line, people in one home cannot listen to calls made to other homes. In a radio system, privacy and noninterference are accomplished by the use of separate radio frequencies. The radio receiver filters out all radio signals except the ones on its designated frequency.

Coming back to the spread-spectrum technologies, which are used in wireless LANs 802.11, frequency-hopping spread-spectrum (FHSS) uses a narrowband carrier that changes frequency in a pattern known to both transmitter and receiver. Properly synchronized, the net effect is to maintain a single logical channel. To an unintended receiver, FHSS appears to be short-duration impulse noise.



Frequency Hopping Spread Spectrum

Direct-sequence spread-spectrum (DSSS) generates a redundant bit pattern for each bit to be transmitted. This bit pattern is called a chip (or chipping code). The longer the chip, the greater the probability that the original data can be recovered (and, of course, the more bandwidth required). Even if one or more bits in the chip are damaged during transmission, statistical techniques embedded in the radio can recover the original data without the need for retransmission. To an unintended receiver, DSSS appears as low-power wideband noise and is rejected (ignored) by most narrowband receivers.


"one" data bit
"zero" data bit
10-chip code word for each "one" data bit
same chip code word but inverted for "zero" data bit

Direct-sequence Spread Spectrum

The Infrared (IR) technology is the last possibility for the 802.11 PHY layer. IR systems use very high frequencies, just below visible light in the electromagnetic spectrum, to carry data. Like light, IR cannot penetrate opaque objects; it is either directed (line-of-sight) or diffuse technology. Inexpensive directed systems provide very limited range and typically are used for PANs but occasionally are used in specific WLAN applications. High performance directed IR is impractical for mobile users and is therefore used only to implement fixed subnetworks. Diffuse (or reflective) IR WLAN systems do not require line-of-sight, but cells are limited to individual rooms.

Infrared is generally considered to be more secure to eavesdropping, because IR transmissions require absolute line-of-sight links (no transmission is possible outside any simply connected space or around corners), as opposed to radio frequency transmissions, which can penetrate walls and be intercepted by third parties unknowingly. However, infrared transmissions can be adversely affected by sunlight, and the spread-spectrum protocol of 802.11 does provide some rudimentary security for typical data transfers.

IEEE 802.11 makes provisions for data rates of either 1 Mbps or 2 Mbps, and calls for operation in the 2.4 - 2.4835 GHz frequency band (in the case of spread-spectrum transmission), which is an unlicensed band for industrial, scientific, and medical (ISM) applications, and 300 - 428,000 GHz for IR transmission. Being unlicensed, the ISM band can be used for operating wireless LAN devices without the need for end-user licenses. In order for wireless devices to be interoperable they have to conform to the same PHY and MAC standards.

6.2.2 MAC layer

The 802.11 MAC layer, supported by an underlying PHY layer, is concerned primarily with rules for accessing the wireless medium. The MAC layer specification has similarities to the 802.3 Ethernet wired line standard. The 802.11 standard uses a protocol scheme known as carrier-sense, multiple access, collision avoidance (CSMA/CA) protocol. This protocol avoids collisions instead of detecting a collision like the algorithm used in 802.3.

Collision detection, as is employed in Ethernet, cannot be used for the radio frequency transmissions of IEEE 802.11. Implementing a Collision Detection Mechanism in a Radio frequency system would require the implementation of a Full Duplex radio capable of transmitting and receiving at once, an approach that would increase the price significantly.

Moreover, in a wireless environment we can not assume that all stations hear each other, which is the basic assumption of the Collision Detection scheme. This means that, even if a station wants to transmit and senses the transmission medium as free, it doesn't necessarily mean that the medium is free around the receiver area.

In CSMA/CA protocol, when a node receives a packet to be transmitted, it first listens to ensure no other node is transmitting. If the channel is clear, it then transmits the packet. The Collision Avoidance (CA) mechanism is used together with a Positive Acknowledge scheme, as follows:

- A station wanting to transmit senses the medium. If the medium is busy then it defers. If the medium is free for a specified time (called Distributed Inter Frame Space), then the station is allowed to transmit.
- The receiving station checks the CRC of the received packet and sends an acknowledgment packet (ACK). Receipt of the acknowledgment indicates to the transmitter that no collision occurred. If the sender does not receive the acknowledgment then it retransmits the fragment until it receives acknowledgment or is thrown away after a given number of retransmissions.

If the channel is not free, the transmitter chooses a random "backoff factor" which determines the amount of time the node must wait until it is allowed to re-transmit its packet. During periods in which the channel is clear, the transmitting node decrements its backoff counter. (When the channel is busy it does not decrement its backoff counter.) When the backoff counter reaches zero, the node transmits the packet. Since the probability that two nodes will choose the same backoff factor is small, collisions between packets are minimized.

Another optimization is also specified in the IEEE 802.11 MAC standard. In order to reduce the probability of two stations colliding because they cannot hear each other, the standard defines a Virtual Carrier Sense mechanism. Whenever a packet is to be transmitted, the transmitting node first sends out a short ready-to-send (RTS) packet containing information on the length of the packet. If the receiving node successfully hears the RTS (as determined by a cyclic redundancy check), it responds with a short clear-to-send (CTS) packet. After this exchange, the transmitting node sends its packet.

All stations receiving either the RTS and/or the CTS, set their Virtual Carrier Sense indicator (called NAV, for Network Allocation Vector), for the given duration, and use this information together with the Physical Carrier Sense when sensing the medium. This mechanism reduces the probability of a collision on the receiver area by a station that is “hidden” from the transmitter to the short duration of the RTS transmission because the station hears the CTS and “reserves” the medium as busy until the end of the transaction.

The duration information in the RTS also protects the transmitter area from collisions during the ACK (from stations that are out of range of the acknowledging station). It should also be noted that, due to the fact that the RTS and CTS are short frames, the mechanism also reduces the overhead of collisions, since these are recognized faster than if the whole packet was to be transmitted.

Fragmentation and reassembly are also addressed in IEEE 802.11 MAC standard. Typical LAN protocols use packets several hundred bytes long (the longest Ethernet packet could be up to 1518 bytes long). There are several reasons why it is preferable to use smaller packets in a Wireless LAN environment, as:

- Due to the higher Bit Error Rate of a radio link, the probability of a packet getting corrupted increases with the packet size.
- In case of packet corruption (either due to collision or noise), the smaller the packet, the less overhead it causes to retransmit it.
- On a Frequency Hopping system, the medium is interrupted periodically for hopping (in our case every 20 milliseconds), so, the smaller the packet, the smaller the chance that the transmission will be postponed after dwell time.

However, it doesn't make sense to introduce a new LAN protocol that cannot deal with packets 1518 bytes long which are used on Ethernet, so the IEEE 802.11 committee decided to solve the problem by adding a simple fragmentation and reassembly mechanism at the MAC Layer.

Power management is supported at the MAC level for those applications requiring mobility under battery operation. Provisions are made in the protocol for the portable stations to go to low power "sleep" mode during a time interval defined by the base station. IEEE 802.11 defines two power modes, an Active Mode, where a wireless client is powered to transmit and receive, and Power Save mode, where a client is not able to transmit or receive, but consumes less power. Actual power consumption is not defined and is dependent upon the implementation.

6.3 Communication phases

When a 802.11 wireless device wants to access an existing BSS (either after power-up, sleep mode, or just entering the BSS area), the device needs to get synchronization information from the Access Point (or from the other stations when in ad-hoc mode) to ensure that all stations are operating with the same parameters. The station can get this information by one of two means:

- Passive Scanning: in this case the station just waits to receive a Beacon Frame from the AP (the beacon frame is a frame sent out periodically by the AP containing synchronization information).
- Active Scanning: in this case the station tries to locate an Access Point by transmitting Probe Request Frames, and waits for Probe Response from the AP.

Both methods are valid and the method used is chosen according to the power consumption and performance trade-off. Beacon frames transmitted from APs contain information about the physical layer, as for instance the frequency-hopping sequence for FHSS or the spreading code for DSSS, as well as and the AP’s clock value and the BSS id. Each station receiving a beacon frame will update its clock and use the indicated frequency-hopping sequence or spreading code.

6.3.1 The Authentication Process

Preventing access to network resources is done by the use of an authentication mechanism where a station needs to prove knowledge of a shared secret. In IEEE 802.11 networks, this process takes place before that a wireless client is associated with an AP.

By default, IEEE 802.11 devices operate in an Open System, where essentially any wireless client can associate with an AP without the checking of credentials. True authentication is done with the use of the 802.11 option known as Wired Equivalent Privacy (WEP), where a Shared Key is configured into the AP and its wireless clients. Only those devices with a valid Shared Key will be allowed to be associated to the AP.

This is very similar to Wired LAN privacy, in the sense that an intruder needs to enter the premises (by using a physical key) in order to connect his workstation to the wired LAN. The authentication process is done once the mobile station has located an Access Point and decided to join its BSS. This consists in an interchange of information between the AP and the station, where each side proves the knowledge of a given password (for possible authentication schemes see section 4.1 in Part 2).

Once the station is authenticated, it then starts the real association process, which is the exchange of information about the stations and BSS capabilities, and which allows the ESS (the set of APs) to know about the current position of the station. This information is needed by the distributed system so it knows which AP to access when delivering messages to a given station.

Each station can only be associated with a single AP, but APs can be associated with multiple stations. A station is capable of transmitting and receiving data frames only after the association process is completed.

7 Security mechanisms for 802.11 networks

IEEE 802.11 networks are often based on spread spectrum technology, which (first introduced around 50 years ago by the military) improves both message integrity and security. Moreover, the IEEE 802.11 standard includes a security technique known as “wired equivalent privacy” (WEP), which is based on the RC4 encryption algorithm (Part 2, section 3.4.3) with a key of 64-bit. Authentication of the mobile device is also included in the 802.11 standard.

Wireless LANs eliminate significant amounts of wire from a given installation and dramatically reduce the number of places for unauthorized users and especially insiders to gain access to the wired physical plant. Moreover, since the access points used in wireless LANs function as bridges, individual wireless users are isolated from perhaps the majority of LAN traffic, thus limiting user access to network packets.

Despite the previous considerations, security in IEEE 802.11 has several drawbacks. RC4 is not considered as a particularly secure encryption algorithm and strong authentication schemes of the users are not specified in the standard. This often creates the need of introducing other means to increase security in wireless LANs. Ericsson high-security solution, for example, is based on dedicated devices named WLAN Guards (WLG), which make use of IETF protocols IKE and IPSec.

7.1 Standard 802.11 security

The IEEE 802.11 standard includes a security technique known as "wired equivalent privacy" (WEP), which is based on the use of an authentication procedure and 64-bit keys with RC4 encryption algorithm. The use of spread-spectrum radio transmission techniques also increases the security in the 802.11 transmission medium.

7.1.1 Security characteristics of Spread-spectrum technology

Most wireless LANs use spread-spectrum radio transmission techniques. Spread spectrum technology was first introduced about 50 years ago by the military with the objective of improving both message integrity and security. Therefore, spread-spectrum systems are designed to be resistant to noise, interference, and especially unauthorized detection. Both direct sequence (DS) and frequency hopping (FH) techniques present unintended receivers with a difficult problem.

In the case of DS, an eavesdropper must know the chipping (spreading) code. Someone trying to intercept an FH transmission must know the hopping pattern. FH is difficult to detect and decode because the signal hops from frequency to frequency in a random, repetitive sequence. For successful communications to take place, both transmitter and receiver must be synchronized, using the same sequence. The short duration of time a transmitter usually stays at a given frequency and the little amount of time the transmitter takes to hop to the next frequency also complicate the task of decoding this signaling. In both cases, the specific frequency band and the modulation techniques in use must also be known.

7.1.2 Wired Equivalent Privacy (WEP)

Security provisions are addressed in the 802.11 standard as an optional feature for those concerned about eavesdropping. The data security is provided with an encryption technique, which is used to accomplish Wired Equivalent Privacy (WEP).

By default, data is transferred “in clear” and any 802.11-compliant device can potentially eavesdrop PHY 802.11-like traffic that is within its range. The WEP option encrypts data before it is sent wirelessly. The same Shared Key used in authentication is used to encrypt or decrypt the data; thus only wireless clients with the exact Shared Key can correctly decipher the data.

The WEP algorithm is a Pseudo Random Number Generator (PRNG) cipher initialized by a shared secret key (Part 2, section 3.4). This PRNG outputs a key sequence of pseudo-random bits equal in length to the largest possible packet, which is combined with the outgoing/incoming packet producing the packet transmitted in the air. The cryptographic algorithm is based on RSA's RC4, which has the following properties:

- Reasonably strong: brute-force attacks to this algorithm are not very easy (although possible) because every frame is sent with an Initialization Vector, which restarts the PRNG for each frame.
- Self-synchronized: the algorithm re-synchronizes for each message. This is necessary in order to work in a connection-less environment, where packets may get lost.

Wired Equivalent Privacy (WEP), when enabled, only protects the data packet information and does not protect the physical layer header so that other stations on the network can listen to the control data needed to manage the network. However, the other stations cannot decrypt the data portions of the packet.

7.2 Ericsson´s high-security solution for WLAN 802.11

RC4 with a 64 bit key (the encryption method specified in 802.11 standard) has been broken multiple times as only 64 MIPS years are needed. Therefore, it is doubtful that this algorithm might be considered sufficiently strong to provide data confidentiality in high-security environments. Strong encryption methods such as RC5 are 3DES should be used to protect the integrity of highly confidential data.

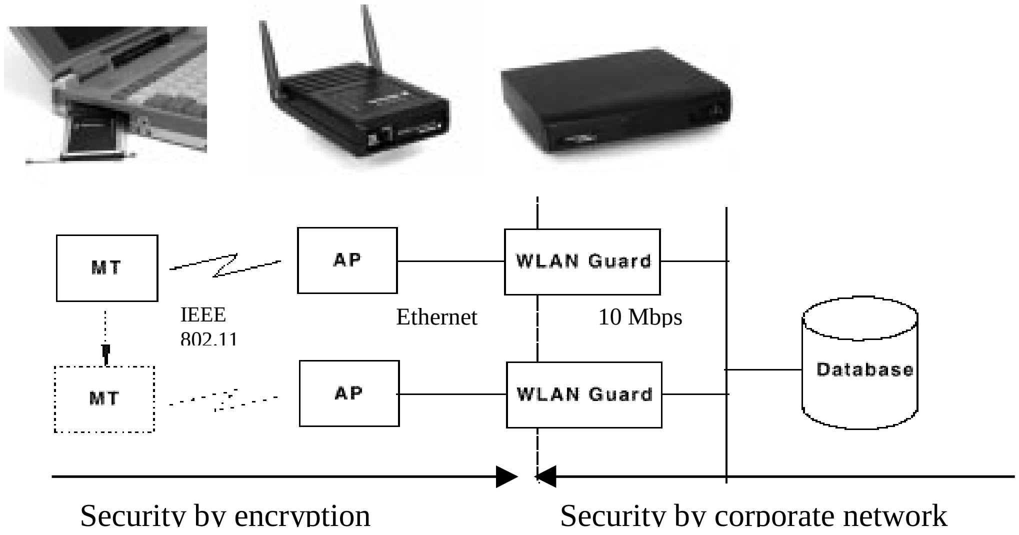
The lack of high security in 802.11 standard has lead to for instance present Ericsson's high-security solution for wireless LAN 802.11, which is based on security-dedicated devices. The latter have been named Wireless LAN Guards (WLAN Guards or, in short, WLG). Guards have been designed mainly for corporate usage. These protect the system, filtering the traffic from and to the air interface and only allowing authorized traffic to pass through.

When other security techniques are deployed, as for Ericsson's WLAN Guards, IEEE 802.11 Wireless Equivalent Privacy (i.e. RC4 and WEP authentication) could be disabled or not deployed at all. This to avoid an excessive and not needed overhead in the protocol.

7.2.1 Basic Architecture

Ericsson`s WLAN Guards use IKE (Part 2, section 6.5) and IPSec (Part 2, section 5) to provide authentication, automatic security association management and confidentiality over the air interface. Digital signatures, together with public-key certificates, for the mutual authentication between MT and WLG, are used. However, it is also possible to use a shared key mechanism.

Ericsson´s high-security solution (Guard)

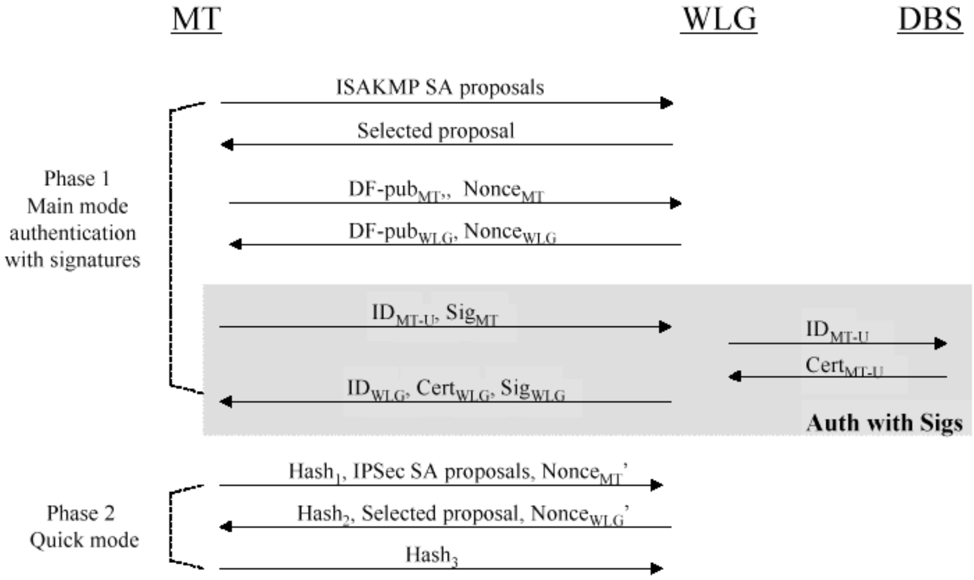


To maximize performance, IKE is implemented without doing an additional Diffie-Hellman key exchange in IKE phase 2 (Part 2, section 6.5.1). Instead, the keying material generated in IKE phase 1 is used to derive keys for IP packet encryption and authentication. This does not provide Perfect Forward Secrecy (PFS) but greatly decrease the time required in accomplishing IKE phase 2.

An IPSec tunnel with MT and WLG as end points is used to provide data origin authentication, integrity, confidentiality and anti-replay for wireless traffic. IPSec Security Associations (authentication algorithm, encryption algorithm, algorithm mode, and session keys) have been already set up in IKE phase 2.

To improve performance in handover, security associations already established are reused by securely transferring necessary attributes between different WLAN Guards by mean of a centralized server. This centralized database is used to store user information and all other parameters that are needed for security and mobility operations.

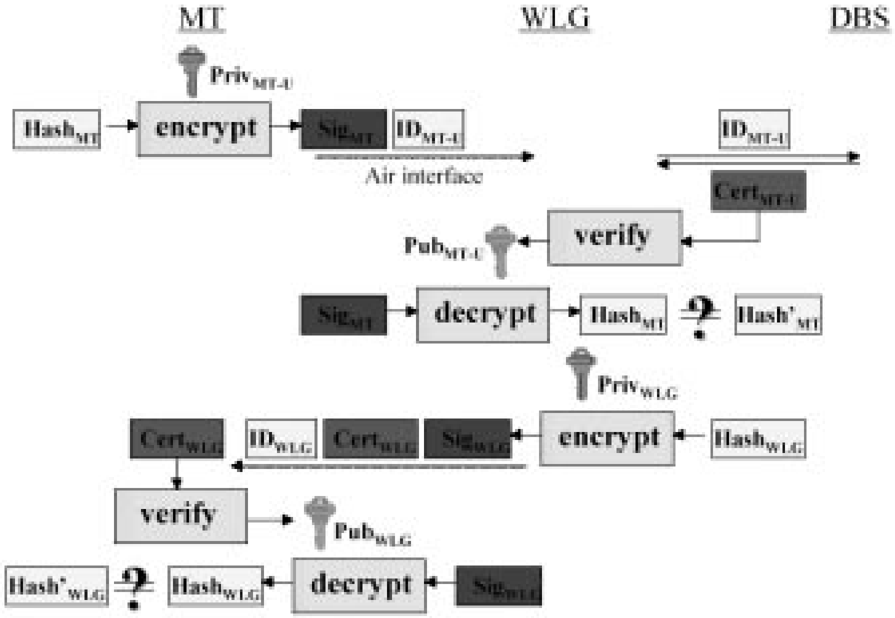
In Ericsson`s WLAN Guards system, the complete IKE negotiation procedure is as described in the following picture:


MT
WLG
DBS
ISAKMP SA proposals
Selected proposal
Phase 1
Main mode
authentication
with signatures
DF-pub$_{MT}$,, Nonce$_{MT}$
DF-pub$_{WLG}$, Nonce$_{WLG}$
ID$_{MT-U}$, Sig$_{MT}$
ID$_{MT-U}$
Cert$_{MT-U}$
ID$_{WLG}$, Cert$_{WLG}$, Sig$_{WLG}$
Auth with Sigs
Hash$_1$, IPSec SA proposals, Nonce$_{MT}$'
Phase 2
Quick mode
Hash$_2$, Selected proposal, Nonce$_{WLG}$'
Hash$_3$

IKE (Phase 1 Authenticated with Signatures)

In IKE phase 1, after Diffie-Hellman key exchange (Part 2, section 6.1), the Mobile Terminal (MT) and the Guard (WLG) compute a hash value over previously exchanged data (HashMT and HashWLG respectively). MT signs (encrypts) HashMT with the user’s private key, and sends the signature SigMT together with the user identifier (IDMT-U) to the WLG over the air interface.

The WLG fetches the user’s certificate CertMT-U from the central database (DBS) by providing the user`s credentials (IDMT-U) to DBS. WLG verifies CertMT-U first to make sure that the contained user public key PubMT-U is authentic. It then calculates Hash’MT in the same way as MT did before, decrypts SigMT with PubMT-U to recover the original HashMT, and compares HashMT against Hash’ MT. If they match the mobile user is proven to be an authorized user. In a similar way the WLG authenticates itself to the MT. The only difference is that the WLG sends its own certificate CertWLG to the MT in addition to the WLG identifier (IDWLG) and SigWLG.



Authentication phases using Public Key encryption with Digital Signatures

7.2.2 Implementation details

Each time MT users attempt to access the wired corporate network, they must authenticate themselves to the serving WLG. Only authorized users are allowed in. On the other hand, users want to be sure that the WLG, which they are talking to, is a valid one. This mutual authentication is carried out during the IKE phase 1 by using digital signatures (Part 2, section 4.2.2).

Each authorized user has a pair of keys: a private key to which only the user has access, and a public key, which is accessible to everyone. In contrast all the WLGs in the system share a pair of keys. There is only one CA in the corporate, and the central ADM runs that CA. The CA issues certificates for all mobile users and WLGs. Because all WLGs share one pair of keys, only one certificate is issued for WLGs.

Present WLAN security solutions don't have a full-blown PKI (Public Key Infrastructure) support (Part 2, section 6.6). Instead it is only employed a very simplified CA that only signs and issue certificates for users and WLGs. CRLs (Certificate Revocation Lists) are not used.

In the IPSec Authentication header (AH), HMAC-SHA1-96 (Part 2, section 4.3.4) with anti-replay checking is the default authentication algorithm. HMAC-MD5-96 (Part 2, section 4.3.2) may be supported as well. In the Encapsulation Security Payload (ESP), Blowfish in CBC mode (with random initialization vector (IV) for each packet) is the default encryption algorithm (Part 2, section 3.4.4). The key length is variable from 40 to 448 bits (present WLGs implement 40, 56 and 128 bits key length). DES in CBC mode may be supported as well (Part 2, section 3.4.1).

7.2.3 Handover between Access Points

IKE phase 1 negotiation requires both Diffie-Hellman key exchange and public-key encryption/decryption operations. This makes IKE exchange computationally very expensive if performed each time a Mobile Terminal (MT) moves from one access point to another. The latency caused by IKE may severely impact the resulting handover performance.

For this reason, established Security Associations (SA) are reused among different WLGs during the lifetime of a SA. When the IPSec SAs expire and the ISAKMP SA is still valid, WLG must restart an IKE phase 2 negotiation with the Mobile Terminal. When the ISAKMP SA expires (IPSec SAs expire consequently) WLG must restart a complete IKE negotiation from phase 1.

In order to reuse both the ISAKMP SA and the IPSec SAs with different wireless Guards (WLG), it is required to have a common IP address for all the wireless Guards. IPSec is based on IP addresses to identify the SA parameters (SPI) and the authenticated entities. It is possible to keep using a SA only as long as IP addresses are not changed. This means that tunneled packets from a MT have this common IP address in the outer IP header as the destination address. The wireless Guard will then strip off this IPSec encapsulation and forward the inner packet.

SA data are stored in a centralized database, named DBS. Once a WLG completes an IKE phase 2 negotiation, it sends the current SA information to the DBS. The SA information includes the ISAKMP SA lifetime, the ISAKMP session keys, their respective algorithms, and keying material for derivation of IPSec session keys**.** The last phase 1 CBC output block for generation of IV for the encryption of the first phase 2 message, IPSec SA lifetime, IPSec session keys and their respective algorithms are also transmitted to the DBS.

It is critical for the overall system security that transmission of SA information to the DBS is very secured. If this security is not provided by a restricted physical access to the network (like in a closed corporate Intranet), the latter should be provided by some encryption mechanism. The wireless Guard encrypts (using Blowfish) the above SA data with a symmetric key shared by all WLGs. After that it calculates a hash (using HMAC-SHA1) value over the MT's MAC address, the IP address (IP MT), the encrypted SA data and the current time T by using another symmetric key shared between all WLGs and the DBS. WLG then sends its IP address, the encrypted SA data, T and the calculated hash value to the DBS.

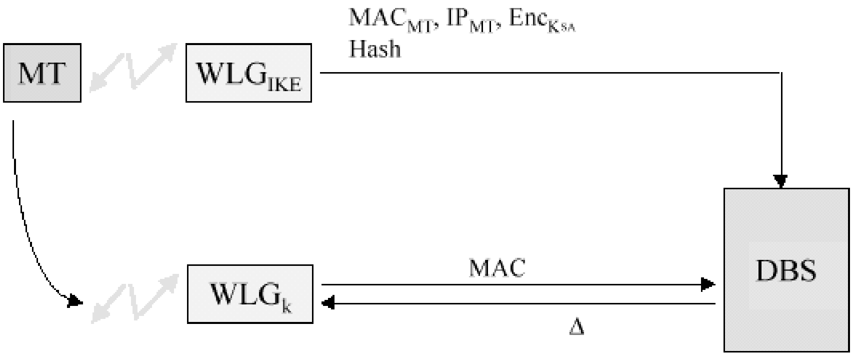
DBS first verifies the hash value to be sure that the message came from a valid WLG and the message has not been altered during transmission. If the message is authenticated, the DBS stores the encrypted SA data, together with the MT IP and MAC addresses, into its database.

When the MT moves from one WLG to another, the new WLG gets the MT's MAC address (MAC MT) from the current AP. WLG contacts the DBS presenting MAC MT. The DBS will first search its database for the IP address associated with that MAC MT, namely IP MT, and then find the encrypted SA data that is associated with IP MT. DBS returns this encrypted SA data back to the WLG. The WLG decrypts the SA data with key KSA and starts to use the decrypted SA information to talk to the MT.

Estimation of AH sequence number value is also required, when handover occurs. The sequence number in the AH is a monotonically increasing counter and its value should equal the order number of the IP packet within the current IPSec session. The current WLG cannot get the value of the sequence number before handover, therefore it has to estimate how many packets have so far been sent by the previous WLGs within this IPSec session and use this estimated number as the starting counter value. The estimation must be sufficiently accurate to avoid false replay detection and meanwhile not to cause the counter value to reach the limit (2^32) quickly.

A good estimation of the sequence number can be achieved by maintaining a central clock at the DBS. When a WLG IKE sends the SA information associated with a particular MT to the DBS, the DBS records the current time T as the starting time of the IPSec session. When another WLG needs the SA information associated with that MT, the DBS returns also the time lapse between the current time and the recorded time T. Based on this time lapse (T' – T), together with the maximum transmission speed, a Guard can have a reasonable estimation of the anti-replay counter value.

The complete handover procedure is shown in the following picture:



The IV for encryption of the first IPSec subsession packet is randomly generated. The IV for subsequent packets within the subsession is the last CBC output block from the previous packet.

7.2.4 Key Management and Distribution

Present Ericsson`s security solution does not have a complete and scalable key distribution infrastructure. It is based on a simple single-CA Public Key Infrastructure (PKI) for distribution of certificates. Moreover, all WLGs share a single key pair.

When a user needs to get a new pair of keys, he or she must go to the system administrator in person and get a floppy containing its private keys and the public key of the Certificate Authority (CA). For a detailed description of keys used in WLG and how they are maintained refers to [126].

8 HiperLAN/2

The High Performance Radio Local Area Network Type 2 [35] (HiperLAN/2) is going under standardization by ETSI as next generation network for wireless LAN systems. HiperLAN/2 has a very high transmission rate (54 Mbit/s) and operates in the unlicensed 5 GHz frequency band, which has been specifically allocated to wireless LANs. In contrast to the IEEE 802.11 wireless Ethernet technology, HiperLAN2 is connection-oriented and supports Quality of Service (QoS) in a native way. It is also included in the UMTS framework.

A good overview of the HiperLAN Type 2 standard is provided in a White Paper written by Martin Johnsson and publicly available at the HiperLAN/2 Global Forum WEB Site [53]. Sections 8.1 and 8.2 are taken from this White Paper, as well as the figures used in these two sections.

8.1 Protocol functionalities

HiperLAN/2 provides a very high transmission rate, which at the physical layer extends up to 54 Mbit/s and up to 25 Mbit/s on layer 3. The standard allows MAC mobility and IP roaming, which means that a mobile terminal in a HiperLAN/2 network can move around freely. The HiperLAN/2 protocols ensure that users always get the best possible transmission performance.

The AP with best Signal to Noise Ratio (SNR) in the radio network is automatically selected. As a MT moves around, it may detect an alternative AP with better radio transmission performance than the AP the MT is currently associated to. If the MT decides to do an handover with this new AP, all established connections will be moved to this new AP and the MT will stay associated to the HiperLAN/2 network, continuing its communication seamless (although during handover some packet loss may anyway occur).

HiperLAN/2 supports two forms of handover, complete reassociation and handover via the support of signaling across the fixed network. Doing a reassociation basically means to start over again with a completely new association procedure, which may take some time, especially in relation to ongoing traffic. The other alternative means that the new AP (to which the MT has requested a handover to) retrieves association and connection information from the old AP by transferring information across the fixed network. The MT provides the new AP with a fixed network address (e.g. an IP address) to enable communication between the old and new AP. This alternative results in a fast handover and minimizes loss of user plane traffic during the handover phase.

In HiperLAN/2 networks, data is transmitted on connections between the MT and the AP (which have been established prior to the transmission using signaling functions of the HiperLAN/2 control plane). There are two types of connections, point-to-point and point-to-multipoint. Point-to-point connections are bidirectional whereas point-to-multipoint are unidirectional in the direction towards the Mobile Terminal. In addition, there is also a dedicated broadcast channel through which traffic reaches all terminals transmitted from one AP. Connections are time-division-multiplexed over the air interface. The connection-oriented nature of HiperLAN/2 makes it straightforward to implement support for QoS.

Connection setup does not result in an immediate capacity assignment by the AP. At the connection setup the MT receives a unique identifier (within the scope of one AP) for each of the established DLC connections. Whenever the MT has data to transmit it initially request capacity by sending a resource request (RR) to the AP. The RR contains the number of pending User Protocol Data Units (U-PDU) that the MT currently has for a particular DLC connection. The MT may use contention slots in the RCH to send the RR message or the SCH (see section 8.2.2). By varying the number of contention slots, the AP could control the actual access delay.

Selective repeat (SR) ARQ is the Error Control (EC) mechanism that is used to increase the reliability over the radio link. EC in this context means detection of bit errors, and the resulting retransmission of U-PDU(s) if such errors occur. EC also ensures that the U-PDU's are delivered in-sequence to the convergence layer. The method for controlling this is by giving each transmitted U-PDU a sequence number per connection. The ARQ ACK/NACK messages are signalled in the LCCH. An errored U-PDU can be retransmitted a number of times (configurable). ARP also eases QoS for delay-critical applications.

HiperLAN/2 defines two different modes of multicasting, N*unicast and MAC multicast. With N*unicast, the multicast is treated in the same way as unicast transmission in which case Automatic Repeat Request (ARQ) is utilized as error control mechanism. Using MAC multicast, a separate MAC-ID (local significance, per AP) is allocated for each multicast group. ARQ can't be used in this case, i.e. each U-PDU is only transmitted once. All multicast traffic for that group is mapped to the same and one DLC connection. HiperLAN/2 allows for up to 32 multicast groups to be mapped to separate MAC identifiers.

A mobile terminal may disassociate explicitly or implicitly from an Access Point (AP). When disassociating explicitly, the MT will notify the AP that it no longer wants to communicate via the HiperLAN/2 network. Implicitly means that the MT has been unreachable for the AP for a certain time period.

HiperLAN/2 standard also provides mechanisms to allow a mobile terminal to save power. The MT may at any time request the AP to enter a low power state and requests for a specific sleep period. At the expiration of the negotiated sleep period, the MT searches for the presence of any wake up indication from the AP. In the absence of the wake up indication the MT reverts back to its low power state for the next sleep period, and so forth. An AP defers any pending data to the MT until the corresponding sleep period expires. Different sleep periods are supported to allow for either short latency requirement or low-power requirement.

HiperLAN/2 networks, as for the IEEE 802.11 case, can be deployed in two different topologies. The Mobile Terminals (MT) might either communicate with an Access Point (AP) connected with a backbone distribution network or be directly connected with other MTs, creating an ad-hoc network.

8.2 HiperLAN/2 Layers

The HiperLAN/2 protocol stack is divided into a control plane part and a user plane part following the semantics of ISDN functional partitioning. The user plane includes functions for transmission of traffic over established connections, whereas the control plane includes functions for the control of connection establishment, release, and supervision.

HiperLAN/2 protocol is structured in three basic layers:

- Physical layer (PHY)
- Data Link Control layer (DLC)

- Convergence layer (CL)

The Physical layer (PHY) and the Data Link Control layer (DLC) will be described in following sections of this chapter. The convergence layer (CL) has two main functions: it adapts service request from higher layers to the service offered by the DLC and converts the higher layer packets with variable or possibly fixed size into a fixed size that is used within the DLC. Padding, segmentation and reassembly functions are also part of the convergence layer functionalities.

8.2.1 Physical layer

The transmission format on the physical layer is a burst, which consists of a preamble part and a data part, where the latter could originate from each of the transport channels within the Data Link Control layer (DLC).

HiperLAN/2 makes use of a modularization method called Orthogonal Frequency Digital Multiplexing (OFDM) to transmit the analogue signals over the air. OFDM is a special form of multicarrier modulation. The basic idea is to transmit broadband, high data rate information by dividing the data into several interleaved, parallel bit streams, and let each one of these bit streams modulate a separate subcarrier.

OFDM also admits great flexibility considering the choice of and realization of different modulation alternatives. This is to both adapt to current radio link quality and to meet the requirements for different physical layer properties. BPSK, QPSK and 16QAM are the supported subcarrier modulation schemes.

Orthogonal Frequency Division Multiplexing (OFDM) has been chosen due to its excellent performance on highly dispersive channels. OFDM is very efficient in time-dispersive environments, e.g. within offices, where the transmitted radio signals are reflected from many points, leading to different propagation times before they eventually reach the receiver.

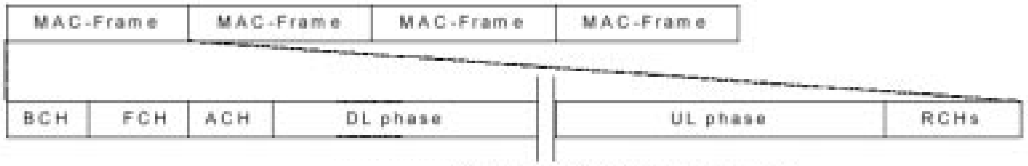
In a HiperLAN/2 network, there is no need for manual frequency planning as in cellular networks like GSM. The appropriate radio channel for transmission within each AP’s coverage area is automatically selected using a DFS (Dynamic Frequency Selection) scheme in the Radio Resource Management protocol. This eases Operation and Maintenance (OAM) in the system, making it rather simple.

8.2.2 Data Link Control layer (MAC layer)

The HiperLAN/2 Medium Access Control (MAC) protocol implements a form of dynamic time-division duplex (TDD) and dynamic time-division multiple access (TDMA) to allow for most efficient utlilization of radio resources. The control is centralised to the AP, which informs the MTs at which point in time in the MAC frame they are allowed to transmit their data, which adapts according to the request for resources from each of the MTs.

In time-division duplex (TDD) and dynamic time-division multiple access (TDMA) schemes, the time-slotted structure of the medium allows for simultaneous communication in both downlink and uplink within the same time frame (called MAC frame). Time slots for downlink and uplink communication are allocated dynamically depending on the need for transmission resources.

The basic HiperLAN/2 MAC frame structure has a fixed duration of 2 ms and comprises transport channels for broadcast control, frame control, access control, downlink (DL) and uplink (UL) data transmission and random access (see following picture). All data from both AP and the MTs is transmitted in dedicated time slots, except for the random access channel (RCH) where contention for the same time slot is allowed. The duration of broadcast control is fixed whereas the duration of other fields is dynamically adapted to the current traffic situation.



Basic MAC frame structure

The broadcast channel (BCH, downlink only) contains control information that is sent in every MAC frame and reaches all the MTs. It provides information about for instance transmission power levels, starting point and length of the FCH and the RCH, wake-up indicator, and identifiers for identifying both the HiperLAN/2 network and the AP.

The frame control channel (FCH, downlink only) contains an exact description of how resources have been allocated (and thus granted) within the current MAC frame in the DL-phase and UL-phase and for the RCH. The access feedback channel (ACH, downlink only) conveys information on previous access attempts made in the RCH.

Downlink or uplink traffic (DL-phase and UL-phase, bidirectional) consists of PDU trains to and from MTs. A PDU train comprises DLC user PDUs (U-PDUs of 54 bytes with 48 bytes of payload) and DLC control PDUs (C-PDUs of 9 bytes) to be transmitted or received by one MT. There is one PDU train per MT (if resources have been granted in the FCH). The C-PDUs are referred to as the short transport channel (SCH), and the U-PDUs are referred to as the long transport channel (LCH).

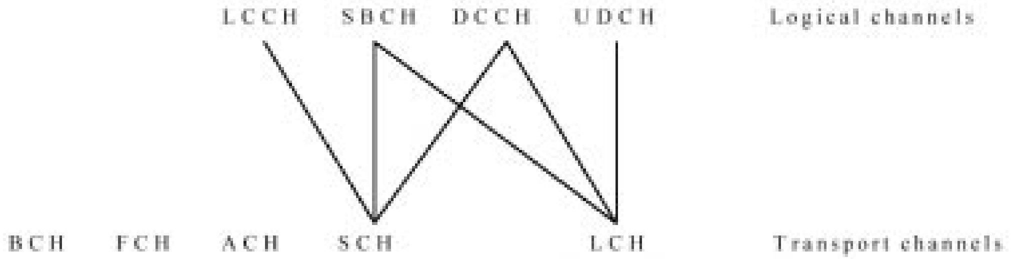
The random access channel (RCH, uplink only) is used by the MTs to request transmission resources for the DL-phase and UL-phase in upcoming MAC frames, and to convey some RLC signaling messages. When the request for more transmission resources increase from the MTs, the AP will allocate more resources for the RCH. RCH is entirely composed of contention slots which all the MTs associated to the AP compete for. Collisions may occur and the results from RCH access are reported back to the MTs in ACH.

The transport channels (SCH, LCH, and RCH) are used as an underlying resource for the logical channels defined in HiperLAN/2 standard and are hereby presented:

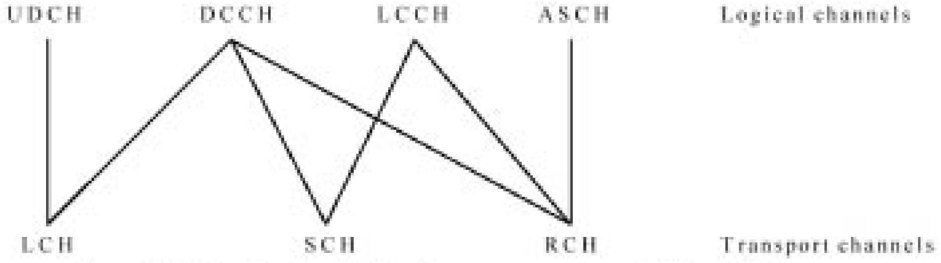
- The **slow broadcast channel** (SBCH, downlink only) conveys broadcast control information concerning the whole radio cell. The information is only transmitted when necessary, which is determined by the AP. All terminals have access to the SBCH.
- The **dedicated control channel** (DCCH, bidirectional) conveys RLC sublayer signals between a MT and the AP. Within the DCCH, the RLC carries messages defined for the DLC connection control and association control functions. The DCCH forms a logical

connection and is established implicitly during association of a terminal without any explicit signalling by using predefined parameters.

- The **user data channel** (UDCH, bidirectional) conveys user data (DLC PDU for convergence layer data) between the AP and a MT. The DLC guarantees in sequence delivery of SDUs to the convergence layer. A DLC user connection for the UDCH is setup using signalling over the DCCH. Parameters related to the connection are negotiated during association and connection setup. In the uplink, the MT requests transmission slots for the connection related to UDCH, and then the resource grant is announced in a following FCH. In downlink, the AP can allocate resources for UDCH without the terminal request.
- The **link control channel** (LCCH, bidirectional) conveys information between the error control (EC) functions in the AP the MT for a certain UDCH. The AP determines the needed transmission slots for LCCH in the uplink and the resource grant is announced in an upcoming FCH.
- The **association control channel** (ASCH, uplink only) conveys new association request and re-association request messages. These messages can only be sent during handover and by a disassociated MT.



Mapping from logical to transport channels in downlink.



Mapping from logical to transport channels in uplink.

8.3 HiperLAN/2 security

8.3.1 The standard

Security mechanisms within HiperLAN/2 standard are defined at the link layer (ISO layer 2). HiperLAN/2 networks support both authentication and encryption. With authentication enabled, both the AP and the MT can authenticate each other to ensure authorized access to the network (from the AP's point of view) or to ensure access to a valid network operator (from the MT's point of view). Authentication relies on the existence of a supporting function, such as a directory service, but which is outside the scope of HiperLAN/2. Both authentication and encryption are optional and may be disabled:

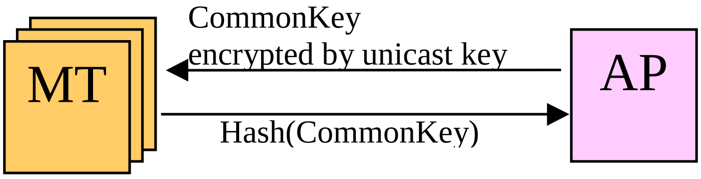
Security options in HiperLAN/2 standard	Encryption	Authentication
No security needed	OFF	OFF
No key distribution supported	ON	OFF
High security desired	ON	ON

The authentication process begun after than a mobile terminal has selected the AP it wants to connect to (the AP with the best radio link quality) and some beacon information has been exchanged. At this stage, the mobile terminal has requested and been given a MAC-ID from the AP. Link capabilities about the supported PHY modes, convergence layers and authentication and encryption procedures & algorithms have been also exchanged.

User data can be encrypted (once the initial link-connection has been established) to protect it from eavesdropping and man-in-middle attacks. If encryption has been negotiated, the MT will start the Diffie-Hellman key exchange (Part 2, section 6.1) to negotiate the secret session key for all unicast traffic between the MT and the AP. In this way, the following authentication procedure is protected by encryption.

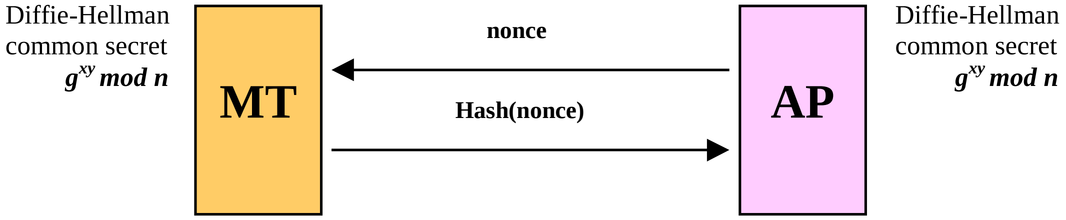
HiperLAN/2 supports both the use of DES and the 3DES algorithms (Part 2, sections 3.4.1 and 3.4.2) for strong encryption. These two algorithms need an Initialization Vector (IV). Periodical transfer of a seed from the AP to the MT supports the generation of the IV in the MT. Broadcast and multicast traffic can also be protected by encryption through the use of common keys (all MTs associated to the same AP use the same key). Common keys are distributed encrypted through the use of the unicast encryption key.

Multicast/Broadcast Keys distribution using Unicast Key



All encryption keys are periodically refreshed to avoid flaws in the security. This is done using nonces that are transmitted in clear on the radio link and used together with the shared Diffie-Hellman key to generate a new session secret key (SSK), as described in sections 6.4 and 6.5 of Part 2:

Generation of a new session secret key


Diffie-Hellman common secret g^{xy} mod n
MT
nonce
Hash(nonce)
AP
Diffie-Hellman common secret g^{xy} mod n

New key (SSK) = ***HMAC-MD5(g^{xy} mod n, nonce)***

There are two alternatives for authentication; one is to use a pre-shared key and the other is to use a public key (Part 2, section 3.2). When using a public key, HiperLAN/2 supports a Public Key Infrastructure (but does not define it) by means of generating a digital signature. The supported authentication algorithms are HMAC (Part 2, section 4.2.1) and RSA (Part 2, section 3.4.6) with MD5 (Part 2, section 4.3.2). Bidirectional authentication is provided for authentication of both the AP and the MT. HiperLAN/2 supports a variety of identifiers for identification of the user and/or the MT, e.g. Network Access Identifier (NAI), IEEE address, and X.509 certificate.

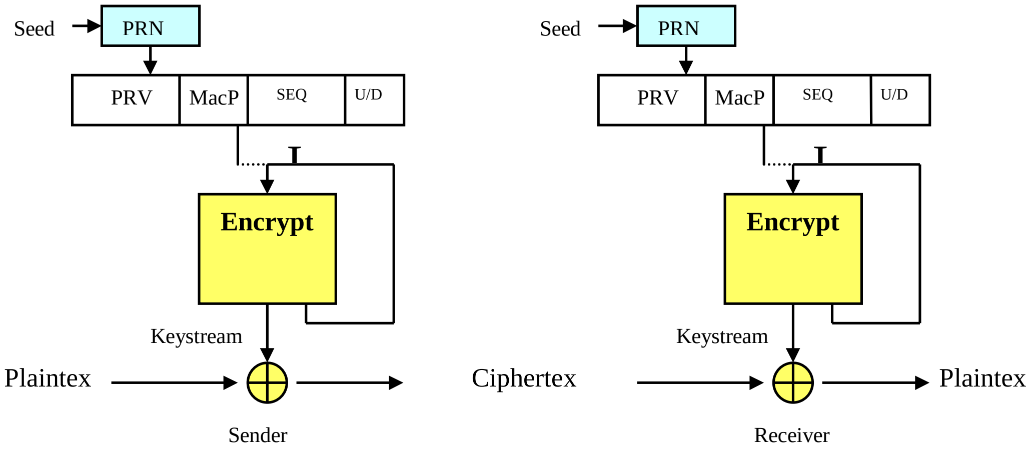
Cryptographic algorithms utilized	IEEE 802.11 Standard	Ericsson's high-security solution for 802.11 networks (Guards)	HiperLAN/2 Standard
Authentication	The hash algorithm is not detailed	HMAC-SHA/RSA	HMAC-MD5 / RSA
Encryption	RC4	Blowfish (DES/3DES)	DES - 3DES

8.3.2 Implementation proposals for future Ericsson's products

Ericsson's security proposals for the first commercial release of HiperLAN/2 networks are mainly focused to some specific application scenarios. This means that only a sub-set of the security techniques defined in the standard will be selected in the security functional description for the first release of Ericsson's HiperLAN/2 products.

As it also resulted from the analysis done in Part 1 of this thesis, the corporate case, for instance, allows for the utilization of pre-shared keys and eases the deployment of a Public Key Infrastructure (PKI). A centralized database can be in this last case utilized to retrieve the public keys without the need of a distributed multiple-CA certificate distribution system.

DES (as mandatory) and 3DES (as optional) in Output Feedback mode (OFB) will be used as cryptographic algorithms for data encryption (as defined in the HiperLAN/2 standard), instead of Blowfish, which was used in IEEE 802.11 Ericsson's Guards. DES (and 3DES) is more standardized and faster in hardware than Blowfish. 802.11 Guards were implemented in software whereas HiperLAN/2 security will be implemented in hardware and this is the reason for this choice in the encryption algorithms. DES, moreover, is license free and exportable (3DES has some restriction in exporting).



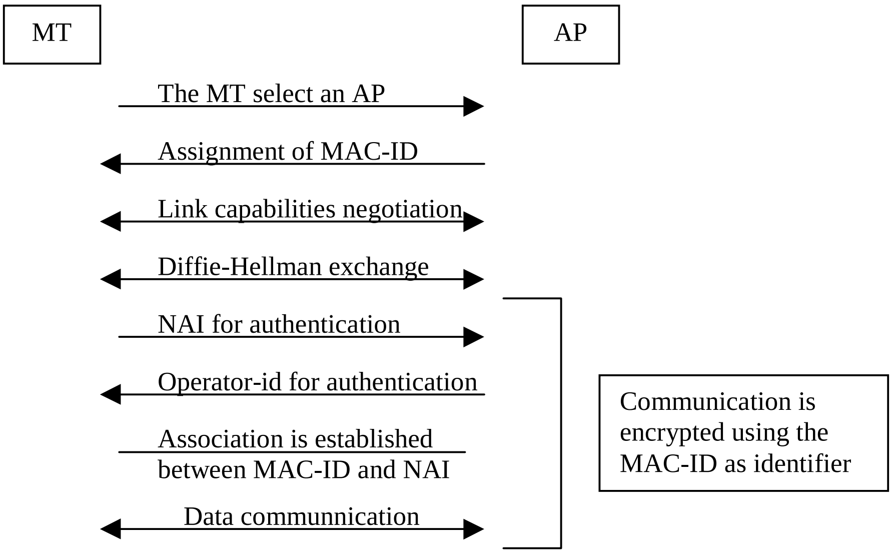
IV = PRV | MacP | SEQ | U/D
PRV: Pseudo random value 51 bits
MacP: 4 least significant bits of current MacId
SEQ: PDU sequence number 10 bits
U/D: Uplink/Downlink 1 bit

AP transmits a ***Seed*** in FCCH in every frame
Seed is used to initialise a pseudo-random number generator (PRNG)
PRNG outputs a different ***PRV*** for each PDU

DES Encryption in Output Feedback (OFB) mode

HMAC-MD5 and RSA-MD5 will be used for authentication, instead of HMAC/RSA-SHA-1, which was utilized in Ericsson's Guards for IEEE 802.11 networks. The key length will be 40-128 bits for pre-shared keys, 128 bits for RSA private keys and 256-2048 bits (512, 768 or 1024 bits as default) for public ones. HiperLAN/2 security is meant to be not too expensive and MD5 is the best solution when high performance is desired, even if at the cost of some less security strength in the algorithm. If users should need a stronger security, IP layer encryption might be provided (e.g. with a new high-security Guard operating with IPSec).

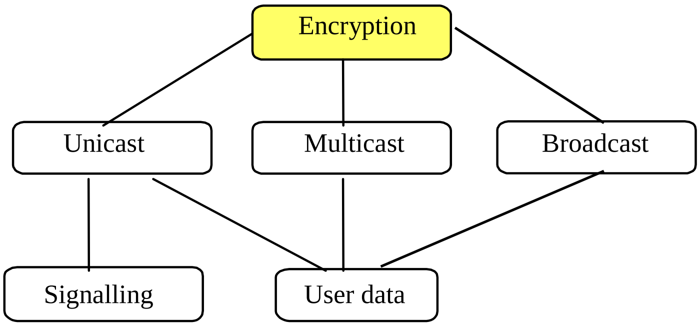
The authentication process is a challenge-response scheme (Part 2, section 4.1.2) and it is done before the association procedure is accomplished and after than the Diffie-Hellman exchange has been done, as for protect user identity. Authentication of the mobile terminal (MT) can be based on the Network Access Identificator (NAI) and authentication of the Access Point (AP) on an operator id (not yet specified in the functional description). Both the mobile terminal and the access point are allowed to disable authentication of the other entity.


MT
AP
The MT select an AP
Assignment of MAC-ID
Link capabilities negotiation
Diffie-Hellman exchange
NAI for authentication
Operator-id for authentication
Association is established between MAC-ID and NAI
Data communnication
Communication is encrypted using the MAC-ID as identifier

Association procedure in HiperLAN/2 Standard

Unicast and multicast traffic will be encrypted as defined in the standard. Unicast traffic shall be encrypted on a per MT basis, while multicast and broadcast communication will be encrypted using a common key for each multicast group or for broadcast transmissions. Unicast, as well as multicast and broadcast are identified for encryption out from the Ethernet MAC addresses assigned during the association phase (before authentication).

Encryption Modules

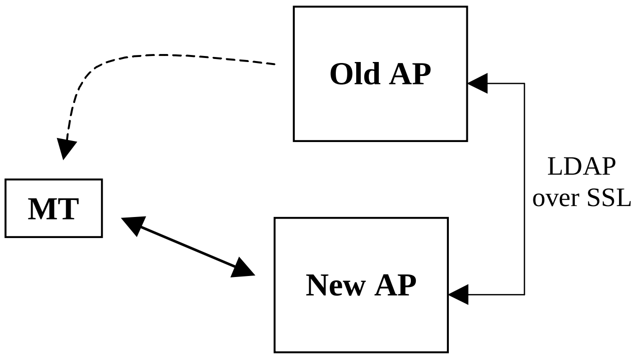


A centralized database is used to store user information and other parameters that are needed for security (both for pre-shared and for public cryptosystems) and hand-over operations. Ericsson plans in its functional description to use Lightly Directory Access Protocol (LDAP) to access this database and retrieve pre-shared or public keys, possibly protecting the data by a security mean (information classified). This resembles a simple PKI system with a single centralized Certification Authority (CA).

The private key of the wireless LAN network is locally stored in each AP, even if a more scalable solution is under investigation. Out-of-band distribution of users' private keys, e.g. manually, will be also possible for very small corporations.

To maximize performance in handover, security associations established between the MT and an AP are reused by securely transferring necessary attributes (especially Diffie-Hellman keys) between different access points. An AP-to-AP signaling protocol is proposed to transfer special tokens, which contains these association data and a mean to authenticate and locate the old AP.

There are two main ways an access point can locate the old AP during handover. There might be static mapping in each AP between the AP-Identificator (transmitted in the MT´s token) and the IP address of each AP. This technique has the drawback to require a "manual" update in the AP when some AP is added or removed. An alternative to this scheme might be to dynamically create and update the bindings existing between AP-ids and IP addresses in the APs using the information contained in the MT's token. Being this process under investigation, further information have been classified and not included in the final version of this thesis.



Token Network Signaling for Hand-over

Finally, a very important issue for the overall security in the network is the presence of a trustable OAM (Operation and Maintenance) environment. Even if authentication and data confidentiality procedures should be highly secure, when an intruder is able to manipulate and modify data in some entities of the system, the security might be completely lost.

OAM management in Ericsson's HiperLAN/2 systems will be based on the Simple Network Management Protocol (SNMP). For security reasons, it will be in particular utilized SNMPv3, which adds security and some administration services to SNMPv2. SNMPv3 defines authentication (using HMAC with hash function MD5 or SHA-1), privacy (using CBC mode of DES) and key management procedures.

Access to the APs using HTTP and Telnet must also be secure. Management Information Base II (MIB-II) is the basic MIB for the AP containing location info, network interface (Ethernet and others) and TCP/IP attributes. A View-Based Access Control Model (VACM) defines which parts of the Management Information Bases in the APs are accessible for different users. Three different kind of users are defined:

- Viewer: a user with read-only access
- Installer: a user with read-write access
- Technician: an Ericsson or reseller technician.

9 Security in public environments

As it has been pointed out in Part 1 of this thesis, hot spots and other public environments are considered as one of the most important application areas for future HiperLAN/2 networks. From analysis presented in the same sections of this thesis it has been motivated also how network and security requirements in these usage cases may differ from those adopted in deploying corporate wireless LANs.

When wireless LAN systems based on IEEE 802.11 standards are desired for use in public networks, present Ericsson wireless security solution (Guards) has several drawbacks in scalability. The functional description that was specified for security in first Ericsson release of HiperLAN/2 networks (studied for the corporate case) must also be modified.

Public environments do not allow the maintenance of users' credentials on a local database, as adopted when deploying corporate networks. Accounting and inter-domain access control mechanisms are also required from public operators to allow for IP roaming. This calls for the use of either AAA solutions, as for instance RADIUS and DIAMETER, or for a distributed multiple-CA Public Key Infrastructure (PKI).

9.1 Public Key Infrastructure

Key distribution and retrieval in corporate networks can be deployed using a centralized server, which contains users` credentials and public keys. This is also the solution adopted in Ericsson's 802.11 Guards products and planned for HiperLAN/2 networks in corporate environments. LDAP is typically used to access the centralized database and the overall structure looks very much like a simple PKI with a single Certificate Authority (CA).

When public hot-spot environments are considered, it is no longer possible to use a single CA to retrieve public keys of users, which may potentially come from everywhere in the world and belong to different original domains. A distributed network of CAs is in this case required (structured either as a web or a hierarchy of CAs). Possible PKI schemes (able to scale over a wide area network and among different domains) will be described in the following of this chapter.

9.1.1 The web of hierarchies

A certificate authority (CA) server issues, manages, and revokes certificates (see section 6.6 in Part 2). The public key for the CA's certificate must be well known and trusted by all the participating end entities. Therefore, having a single Certificate Authority (CA) for the entire world would not be scalable to the public case. A distributed PKI is required, where Certificate Authorities are allowed to certify other CAs. The CA can delegate its authority to a subordinate authority by issuing a CA certificate, creating a certificate hierarchy. In effect, one CA will tell its users that they can trust what a second CA says in its certificates. Different CAs would serve users belonging to different network domains. This both for administration (e.g. different domain policies) and performance reasons (e.g. single point of failure and network congestion).

The certificates issued for end-users are called *end-user certificates* while the certificate issued for validation among different CAs are called *CA-certificates*. The ordered sequence of certificates from the last branch to the root is called a certificate chain or *certificate path*. Each certificate contains the name of that certification's issuer (i.e., this is the subject name of the next certificate in the chain).

In general, there may be an arbitrary number of CAs on a path between two end-users. To obtain the peer’s public key, a user has to verify the certificate of each CA in turn until the peer end-user certificate is obtained. A first public key is used to validate a certificate from the first CA in the certification path. This certificate will then contain a second public key used to validate a third certificate from the third CA in the path. An additional public key is in this way obtained and this might be the user’s public key or the key for one more CA, continuing along the certification path validation. This process is called *certification path validation*. The length of the certification path is the number of CAs between the peer entities.

When several CAs are utilized, it is very important the way these CAs are organized to build the overall PKI. Some PKIs use a hierarchical model (called general hierarchy), where each CA certifies its parent and its children. *Cross-certificates* (which are certificates that do not follow the basic hierarchy) may also be integrated in this architecture. Some PKIs use a variant of the general hierarchy known as a top-down hierarchy, in which CAs only certify their children and the top-level CA is the source of all certification paths.

Other PKIs have no structure at all and in effect each CA is its own root CA and has full authority over how its trust is assigned. An example of this unstructured PKI is the Pretty Good Privacy (PGP) model. In the latter, each CA bases its trust on several certificates of other CAs. If enough of the other CAs issue certificates that bind a particular name to a particular key, then the CA can itself accept that binding with some confidence. This structure is called a *web of trust*.

The web of trust model does not scale well. Moreover, there is also the problem of how trust is delegated, which may differs among different Certificate Authorities (CAs). The hierarchical model scales well, but poses problems on finding an agreeable structure for different entities that may not desire to delegate too much their security to other third-part entities. A strict hierarchical model is adequate for a corporation but is too restrictive when considering the needs of multiple interconnected corporations. Flexibility is important when considering the manner the certification path takes. Strict hierarchical certification paths will most likely exist in the government spaces and in some public networks.

A general hierarchy lets any CA be the root CA of a certification path. However, the structure still relies heavily on the upper-level CAs, especially the top-level CA. A large number of certification paths in a general hierarchy pass through a high level CA. This forces the PKI’s users to implicitly trust the latter. If its private key were ever compromised, it could be used to forge messages between entities which rely on a certification path that includes A. Since so many paths do pass through A, it becomes a very tempting target for attacks. Certification paths in a general hierarchy also run the risk of becoming too long, resulting in problems similar to the web-of-trust. Cross-certification helps to reduce path lengths, at the risk of complicating path discovery.

In a top-down hierarchy, all users must use the top-level CA as their root CA. This requires all users to obtain a copy of the top-level CA's public key *prior* to using the PKI. Also, all users must fully trust the top-level CA for all purposes. This makes a top-down hierarchy impractical for a worldwide PKI.

A web of hierarchies, which is a combination of a strict hierarchical and a web of trust certificate path, will most likely be the solution for PKI in inter-domain roaming.

Another important choice in deploying a PKI is the format of the digital certificates utilized. The most widely accepted formats are International Telecommunication Union's X.509 and PGP signed keys. It is suggested the use of X.509 version 3 (x.509v3) in WLANs hot-spot systems. One of the major improvements to version 3 of X.509 is the ability to allow flexible extensions to the certificate structure. These extensions include additional key and policy information, user and CA attributes, and certification path constraints.

9.1.2 Interaction between PKI and present corporate solutions

A worldwide Public Key Infrastructure (PKI) that supports international, government, and state policies/regulations will probably not be available for a long time. In the meantime, corporations should build their own public key infrastructures upon present security solutions to satisfy current business needs (and support their own security means). Many companies prefer to manage their own Certificate Authority (CA) instead of outsourcing this function to a third party. However, to allow interoperability outside the corporation, one must register the corporate CA server in an overall Public Key Infrastructure.

The LDAP directory server (used for corporate security) provides a single point of administration for corporation and user credentials. Lightweight Directory Access Protocol (LDAP) is used to make it possible for applications running on a wide array of platforms to access X.500 directories. It creates a standard way for applications to request and manage directory information and may be still used to retrieve public key within the corporation. LDAP server might be also used as CA server, which would then interconnect the system with other external CAs, creating an inter-domain PKI as described in the previous section.

The security requirements for the directory server would be mainly to supports network authentication through IP address/DNS name, and user authentication through LDAP user name and password (or a X.509 version 3 public-key certificate). The database server should also control users' credentials to perform read, write, search, or compare operations down to the attribute level on behalf of the CA server. SSL should still be used to provide message privacy and message integrity for LDAP communications. The server should also maintain an audit journal of all key management operations it performs.

The amount of trust that has to exist between distinct entities of a PKI is also important. Some PKIs require that users place, or *delegate*, all their trust in a single CA (for example, the top-level CA of a top-down hierarchy) while others allow users to decide which CAs to trust. Sometimes it is the CAs who must place their trust in other CAs or in their own subjects. A few PKIs allow their entities to refine the kind of trust they delegate. A CA could refine its trust in another CA such that the second CA would only be trusted for certain kinds of certifications. For example, the first CA could have a policy stating "I only trust this other CA to issue certificates that relate an email address to a public-key value" and this could be expressed in the PKI in a way that makes conformance to the policy automatic.

A Certificate Revocation List (CRL) should be added to present LDAP server functionalities. Certificates that are expired will not be included in this list. The CRL list should contain certificate serial numbers and associated revocation status values (e.g. the specific times and reasons). These values are signed and time stamped by the issuing CA. The CRL latency period is the period between CRL updates. This value will be set by each CA security policies.

9.2 Authentication, Authorization and Accounting (AAA)

In traditional fixed networks, users obtain access to network services by negotiating access with a "home domain", generally an ISP or a private corporate network. Authentication, Authorization and Accounting (AAA) are usually required, either to charge the user or protect network resources.

With the increasing demand for mobile devices (as it is often the case for wireless LAN terminals), a new need has been generated to allow users to attach to any domain that would be convenient to their current location. In this way, a client needs access to resources being provided by an administrative domain that is not the same as its home domain (usually called "foreign domain"). The visited domain must be able to verify the user’s identity, possibly through user’s home domain. It should also be assured of receiving the proper revenues for the services provided to the visiting mobile users.

9.2.1 The AAA Infrastructure

Authentication, Authorization and Accounting (AAA) verification infrastructure defines a large set of security services [36]. Only the most important ones, regarding the wireless LAN case, will be described.

One of the basic AAA goals is to deploy a protocol able to transport authorization information in a secure and trusted way. Mechanisms that preserve authenticity, integrity and privacy for this information are provided. Moreover, a service administrator in the middle of a delivery chain can not read or change authorization information being sent between other AAA parties. AAA infrastructure also specifies mechanisms for updating the rules that are used to control authorization decisions and to support failure recovery, when one AAA entity fails.

AAA allows security requirements to differ between different network domains. It provides timed security associations, which means all authorization information must expire after a given time. It also should be possible to revoke authorization permissions before this time expires.

Mechanisms that prevent security attacks from intermediate entities or domains are provided. This is a basic requirement to prevent man-in-the-middle attacks, where an intermediate administrator might change AAA messages to gain some advantage or make fraudulent charge [36].

Protection from replay and flooding attacks is also supported. Non-repudiation and roaming are both implemented to make access services and billing possible for visited domains. Some mechanism that allows visiting nodes to discover the AAA server for the visited network is always provided.

In AAA, it is possible to base authorization decisions on information about users, mobile terminals, requested services, and operating environment. Authorization based on group membership (e.g. in case of multicast) and geographic location is also supported. Wireless LAN public networks should at least support authorization decisions based on users` identity, visited domain, and required services (as derived from Part 1).

Different levels of Authentication, Authorization and Accounting (AAA) security should be provided, as well as both highly secure and less secure mechanisms for data integrity and confidentiality.

9.2.2 Remote Authentication Dial In User Service (RADIUS)

The Remote Authentication Dial In User Service (RADIUS) protocol has for a long time been the most important AAA verification protocol implementation. It is the dominating legacy authentication protocol on the market and it is often used by ISPs to authenticate dial-in users. It has been mainly designed to carry authentication, authorization, and configuration data between several distributed Network Access Servers, which desire to authenticate their users, and a shared Authentication Server [14].

The main idea of the RADIUS protocol is that, even for a distributed architecture, it is far easier and more efficient to have a single database containing users' secrets, authorizations and accountings data. RADIUS defines how distributed servers can access this database to provide authentication, authorization and accounting services to end-users.

RADIUS is implemented with a client/server architecture. Traditional Network Access Servers (NAS) and other Edge devices operate as clients of the RADIUS protocol. These are responsible for forwarding user information to RADIUS servers and interacting with end-users. RADIUS servers are responsible for receiving user authorization requests, authenticating the user, and returning all necessary configuration information to RADIUS clients. RADIUS servers can also act as proxy clients towards other RADIUS servers.

If a mobile device needs access to a network that uses RADIUS for authentication, it must present some authentication information (most likely the user's identifier and password) to the RADIUS client (the edge device). The RADIUS client might chose to authenticate the user against the RADIUS server by sending it an Access-Request. When the RADIUS server receives the request it first validates the sending client by checking the client ID. If this is a valid client, the server and the client have a shared secret enabling secure communication. The RADIUS server has available (locally or in a remote way) a database with users` credentials.

Authorization data exchanged between RADIUS clients and servers are always authenticated through the use of shared secrets. These are never transmitted over the network and are periodically changed. It is of great importance to protect the confidentiality and integrity of these keys, which are used to protect any user specific secrets.

User passwords are sent encrypted between end-users and RADIUS servers to avoid RADIUS clients and RADIUS proxy having access to users' secrets. This also eliminates the possibility that someone snooping on an unsecured network could determine a user's password. Methods based on RSA Message Digest Algorithm MD5 are usually used (Part 2, section 4.3.2).

RADIUS can support many different mechanisms to authenticate users. Challenge/Response procedures are, for instance, used by RADIUS. These are among the most important mechanisms of RADIUS and are implemented through the *Access-Challenge* RADIUS message. Clear text Password Authentication (PAP) and Challenge Handshake Authentication Protocol (CHAP) [108] are also supported, as long as the visited Network Access Servers provide these mechanisms.

RADIUS Servers answer verification requests either with *Access-Accept* or *Access-Reject* messages. Access-Accept packets provide specific configuration information necessary to allow RADIUS Clients to begin secure connection with users.

RADIUS uses UDP messages to carry its data. This was decided upon several technical reasons. Retransmission to a secondary authentication server, for example, is used in RADIUS every time the primary server is not available. This requires retransmission timers that differ from those defined by the TCP protocol. Aggressive TCP retransmission and reliable TCP data delivery are also not useful in RADIUS [14]. Finally, the stateless nature of RADIUS and the ease of UDP server implementation suggested the use of UDP messages instead of TCP.

RADIUS provides 63 basic attribute extensions, although it is open for additional extensions (e.g. for the Extensible Authentication Protocol (EAP) [10]). Some of these are used, for example, to encode user-name and passwords, to support CHAP, to provide information about the actual NAS to be used, routing parameters, and callback numbers. For a complete description of all attributes supported by RADIUS refers to [14].

9.2.3 DIAMETER

DIAMETER [16] can be considered as a next generation RADIUS protocol. It has been developed to address RADIUS flaws in inter-domain roaming support and to provide a much more scalable architecture. Its framework consists of a base protocol [17] and a set of protocol extensions (e.g. end-to-end security, PPP, Mobile-IP and Accounting). The base protocol provides all the basic functionalities that must be provided to all the services supported in DIAMETER, while application-specific functionalities are provided through extension mechanisms.

The most important difference between DIAMETER and RADIUS is that DIAMETER is based on a peer-to-peer architecture, instead of client/server model. This easily allows service providers to cross-authenticate their users and to support mobility between many different domains.

DIAMETER defines specific UDP retransmission procedures and failure recovery schemes. This is mainly accomplished through end-to-end message acknowledgements. An AAA Client is in this way enabled to know a priori if a specific server is available, allowing for quick transition to back-up servers. DIAMETER requires that each node in a proxy chain acknowledge a request, or response message, either with a successful response or one that contains an error code.

It uses sequence numbers and acknowledgement numbers to provide reliable UDP transmission and allow each node in a chain to retransmit unacknowledged messages. Retransmission detection based on this sequence number simplifies DIAMETER server implementations and make it possible to support a much larger number of transactions per second.

DIAMETER, opposite to RADIUS, allows a server to send unsolicited messages to a NAS, used for instance to allow a server to inform a NAS that a session should be terminated. This is possible because DIAMETER is peer to peer. While a NAS, for dependability, can authenticate with different DIAMETER servers, watchdog message are used, which are sent to a peer entity to detect reachability when no traffic has been exchanged for some time. RADIUS had to retransmit a message many times before determining that an AAA server was down and try a secondary Server.

The base protocol also defines a windowing scheme, which requires that each peer advertise its receiver window. It allows AAA Servers to control the flow of incoming requests and can also be used to distribute the load across multiple servers or different paths of an AAA authorization chain.

DIAMETER removes RADIUS's inherent limitation on Attribute-Value-Pairs (AVP) length, making the protocol much more flexible and effective [16]. For compatibility sake it supports all RADIUS attribute extensions, at the same time extending the set of possible attributes with many application-specific extensions.

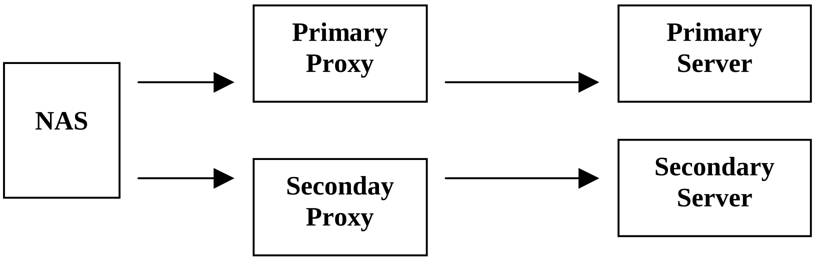
End-to-end integrity, confidentiality and non-repudiation of Attribute-Value-Pairs (AVP) are provided in DIAMETER, by mean of a Digital Signature added by DIAMETER servers. It is even possible to secure only some of the attribute data, leaving remaining attribute extensions not secured. Secured Attribute-Value Pairs are called Protected AVPs, whereas the non-secured are named Unprotected AVPs. DIAMETER brokers can modify only the unprotected Attribute-Value-Pairs (AVPs) contained in a message. End-to-end security extension can only be provided in networks where AAA peers share secrets. This is usually accomplished through a broker that issues roaming certificates to all DIAMETER servers.

DIAMETER makes efficient use of third party proxy servers, called Brokers, which act as intermediate servers to set up roaming agreements between different domains. Brokers also provide message routing functions and redirect services, by returning contact information to a requesting DIAMETER server. This enables to communicate directly with other servers, without the need of further involving the DIAMETER Broker.

A collection of ISPs supported by a broker is called a "roaming consortium". Roaming consortia reduce configuration information that would otherwise be required between all servers participating in the Consortium. Brokers can either forward data between end-to-end AAA destinations or redirect traffic between these entities. When DIAMETER Brokers do forwarding, DIAMETER Servers do not need to validate any certificates or security association with other peer entity.

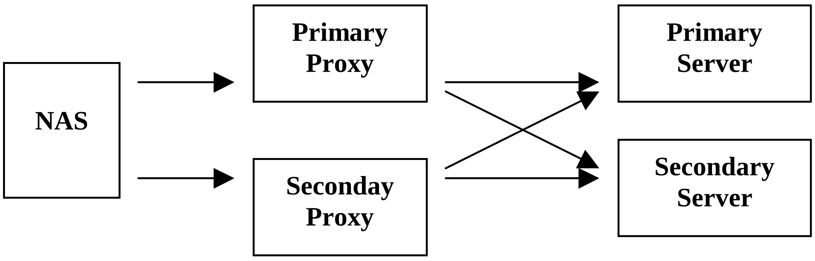
Certificate distribution is used to uphold redirection of messages within a consortium, as well as to implement end-to-end security, where security is maintained between requestor and responder even if a request is handled by intermediate nodes. Brokers, acting as Certificate Authority Entities, issue certificates to all DIAMETER servers within the consortium. All this is provided in a much more efficient way than in RADIUS. Long-time secrets between peers are not needed and protocols such as IPSec can be used to generate and refresh short-time session keys.

As already described, RADIUS Servers can not know a priori if a downstream RADIUS server is reachable. Therefore, most RADIUS networks create parallel links between primary and alternate servers. This solution, obviously, does not scale well. When dependability is important, many parallel links have to be established and many retransmissions are required.



RADIUS Proxies chain

The DIAMETER protocol has instead been designed to specifically support roaming between different networks. Every node in the network is responsible for its own packet retransmissions and each node knows a priori if a peer entity is reachable, by the use of watchdog messages. This allows the deployment of much more resilient network architecture, where each node in an authentication chain is connected with several upper entities, without the need for the far less efficient RADIUS parallel paths.



DIAMETER Proxies chain

Another important feature of DIAMETER is that it incorporates the Extensible Authentication Protocol (EAP) [10] from the beginning. EAP is a general protocol for PPP authentication support of multiple authentication mechanisms. It does not select any specific authentication mechanism during the PPP Link Control Phase (LCP), postponing this to the Authentication Phase. Authentication methods based on MD5-Challenge [102], One-Time Password (OTP) and Generic Token Card are also supported by EAP.

EAP allows the authenticator to request more information before determining the specific authentication mechanism and permits the use of a "back-end" server which actually implements the various mechanisms, as could be the case for RADIUS or DIAMETER Servers. Since different administrative domains may own the NAS and RADIUS servers, this is a very necessary feature. The fact that non-trusted NASs generate a challenge may otherwise provide the ability for authentication replay attacks.

When inter-domain roaming is required, DIAMETER should definitively be preferred to RADIUS. Many reasons for this have been here presented. Moreover, NAI information and X.509 certificates should be used to identify users during inter-domain AAA verification procedures.

A final issue that should be considered is that many RADIUS servers have already been implemented in public and private networks. This creates an existing "legacy" system that should be able to inter-operate with new DIAMETER implementations. A recent Internet draft [127] describes a protocol that enables DIAMETER servers to inter-operate with current RADIUS servers using PPP CHAP verification. It is especially designed to allow DIAMETER servers on visited domains to inter-operate with RADIUS servers located on the home network. This is the most needed level of RADIUS/DIAMETER hybrid interoperability. Brokers have an important role in the structure presented in this draft. DIAMETER brokers are in fact required to interact with RADIUS servers on the Home network to obtain user data.

9.3 Hybrid Authentication with Extended ISAKMP

The multiple-CA Public Key Infrastructure (PKI) described in section 9.1 and integrated with an accounting management protocol for inter-domain roaming (possibly based on digital signatures) might solve scalability in present security solutions for wireless LANs. This would allow extending the use of 802.11 Guards and HiperLAN/2 standard security for usage in public environments.

However, whereas it could be profitable to deploy a worldwide PKI infrastructure for network domains, it could be rather expensive (at least in the close future) to do it for mobile terminals, being the number of mobile users much greater than the number of different domains. Moreover, many organizations and ISPs have already deployed some legacy authentication system, such as RADIUS, and wish to use this to authenticate users.

It may be in all these cases worth (until a PKI for users is available) the use of some unidirectional authentication mechanism such as RADIUS or DIAMETER (for authentication of users and accounting) in combination with ISAKMP (IKE) or some other PKI-based protocol (for authentication of network domains). The Hybrid IKE Authentication Mode [69] could be used in IKE Phase 1 to establish a unidirectional authenticated IKE SA. This would be used to authenticate only the visited domain (which could use a technique based on some public-key mechanisms). Users will thereafter be authenticated using RADIUS or DIAMETER, as it will be later described. Mutual authentication is important both for user identification and to prevent man in the middle attacks. When a PKI for digital signatures and e-commerce will be finally deployed, a pure PKI solution should be anyway preferred.

The Hybrid IKE Authentication Mode describes a set of new authentication methods to be used in Phase 1 of IKE. These new schemes are based on an asymmetry between the authenticating entities where the two peers use different methods to authenticate each other. The network domain is authenticated to the users with standard public key techniques, while users are authenticated by some other mean. Either the Edge Device or the User can initialize the first stage and either Main Mode or Aggressive Mode can be used.

Using Hybrid IKE, only the visited domain is authenticated. This Phase 1 will be immediately followed by an Extended Authentication [88] to authenticate the user and make the IKE SA bidirectionally authenticated. The Extended Authentication Phase is handled using a new IKE exchange called Transaction Exchange [87] and initiated by the network operator.

In the Transaction Exchange, the negotiated IKE SA is utilized to securely transfer configuration information between the two peers. This new IKE exchange can be used to retrieve certain information from the other peer before that the non-ISAKMP SA is established. Status, configuration and error messages may be in this way exchanged after IKE Phase 1.

Transaction Exchanges can be done using either a Request/Reply or a Set/Acknowledge paradigm. The initiator can send either a Request or a Set message and this is followed by the other peer that sends either a Reply or an Acknowledge message. An identifier is used to map a Request to a Reply or a Set to an Acknowledge. Request/Reply allows a host to request information from another host, e.g. a configuration manager. Set/Acknowledge allows a host to send configuration information it wants another peer to alter, most often used by a configuration manager.

Transaction Exchanges greatly extend the communication services offered by the Information Exchange mode defined in standard IKE. Informational Exchange messages are in fact not acknowledged and, because of the nature of UDP, not even guaranteed to arrive. Therefore, the latter can not be used to negotiate configuration parameters, as it is possible with Transaction Exchanges.

Utilising Transaction Exchange messages to carry data, the Extended Authentication ISAKMP Method [88] (often called X-Auth) will be used to authenticate the mobile user. It allows IPSec's ISAKMP/Oakley protocol to support remote authentication methods based on unidirectional authentication mechanisms like RADIUS and DIAMETER. It supports several different verification schemes, such as Simple Authentication, Challenge/Response, Two-Factor Authentication, One-Time-Password (OTP), and Pre-shared Keys. A Challenge/Response scheme used with DIAMETER is suggested in this thesis project.

The visited domain initiates the Extended Authentication ISAKMP Method after than the Hybrid Phase 1 ISAKMP exchange is completed. This must be done before any Quick Mode ISAKMP scheme is begun. Only ISAKMP Info mode exchanges may be allowed between the Hybrid and the Extended Authentication schemes. IKE Phase 1 is not affected by the use of Extended Authentication ISAKMP Method.

During X-Auth, user authentication credentials required by the legacy authentication system are transferred to the network domain in a secured ISAKMP packet using the negotiated IKE SA. The network domain then authenticates the user towards the legacy authentication system by sending the user authentication information to a RADIUS or DIAMETER server. The server responds to the edge device in the visited domain, which puts the response in an ISAKMP packet and sends it to the User. In this way the User is informed whether the authentication succeeded or not. Upon failure, the negotiated IKE SA must be discarded.

In the mechanism described here, ISAKMP protects sensitive information, such as user-name and user-password, which are IPSec protected across the access link. The destination of these data, i.e. the visited domain, is also authenticated, during the Hybrid ISAKMP Phase 1 exchange. Then the Extended Authentication ISAKMP Method is used to authenticate the remote User, using RADIUS or DIAMETER and allowing the system to scale very well.

If both Hybrid and X-Auth exchanges success, the participants can use the IKE SA for setting up other SAs. If the Extended Authentication ISAKMP Method fails, the ISAKMP Phase 1 SA is deleted. This should also be logged as a security exception and used for auditing and possibly prevent denial of service attacks.

9.4 Users anonymity

Anonymity, which may be defined as the ability to use a PKI while only revealing the information pertinent to the situation, might be very important in many PKI operations. An irrefutable signature seems to imply that the signer should be readily identifiable. Yet there are many situations where anonymity would be preferred, most notably in the area of shopping and remote access to corporations.

PKI and AAA verification infrastructures necessarily involve some degree of intrusiveness in order that sufficient quality can be achieved. This is especially true for the PKI hierarchical model (section 9.1.1). Having little or no choice in the manner in which key-pairs are generated and the presence of an external issuer as ownership of the key-pair (with individuals merely licensed to use it) may be not accepted by many organizations or individuals.

Current X.509 v.3 certificates go so far as to permit an agent of an organisation to protect their personal identity through the use of a role-title, but they preclude an individual from having that capability. Moreover, some implementations may preclude a private person from possessing multiple personal key-pairs, even though the same person is permitted to possess multiple key-pairs for organisations that they represent. Some schemes even involve the key-pair generation process being compulsorily performed by some organisation on behalf of individuals, and compulsory storage of the private key.

Individuals, including not only consumers, but also employees and contractors (especially in sensitive occupations) will not be confident in a system that does not guarantee privacy and, finally, real confidentiality of transmitted data. To avoid such seriously negative implications, the verification infrastructure should allow users whether to use a key-pair in connection with all organisations, some organisations or a single organisation. It should also be possible to select an anonymous identifier to be used in connection with a key-pair for certain applications.

The question of role-related key-pairs is especially critical. Public concerns about privacy-invasive behavior among governments and corporations are amplified when organizations are able to combine information about an individual from multiple sources. From the individual's perspective, the trustworthiest approach is not to rely just on legislated privacy protections, with their manifold exceptions, but also to deny organizations the ability to correlate information. The interests of individuals would be well served by role-specific and/or organization-specific identities and key-pairs, and are seriously threatened by any requirement that they use a single key-pair with all organizations, or with all organizations of a particular category.

A good PKI should provide both strong, irrefutable authentication and a high degree of privacy through anonymity. Several new PKI proposals use credential authentication to that end. Conventional, hierarchical PKI may still be appropriate in contexts where authority is clearly defined. This includes defense organisations, internal communications within organisations, and communications within a layer of government. In particular, federal government agencies may find such an approach appropriate as a means of authenticating messages between agencies, and to confirm the eligibility of an individual to communicate on behalf of an agency. Initiatives in this area may be reasonably described as Government Public Key Infrastructure (GPKI).

A different approach for accounting in public networks might be the use of e-cash, which would provide full anonymity to the mobile user [55]. The e-cash might be used both to pay the network connection to the public operator and to pay local or remote services. An example of a local service could be paying the entrance to a museum whereas a remote one might be downloading music from a remote web-site.

10 Conclusions

IEEE 802.11 and HiperLAN/2 standards for wireless LAN networks provide their own security mechanisms for authentication and data protection at the link layer. Whereas HiperLAN/2 standard security is sufficient for most of the application scenarios defined in Part 1, this is not the case for the security means defined in IEEE 802.11 standard.

The lack of strong security in IEEE 802.11 networks has lead to the development of high-security solutions to be used with 802.11 wireless connections, as Ericsson’s Guard product. Guards work at the IP layer and are based on IPSec and IKE technologies. Working at the IP layer means that they are somehow independent from the physical medium and might be easily extended for use in HiperLAN/2 or 802.3 Ethernet networks.

When the public environment is considered, present 802.11 Guard products and security mechanisms planned for corporate usage of HiperLAN/2 networks are sufficient. These also provide the flexibility needed to adapt to different public scenarios. The main drawback in both the systems is the lack of a mean to manage and distribute key-pairs. Moreover, IP roaming in public domains require accounting functionalities to allow public operators to charge users for connection and services provided.

Security in public hot spot areas might be achieved by integrating 802.11 Guards and HiperLAN/2 link layer security with a complete PKI (and possibly with an accounting management protocol). AAA protocols, as RADIUS or DIAMETER, could be used for accounting and for users authentication (if a PKI should not yet be available for end-users) as described in chapter 4. The result would be a method based on the asymmetry between the authenticating entities and divided into two authentication phases. A PKI for network domains provides users with public keys for authentication of the visited domain. The visited domain uses then some other technique, such as RADIUS or DIAMETER, to authenticate the user.

- PART 3 -

Advanced Services

Security framework for Mobile IP
Support for inter-domain accounting
Access to VPN services
Secure assignment of IP addresses

11 Secure Mobility

IP layer mobility refers to the ability of moving among different IP sub-networks without loosing network connectivity. It refers both to session and user mobility, even if session mobility is currently the designers' main goal. Session and user mobility will both be addressed in this paper.

Many requirements are applicable to macro mobility. Routing of IP datagrams to mobile nodes, for instance, should be transparent to applications and end-users. Mobile nodes should be able to communicate with other nodes that do not implement mobility functions. Protocol overhead should be as low as possible and only a small additional latency should be introduced in transmission of data.

Mobile Security is surely one of the hottest areas in the Internet community and this chapter will provide a broad overview of current state-of-the-art proposals.

11.1 Security in traditional Mobile IP

If security mechanisms were not used, networks based on Mobile IP could be easily compromised through remote redirection attacks. For this reason, all messages exchanged between mobile nodes and home agents must be authenticated. A Security Association (SA) [60] is used for this purpose. This defines the security mechanisms that have to be applied to all messages exchanged between home agents and mobile nodes. Each SA, indexed by a Security Parameter Index (SPI) [113] between the home address and the mobile node, indicates an authentication algorithm, a secret key (either shared or part of a public/private pair), and a style of replay protection to be used.

Authentication is provided through the use of the Authentication Header, as defined in IPSec [58]. The default authentication algorithm is based on keyed-MD5 Message Digest [102] in "prefix+suffix" mode. Any other algorithm may also be supported. It is also worth noting that the authenticator field itself and the UDP header are not included in the computation of the message digest. All other fields are used in computing the Message Digest.

Authentication mechanisms between mobile node and home agent are defined as compulsory. They are required to protect the mobile node and the home agent against naive attacks that result from the uncontrolled propagation of remote redirection messages. Authentication between the mobile node and the foreign agent, as well as between foreign and home agent, may also be included in registration requests and replies. However, the use of the latter is not specified by the protocol.

Protection from replay attacks against registration requests is implemented through an identification field, which is present in the authentication header. The home agent uses this field to verify that a registration message has been freshly generated by a mobile node and not replayed by an attacker. There are two possible replay protection methods described in Mobile IP [89]. The first is based on the use of timestamps and it is defined as mandatory. The second is based on the use of nonces and is optional.

The Mobile IP Protocol also states that Mobile nodes, foreign agents and home agents should log any security exceptions and errors for auditing and incident handling purpose (see Part 2 of this Thesis for auditing and logging).

Users who wish to provide confidentiality and integrity to exchanged data should use additional security mechanisms that are outside the scope of Mobile IP. End-to-end encryption is usually suggested to provide appropriate protection. Users concerned about traffic analysis should also consider the use of appropriate link-layer encryption. If absolute location privacy is desired, mobile nodes could create a tunnel to the home agent and encrypt all the tunneled traffic. Mixers, traffic padding mechanisms and routing control procedures (as described in section 2.3 of Part 2) may also be used.

11.1.1 Threats

Remote fraudulent redirection could have been the main security threat for Mobile IP. However, authentication between the mobile node and its home agent provides a basic protection from this kind of attack. Other security flaws have not been addressed in the Mobile IP specification.

Address Resolution Protocol messages are not authenticated and can potentially be used to steal another host's traffic. The use of Gratuitous ARP brings with it all of the risks associated with the use of ARP and some unauthorised host on the home network could, by these means, realise a masquerading attack. There is currently no method available to provide secure ARP. However, this is a link layer issue and should be addressed by a link layer protocol, not by Mobile IP (which works at the network layer).

Mobile IP allows for an authentication extension between Mobile Nodes and Foreign Agents (the Mobil-Foreign Authentication extension). A mobile node could use this extension to authenticate itself to a foreign agent. Unfortunately, this extension does not provide secure replay protection. It also does not specify how to maintain key distribution architecture when a mobile node could access any arbitrary foreign agents. This is a severe threat when inter-domain roaming is supported and non-repudiation is required for the offered services.

No authentication is required for Agent Advertisement and Agent Solicitation messages. Therefore, it could be possible for a fake foreign agent to send advertisements to get a new and never used registration request from the mobile node. This could later be used for Denial-of-Service or man-in-the-middle attacks against the home agent.

Mobile IP specifies that an *agent advertisement from a new Foreign Agent should not cause a mobile node to attempt a new registration if its current registration has not expired and it is still also receiving Agent Advertisements from the foreign agent with which it is currently registered* [89]. A fake foreign agent could be effective against new arrived mobile nodes that require a new foreign agent, having lost connectivity with their previous FAs.

Finally, Mobile IP protocol does not specify how to provide data confidentiality and integrity. These are supposed to be handled by the IPSec Suit [113] but a detailed specification of how IPSec should interact with Mobile IP has not yet been provided.

11.2 Authentication Extensions

A possible approach to address most of the security threats within Mobile IP is the use of other independent security protocols, either at the network or the transport layer. Service providers could use strong authentication techniques (e.g., CHAP [108]) to prevent theft-of-service and, to some extent, Denial-of-Service attacks. Users requiring confidentiality could use either link layer encryption, IP-layer encryption [59], or application-layer encryption, depending upon their individual requirements [110].

This approach, while effective, needs to be integrated with Mobile IP characteristics. It also leaves partly unsolved the problem of Denial of Service attacks when IP addresses are dynamically assigned. In the following sections, state-of-the-art proposals to secure Mobile IP will be presented and analysed. Following chapters will then address secure DHCP mechanisms (chapter 12) and secure access to VPNs (chapter 13).

There are three main possibilities to authenticate the entities involved in Mobile IP protocol:

1. Authentication between the Mobile Node and the Home Agent
2. Authentication between the Mobile Node and the Foreign Agent
3. Authentication between the Foreign Agent and the Home Agent

11.2.1 Mobile Node – Home Agent Authentication

Mobile IP provides a security mechanism to authenticate a Mobile Node with its Home Agent (the MN-HA Authentication extension). It also provides protection from replay attack, through the use of either timestamps or nonces. This security mechanism is very valuable, protecting both from address spoofing and redirection attacks. Mobile IP requires strong authentication between the mobile node and its home agent and this authentication should be implemented within any IP mobility management architecture.

However, if a (bidirectionally authenticated) Security Association between the Mobile Node (MN) and the Foreign Agent (FA) and another (mutually authenticated) SA between the Foreign Agent (FA) and the Home Agent (HA) are provided, it may not be necessary to have a Security Association between the MN and the HA. This is true as long as the visited foreign agent, which has been authenticated to prevent man-in-the-middle attacks, is a trusted entity. Moreover, if a MN-AAA authentication extension [91] is used, the latter might as well take the place of the MN-HA authentication procedure.

If the MN needs to redirected its traffic through the HA, either as a result of triangle routing problems or source based filtering when it is using a co-located care-of address, a SA between the Mobile Node and the Home Agent should also be established [85].

11.2.2 Mobile Node - Foreign Agent Authentication

Mobile IP provides a basic authentication mechanism between Mobile Nodes and Foreign Agents (the Mobile-Foreign Authentication). However, it may be worth to extend this procedure to support more secure techniques, based for example on a mutually authenticated challenge response scheme. This could be especially useful for portable computer devices accessing the network through a wireless link.

Some extensions to Mobile IP Agent Advertisements and Registration Requests could be introduced to enable foreign agents to use challenge/response mechanisms, when authenticating the mobile node, as it have been recently proposed in an Internet draft by Perkins and Calhoun [91].

In this case, a new extension to the ICMP Router Discovery Messages [28] used as Foreign Agent Advertisements is needed to issue the nonce needed to challenge the mobile node. The foreign agent utilizes this challenge to verify that the mobile node is not replaying any earlier registration request. The same nonce might be used also from the Mobile Node to authenticate the FA, computing some Keyed-Hash over this random progressive number.

If no other security association exists between the mobile node and the foreign agent, the mobile node must include in the registration request an authentication digest computed over the nonce, which was received in the foreign agent advertisement. This authentication field can be either also forwarded to the home agent or stripped off by the foreign agent when it forwards registration request and replay messages from and to the home agent.

Some delay might also be allowed in the use of a valid nonce. This would accommodate mobile nodes that might not have the last nonce or that might be not able to calculate the message digest in a fast enough way. The last case refers to the situation where a new message advertisement (with a new nonce) is sent before than the mobile node might have been able to compute the digest over the old nonce.

The use of challenge/response mechanisms between the mobile node and the foreign agent provides protection against replay or theft-of-the-service attacks. However, it is necessary to specify how to maintain the key distribution infrastructure that allows the foreign agent to retrieve the mobile node’s secret key.

When challenge/response mechanisms are used, secret keys should not be statically assigned. In the case where the foreign agent does not have any security association with the mobile node and the mobile node's home agent, the foreign agent could get assistance from external administrative systems, as a PKI or an AAA verification infrastructure (as described in section 11.5). A Security Parameter Index (SPI) field in the authentication extension could specify the particular secret keys and algorithm (shared between the Mobile Node and the verification infrastructure) that must be used to perform the authentication. The challenge value must be always included in the registration request.

To support authentication schemes where the Mobile Node can belong to an administrative domain other than the one it is visiting, the Network Access Identifications [2] (already proposed in Part 3 to identify wireless LAN devices) could be used to map mobile node ID with its administrative domain. Allowing mobile nodes with different IP addresses or NAIs to use the same challenge value does not create any security flaw. The authentication provided by the mobile node will in fact be always computed over different data, there being in any case different mobile nodes IDs. For a complete architecture able to provide global roaming refer to chapter 11.5.

Message advertisements containing the challenge/response nonces are not authenticated and transmitted in clear. Therefore, an attacker could try a chosen-plaintext attack, having access both to the ciphertext and the associated plaintext for several registration requests sent from the mobile node. It could even sends fake message advertisements to get associated ciphertext to analyse. To avoid this threat, message advertisements could be authenticated with a private key associated with the foreign agent and with timestamp-based replay protection.

If the Foreign Agent is able to establish an IPSec Security Association with the Mobile Node, agent advertisements could also be expanded to implement IPSec. The Mobile Node should initiate the establishment of the IPSec Security Association (SA) with the FA and it is recommended that the Aggressive Mode of IKE in Phase One is used in order to reduce the number of exchanged messages between the MN and the FA [85]. A new SA should to be negotiated each time the MN associates to a new FA (even if it might be possible to reuse SAs by transmitting data FA-to-FA within a single network domain, similar to the way the AP-to-AP channel is used in HiperLAN/2 standard for IPSec SA).

If the FA only wants to authenticate the Mobile Node without providing data confidentiality (for protection from eavesdropping see section 11.3), it is suggested the use of MN-FA authentication schemes before any MN-HA authentication. If the HA is required to process MN registration requests before any other authentication has been provided, it could be object of a Denial of Service attacks. Moreover, if the MN is accessing the network through a wireless link, eavesdropping could be a threat. Useful data, such as the user ID and user location, could be inferred from a HA Registration Request sent in the clear over the air.

If the Foreign Agent is in a different domain than the Home Agent, the visited network might desire to authenticate the Mobile Node also for accounting purposes. MN-FA authentication extensions might be used for the visited domain to be able to make claims for the offered services, i.e. for billing, when accounting is not provided during the link-layer association phase. In this case, it would not be enough to use simple encryption with shared secret keys. Some digital signature would be required, as is the case for the AAA infrastructure (chapter 11.5).

Mobile-Foreign Authentication extensions should be used before than any other security associations, if they are meant to provide AAA verification services. The MN-FA authentication should in this case offer local assurance about non-repudiation and replay protection that is verifiably by the foreign agent.

Finally, when user identity and location must be protected, the registration request sent from the roaming user to the foreign agent might be somehow encrypted. This does not make it necessary to establish a secure channel with the FA before than the MN has successfully registered with the HA. The MN may encrypt the registration request (with a MN-HA encryption) and have it made available to the FA only in the response message that comes back from the HA. Since the HA is the one who can decrypt this information, it can send this information back to the FA in the response message and authenticate the MN to the FA.

11.2.3 Home Agent – Foreign Agent authentication

In a commercial environment it might be important to authenticate all messages between the foreign agent and the home agent, so that billing is possible, and so that service providers do not provide service to users that are not legitimate customers [89]. This is true only when access control and accounting is not performed during link-layer association (when connection is established). It is indeed profitable to implement accounting before than any IP mobility protocol is considered. IP mobility is often an optional service and accounting should be provided during the association phase, integrating it in the IP roaming structure (NOT in the IP mobility service).

A foreign network is likely to have multiple Foreign Agents that can be assigned to a Mobile Node. Similarly, there could be multiple Home Agents that can potentially serve the

Mobile Node in its home network. It would be a waste of resources to have a Security Association between each possible pair of Foreign Agents and Home Agents.

Foreign Agents and Home Agents should be connected with some high level agent, which would establish the needed inter-domain tunnels. All the traffic between the domains can be in this way secured through a common tunnel and only a single ISAKMP Security Association [72] has to be defined. This high level agent might in some cases be physically placed within a Firewall to provide it (and its security association data) with high-level host security.

Moreover, it is not possible for a network to establish a Security Association with every other network, in order to provide roaming on a large scale. Authentication between different domains should be upheld through some AAA based protocol, as it would be the case of using DIAMETER Proxy, known also as Service Broker [17].

A Service Broker creates a consortium of networks and service providers with agreements to allow roaming of their users within each other's network. When a network joins a service broker consortium, it receives instant roaming capabilities with all the other networks that are part of that consortium. In this way only one ISAKMP Security Association, between the visited network and the Service Brokers provider, is needed [85]. Each time the visited network needs to establish a session key with any home domain within the Broker consortium, it gets the session keys from the Service Broker, without the need for establishing a new ISAKMP SA with the home domain. Chapter 7 further describes the third party Broker architecture.

11.3 Data confidentiality - IPSec Security Associations

Several Internet draft proposals support the use of IPSec [113] as a basis for introducing security mechanisms within the Mobile IP protocol. This is mainly due to the consideration that the IPSec protocol is becoming the standard security protocol both in the public Internet and in most of private corporate networks. All these proposals claims that *Mobile IP should leverage the developments of IP Security and Internet Key Exchange (IKE) rather than developing it's own security mechanisms* [85].

The security characteristics of IPSec can be used in Mobile IP networks to establish secure communication links. IPSec provides all the security services required within a Mobile IP implementation, i.e. connectionless integrity, data origin authentication, anti-replay protection and data confidentiality. A great advantage in using IPSec is its standardisation and flexibility, established during the ISAKMP security association phases [72].

Four different security associations (SAs) could be utilized for securing Mobile IP communication with IPSec, even if not all of these are always required:

4. A Security Association between the Foreign Agent and the Home Agent
5. A Security Association between the Mobile Node and the Foreign Agent
6. A Security Association between the Mobile Node and the Home Agent
7. An end-to-end Security Association between the Mobile Node and the peer host

The Security Association between the Mobile Node and the Foreign Agent may be optional, if some link layer security mechanism or external authentication scheme between the MN and the FA is provided. A Security Association between the Mobile Node and the Foreign

Agent is instead necessary when data confidentiality and integrity are required, but they are not implemented in the link-layer.

An important choice is when a mobile node should establish the Mobile Node-Foreign Agent security association; i.e. either before or after the Mobile Node has authenticated with its home agent. A possible choice is that, after the Mobile Node has registered with its Home Agent, it initiates a security procedure with the serving Foreign Agent. This approach has been suggested when IPSec and IKE are used as security mechanisms between MNs and FAs [85]. It implies that the initial registration request messages and response messages are in clear. The authors of this Internet draft state that this is not expected to be a security threat, even in the case where there could be malicious nodes eavesdropping on the link.

Performing IKE negotiations is expensive. A MN should not initiate IKE negotiations with a FA in a visited network until the FA has been assured about the authenticity of the MN by its home network. By doing so, the FA would not have to be involved in performing IKE negotiations with MNs that may be fraudulent. This would also avoid keeping the FA tied up with trying to authenticate the MN on it's own.

In wireless networks, e.g. Wireless LAN 802.11a, link layer security is often provided for communications over the air interface. Having a security association used for data confidentiality between mobile nodes and foreign agents adds an additional and perhaps unnecessary layer of security to the system. The delay introduced by this additional security association may be too high. This is especially true in the case when the Mobile Node moves to a network with a new Foreign Agent and the Mobile Node has to establish a new Security Association with the new Foreign Agent. To effectively utilise authentication schemes between foreign agents and mobile nodes in wireless networks, some optimized mechanism for key exchange between MNs and FAs is required.

A Security Association between the visited foreign network and the home network might be provided through static IPSec secure tunnelling between the foreign domain and the home domain. In effect, a Virtual Private Network (VPN) could be created between the foreign domain and the home domain of the mobile user. Foreign Agents and Home Agents should route the traffic generated by mobile nodes to some local security gateway, e.g. a Firewall, which concentrates the transmitted data through a secure inter-domain tunnel.

A Security Association used for data encryption between the Mobile Node and its Home Agent is usually not required. This is especially true if there is already a Security Association (SA) both between the Mobile Node and the Foreign Agent and between the FA and the Home Agent and if all the entities have been authenticated (so to protect against man-in-the-middle attacks).

Any end-to-end Security Association between the Mobile Node and an end-to-end peer host is optional and it is established only if the Mobile Node requires it. It is a private security policy choice and it is mainly required only for transmission of sensitive data or for access to Virtual Private Networks (chapter 8).

The main drawback in the use of IPSec is the required computational cost, especially when wireless devices are considered. However, compression schemes and ad-hoc optimisations can decrease the protocol overhead introduced by IPSec [21].

If IPSec SAs are established using public key authentication in IKE, the existence of a Public Key Infrastructure (PKI) is required both for mobile terminals and for the network agents. Whereas it could be quite easy to deploy such an infrastructure for network domains (i.e. for the mobility agents), it could be rather expensive to do it for mobile nodes, as already describe in Part 3.

The Hybrid IKE Authentication Mode [69] may be used to establish an IKE Security Association unidirectional authenticated at the end of ISAKMP Phase 1. This would be used to authenticate only the mobility agents, which could use a technique based on public-key scheme. A unidirectional IKE exchange (using the Extended Authentication ISAKMP Method), either Main Mode or Aggressive Mode will then be used as described in [69].

11.4 ISAKMP Extension for Mobile IP

The IPSec protocol suit was developed to provide data confidentiality, integrity and authentication between non-mobile hosts, i.e. when IP addresses are statically allocated during a communication session. It is much less suited for roaming nodes that use Mobile IP protocol. If new Security Associations have to be negotiated each time a mobile node gets a new IP address, IPSec might introduce a not acceptable delay in Mobile IP handover.

11.4.1 Overall Scheme

With traditional IPSec protocol, when a mobile client moves to a new IP network and gets a new IP address, previous IPSec Security Associations (SAs) are no longer utilizable (being SA parameters in the SAD and SPD and the ISAKMP cookies associated with the old IP address). This requires the establishment of additional SAs, which means a new ISAKMP/IKE Phase One exchange for each ISAKMP SA and a new ISAKMP/IKE Phase Two exchange (Quick Mode) for each non-ISAKMP SA. These additional ISAKMP operations might introduce a high delay in network handover, which is likely to compromise session mobility.

This draft presents a new ISAKMP Payload that is able to extend usage of ISAKMP exchanges to support IP mobility without decreasing network performances. With this new method the mobile node is able to update any ISAKMP SA by sending a single ISAKMP Informational Exchange messages to the correspondent node. Although the message is sent in clear, an ISAKMP Hash payload authenticates message origin and assures data integrity. No computationally expensive operation is required and no sensible data is exposed.

The scope of Security Associations (SAs) in IPSec suit can be easily leveraged for usage even when the IP address of one of the peer host may change, as it is the case in Mobile IP protocol. The ISAKMP scheme for IP mobility handover would be as following:

1. The mobile node (MN) moves to a new network and gets a new IP address (an authenticated mobile-home registration procedure MAY also be done).

2. MN sends (in clear) an ISAKMP message to the correspondent node (CN), e.g. the Mobile IP home agent, which contains its old IP address in a new ISAKMP payload (called “IP Address Update”). A Hash Payload [ISAKMP] is included to assure data integrity.

3. CN uses secrets associated with the old address of the MN (contained in the “IP Address Update” payload) to authenticate the Hash payload. If authentication succeeds,

all SA entries for the old IP address in SAD and SPD databases are updated with the new IP address.

4. CN sends back to the MN a message acknowledging the establishment of the new SAs.

5. The mobile node can utilize the “renewed” SAs to communicate with the correspondent node without the need of any additional ISAKMP/IKE exchanges.

11.4.2 Exchange Type

Any message transmitted in this protocol is based on ISAKMP format, which can be implemented over any transport protocol or over IP itself. The Internet Assigned Numbers Authority (IANA) has assigned UDP Port 500 to ISAKMP. Implementations MAY additionally support ISAKMP over other transport protocols or over IP itself.

Messages transmitted in steps 2 and 4 are based on the “Informational Exchange” mode defined in ISAKMP and IKE. The latter is designed as a one-way transmittal of information that can be used for security association management. The Informational Exchange usually transmits ISAKMP Notify or Delete payloads. In our case its usage is extended to carry a new ISAKMP payload, called “IP Address Update” payload (section 11.4.3).

The Informational Exchange message in step 2 is sent authenticated. This means that the A(uthentication) Only Flag in the ISAKMP header is set to 1. This bit is intended for use with the Informational Exchange with a Notify payload and allows the transmission of information with integrity checking but no encryption. Although Section 4.8 in the ISAKMP RFC [ISAKMP] states that *a Phase 2 Informational Exchange MUST be sent under the protection of an ISAKMP SA*, an exception to that policy is allowed when the A Flag is used, as it is done for “IP Address Update” messages. When the Authentication Only bit is set (1), only authentication security services will be applied to the entire payload of the Informational Exchange and no encryption is applied to the message.

If an Informational Exchange occurs prior to the ISAKMP Phase 1 negotiation, there will be no protection provided for the Informational Exchange. Once an ISAKMP SA has been established, the Informational Exchange MUST be transmitted under the protection provided by the keying material. In step 2 the ISAKMP security association with new IP address has not yet been established and the exchange is done in cleartext (the Encryption bit is set to zero in the ISAKMP header). Nevertheless, data exchanged in step 2 are not sensitive and no security flaw is introduced in the protocol. In step 4 an IPSec Security Association has been established and it may protect the ISAKMP Informational Exchange, thus the Encryption bit (E Flag in the ISAKMP header) may be set to one.

11.4.3 ISAKMP Header and Message Payload

A new ISAKMP payload is defined for usage in message exchanged in steps 2, which contains the old IP address for the mobile node. The Initiator Cookie and the Responder Cookie contained in the ISAKMP header are those utilized for the old ISAKMP SA. These data are sufficient to identify and update any SA existing between the mobile node and the correspondent node.

<table>
<tr><td colspan="5">Initiator Cookie</td></tr>
<tr><td colspan="5">Responder Cookie</td></tr>
<tr><td>Next Payload</td><td>MjVer</td><td>MnVer</td><td>Exchange Type</td><td>Flags</td></tr>
<tr><td colspan="5">Message ID</td></tr>
<tr><td colspan="5">Length</td></tr>
</table>

ISAKMP Header

Initiator Cookie (8 octets) - Cookie of entity that initiated the SA notification exchange to update the IP address. It is the cookie used from the mobile node (in the old IP network).

Responder Cookie (8 octets) - Cookie of entity that is responding to the SA update request. It is the cookie from the home agent/correspondent node.

Next Payload (1 octet) - Identifier for the payload type of the next payload in the message. If the current payload is the last in the message, then this field will be 0. A value of 16 is used in the Next Payload Type to identify the "IP Address Update" payload.

Major Version (4 bits) - indicates the major version of the ISAKMP protocol in use.

Minor Version (4 bits) - indicates the minor version of the ISAKMP protocol in use.

Exchange Type (1 octet) - indicates the type of exchange being used. A value of 5 is in our case utilized to identify the "Informational Exchange" mode.

Flags (1 octet) - indicates specific options that are set for the ISAKMP exchange. For the message exchanged in step 2 the E(ncryption) bit is set to 0, the C(ommit) bit is set to 0, and the A(uthentication Only) bit is set to 1. For the message exchanged in step 4 the E(ncryption) bit may set to 0 or 1, the C(ommit) bit is set to 0, and the A(uthentication Only) bit is set to 0.

Message ID (4 octets) - Unique Message Identifier used to identify protocol state during Phase 2 negotiations. During Phase 1 negotiations and Informational exchanges the value is set to 0.

Length (4 octets) - Length of total message (header + payloads) in octets.

<table>
<tr><td colspan="2">Next Payload</td><td colspan="2">RESERVED</td><td>Payload Length</td></tr>
<tr><td>Type</td><td colspan="2">Counter</td><td colspan="2">RESERVED</td></tr>
<tr><td colspan="5">Old IP Address</td></tr>
</table>

IP Address Update Payload

Next Payload (1 octet) - Identifier for the payload type of the next payload in the message. If the current payload is the last in the message this field will be 0.

RESERVED (1 octet) - Unused, set to 0.

Payload Length (2 octets) - Length in octets of the current payload, including the generic ISAKMP payload header.

Type: Specifies the type of IP address contained in the “Old IP Address” field (set to 0 for IPv4 addresses and set to 1 for IPv6 addresses).

Counter: Indicates the number of times that an ISAKMP SA has been re-associated with a different IP address. The re-association procedure should not be done “too many” times for not exposing the ISAKMP secret to plaintex-known attacks, where the enemy knows both the plaintext and the message digest and thus might try to attack the secret key.

Old IP Address: The IP address belonging to the mobile node in the old network. This is the IP address used by the correspondent node to identify previous IPSec SAs for the mobile node.

11.4.4 Security Associations

In each IPSec implementation there is a nominal Security Association Database (SAD), in which each entry defines the parameters associated with one SA. For outbound processing, entries in the SAD are pointed to by entries in the Security Policy Database (SPD). For inbound processing, each entry in the SAD is indexed by a source IP address, IPSec protocol type, and SPI.

The ISAKMP message transmitted in step 2 is sufficient to identify the ISAKMP SA and all entries in SAD and SPD databases associated with the mobile node. This makes it possible to update them for usage with the new IP address (section 11.4.5). The following Security Association Database fields (used in doing IPSec processing) are kept unchanged when the new SAs are established in SAD and SPD databases to ease mobility with IPSec:

- Sequence Number Counter: used to generate the Sequence Number in AH or ESP headers.
- Sequence Counter Overflow: the flag indicating whether overflow of the Sequence Number Counter should generate an auditable event.
- Anti-Replay Window: the counter and a bit-map (or equivalent) used to determine whether an inbound AH or ESP packet is a replay.
- AH Authentication algorithm, keys, etc.
- ESP Encryption algorithm, keys, IV mode, IV, etc.
- ESP authentication algorithm, keys, etc.
- Lifetime of this Security Association: a time interval after which a SA must be replaced with a new SA (and new SPI) or terminated.
- IPSec protocol mode: tunnel, transport or wildcard. Indicates which mode of AH or ESP is applied to traffic on this SA.
- Path MTU: any observed path MTU and aging variables.

11.4.5 Payload processing in the correspondent node

The ISAKMP cookies and the IP address previously used by the mobile node (transmitted in the “IP address update” payload within the ISAKMP message) are used in the correspondent node to index previous ISAKMP and non-ISAKMP Security Associations. When the correspondent node receives the ISAKMP message from the mobile node (step 2), it uses the secrets associated with the old IP address (SA) to verify data origin and integrity (computing the HASH payload).

If message authentication succeed, the SAD and SPD databases are update with new SA entries, where the old IP address is substituted by the new one and all other data are left unchanged (section 11.4.4). The old IP address of the mobile node is in fact sufficient to index any non-ISAKMP SA entry in SAD and SPD (SPI and the Security Protocol identifier are left unchanged).

Moreover, the two cookies included in the ISAKMP header together with the old IP address contained in the “IP Address Update” payload (section 11.4.3) uniquely identify the ISAKMP SA (an ISAKMP SA is identified by the couple of cookies included in the header of any ISAKMP message). ISAKMP cookies (ISAKMP SA) are also updated, according to the procedure defined for SAD and SPD databases.

11.4.6 Configuration Transaction

When an ISAKMP implementation supports the “ISAKMP Configuration Extension”, as defined in [CFG], the latter could be used to transport the “IP address update” payload (section 11.4.3) and the acknowledgement message transmitted in step 4 (section 11.4.1). [CFG] specifies a new ISAKMP exchange mode, which is called "Transaction Exchange" and has assigned an Exchange Type value of 6 in the ISAKMP header.

A "Configuration Transaction" is defined as two configuration exchanges, the first being either a Set or a Request and the second being respectively an Acknowledge or a Reply message. A common identifier is used to identify the transaction between exchanges. The "Set/Acknowledge" mode, which allows a host that wishes to send information to another host to start a configuration transaction, is suggested for “IP Address Update” usage.

The “IP address update” payload should also conform to the payload format defined in [ISAKMP_CGF], with the Attribute Message Type set to ISAKMP_CFG_IP_UPDATE and a value of 6 assigned.

11.4.7 Security Considerations

Security for the ISAKMP extension described in this draft is consistent with the security provided in the ISAKMP protocol. No secret material is exposed to possible enemies and no Denial of Service threat is introduced. No secret material is exchanged in the plaintext message transmitted in step 2. The procedure described in section 11.4.5 protects the Informational Exchange from Denial of Service (DoS) or other kind of attacks.

A hostile user might send fake ISAKMP messages with “IP address update” payloads but never authenticate the messages (it does not know the secrets used to compute the HASH payload). If a fake “IP Address Update” message is received, authentication procedure fails; thus SAD and SPD databases are not updated, traffic is not redirected, and no computationally expensive operation is required in the correspondent node.

11.5 Mobile IP and (AAA) Accounting

A network domain is likely to require some user’s credentials to be verified before permitting access to the domain’s resources. AAA inter-domain management might be provided within the mobility protocol, even if this architecture is not suggested from the author of this thesis project. Accounting is a basic requirement for public operators and it should be provided even when mobility is not implemented or required, i.e. AAA functionalities should NOT be deployed as a sub-set of the mobility protocol.

AAA verification functionalities are mainly required by the visited network. Therefore, it should be up to the visited domain to implement accounting, rather than relying on the mobile node's mobility protocol. Whereas the MN has to authenticate with the HA to provide communication security, AAA procedures aim to protect the visited domain from theft-of-service attacks. Thus the visited domain needs to obtain some non-reputable signature about the acceptance of offered services, so to be sure that the visiting user will "pay" for services that have been used.

11.5.1 Interaction between Mobile IP and AAA protocols

Visited domains have an active role both in Authentication, Authorization and Accounting (AAA) procedures and in IP mobility implementation. This means that it could be possible to improve the performance of the overall system by integrating accounting and mobility messages, having the (optional) mobility architecture takes advantage from the AAA verification infrastructure.

Foreign agents might be able to convert the Mobile IP Registration Request into a DIAMETER AAA-Mobile-Node-Request, as described in the DIAMETER Mobile-IP Extensions [18]. The MN may request specific security policies for the MN-FA or MN-HA connections, whereas the FA may add extensions that indicate the security services that it requires.

However, if Foreign Agents would contact the AAA server by themselves, acting as autonomous AAA clients, authentication would be too expensive. Too many FAs could be present in a network domain. Therefore, some more hierarchical scheme is required. This hierarchy is usually implemented through Attendant Agents, which act as AAA Clients, AAA Local Authorities (AAAL), which act as AAA Proxies or Servers, and AAA Home Authorities (AAAH), which act as AAA Servers [40].

AAA registration is likely to require a longer time than normal MN-HA registration procedure. Integration of the AAA functions within initial Mobile IP registration scheme is therefore encouraged. If Registration Request and Replies are included in the AAA messages, it is possible to authenticate the user, perform authorization, and process the Registration Request at the same time, saving round trips time.

The mobile node could use, for instance, authentication extensions based on RADIUS [14] or DIAMETER [17] AAA protocols within the Registration Request message. In addition to these extensions, the Mobile Node should also include a Network Access Identification extension [2] to enable the foreign agent to make use of any available AAA verification infrastructure. MN-FA messages extensions described in section 11.2.2 may be used for this purpose. Foreign agents' advertisements could also be extended to provide a nonce that may be used by the MN in the MN-AAA verification request, as prevention from replay attacks.

Moreover, after the initial registration of the MN on a visited domain, the complete AAA verification infrastructure should be no longer needed for subsequent Mobile IP registrations, as long these are done within the same administrative domain (same visited AAA server). It is also suggested, when possible, to package authorization information so that multiple service's authorization requests and replays are carried in a single message.

In the case of a verification request coming from a mobile node, the Authenticator value should be unpredictable and unique over the lifetime of a secret (the password shared between the user and the RADIUS server). Repetition of a request value in conjunction with the same secret would permit an attacker to reply with a previously intercepted response. For ease of implementation, it is expected that the same secret be used for authentication in different network domains. Therefore, Authenticator field in verification requests should exhibit global and temporal uniqueness. The key should also be large and not predictable enough to provide protection against exhaustive search attacks.

Request Authenticator value in an Access-Request packet should also be unpredictable. If not, a fake AAA Client may trick an AAA Server into responding to a predicted future request. AAA Server response could be later used to masquerade as AAA Server versus that mobile user that sends the predicted Access-Request packet.

Any Access-Request message, in AAA, should contain a User-Name attribute and either NAS-IP-Address attribute or NAS-Identifier attribute (or both, although that is not recommended). It must also contain either a User-Password or CHAP-Password attributes. NAI is used both as the User-Name attribute and to identify the AAA Home Server, which will be contacted by visited domains for authentication of the mobile users.

11.5.2 AAA support for Key Distribution

The Authentication, Authorization and Accounting (AAA) verification infrastructure presented in Part 2 (possibly based on DIAMETER) might be extended (as described in the following of this section) in order to support:

- Efficient scaling of security associations for VPN and End-to-end Privacy
- Roaming across administrative domain boundaries
- Key Distribution
- Support for dynamic home agent and address assignment (DHCP)

Each valid MN possesses a NAI that can be used to find a home AAA server, possibly via some broker servers that implement a large-scale roaming agreement. Each mobility device shares a security association (SA) with an AAA server (e.g. a DIAMETER server) within its own home network, identified by the NAI. The AAA servers in visited domains can either share a direct security association with the Home Network or use an intermediate broker. In both cases, they use AAA to get all the secrets needed from the home domain, for Accounting, Access Control and Authorization of mobile users. Each mobility agent also shares a security association with its local AAA server.

The AAA server on the Home Network is responsible for authentication of the user and generation of dynamic session keys to be distributed among all mobility agents. It also helps in the dynamic assignment of home agent and home address to the mobile node, providing for instance the keys utilised by DHCP Servers for message authentication (section 12).

When the mobile node wants to authenticate itself with a visited domain, it includes a MN-AAA Authentication extension in its registration request. It might not even have a Security Association with its home agent. Mobile IP requires the mobile node to have a security association with its home agent, but this authentication could be based on the NAI and provided through the AAA Server.

The AAA home server, using the information provided in the MN-AAA Authentication extension, can authenticate the mobile user. It also generates the keys for the mobile node and for the other entities that required it. Visited Domains should be allowed to add requests for other keys, such as those used by DHCP servers and FAs.

The AAA verification infrastructure might be used to distribute keys to be shared between the Mobile Node and the Home Agent (section 11.2), the Mobile Node and the Foreign Agent (section 11.2.2) and between the Foreign Agent and the Home Agent (section 11.2.3). Keys to be used in VPN remote access (section 13.3) are also provided. IPSec and IKE use these keys as pre-shared keys to establish other session keys. Section 0 details how the AAA infrastructure is used to distribute these keys.

Both the Mobile IP Key Extensions, as defined in [90], and the Mobile IP Transform Policy Extension, as defined in [74], specify extensions to Mobile IP Registration Reply messages that could be used to distribute the security information to mobile nodes. Being user authentication based on the Network Access Identifier (NAI) it is possible for a mobile node to use these extensions even without having a home address. This means that, when IP addresses are dynamically assigned (as described in section 12) it is still possible to use the latter. All keys must be encoded according to security associations existing between the AAA home server and the related entities. MN-HA and MN-FA Key Extensions or Mobile IP Transform Policy Extension are inserted in Mobile IP Registration Replay packets. DIAMETER Transform Policy Extension [74] may be also used to provide keys to visited domain or home agents.

All of these extensions contain both the necessary secret keys and a way to identify the cryptographic algorithms to be used. Security Parameter Index (SPI) is used for this purpose. The lifetime of each SA must also be distributed with the SPI. SPI values should conform to those defined in Mobile IP [89].

The mobile node, using the secrets shared with its AAA home server and the information contained in the MN-AAA Authentication Extension, authenticate the Registration Reply packet and the enclosed Keys (MN-HA and MN-HA Key Extensions). The same can be done by other entities, such as the visited domain or the home agent, using their DIAMETER security associations with the AAA mobile user' home server. Brokers may be used to further scale the system.

Keys destined to the Mobile Node and to the visited domain may be sent within the same DIAMETER message [74]. MN-FA and FA-HA keys to be used by the visited domain may be encrypted on a hop-by-hop basis so that the visited domain can decode them. MN-FA and MN-HA keys to be used by the mobile node may be encrypted with the MN's original shared secret. In this way only the MN can decode them, even if the message is managed by other AAA entities.

The IPSec document roadmap [60] states that "The default automated key management protocol selected for use with IPSec is IKE. Other automated SA management protocols MAY be employed". This means that Key distribution based on the AAA infrastructure conforms to ISAKMP specification as an "other automated SA management protocol".

DIAMETER servers on visited domains maintain a copy of the keys distributed during the AAA authorization phase with the Home DIAMETER Server. As long as a Mobile Node moves within the same foreign domain, DHCP servers and Mobility Agents may request keys from the local DIAMETER server, without the need for further communication with the AAA Home Server and the Home Network.

Scalability is of major importance in any network architecture. This means that each Foreign Agent, DHCP Server, or Mobility Enabled Router shares a security association with a local DIAMETER server. This has a security association with other DIAMETER servers and possibly with several DIAMETER Brokers, creating a hierarchical and flexible AAA architecture, as described in section 11.5.1.

Under certain circumstances, mobile nodes may boot on subnets that are technically foreign subnets, but the services they need are all local. This could be the case for airports and other common areas where business clients are likely to spend time. In these cases, communication with the home domain might be not necessary. It could be worth only to use the AAA verification infrastructure, i.e. MN-FA-AAAL, without establishing any authentication with the home agent.

The DIAMETER Accounting extension [4] may be used to provide information on resource consumption to mobile nodes, visited domains and home network. It is also suggested that the Accounting Data Interchange Format (ADIF) [1] be used to encode this kind of information.

12 Dynamic Allocation of IP Addresses

The Dynamic Host Configuration Protocol (DHCP) [31] provides an extensible framework through which a host can acquire various configuration parameters from a centrally managed server. Configuration parameters that may be obtained through DHCP include (among many others) the host's IP address, subnet mask, default router, DNS domain, DNS server and NTP servers.

When Dynamic Host Configuration Protocol (DHCP) is used to assign temporary IP addresses, several security problems have to be addressed. Network administrators may wish to allocate IP addresses only to authorised hosts. Additionally, in some network environment, e.g. wireless networks without link-layer authentication, it would be necessary to provide authentication for each exchanged DHCP message.

12.1 Security threats against DHCP

Several security threats arise from the use of DHCP. If DHCP messages are not authenticated, for instance, clients may be subject to Denial of Service or man-in-the-middle attacks through the use of fake DHCP servers. These could provide incorrect configuration information to the client. Any non-authorised user could, using DHCP, masquerade as a legitimate user and do some theft-of-service attack. Client receiving bad DHCP settings might for example get false DNS address or sends traffic to a sniffer and not a gateway. Many other attacks based on address spoofing could also be eased.

Anyway, the most severe threats could be Denial-of-Service (DoS) attacks against the DHCP server. These attacks typically involve the exhaustion of the valid IP addresses, network bandwidth or CPU capacity. An attacker could, for instance, request all available IP addresses from a DHCP server by sending requests with fabricated client MAC addresses.

Another possible threat to Denial of Service might be to have a legitimate or fraudulent user responding to any DHCP´s check for duplicate address assignment (i.e. duplicate address detection through ICMP echo request from the server or ARP from the clients [31]). All these threats could be especially severe in "hostile" environments, where the network medium is not physically secured, as is the case for wireless networks.

From security threats hereby presented, it results that the most important security needs that arise when using any protocol (e.g. DHCP) for dynamic assignment of IP addresses are:

- Clients and servers need to authenticate each other.
- Server needs to check client's privileges.
- Client checks integrity of the server's message.
- Client and server agree on further security.
- Migration issues must be addressed.
- Servers should be able to support secure and non-secure DHCP clients.

If access to the network is protected either by physical means or through authentication procedure implemented at the link layer, threats to DHCP are greatly reduced and the threat to DHCP becomes inherently an insider problem.

12.2 DHCP Authentication Schemes

Some authentication mechanism should be introduced or added to the DHCP protocol to protect the network from security threats in IP addresses assignment. This authentication scheme should also provide replay protection, through the use of either nonces or timestamps.

12.2.1 Authentication of DHCP messages

A recent Internet draft by Droms and Arbaugh [32] outlines a mechanism (Protocol 1) for adding authentication information to DHCP messages. The mechanism guards against source spoofing attacks, message alteration, and replays. It assumes that the entities exchanging authenticated information share a secret key not known to anyone else. The sender uses this key to compute a keyed hash (or MAC) over the information to be protected and a replay detection field.

Each DHCP message includes the Message Authentication Code (MAC, described in section 4.2 of Part 2) to provide authentication over the message. To provide complete authentication, a client must authenticate when sending a DHCPDISCOVER message. The DHCP server must authenticate when it replays to the client with an extended DHCPOFFER message that includes a nonce and the Message Authentication Code. In the case where it is either required only authentication of the mobile node or the DHCP server, one of these authentications could be omitted. If authentication is desired, the receiver recomputes the MAC over the same fields using its copy of the shared key. It compares then the result against the MAC value received with the incoming message. A successful match authenticates the sender.

If Relay agents are used, special attention should be paid to the fields over which to compute the hash function. Because a DHCP relay agent may alter the values of the 'giaddr' and 'hops' fields in the DHCP message, the contents of those two fields must be set to zero for the computation of any hash function over the message header [32].

A Message Authentication Code is usually easy to compute and do not introduce long delay in the network procedures. However, the use of shared secrets requires an out-of-band distribution of keys to clients. This might be easy to manage within an organization but does NOT scale well across different network domains. User identification should be based on some permanent user ID. It is not possible to use IP addresses as identifiers, as these are dynamically assigned. A Network Access Identificator (NAI) could be considered as a valid alternative, as well as local IDs to be used only within local networks.

Another protocol (protocol 0) is also described in [32]. This carries an authentication token (a password) in the clear and only offers weak authentication protection from network errors. The latter can not be used when dealing with fraudulent attacks.

12.2.2 DHCP authentication based on IPSec

It might be desired to leverage security for dynamic assignment of IP addresses on the use of the IPSec protocol, without having to define a new security protocol or modify DHCP protocol. A problem with this approach is that the client/host has no ability to interact via IP before a valid DHCP interaction is completed. To address this problem, an non-authenticated DHCP exchange might be followed by a (IPSec-based) authentication exchange between the DHCP client and server.

The IP address assigned before the authentication phase is leased for a short time and can be used in the network only after the authentication exchange. The overall procedure to assign IP addresses and configuration information is divided in three phases:

- Phase 1: Minimal networking functionalities are provided using un-trusted DHCP.
- Phase 2: A (IP-sec) security association is established between DHCP client and server.
- Phase 3: DHCP with security is used to renew leased address and get parameters.

During phase 1 a not-trusted configuration is provided to the client to allow this to use IP networking. At end of Phase 1, client does not trust received configuration parameters and server does not trust the client.

In phase 2 it is established a secure (IPSec) communication channel, using parameters temporary assigned in phase 1 to communicate with the DHCP server. A secure channel (e.g. by ISAKMP and IKE) is created. Authentication, key management and negotiation of other parameters may be provided during this phase.

Finally, in phase 3 configuration parameters are renewed using DHCP and protecting the communication with the security association established in phase 2. This is done both to renew lease of the IP address and get fresh set of parameters.

The technique hereby presented works especially well with legacy devices and devices that simply can't support security, as long as they are run in environments that can support them. Either the client or the server can choose to go from phase 1 to phase 2 and if none of them choose to do this, the DHCP protocol runs just like it does today, without authentication. A drawback with the use of IPSec to authenticate DHCP exchanges is that a system using this technique is still susceptible to Denial of Service attacks and that the computational load for IPSec could be high.

Another problem with IPSec is the state maintenance in DHCP clients and servers. Common implementation of ISAKMP involves state maintained in memory. If the server crashes and the client doesn't, then when the client tries to renew, then the packet will go nowhere. The client will think it is sitting on top of an active SA and have no way to determine that the SA is not any longer valid. If this should happen, then the client will either timeout, break the connection and rediscover, or it will rediscover right away, either way breaking the existing TCP connections. This is a more general problem in IPSec.

12.2.3 Dynamic Registration and Configuration Protocol (DRCP)

An alternative approach than securing the DHCP protocol might be the development of a new protocol for dynamic IP address assignment and configuration of network clients. The Dynamic Registration and Configuration Protocol (DRCP) [73] is an example of such a new protocol, which is an enhanced version of DHCP and is optimized for roaming users.

The Dynamic Registration and Configuration Protocol (DRCP) is based on DHCP and is compatible with DHCP protocol if only DHCP servers are present in the network. The most important differences between DRCP and DHCP are that DRCP allows clients to know when to get a new address (independent of the access technology) and agents to be either a relay agent or a server depending on client's authorization.

In DRCP clients are identified by NAI rather than hardware address (as it is in DHCP). This means that DRCP authenticates users more than hosts or physical interfaces. It is also possible to associate a priority class with a NAI address. Finally, DRCP adds options for QoS negotiation, service activation, and authentication.

Another important change in DRCP respect to DHCP is that DRCP does not require an ARP check before an assigned address is used. DHCP specification says in fact that a client SHOULD do an ARP check before assigning an IP address as a consistency check. This checking result in long delay before communication can resume after a move but protect the network from double assignment of IP addresses (due for instance to DHCP server crash). DRCP eliminates this operation to allow rapid configuration (milliseconds rather than seconds), which is especially useful for roaming and for mobile users.

DRCP messages are a superset of DHCP messages. Therefore, any DRCP implementation support all DHCP messages defined in the DHCP specification [31]. DRCP also re-defines or better extend DHCP messages, as for the DRCP Message Advertisement (a superset of DHCPNAK), which informs about available IP addresses, or the DRCP Discover Message (a superset of DHCPDISCOVER), which can be both broadcast or unicast.

DRCP is strongly integrated with AAA protocols, as RADIUS and DIAMETER. DRCP servers are in fact able to inter-work with AAA protocols to support inter-domain (AAA) functionalities. It could for instance be possible to use DRCP in combination with AAA to retrieve both IP addresses used in SMIP (see appendix F). Also, DRCP requires AAA support only during the first DRCP interaction in a network domain. Finally, several Security Associations are defined between DRCP agents (Clients, Proxies, Servers and Public verification Agents).

12.3 Key Distribution and Roaming

Key distribution is not trivial with respect to DHCP clients and servers. If shared key mechanisms are used the secret keys must be shared between each client and all possible DHCP servers. This is especially difficult when inter-domain roaming has to be provided, as it is for instance often the case when Mobile IP is used. Actual proposals mainly focus on solving the intra-domain problem, where the out-of-band exchange, i.e. often manual distribution, of shared secrets is feasible [32].

Two different solutions are possible to address the case of inter-domain roaming. The first consists in having a different pre-shared secret with any DHCP server in any network domain. A client would have multiple shared secrets for use in different domains, requiring one time cost to configure a new secret for each domain use. Therefore, this solution is not scalable to a global system. Moreover, a client might have difficulty in deciding which key to use in different domains. Either a manual intervention to change domains or a domain negotiation protocol would be required.

Making use of secrets shared among many clients, e.g. all those belonging to the same home domain, could ease the previous problem. However, the sharing of keys is strongly discouraged. It greatly eases unauthorised clients' access to the secret key and enables them to masquerade as an authorised user.

An alternative solution might be using a Public Key Infrastructure (PKI) or some AAA-based scheme, e.g. RADIUS or DIAMETER as described in section 11.5.2, to distribute keys among different domains. This may introduce a high delay in the DHCP procedures but still could be the only valid solution, when global roaming is required (section 6.4). Vipul Gupta for instance proposes a new protocol (Protocol 2) [41] based on public key cryptography for DHCP authentication within the general framework outlined in [32].

This new protocol supports the use of digital signatures and multiple forms of replay detection in authenticating DHCP messages. The use of digital signatures simplifies key management and also allows authentication of mobile clients that roam between different administrative domains.

The general format of the DHCP authentication option in this draft closely resembles the format defined in [32] with minor exceptions. The major changes introduced are:

- The "Global Replay Counter" field in [32] has been renamed "Replay Detection Field" to indicate support for multiple forms of replay detection.
- A Key ID field is defined as a generalisation of the "secret ID" field in Protocol 1 [32] and identifies the public-key or shared-key needed to verify the authenticator.
- An Authenticator field is introduced as a generalisation of the "MAC" field in Protocol 1 [32]. This can contain either a MAC or a digital signature depending on whether the authentication algorithm uses symmetric or asymmetric key cryptography.

If only intra-domain roaming is required, some lightweight PKI or LDAP architecture could be used [9]. When inter-domain roaming is supported, distributing secret/public keys using either the AAA infrastructure, as described in section 11.5.2, or a complete PKI, could be the only possible choice. An authentication scheme for DHCP messages with roaming might be as following:

1. The roaming client establishes link connectivity (e.g. completing an 802.11 association procedure) and sends out a DHCPDISCOVER request (with a request for authentication). The DHCP client identifier is set to contain the roaming user's Network Access Identifier (NAI).

2. By looking at the NAI, a DHCP server on the visited network can determine the network domain to which the client belongs. It checks existing roaming agreements and responds with a DHCPOFFER message containing an authentication option. In this option it is specified the authentication algorithm and the key distribution technique used (e.g. X509_CERT_CHAIN). The DHCP public key is also included, authenticated by mean of a certificate chain signed by the network domain the roaming user belongs to.

3. Using a locally available copy of its network domain's public key, the client verifies DHCP server’s public key and signature and authenticates the offer. If authentication is successful, the client sends out a DHCPREQUEST message. In the authentication option, the Key ID Type is set to X509_CERT_CHAIN, the Key ID Value is set to include the client's certificate issued by ISP-A and the authenticator contains a digital signature computed by the client using its private key.

4. The server first verifies the client's certificate (this may require it to interact with another entity such as a certificate repository) and uses the public key it contains to verify the client's signature. If verification succeeds, it sends back a DHCPACK message completing the sign-on process, otherwise it sends back a DHCPNAK

A drawback in the use of public key cryptography is the high computational power that public cryptography operations require. Therefore, it might be convenient to limit the use of public key only to initial authentication when entering a new network domain. During this initial authentication phase, a shared secret might be established for future authentication within this domain (based on some of the techniques described in section 12.2).

It has also been recently proposed, for security reasons, to have circuit access equipment and DCHP Relay Agents, which are trusted component, add information to the DHCP client requests that are forwarded to DHCP servers [86]. This helps when some mapping is possible between Users’ ID and points of attachment. Unfortunately, this is not the case for mobile devices and not at all for roaming users.

13 Virtual Private Networks

Secure access to Virtual Private Networks (VPNs) is a special case of end-to-end security. A VPN uses a combination of authentication, data encryption, and tunnelling to create a secure channel between a user and a corporate network or between two networks. In a typical remote access situation, the users dial in to the local access provider’s POP, establish a connection to the Internet, and then identify themselves to the corporate VPN’s authentication system.

The VPN verifies the identity of a user either on the basis of username and password, token card and PIN number, or some other mechanism. Upon successful authentication, tunnelling and/or encryption is set up for all traffic between the VPN client and VPN server.

VPNs offer direct cost savings over other communications methods (such as leased lines and long-distance calls), they can also offer other advantages, including indirect cost savings as a result of reduced training requirements and equipment, increased flexibility, and scalability.

13.1 Remote access from the Internet

Secure access to remote private networks (i.e. Virtual Private Networks) is one of the main challenges for the deployment of roaming and mobility services in wide area networks. Rather than depend on dedicated leased lines or frame relay's permanent virtual circuits (PVCs), an Internet-based VPN uses the open, distributed infrastructure of the Internet to transmit data between corporate sites.

Using Internet for VPN solutions is cost savings, when compared to traditional distributed corporate networks (built using leased T1 (1.5 Mbps) links and T3 (45 Mbps) links). The latter must in fact deal with tariffs that are structured to include an installation fee, a monthly fixed cost, and a mileage charge. These costs are added up to monthly fees that are greater than typical fees for leased Internet connections of the same speed. Another cost-advantage in using Internet-based VPN is the possibility of paying only for transmitted data, without the need to be charged for a T1 or T3 connection that is fully used only during busy times of the day.

Companies using Internet to deploy VPN solutions usually set up connections to the local connection points of their Internet service provider (ISP) and let the ISP ensure that the data is transmitted to the appropriate destinations via the Internet. Thus, the rest of the connectivity details are left to the ISP's network.

Because the Internet is a public network with open transmission of most data, Internet-based VPNs must include measures for encrypting data passed between VPN sites, which protects the data against eavesdropping and tampering by unauthorised parties. Authentication must also be provided. Various password-based systems, hardware-based tokens, digital certificates or challenge-response systems, such as the CHAP protocol (section 4.1.2 in Part 2) or the Remote Authentication Dial-In User Service protocol (section 4.2.2 in Part 3), can be used to authenticate users on a VPN (and control access to network resources).

Security services for VPN can be provided between two different kinds of end-points, either an individual computer or a LAN with a security gateway, which might be a router or firewall. Only two combinations of these end points, however, are usually considered in designing VPNs. In the first case (LAN-to-LAN) a security gateway at each end point serves as the interface between the security channel and the private LAN. Users on either LAN can transparently use the security service to communicate with each other.

The second case (client-to-LAN security) is usually used for a mobile user who wants to connect to the corporate LAN. The client, i.e. the mobile user, initiates the establishment of the security channel in order to exchange traffic with the corporate network. To do so, special security software must be provided to the mobile client to communicate with the gateway and protect the destination LAN.

13.2 Point-to-Point Protocol (PPP) –based solutions

The traditional way of providing secure access to remote corporate network is the establishment of an authenticated and confidential link-layer connection (usually based on Point-to-Point Protocol, PPP) between the mobile node and the visited network. At present, however, link-layer-tunneling protocols are insufficient to be VPN solutions on their own. They offer no protection against eavesdropping and tampering at the network layer where intruders work and, in general, they do not provide the data encryption, authentication, or integrity functions that are critical to maintain VPN privacy.

When PPP [109] is used alone, the remote user initiates communication with the corporate LAN by establishing a PPP connection to the edge device (e.g. an Internet Server Provider). Both the edge device and the corporate LAN network are connected to the Internet. Therefore, the PPP client encapsulates the native packet, which then traverses the Internet in a valid IP packet. The PPP header contains an encapsulated protocol identifier and a frame checksum that is used to verify data integrity upon receiving the PPP frame. The PPP server at the corporate site strips the PPP framing and delivers the native packet to its destination.

New IP Header	**PPP Header**	**Native Packet**	**Frame Checksum**

PPP encapsulation

There are four steps to establish a PPP connection:

- Link Control Protocol (LCP) is used to establish, maintain, and close the actual links.
- User authentication occurs using the method selected in step 1. The level of security that can be negotiated ranges from clear text password authentication (PAP) to encrypted challenge-response authentication (CHAP) [108].
- Various Network Control Protocols (NCP) are used to configure the protocols used by the remote client. Dynamic IP addressing takes possibly place in this step.
- PPP-encapsulated datagrams are transmitted between the two endpoints. PPP encapsulation prepends a PPP and a new IP header. At the exit point of the tunnel, a PPP server strips off the prepended headers and forwards the original protocol to the addresses.

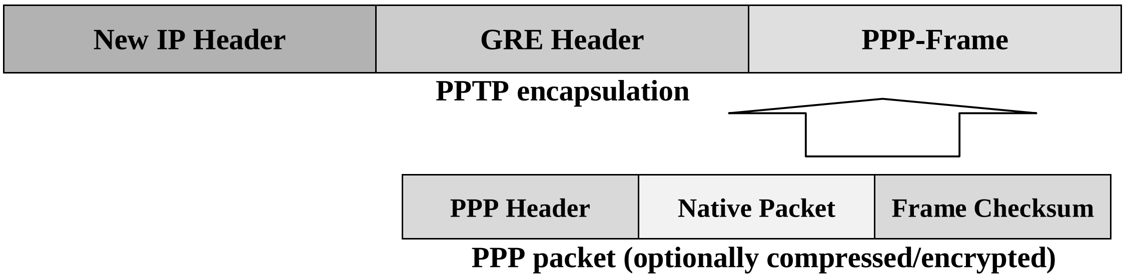
No security, other than limited user authentication is provided in PPP protocol. However, since any valid IP protocol can be tunnelled by it, running PPP over IPSec can enhance security. Moreover, other security mechanisms are often used between foreign domains and the home domain, i.e. the corporate network.

Three different protocols can be used for creating VPNs over the Internet as extensions to the point-to-point protocol (PPP) connection as dial-in protocol; point-to-point tunneling protocol (PPTP), Layer 2 forwarding (L2F), Layer 2 tunneling protocol (L2TP), and IP security protocol (IPSec).

13.2.1 Point-to-Point Tunneling Protocol (PPTP)

Point-to-Point Tunneling Protocol (PPTP) [46] is a widely deployed solution for dial-in VPNs. It builds on the functionality of PPP to provide remote access that can be tunnelled through the Internet to a destination site. PPTP was designed to provide authenticated and encrypted communications between a client and a gateway or between two gateways without requiring a public key infrastructure (by using a user ID and password). It was first delivered in 1996, two years before the availability of IPSec and L2TP. The design goal was simplicity, multiprotocol support, and ability to traverse a broad range of IP networks.

As currently implemented, PPTP encapsulates PPP packets using a modified version of the generic routing encapsulation (GRE) protocol, which gives PPTP the flexibility of handling protocols other than IP, such as Internet packet exchange (IPX) and network basic input/output system extended user interface (NetBEUI). The visited domain encapsulates PPP data into GRE that are forwarded to the corporate domain. PPP data, once received by the corporate domain, are stripped out of the GRE packets and, decrypted, and verified.



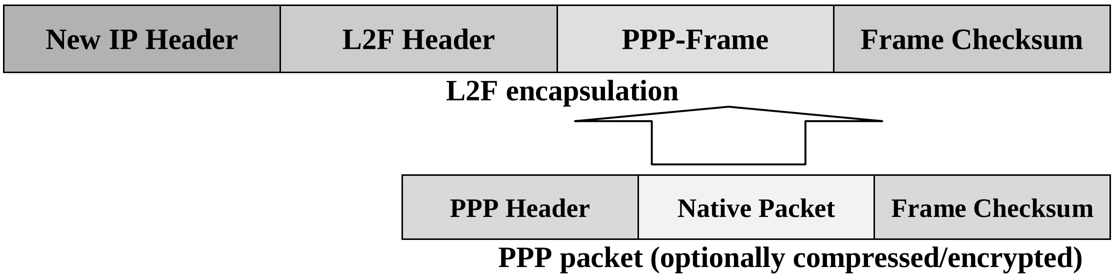
Because of its dependence on PPP, PPTP relies on the authentication mechanisms within PPP, namely password authentication protocol (PAP) and CHAP. Because there is a strong tie between PPTP and Windows NT, an enhanced version of CHAP, MS–CHAP, is also used, which utilizes information within NT domains for security. Similarly, PPTP can use PPP to encrypt data, but Microsoft has also incorporated a stronger encryption method called Microsoft point-to-point encryption (MPPE) for use with PPTP.

Aside from the relative simplicity of client support for PPTP, one of the protocol's main advantages is that PPTP is designed to run at OSI Layer 2, or the link layer, as opposed to IPSec, which runs at Layer 3. By supporting data communications at Layer 2, PPTP can transmit protocols other than IP over its tunnels. PPTP does have some limitations. For example, it does not provide strong encryption for protecting data nor does it support any token-based methods for authenticating users.

13.2.2 Layer 2 Forwarding (L2F)

Like PPTP, Layer 2 Forwarding (L2F) [122] was designed as a protocol for tunnelling traffic from users to their corporate sites. One major difference between PPTP and L2F is that, because L2F tunneling is not dependent on IP, it is able to work directly with other media, such as frame relay or asynchronous transfer mode (ATM).

The L2F header contains a packet key that is part of the authentication process and (optionally) a checksum that is used for testing data integrity. Like PPTP, L2F uses PPP for authentication of the remote user, but it also includes support for TACACS+ and RADIUS for authentication. L2F also differs from PPTP in that it allows tunnels to support more than one connection.



There are also two levels of authentication of the user, first by the ISP prior to setting up the tunnel and then when the connection is set up at the corporate gateway. Because L2TP is a Layer 2 protocol, it offers users the same flexibility as PPTP for handling protocols other than IP, such as IPX and NetBEUI.

13.2.3 Layer 2 Tunneling Protocol (L2TP)

Layer 2 Tunneling Protocol (L2TP) [114] is being designed by an IETF working group as the heir apparent to PPTP and L2F, designed to address the shortcomings of these past protocols and become an IETF–approved standard. It intends to combine the best features of L2F and PPTP, at the same time trying to provide interoperability between diverse vendors.

L2TP uses PPP to provide dial-up access that can be tunneled through the Internet to a site. However, L2TP defines its own tunneling protocol, based on the work done on L2F. L2TP transport is being defined for a variety of packet media, including X.25, frame-relay and ATM. L2TP over IP uses UDP port 1701 and includes a series of L2TP *control* messages for tunnel maintenance. L2TP also uses UDP to send L2TP-encapsulated PPP frames as the tunneled *data*. The encapsulated PPP frames can be encrypted or compressed. To strengthen the encryption of the data it handles L2TP uses IPSec's encryption methods.

Tunnel and session management is performed via control messages containing Attribute-Value Pairs (AVPs). Control messages use a reliable control channel within L2TP to guarantee delivery. A tunnel consists of a control connection carrying the control messages and any number of L2TP sessions carrying encapsulated PPP datagrams. Having different tunnels between the same connection also allow the creation of tunnels with different Quality of Service.

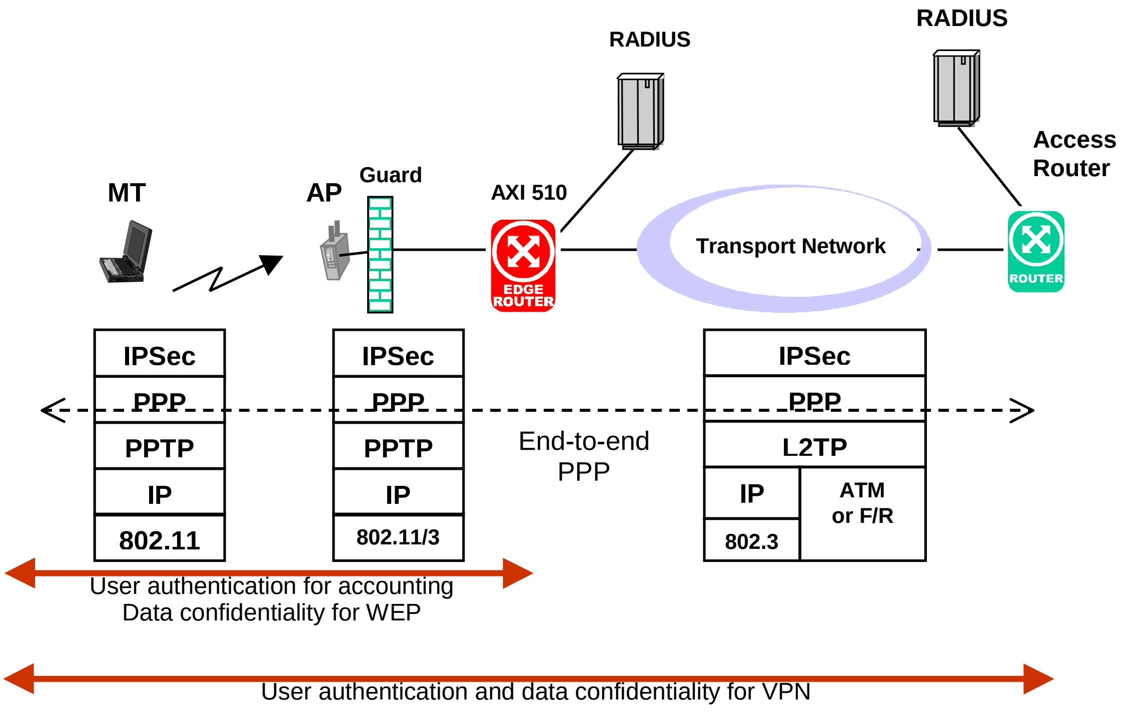
Because it uses PPP for dial-up links, L2TP includes the authentication mechanisms within PPP, namely PAP, CHAP and the Extensible Authentication Protocol (EAP), which supports token card and smart card authentication mechanisms. Similar to PPTP, L2TP supports PPP's use of the extensible authentication protocol for other authentication systems, such as RADIUS. PPTP, L2F, and L2TP all do not include encryption or processes for managing the cryptographic keys required for encryption in their specifications.

The current L2TP draft standard recommends that IPSec be used for encryption and key management in IP environments, especially between the two network domains. The main advantage of this solution is that it introduces little overhead on the network access link. In fact, applying IPSec only between the visited gateway and the corporate network does not add any overhead to the last-link between the visited domain and the client. The latter might be low bandwidth and maybe protected by other link-layer authentication techniques, as it is for example the case for most cellular networks.

13.3 Proposals for WLAN networks

From analyses done in the previous sections L2TP appears to be the best solution for high sensitive data and for multi-vendor interoperability, both for client-to-gateway and gateway-to-gateway applications. However, its usage of IPSec does require a PKI to be scalable and, because of incompatibilities between IKE and NAT, L2TP (used with IPSec) can not pass through typical NATs. For this reason PPTP should be considered as an interesting choice for all scenarios that do not require the sophistication of IPSec-based communications, who do not want to deploy a PKI, or who require an NAT-capable VPN protocol.

PPTP might be used in both client-to-gateway and gateway-to-gateway scenarios for low confidential data communication. With mutual client/server authentication based on users' passwords and encryption keys seeded by the authentication process, PPTP is easy and inexpensive to set up and simple to administer. By virtue of its design, PPTP can also be passed through Network Address Translators (NAT) and this eliminates the requirement that each PPTP end-point have a registered IP address when used across the Internet.



L2TP is the best solution when data have to be delivered over different network channels (both IP-based and not IP). The use of IPSec protocol on top of PPP protocol (as in L2TP, section 13.2.3) introduces a high overhead and, finally, might be redundant when PPP is not necessary for establishing a dial-up connection and when data are known to be transmitted only over IP networks. In wireless LAN networks, for instance, link layer connection is provided without the use PPP negotiation. Therefore, IPSec could be used directly over the native packet and without the need for establishing any PPP authenticated tunnel (as long as dial-in access is done without PPP protocol and data are delivered over pure IP networks).

A drawback with the use of a pure IPSec solution is that in most client-to-gateway VPN situations user authentication and internal address configuration are among the critical aspects of security and management. Unfortunately, the latter are not yet well supported in IPSec. Multicast support and defined methods for carrying multi-protocol traffic might also be essential, particularly in gateway-to-gateway scenarios. Basic IPSec tunnel mode does not have defined standard methods for extensible user-based authentication and address assignment for accomplishing these aspects, thus IPSec tunnel mode would be unsuited to most client-to-gateway VPN situations if new extensions will not be provided. Additionally, IPSec Remote Access solutions should be integrated with existing network infrastructures, such as DHCP, and with IETF standards for extensible authentication, such as EAP.

It is especially for the previous drawbacks that L2TP is often used. It is an interoperable protocol that addresses the current shortcomings of IPSec-only client-to-gateway and gateway-to-gateway scenarios (user authentication, tunnel IP address assignment, and multiprotocol support). L2TP has broad vendor support, particularly among the largest network access equipment providers, and has verified interoperability. By placing L2TP as payload within an IPSec packet, communications benefit from the standards-based encryption and authenticity of IPSec, while also receiving a highly interoperable way to accomplish user authentication, tunnel address assignment, multiprotocol support, and multicast support using PPP. This combination is commonly referred to as L2TP/IPSec. Lacking a better pure IPSec standards solution, L2TP/IPSec provides the best standards based solution for multi-vendor, interoperable client-to-gateway VPN scenarios.

13.3.1 Internal address configuration with IPSec

One of the problems that have been pointed out with usage of standalone IPSec to provide VPN solutions is the lack of a method to support internal address configuration. In many remote access scenarios, a mechanism for making the remote host appear to be present on the local corporate network is quite useful. Assigning the host a “virtual” address from the corporate network, and then tunneling traffic via IPSec from the host’s ISP-assigned address to the corporate security gateway may accomplish this. A possible schema for dynamic DHCP configuration of an IPSec Tunnel Mode (taken from a recent Internet draft of B. Patel et al [84]) might be as described in the following of this section.

The remote host on the Internet connects to the VPN server and establishes an IPSec tunnel to it. It then interacts via the IPSec tunnel with a DHCP agent, which provides the remote host with an address from the corporate network address space. The remote host subsequently uses this as the source address for all interactions with corporate resources. This implies that the corporate security gateway continues to recognise the host's original, routable IP address as the tunnel endpoint. The virtual identity assumed by the remote host when using the assigned address appears to the corporate network as though it were situated behind a security gateway bearing the original routable IP address. All the traffic between the remote host and the corporate network will be carried over the IPSec tunnel via the VPN server as shown below. It is also important to point out that, obviously, the IP address can not be used as user ID, being it dynamically allocated. The Network Access Identificator (NAI) [20] may instead be used.

Since the DHCP server will typically not reside on the same machine as the VPN server, it is necessary for the VPN server to act as a DHCP relay, as well as an IPSec security gateway between the Internet and the Intranet. A typical configuration of the remote host in this application would use two addresses; one as an interface to connect to the Internet (Internet interface), and a second one as a virtual interface to connect to the Intranet (Intranet interface). The IP addresses of the Internet and Intranet interfaces are used in the outer and inner headers of the IPSec tunnel mode packet, respectively. The configuration of the Intranet interface of the IPSec tunnel mode host is accomplished in the following steps:

1. The remote host establishes an IKE security association with the VPN server in main mode or aggressive mode (this IKE SA serves to secure additional quick mode SAs).
2. The remote host establishes a DHCP SA with the VPN server in a quick mode exchange. The DHCP SA is an IPSec tunnel mode SA established to protect initial DHCP traffic between the VPN server and the remote host.
3. DHCP messages are sent back and forth between the remote host and the DHCP server, using the VPN server as a DHCP relay. This traffic is protected between the remote host and the VPN server using the DHCP SA established in step 2. After the DHCP conversation completes, the remote host’s Intranet interface obtains an IP address as well as other configuration parameters.

4. The remote host MAY request deletion of the DHCP SA since future DHCP messages will be carried over a new VPN tunnel. Alternatively, the remote host and the security gateway MAY continue to use the same SA for all subsequent traffic.
5. The remote host establishes a tunnel mode SA to the VPN server in a quick mode exchange.

13.4 Key Distribution Support (PKI and AAA)

All previous protocols for Virtual Private Networking need some infrastructure able to provide and maintain secret keys between roaming nodes and corporate networks. IKE, for instance, requires some independent way of securely establishing the identity of the peer entities, such as a Public Key Infrastructure (PKI) or the usage of pre-shared keys.

Unfortunately, a Public Key Infrastructure used to distribute keys to users desiring to remotely access the corporate network may be too expensive, if the same were not used also for e-commerce or some other mean. The use of pre-shared secrets does not scale at all. It also introduces several security threats, finally requiring some other secure way to assign these pre-shared secrets. Therefore, the PKI key distribution system for VPN should try to reuse as mush as possible already deployed Public Key Infrastructures (or be leveraged on usage of the foreseen future general-purpose PKI).

As an alternative to a PKI system, the (DIAMETER) AAA infrastructure could also be used to dynamically distribute security policies and keys, as described in section 11.5.2. This obviates the need for a parallel public key infrastructure (PKI) or other key distribution mechanisms. In this way, the cost of the AAA infrastructure is shared between different services. Moreover, an AAA verification system will be already available for inter-domain roaming support and for accounting services (as described in the second part of the thesis) and in this case only a little additional cost would be introduced for key distribution.

Using the AAA infrastructure for key distribution, secrets would be stored on the home AAA server (on the corporate network) and might be dynamically distributed to remote nodes within AAA messages (section 11.5). DIAMETER messages will be used to carry secret information, protected on a hop-by-hop basis [74]. In case Mobile IP protocol were utilized, keys might also be distributed to remote user (Mobile Node, MN) using the Key Extensions (defined in [90]) within the Mobile IP Registration Reply, protected by the secret shared between the MN and the home AAA server.

14 An integrated architecture

Wireless LAN networks based on IEEE 802.11 and HiperLAN/2 standards were introduced in the second part of this thesis. Security techniques for data confidentiality, integrity, and authentication over the radio link were thereby described and extended to the public environment. The overall architecture of the proposed solution also defined accounting and inter-domain roaming support to allow for correct charging in public scenarios, where security needs to be integrated with some (AAA) verification infrastructure (see Part 1).

In Part 3 (sections 11-13) three advanced services (IP mobility, dynamic assignment of IP addresses, and VPN schemes) have been presented and secured. These may be provided over the infrastructure presented in Part 2 and can be considered as an extension to the basic wireless LAN security framework. A suit of protocol proposals has been proposed to provide security in user and session (IP) mobility, secure Virtual Private Network functionalities, and secure assignment of IP addresses. It operates at the IP layer and is built assuming the security means described in Part 2 for over-the-air wireless security. The latter assumption means that none of the services hereby presented is required for having wireless-security and access to network resources (both within corporate networks and public hot spots).

IP Mobility was defined as the ability to move among different networks without loosing network connectivity. Data confidentiality and inter-domain mobility protocols (e.g. Mobile IP) should be integrated to provide data confidentiality and integrity over public networks, i.d. the Internet. Several threats have been found in most of the current macro-mobility solutions and several proposals were presented to address these threats (section 11). Security mechanisms adopted by the IETF Mobile IP protocol [89] were first introduced (section 11.1) together with related security flaws. Several authentication extensions and data confidentiality mechanisms have then been valued to address these security flaws (sections 11.2 and 11.3). A new ISAKMP Payload was defined (named IP Address Update Payload) to optimize usage of IPSec Security Associations with mobile terminals. Interaction between Mobile IP and (AAA) accounting protocols (as RADIUS and DIAMETER) was finally investigated (section 11.5).

Section 12 presented possible threats in dynamic assignment of IP addresses and introduced several proposals to cope with them. Secure dynamic IP addresses assignment (section 12), through usage of Authenticated DHCP messages (section 12.2.1), IPSec protected DHCP (section 12.2.2) or DRCP (section 12.2.3) protocol, has been provided. These security proposals might easily be integrated in the security framework for mobility, when the latter is provided using co-located care-of-addresses. For instance, the ISAKMP “IP Address Update” payload presented in section 11.4 could also be used for the DHCP with IPSec security solution (when the same DHCP server serves several subnetworks).

Secure access to Virtual Private Networks (VPN) is the last argument analysed in the third part of the thesis. It represents a special and particularly important case of security service, whose utility in the case of wireless LAN networks has been introduced in the first part of this thesis. VPNs use a combination of methods for authentication, cryptography, and tunneling, in such a way to create a secure connection between a customer and a business network, or between two different networks. In a typical usage scenario, users associate to the local access network (authenticating themselves), establish an Internet connection, and finally are perform some authentication procedures with the remote system.

Section 13.1 provided a general description of possible solutions for VPN, either based on dedicated lines or on public networks (e.g. Internet). Various techniques based on Point-to-Point (PPP) protocol have been subsequently introduced in sections 13.2.1 (Point-to-Point Tunneling Protocol, PPTP), 13.2.2 (Layer 2 Forwarding, L2F) and 13.2.3 (2 Layer Tunneling Protocol, L2TP). From a comparison between the previous protocols it resulted L2TP to be the best solution when interoperability between different vendors is requested. PPTP resulted instead to be the best choice for clients who do not demand IPSec security or than do not wish to implement any PKI architecture. After having analyzed traditional VPN solutions (based on PPP), a possible solution based exclusively on IPSec and designed for systems that do not need PPP for link-layer connectivity (as it is the case for 802.11 and HiperLAN/2 standards) was introduced. Usage of (AAA) accounting schemes with VPN was finally analyzed in section 13.4.

14.1 Future Work

This thesis works at a rather theoretical level of abstraction. Several other Master of Science degree projects might be done to deepen some of the issues that have been here studied or to implement some of the protocol proposals here produced. Following sections provide some possible thesis proposals, even if many other subjects might be as well derived from this work, especially regarding the mobility issues:

14.1.1 Implementation of (Anonymous) Access with Accounting

Wireless LAN technology offers high-speed, wireless connectivity that enables mobile computing in many different environments. Security is one of the hottest areas within the internet-working world and it is of utmost importance when deploying wireless networks. When wireless networks are desired for usage in public hot spots, such as airports or train stations, connection might be offered either from public telecom operators or from Internet Service Providers (ISPs). Some mean to guarantee access providers for correct charging of the users must be in these cases provided.

Several solutions that combine accounting, authentication (and possibly anonymity) of users have been lately investigated in many Ericsson departments (and in this thesis). Most of them are based on RADIUS protocol for accounting, IPSec protocol for data confidentiality and authentication, and X.509 certificate for key distribution. IP mobility (i.e. the ability of moving between different IP networks without losing network connectivity) is also on the verge of hitting the market on a broad scale and should be integrated in the overall structure.

A Master of Science thesis might consist in developing a prototype having special focus on anonymous access with accounting and based on the theoretical analyses done in this thesis. It might be also investigated some co-operation with the Ericsson NomadicLab department in Finland, which has developed (using IPv6) a complete prototype of the architecture defined in this project.

14.1.2 ISAKMP Extension for IPSec

The IPSec protocol suit has been developed to provide data confidentiality, integrity and authentication between non-mobile hosts, i.e. when IP addresses are statically allocated during a communication session. It is much less suited for roaming nodes that use Mobile IP protocol. If new Security Associations have to be negotiated each time a mobile node gets a new IP address, IPSec might introduce a not acceptable delay in Mobile IP handover.

An extension to ISAKMP has been introduced in this thesis, which is able to optimize handover performances when IP addresses are dynamically assigned on moving between different network domains. This is able to support IP mobility without decreasing network performances. With this new method the mobile node is able to update any ISAKMP SA by sending a single ISAKMP Informational Exchange messages to the correspondent node. Although the message is sent in clear, an ISAKMP Hash payload authenticates message origin and assures data integrity. No computationally expensive operation is required and no sensible data is exposed.

A Master of Science degree project might investigate this new protocol proposal and implement a mobile IP system with IPSec security to value the improvement in performances obtained using this the new ISAKMP exchange. Difference in handover latency between traditional IPSec and IPSec with the new “IP Address Update” payload should be especially measured. A value for the Counter Field in the “IP Address Update” payload (section 11.4.3) might also be proposed.

14.1.3 Security in Ad-hoc Networking

Two network architectures are defined for wireless LAN networks: the Infrastructure Network and the Ad-hoc Network. An Infrastructure Network provides communication between wireless clients through wired network resources. The transition of data from the wireless to the wired medium is via an Access Point (AP). An Ad Hoc network is an architecture that is used to support direct communication among wireless clients. Typically created spontaneously, ad-hoc networks do not support access to wired networks, and do not need an AP to be part of the network. In the ad-hoc network, computers are in this way brought together to form a network “on the fly”. There is no structure into the network, no fixed points, and usually every node is able to communicate with every other node.

Security is one of the hottest areas within the internet-working world and it is of utmost importance when deploying wireless networks. This thesis has only been focused on the Infrastructure Network architecture. However, security in Ad-hoc networks is planned to have an important role in future wireless technology and much work has still to be done. Military tactical operations are still the main application of ad-hoc networks today but, since an ad-hoc network can be deployed rapidly with relatively low cost, it will soon become an attractive option for commercial uses such as sensor networks or virtual classrooms.

In many of the Ad-hoc application scenarios (e.g. the battlefield communications architecture) the environment might be hostile and there could be relatively poor physical protection for mobile devices. This creates the need to protect the wireless network not only from external attacks but also from compromised or damaged nodes (still authenticated but no longer trustable). A degree project might consist in an investigation of business opportunities for ad-hoc networking in hostile environments. A technical proposal for Ad-hoc security should also be given, at least in a theoretical form. A prototype might be developed, if the time will allow this.

14.1.4 Secure schemes for dynamic assignment of IP addresses

When Dynamic Host Configuration Protocol (DHCP) is used to assign temporary IP addresses, several security problems have to be addressed, especially for the case of mobile users. Network administrators may wish to allocate IP addresses only to authorised hosts. Additionally, in some network environment, e.g. wireless networks, it would be necessary to provide authentication for each exchanged DHCP message.

Some authentication mechanism should be introduced or added to the DHCP protocol to protect the network from security threats in IP addresses assignment. This authentication scheme should also provide replay protection, through the use of either nonces or timestamps. Three possible schemes have been described in this thesis:

- Authenticated DHCP messages
- DHCP with IPSec
- Dynamic Registration and Configuration Protocol (DRCP)

A Master of Science project might investigate these three alternative schemes and defines in which network scenarios each of these methods should be preferred. Some implementation is likely to be required to provide concrete results, even if theoretical analyses will have as well a major importance in the overall study.

- APPENDIXES -

Security Framework
Data Confidentiality
User Authentication
IP Security Protocol (IPSec)
Key Management Protocols
IP Mobility

15 Appendix A – Security Framework

Any link between computers may potentially be insecure, as it can be any of the computers through which data flows. Therefore, the need of secure networks makes it necessary to protect both the channels where data are transmitted and the resources where information are stored or elaborated. This thesis has mainly been focused on network and communication security. However, it must always be kept in mind that a network can be effectively secured only when hardware and operative systems are secured as well. The latter security issues are usually referred as Host Security mechanisms.

This thesis project has also been characterised by a technical approach to the security themes. This is only a partial analysis of a complete secure system. Important decisions have to be taken even before technical solutions are considered. It is often not easy to specify what it is going to be protected. Identifying the threats, and determine how likely they are, is another subtle problem. Any implementation choice must then be taken in a cost-effective manner. The cost of protecting a network against a threat should be less than the cost of recovery if the threat should happen. Moreover, it must always be considered that the security of a complete network is equal to the security of its weakest part.

15.1 Dependability

Dependability is the trustworthiness of a system and can be seen as the Quality of Service (QoS) that a system offers. Security and dependability are two closely connected areas [66]. Security can indeed be seen as a specific theme of the more general topic of how to obtain a dependable computing system. In this thesis, security is treated as one characteristic of dependability, on the same level as availability, reliability and safety. This approach includes into the security scope protection against both information threats and illegal use of resources.

By far and large security can be defined as *the probability for a system to protect objects with respect to confidentiality and integrity during a specified time period* [80]. This allows security to conform to the classical taxonomy used in the dependability area. Moreover, this definition of security points out that it is possible to deploy only relative “secure system”. Secure is simply a system where an intruder has to spend an unacceptable amount of time or money in order to make an intrusion. Moreover the risk an intruder has to take may be considered to be too high.

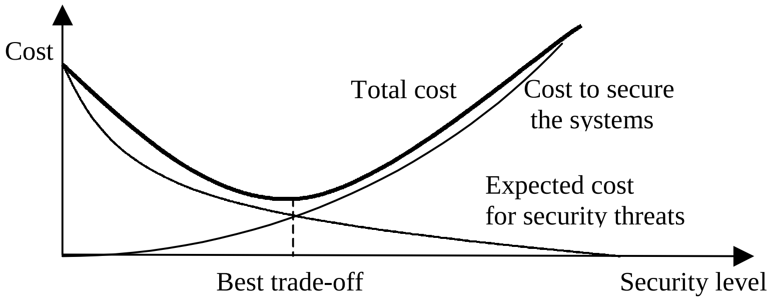
15.1.1 Policy Tradeoffs

As mentioned above, a corporation should not spend more than it is actually worth to protect a network. This creates the need to take cost-effective decisions on what and how it deserves to be protected. There are several tradeoffs between the cost for increasing system security and the potential cost incurred as a result of successful security violations. Any security decision will be largely determined by the following tradeoffs:

- Services offered versus security: each service carries its own security risks. For some services the risk is bigger than possible benefits. Therefore, it could be worth to eliminate the service rather than try to secure it.

- Ease of use versus security: to secure a system it is often required some co-operation by its users. Requiring passwords, for instance, makes the system less convenient for the users. Using one-time passwords makes the system even more difficult to use, but at the same time more secure. Tradeoffs in this area are a very important policy decisions and should be taken consulting both security experts and end users.
- Cost versus security: to build a secure system, some cost must be undertaken. There are both technical costs (e.g. security hardware, software, firewalls and one-time password generators) and administrative cost (e.g. the cost of train the users or change some operative procedure).
- Performance versus security: Security procedures, as encryption and decryption algorithms, take time. This can sensibly decrease the network throughput and even make some procedures no longer utilisable.

The total cost to secure the system can be defined as the combination of all the previous factors, i.e. cost for decreased system performance, cost for increased system complexity, cost for decreased usability of the system and increased operation and maintenance costs, etc.



The expected cost for security threats can be valued as the cost for a single security violation times the frequency of violations. It is very hard to calculate and it is related to what a company considers to be more important for its goals. The total cost of the system is given by both the total costs to secure the system and expected cost for actual security threats.

Any decision must also be taken according to the levels of risk connected to a certain service or data. Each type of cost must be weighed against each type of attack. Information of various sensitivity levels may be carried over a single network and they could require different security services. Information labels [121] are often employed to distinguish such information. Systems that support different kind of security levels are often referred to as multilevel security systems (MLS).

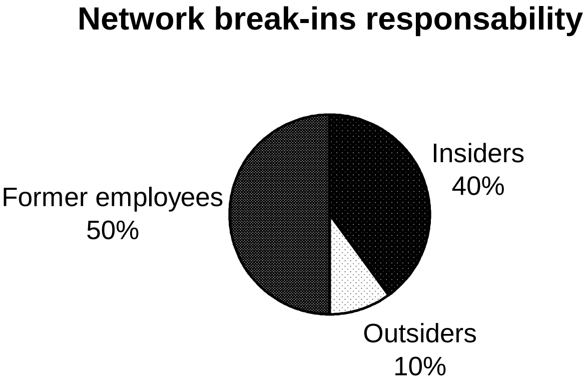
Another consideration is the usually called "Garbage Truck Syndrome." This refers to what would happen to a corporation if a person should be unavailable (e.g. suddenly ill or left the company unexpectedly). While the greatest security resides in the minimum dissemination of information, the risk of losing critical information increases when that information is not shared. It is important to determine what is the proper balance for a company.

15.1.2 Social Engineering

The general impression that people have of security problems is that they are the result of technical flaws that an intruder has exploited. It is, in the same way, quite common to think that network security is mainly a technical problem. The truth is instead that the human factor is of utmost importance for a secure system [27].

Social engineering is an approach that many intruders use to gain accesses to computer systems or at least help them to bypass security barriers. Social engineering may be defined as the act of gaining the trust of legitimate computer users to get confidential information. Users could for instance help someone, unintentionally, to gain unauthorized access to the network. When building a secure environment, users should be trained to behave in a secure way, in accordance with the corporate security policy. The users' Security Handbook [44] can be used, for instance, to define the rules that all users should conform to.

It is worth to remember that most of the security problems for a company came from insider legitimate users and not from technical flaws. Statistics say that only about 10% of all computer crimes are performed as outside break-ins, 40% are committed by insiders and about 50% by former employees as an act of revenge [51].



IP tunnelling within a Firewall is an example of a service that, even if theoretically innocuous, can be actually dangerous because of malicious or careless legitimate users. Tunnelling refers to the technique of encapsulating a data unit from a protocol in another and using the facilities of the second protocol to traverse part of the network. At the destination point, the encapsulation is simply removed and the original data retrieved. An insider and an outsider who dislike a Firewall could build a tunnel between an inside host and an outside host to bypass it. This is typically an insider problem, where computer techniques can be weakly effective and other security policies have to be considered.

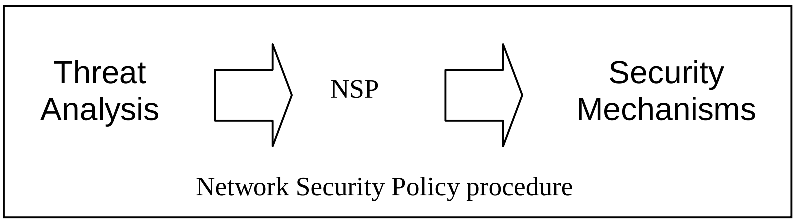
15.2 Network Security Policy

A Network Security Policy (NSP) is a formal statement that describes an organisation's security concerns and specifies the way network security should be achieved in that organisation’s environment. It exists to protect a site's hardware, software and data. The task to define a proper security policy is often a political decision to be taken by corporate management, but it usually affects the work of any users. Therefore, in order for a security policy to be appropriate and effective, it needs to have the acceptance and support of all levels of employees within the organisation.

Parts of the NSP must define the rules by which the users who are given access to a site's technology and information assets must conform. This is usually defined as Service Access Policy (SAP) and establishes the services and the protocols that should be accessible for internal and external use. The main purpose of a SAP is to inform users, staff and managers of their obligatory requirements for protecting technology and sensible information. The policy should specify the mechanisms through which these requirements can be met.

It is important for a SAP to be as explicit as possible to avoid ambiguity or misunderstanding. Users should always conform to the corporate SAP. It could even be worth to use sanctions for individuals who fail to comply with the organization’s computer security policies, when technical prevention is not feasible. When establishing compliance structures, it must also be considered that violations of policy could be unintentional on the part of employees. For example, non-conformance can be due to a lack of knowledge or training.

Another purpose for a NSP is to provide a baseline from which to acquire, configure and audit computer systems and networks for compliance with the policy. An attempt to use a set of security tools in the absence of an implied security policy is meaningless.



Network Security Policy procedure

Firewalls are typically based on network security policies and often used for auditing and logging of ongoing activities. They are one of the main instruments to impose a NSP both to external users but also to internal ones. IPSec [113] has also been developed on the idea of being a flexible framework to implement different Network Security Policies. It is designed to provide interoperable, high quality, cryptographic-based security services and it can be used in many different contexts. The ways in which it is implemented is determined by the specific NSP. This has been possible because of IPSec is algorithm-independent and it is structured in almost independent modules. This modularity permits the choice of different sets of algorithms without affecting any other implemented part. It represents a very good example of NSP-based architecture.

It is important to define a security policy that enables the design of both a correct but also a practical, usable system. A NSP is usually structured in several levels, from the general corporate security statements to specific users’ security rules or application policies. IPSec can be used also in this case as a Best Current Practice reference. The services it offers are based on requirements defined by the Security Policy Database (SPD). This can be considered as a formal codification of low-level security policies. Another database, called Security Association Database (SAD) provides a matching between the SPD and Security Associations (SA) that define the type of security services offered to any specific traffic flow [60].

A NSP should also define the actions to be taken when an incident possibly occurs. Any company should have special plans for incident handling. These planes should address how various incidents will be handled. Certain steps are necessary to be taken during the handling of an incident. Without clearly defined policies, any undertaken activity will remain without focus.

Management and technical administrators should define the goals of any incident-handling plane in advance. First of all, any notification to either local or off-site personnel must be explicit. This requires that any information about the incident should be clear, concise, and fully qualified. Attempting to hide aspects of the incident by providing false or incomplete information may worse the situation. Notably, only half of all organizations carry out formal reporting of security incidents and only half of these take action against the offender [62].

Moreover, 38% of the incident recovery planes have never been tested. A plan that has not been tested is unpredictable and can not be relied upon when needed. It could also be useful to have some co-ordination among different trusted networks to share knowledge and experience about incident handling. Computer Security Incident Response Teams (CSIRTs) are an example of this co-operative structure. RFC 2350, Expectations for Computer Security Incident Response [13], defines the general requirements expected by any CSIRT.

Finally, In order for a security policy to be viable for the long term, it is required to be flexible. A security policy should be as independent as possible from specific hardware or software. It should clearly distinguish between policy and its implementation. It could also be worth to have a distributed Network Security Policy system, when several local sub-networks constitute a corporate network. These have specific local policies but must conform to a high-level corporate security policy. Several papers present distributed security policy information systems, called Security Policy System (SPS). These provide the mechanisms needed for discovering, accessing and processing security policy information of hosts and sub-networks in a common security domain. They are usually realized through the use of specific protocols, e.g. the Security Policy Protocol [104].

For a formal and complete definition of what must be included in a NSP and how this should be detailed and implemented refers to [39],[26].

15.2.1 Well-known Attacks

Security Threats are often related with well-known attack techniques. Each of these techniques exploits different system flaws or bugs and has its own specific goals. In this section it is provided a brief overview of the most common attacks a system can be subjected to:

- **Address spoofing attack**: a type of attack in which the attacker steals a legitimate network address (e.g. IP address) and uses it to impersonate the users that owns the address.

- **Birthday attack**: a form of attack in which it is necessary to obtain two identical values from a large population. The "birthday" part refers to the fact that it is far easier to find an arbitrary matching pair than to match any particular value.
- **Brute-force attack**: an attack where every possible key is tried until the recovered plaintext is meaningful. The complexity of this kind of attack is commonly very high, especially for well-known algorithms.
- **Chosen-plaintext attack**: an attack in which the enemy has accesses to ciphertext and associated plaintext for several messages, and he can use both the information to find out the keys.
- **Verifiable password attack**: an attack where the enemy tries to guess the user password, often using a dictionary attack on several eavesdropped messages.
- **Denial of service/Resource-clogging attack**: an attack where an attacker floods the server with bogus requests, or tampers with legitimate requests. Though the attacker does not usually benefit from it, service is denied to legitimate users.
- **Redirection attack**: the attacker redirects the traffic from a legitimate user to its own address.
- **Man-in-the-middle attack**: an attack in which an attacker interposes itself between two parties and pretends to be the other party for both the legitimate entities
- **Trojan horses/Malicious software attack**: an attack where software legally installed on a user system performs indeed some illicit actions to facilitate an intruder or damage a system.
- **Replay attack**: an attack in which an attacker captures a message and replays it at a later time.
- **Bidding-down attacks**: an attack where an attacker forces the negotiating parties to choose the weakest available option.

15.2.2 Security threats

The basic goals of security can be identified as availability, confidentiality, and integrity. In order to control the risks of operating an information system, managers and users must know the vulnerabilities of the system and the threats that may exploit them. Threats can be seen as potential violations of security and exist because of vulnerabilities of the system, i.e. its weaknesses. There are two basic types of threats, accidental and intentional threats. Accidental threats either result in an exposure of confidential information or cause an illegal system state to occur. Intentional threats are external or internal attacks to the network.

Knowledge of the threat environment allows the system manager to implement the most cost-effective security measures. In some cases, managers may find it most cost-effective to simply tolerate the expected losses. Each threat should be examined with special attention to how the threat could affect the users. Depending on the specific network, it will be possible to define more specific threats that have to be addressed. The following are classic threats that should be considered for any possible environment:

- Unauthorized access to resources and/or information (hackers, industrial espionage, etc.)
- Unintended and/or unauthorized disclosure of information (insiders crimes, users errors, etc.)
- Denial of service (sabotage, hackers, etc.)
- Traffic flow analysis (threats to personal privacy, industrial espionage)

Error and Omissions could seem not to be security threats. Nevertheless, unintentional errors often contribute to security problems, directly and indirectly. Sometimes the error is the threat, such as a data entry error or a programming error that crashes a system. In other cases, errors create vulnerabilities that an intruder can exploit.

Passwords written down on some physical support and then used by unauthorized users are a typical case of security flaw due to the end users. Normal users have often to remember an average of five passwords [63]. This eases security violations from a user. Key recovery and backup systems should always be provided to help end-users that have forgotten their passwords or destroyed the file or smart card that contained the secret key [124].

Insiders, as it has been pointed out when Social Engineering was addressed, are typically the main security threat. Authorized users perpetrate the most of the frauds on computer systems. Insiders have both access to and familiarity with the victim computer system, including what resources it controls and where the flaws are. Therefore, they are in a better position to commit crimes. *Sabotage* is another possible insider threat. Disgruntled employees cause more sabotage damages than former employees do. Covert channels are another possible threat with insider.

Covert channels are unprotected channels that can be used by an insider to send confidential information to unauthorized entities. It is in general very hard to identify covert channels in a system since they can be of many different types, e.g. message length variations during transmissions, time and length of transmissions, presence and size of files, creation time for objects, modulation of disk usage, CPU time usage, etc.

Hackers, sometimes called crackers, are a real and present danger to most organizational computer systems linked by networks. Although insiders cause more damage than hackers do, the hacker problem remains serious and widespread. The hacker threat often receives more attention than more common and dangerous threats. This thesis will try to give a complete overview of the security framework, although being detailed and complete on the hacker problem.

Industrial espionage is a special case of the hacker threat. It involves collecting proprietary data from private corporations or government agencies for the benefit of another company or organization. Industrial espionage can be perpetrated either by companies seeking to improve their competitive advantage or by governments seeking to aid their domestic industries. Foreign industrial espionage carried out by a government is known as economic espionage. The main objective of industrial espionage is often to obtain information related to technology. Industrial espionage is often able to perform both direct and indirect attacks to data confidentiality. A *direct attack* aims directly the data. In an *indirect attack*, information is received from or about an object without attacking the object itself. For example, it may be possible to derive confidential information without accessing an object at all, by gathering statistics on it and from them derive the desired information.

Threats to Personal Privacy must also be considered, when a public network is analyzed. The accumulation of vast amounts of electronic information about individuals by the government, credit bureau, and private companies combined with the ability of computers to monitor, process, aggregate, and record information about individuals have created a very real threat to individual privacy. As more of these cases come to light, many individuals express increased concern about threats to their personal privacy. A lack of attention in this area could mine the successful introduction of any new public access networks.

Finally, many threats are impossible to be technically analyzed and prevented. *Data-driven attacks*, for example, are difficult to prevent in a structured way. The only solution is often use dedicated software for antiviral control, when it is referring to a virus problem, or appropriated policies, when it is considered the general case of Trojan Horses or Worms, e.g. based on Java applets or ActiveX controls. Malicious code can often compromise the security of a complete system.

Passive attacks are also very difficult to avoid. A passive attack is done by monitoring a system performing its tasks and collecting any kind of information on it. Passive attacks do not interact or disturb normal system functions. Examples of passive attacks are monitoring network traffic, CPU and disk usage. Encryption of network traffic can only partly solve the problem since even the presence of traffic on a network may reveal some information. This kind of attack is normally named *traffic flow analysis*. It can be especially valuable to detect unusual activities (Rumours say that prior to the US Panama invasion, Domino's pizza deliveries to the Pentagon jumped 25%, a situation in which an external observer could detect that some-thing unusual was going on).

To analyze and value the threats in the right way, it is always suggested to assume that enemies have complete access to the communication between all the entities. The book "Internet and Intranet Security", by R. Opplinger [82] provides a good overview of how to cope with the various possible security threats under the above statement.

15.3 Security Mechanisms

Many techniques have been developed to provide the different services necessary to operate the network in an efficient and secure manner. In the following of Part 2 we are going to provide a general overview of the most important of these security mechanisms, regarding the networking case and with special focus on data confidentiality and integrity.

The quality of the security provided by these mechanisms depends completely on the strength of the mechanism itself, the correctness of that mechanism's implementation, the security of the key management mechanism and implementation, and upon the correctness of the implementations in all of the participating entities.

Cryptography is one of the main instruments used to send information between users in a way that prevents others from reading it, i.e. confidentially. Authentication is the process of reliably verifying the identity of someone (or something). Some key distribution infrastructure is needed to support both of these. Authentication, cryptography and key distribution will be further detailed in following appendixes.

Firewalls are used to filter traffic from and to high-security systems, e.g. corporate or military networks. These are also able to provide some protection against malicious code, e.g. Trojan horses, virus and worms. Other mechanisms may be used to specifically protect the network from data attacks, e.g. anti-virus software.

15.3.1 Firewalls

A firewall is a system or group of systems that enforces an access control policy between two networks. Firewalls are used to filter traffic from and to high-security systems, e.g. corporate or military networks. These are also able to provide some protection against malicious code, e.g. Trojan horses, virus and worms [82].

A firewall can be thought of as a pair of mechanisms: one, which exists to block traffic, and the other, which exists to permit traffic. Some firewalls place a greater emphasis on blocking traffic, while others emphasize permitting traffic.

Probably the most important thing to point out about firewalls is that they are used to implement access control polices, described in section 15.2. Without a good idea of what kind of access it is desired to permit or deny, or simply permitting someone or some product to configure a firewall based on what others may think, may not be coherent with the security policy needed for an organization.

Some firewalls permit only Email traffic through them, thereby protecting the network against any attacks other than attacks against the Email service. Other firewalls provide less strict protections, and block services that are known to be problems. Generally, firewalls are configured to protect against unauthenticated interactive logins from the "outside" world. This, more than anything, helps prevent vandals from logging into machines on your network. More elaborate firewalls block traffic from the outside to the inside, but permit users on the inside to communicate freely with the outside. The firewall can protect you against any type of network-borne attack if you unplug it.

Firewalls are also important since they can provide a single "cheek point" where security and audit can be imposed. Unlike in a situation where a computer system is being attacked by someone dialing in with a modem, the firewall can act as an effective "phone tap" and tracing tool. Firewalls also provide important logging and auditing functions (section 15.3.4). They often provide summaries to the administrator about what kinds and amount of traffic passed through it, how many attempts there were to break into it, etc.

Firewalls can't protect against attacks that don't go through the firewall. Many corporations that connect to the Internet are very concerned about proprietary data leaking out of the company through that route. Unfortunately for those concerned, a magnetic tape can just as effectively be used to export data. There are incredibly many corporations buying expensive firewalls and neglecting the numerous other backdoors into their network.

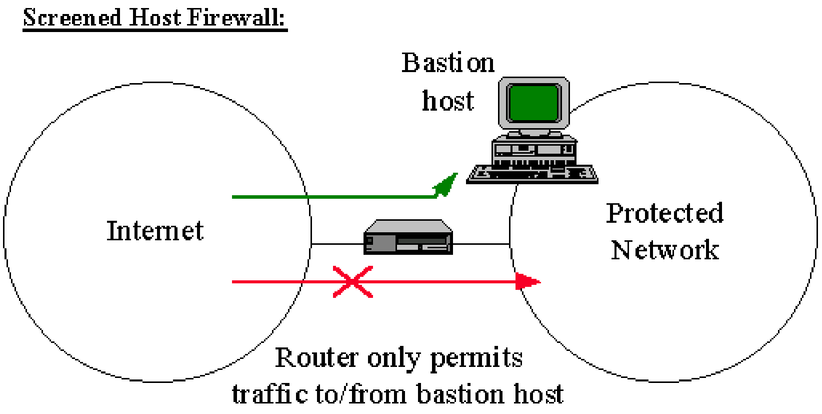
Another thing a firewall can't really protect from is sabotage or idiots from inside your network (section 15.2.2). While an industrial spy might export information through your firewall, he's just as likely to export it through a telephone, FAX machine, or floppy disk. Firewalls also cannot protect a network against insiders’ stupidity. Users who reveal sensitive information over the telephone are good targets for social engineering; an attack already described in section 15.1.2.

Firewall policies must be realistic, and reflect the level of security in the entire network. For example, a site with top secret or classified data doesn't need a firewall at all: they shouldn't be hooking up to the Internet in the first place, or the systems with the really secret data should be isolated from the rest of the corporate network.

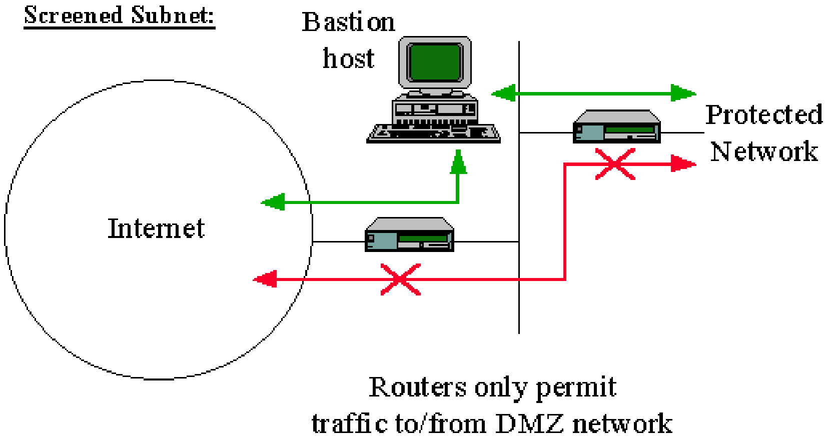
There are two different types of firewalls, i.e. Packet filters, which work at ISO network level (e.g. IP layer) and application gateways, which work either at transport or application ISO levels.

Network level firewalls generally make their decisions based on the source, destination addresses and ports in individual IP packets. A simple router is the "traditional" network level firewall, since it is not able to make particularly sophisticated decisions about what a packet is actually talking to or where it actually came from.

A special example of network level firewall is the "screened host firewall". In a screened host firewall, access to and from a single host is controlled by means of a router operating at a network level. The single host is usually a bastion host, which is a host provided with specially secured operating system and able to resist most attacks.



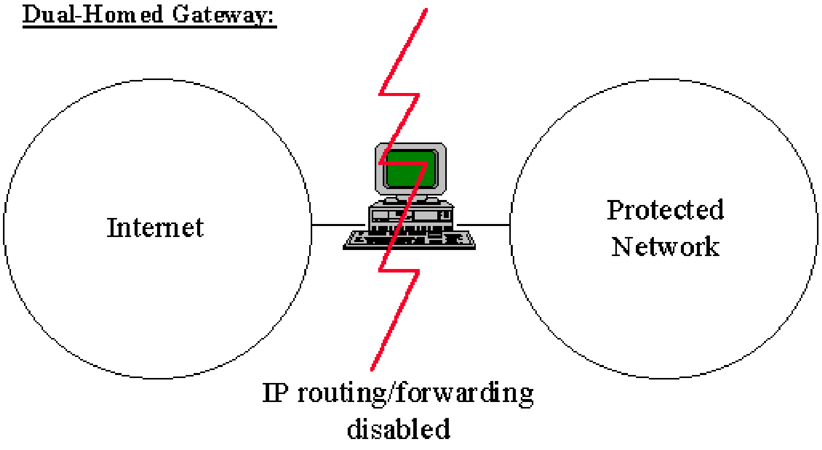
Another special case of packet filters (network layer firewalls) is the “screened subnet firewall”. In a screened subnet firewall, access to and from a whole network is controlled by means of a router operating at a network level and filtering the packets. This is similar to a screened host, except that it is effectively a network of screened hosts.



Modern network level firewalls have become increasingly sophisticated, and now maintain internal information about the state of connections passing through them, the contents of some of the data streams, and so on. One thing that's an important distinction about many network level firewalls is that they route traffic directly though them, so to use one you usually need to have a validly assigned IP address block. Network level firewalls tend to be very fast and tend to be very transparent to users.

Application gateways (transport and application level firewalls) are generally hosts running proxy servers, which permit no traffic directly between networks, and which perform elaborate logging and auditing of traffic passing through them. Since the proxy applications are software components running on the firewall, it is a good place to do lots of logging and access control.

Application level firewalls can also be used as network address translators (NAT), since traffic goes in one "side" and out the other, after having passed through an application that effectively masks the origin of the initiating connection. Having an application in the way in some cases may impact performance and may make the firewall less transparent.



In previous picture it is showed a special application of application level firewalls, called "dual homed gateway". A dual homed gateway is a highly secured host that runs proxy software. It has two network interfaces, one on each network, and blocks all traffic passing through it.

15.3.2 Protection from Trojan horses

A Trojan Horse is a program performing a legal action but at the same time secretly performing also an illegal one. Viruses and worms are special cases of Trojan horses, as well as logic bomb and back door procedures.

An example of a Trojan horse is a text editor searching documents for special keywords and, if a keyword is found, making a copy of the document available to someone else. A Trojan horse may also create new user accounts, modify the system into accepting users secretly or to modify encryption algorithms.

The protection mechanisms in most systems have problems protecting information against such an attack. A document may be protected using dedicated software, e.g. anti-virus software, but when entities have the possibility to select protection of objects at their own will, it is very hard for a system to stop a Trojan horse from requesting a change of protection for an object. In fact, most actions an entity may perform can be performed secretly by a Trojan horse, since the Trojan horse normally executes with the same privileges as the entity using it.

A special type of Trojan horse is a logic bomb. It is a program with a “feature” incorporated into it and this feature often consists of the destruction of objects. The bomb is programmed to go off at a specific time or when a specific event occurs. The idea behind a logic bomb is often to cause as much damage to a system as possible.

A Trojan horse can enter a system in many ways, even if the most typical threat is to enter a system from the network (like viruses and worms). It may anyway also be installed with any piece of new software. The best way to prevent Trojan horse attacks is to check incoming data through the network and new installed software, possibly using up-to-date dedicated programs.

Since the most vulnerable target for a Trojan horse is an entity with high privileges and the Trojan horse can do whatever action allowed to that user, entities should also be given the least possible amount of privileges, as long as they can fulfill their working tasks.

15.3.3 Routing Control and Traffic Padding

Introducing mechanisms that allow entities to influence traffic routing can reduce the problems with information collection (passive attacks) and denial of service. If entities have the possibility to bypass certain hosts or parts of a network, an entity refusing to forward messages or an unreliable part of a network can be bypassed.

Routing control can be used to ensure that sensitive information is not sent through unreliable hosts or networks. It is also possible to use routing control in combination with fragmentation of messages, where fragments are transmitted through different networks and possibly in random order.

To make it harder to collect information from network traffic, traffic padding can be added to the system. Traffic padding is a mechanism that adds dummy messages that are transmitted between randomly selected hosts at random intervals. When messages need to be sent, these dummy messages are substituted with real messages, thus an attacker can not differentiate real traffic from dummy messages on the network. Also, real messages can be sent indirectly to their final destination (i.e. through intermediate hosts) to make use of dummy messages sent between other hosts. This mechanism is often combined with padding messages to a fixed message length to give even less information to an observer.

The disadvantage of the mechanisms here presented is the possible loss of network bandwidth due to the transmission of dummy messages and increased message lengths.

15.3.4 Violation detection and Recovery mechanisms

Not all mechanisms are used to prevent security violations. An important group of security mechanisms are those used after a security violation has taken place. Detection mechanisms detect a security violation, and recovery mechanisms restore the system to its state prior to the violation.

Ideally, detection mechanisms should detect a security violation immediately and it should give enough information to enable the tracery of the violation a specific entity or user. In many cases, the knowledge of the existence of such mechanisms can be enough to prevent attacks. If an attacker, for example, knows that the attack will be detected (even if it is detected at a later date) and traced back to him/her, this may be reason enough for not trying that specific attack.

An audit trail is a log containing security-related events and transactions. It contains information about when, how and by whom a transaction was ordered, thus it is a valuable tool for protecting both objects and the integrity of entities. Audit trails should be used to monitor all sensitive actions, especially those actions that affect the security in the system.

The audit trail should be detailed enough to make it possible to trace security violations back to individual users. It can, for example, contain digital signatures from entities ordering transactions, which can be used at a later time to verify (prove) that an entity actually did order a transaction.

16 Appendix B – Data Confidentiality

Communication security deals with protection of information during transportation. When objects are transported, either between computers or locally within a computer, an active attack may be undertaken in order to interact with the communication process, for example to modify, retransmit, reorder or destroy information. Also, objects need to be protected against passive attacks and exposures during transmission.

16.1 Secure Protocols

A secure protocol is a protocol that is especially designed to protect the integrity and/or confidentiality of all objects that are transported. In general, it is not necessary to preserve each individual object, but merely to preserve integrity and/or confidentiality of the whole communication process. Also, provisions may be taken to limit the possibilities of passive attacks.

Authentication mechanisms must be provided to ensure that communication at all times is performed between the correct entities, and this authentication mechanisms need to be protected against replays from the network. Most likely, a method for distributing encryption keys to entities opening new connections is needed.

A Secure protocol also needs to be protected against modification of packets, replays of old packets and against lost packets. Replays can be especially cumbersome, since a replay of an old message is performed with a completely valid packet (i.e. a valid checksum and correct encryption). Therefore, some kind of time stamp or sequence number that cannot be forged must be contained in all messages.

The encryption algorithm must also be able to deal with retransmission of old packets and to deal with duplicates, i.e. it must have a mechanism to handle resynchronization. Finally, the protocol needs to be protected against denial of service attacks and loss of packets. It is necessary for the communicating parties to exchange messages at regular intervals to make sure that all messages are received.

The mathematical functions used to encrypt the message may be either based on the secret of the algorithm or on some shared secret key. When security relies on the fact that the algorithm is secret, the scheme is named restricted cipher. Restricted ciphers have historical interest but are nowadays used only for very low security applications.

Present encryption solutions are almost entirely based on the use of secret keys. This means that the algorithm is given as public and the security relies only on this secret key. The key should be a large number and may be in various ways involved to produce the ciphertext.

An important choice in securing a communication channel is to determine the most suitable layer for encryption within the ISO/OSI model. This must be done in order to proceed as efficiently and securely as possible. Each layer gives some advantages but also has some disadvantages.

Encryption at the physical layer is used when the physical medium is particularly not trustable, e.g. in wireless communication. Hardware encryption devices are often utilized, as low computational time at this layer is generally required. Traffic-flow security may also be provided at this ISO layer. A different encryption may be used between each link pair and no entity is in this way able to analyze the complete routing path, i.e. it is not feasible to know where and which the communicating entities are.

If data are end-to-end encrypted, i.e. from the source to the final destination, security can be provided even if intermediate router are not trusted. The secrecy level provided may be higher than for link-layer encryption and completely resides only on the two peer entities. A problem with end-to-end encryption is that it allows traffic analysis. An enemy may get useful information simply from which parties are communicating, length and rate of messages, and so on (passive attacks).

If encryption takes place at high ISO layers, i.e. presentation and application layers, it is possible to provide different security levels for different sensibility of data. Security can in this case be completely independent of the kind of network used and interact with the user’s software. Performances are anyway much decreased, having each security hop-to-hop association be computed at a high layer.

It also must be noticed that it is often possible for an attacker to guess some parts of a message. This is important to know, since it eliminates several encryption algorithms because messages often include predictable information such as frame headers, sequence numbers, etc.

16.2 Public-key and symmetric cryptosystems

Two different kinds of key-based algorithms are available, i.e. symmetric and public-key cryptosystems. Their structure and properties are very different but none of them can be considered better than the other one, if not for specific applications. A common solution is often to use a hybrid cryptosystem. A public-key protocol is first used to secure and distribute session keys. Then, these session keys are used to protect data through some symmetric algorithm.

16.2.1 Symmetric cryptosystems

Both confidentiality and integrity of objects can be preserved by encryption. If a symmetric cryptosystem such as DES or Blowfish (section 16.4) is used, the same key is used for both encryption and decryption.

Symmetric cryptosystem are usually quite fast and may be very effective as computational cost. Distribution of keys is the hard point of symmetric cryptography. With these encryption algorithms, it is a major problem to find a way to distribute keys to entities, since all entities need one unique encryption key for each entity it needs to communicate with.

When symmetric cryptosystems are used, the number of keys inside a community increases quadratically with the number of users. Assuming one secret key per users pair, a network of n users needs n (n-1)/2 keys.

16.2.2 Asymmetric Cryptosystems

In an asymmetric cryptosystems, usually named public-key system, each entity has two keys: a secret one and a public one. The secret one is never transmitted over the network and it is known only to a specified entity. The pubic one is not kept secret and can be stored in a public database. It is also required to be computationally not feasible to deduce the private key from the public one.

If we only care for integrity or confidentiality but not both of them, it is possible to use an asymmetric cryptosystem. By keeping the encryption key secret and the decryption key public, it is possible to preserve object integrity. If, on the other hand, the decryption key is secret, it is possible to enforce object confidentiality.

Public-key algorithms are usually based either on the difficulty of computing discrete logarithms or on the factorizing problem. Schemes based on elliptic curves have been also studied.

The great advantage of public-key cryptography is that it solves the key management problem of symmetric cryptosystems, since there is no need of prior agreement on the secret key. Public-key algorithms are very effective for end-to-end communication between a large number of hosts. With an asymmetric cryptosystem it is enough to give only one private key to each entity, and still allow all entities to communicate. If both integrity and confidentiality are desired, it may also be possible to do two asymmetric encryption phases, using both the sender’s private and the receiver’s public keys.

Public-key cryptosystems are mainly used to encrypt session keys or cipher “compressed” with a previous secret-key algorithm, not to encrypt plaintext data. Public-key cryptosystems are not a substitute for symmetric algorithms. They are definitely slower than symmetric ones and are vulnerable to chosen-plaintext attacks (section 15.2.1).

Any key-management scheme that need to scale to the number of nodes possible in the Internet must anyway be based on an underlying authenticated public-key infrastructure. Public keys can be authenticated using a variety of mechanisms, such as public key certificates, a secure directory server, and so on.

For an asymmetric public-key cryptosystem it is also necessary to grant authenticity and integrity for public keys (see section 16.4.5). ITU standard X.509 (section 19.6) provides an example of authenticating public keys using certificates. Secure Domain Name Security (DNS) provides an example of authenticating public keys (and other resources) using a secure directory service. Still another example of authenticating public keys is the web-of-trust "introducer" model, best exemplified in the Pretty Good Privacy (PGP) secure e-mail software package.

16.3 Encryption Modes

It is possible to apply symmetric ciphers in two different ways, named block and stream cipher techniques. Block ciphers operate on blocks of plaintext, while stream ciphers operate on one by one bit or byte. An encryption “mode” is defined as a combination of theses two basic cipher techniques, some feedback procedure and simple operations.

A stream cipher combines each bit of plaintext with one bit of a key. A key-stream generator is used to produce a stream of bits, which is xored with the plaintext, bit to bit. To recover the message, the ciphertext bits are xored with the same key-stream. If key-streams generated were completely random, the resulting security would be perfect. It is however the case that actual key-stream generator produces some deterministic stream.

On the other side, a block cipher operates on data blocks of a given size and encrypts, proceeding on chunk packets. DES and Blowfish are often used with this algorithmic procedure (sections 16.4.1 and 16.4.4). Stream ciphers are usually more suitable for hardware implementation, whereas block ciphers are usually good for software implementations since they operate on data in computer-sized blocks avoiding time-consuming bit manipulation.

The Electronic Codebook (ECB) mode is the simplest block cipher mode. It simply encrypts a block of plaintext in a block of ciphertext. ECB mode can be implemented in parallel (each plaintext block is independent), but it is weak. Two identical blocks encrypted with the same key produce the same ciphertext. The beginning and the end of messages are often stereotyped, so a cryptoanalytic attack may get very useful information.

The Cipher Block Chaining (CBC) mode is an enhanced mode of ECB, which chains together blocks of cipher text. The plaintext is xored with the previous ciphertext block and then encrypted. In this way, the encryption of each block depends on the previous blocks. Since two identical messages will still produce the same ciphertext, random data is encrypted as first block. This initial block (Initialization Vector, IV) makes each encrypted message different, even when the plaintext should be the same.

Cipher Feedback (CFB) mode uses previously generated cipher text as input to generate pseudorandom outputs, which are combined with the plaintext to produce cipher. The Output Feedback (OFB) mode is identical to CFB except that the previous cipher output is used as input in OFB while the previous cipher input is used as input in CFB.

16.4 Analysis of cryptographic algorithms

When data confidentiality is desired, cryptography mechanisms may be used to keep messages private and confidential even over an insecure medium. Cryptography consists in encoding a message using some shared secret and in such a way that a not authorized entity is unable to retrieve the original messages from the encrypted one.

16.4.1 The Data Encryption Standard (DES)

The Data Encryption Standard (DES) [116] is a block cipher based on the algorithm known as Lucifer and designed by IBM. DES is a symmetric algorithm (the same key and algorithm are used for encryption and decryption), which uses a 56-bit key, and maps a 64-bit input block into a 64-bit output block. Although the key has a 64-bit length, only 7 bits in each 8 bits are used since 1 bit is used for odd parity on each byte. The issue of key length is very important since advances in chip speeds have made key breaking by a bit of cleverness and exhaustive search easier.

The Data Encryption Standard defines four different Modes of Operation. These four modes are called the Electronic Codebook (ECB) mode, the Cipher Block Chaining (CBC) mode, the Cipher Feedback (CFB) mode, and the Output Feedback (OFB) mode. ECB is a direct application of the DES algorithm to encrypt and decrypt data. CBC is an enhanced mode of ECB, which chains together blocks of cipher text. CFB uses previously generated cipher text as input to the DES to generate pseudorandom outputs, which are combined with the plaintext to produce cipher. OFB is identical to CFB except that the previous output of the DES is used as input in OFB while the previous cipher is used as input in CFB.

The 64-bit input is subjected to an initial permutation to obtain a 64-bit result (which is just the input bits shuffled). The 56-bit key is used to generate sixteen 48-bit per-round keys, by taking a different 48-bit subset of the 56 bits for each of the keys. Each round takes as input the 64-bit output of the previous round, and the 48-bit per-round key, and produces a 64-bit output. After the 16th round, the 64-bit output is subjected to another permutation, which happens to be the inverse of the initial permutation

Decryption works by running DES backwards. To decrypt a block, it is first necessary to run it through the initial permutation to undo the final permutation (the initial and the final permutations are inverses of each other). The same key generation process is done, though keys are used in the opposite order. After 16 rounds of decryption, the output will be subjected to the final permutation (to undo the initial permutation). When implementing DES, it is also important not to select any weak keys, although the odds of picking one of these at random are low.

Security strength in using DES has been for along time studied. Several theoretical studies have estimated costs and times for attacking and breaking sixteen-round DES. In 1993, cryptanalysis studied estimated that a brute-force DES-cracking machine could find a key in an average of 3.5 hours at the cost of only one million dollars. The cost drops by a factor of 5 every 10 years. Recent work in the cryptographic community, particularly the construction of "Deep Crack" by the Electronic Frontier Foundation has made even clearer that a 56-bit cipher is no longer acceptable for strong security. The Deep Crack engine can recover a DES key in as little as 24 hours.

16.4.2 Triple DES

The Triple DES (3DES or TDES) [117] algorithm is a variant of the US Data Encryption Standard (DES) algorithm. 3DES provides greater security than traditional DES at the cost of some more computational cost. Both input and output result in the same number of octets and this facilitates in-place encryption and decryption.

Triple DES (3DES) operates on blocks of eight octets. This may require some padding to the input plaintext. It processes each block of the plaintext three times, using DES for each of these three steps. More in detail, a DES encryption followed by a DES decryption followed by a DES encryption are performed, using three different keys.

The secret key used in 3DES is 192-bits long and consists of three independent 64-bit length keys. As in DES, only 56 bits of each 64-bits quantity are effectively used for keying. One bit on each octet is used as a parity bit. Each of the three DES steps done in 3DES round is done using one of these three keys. The three DES steps implementations may also be pipelined in series to provide parallel computation.

The Triple Data Encryption Algorithm defines seven different Modes of Operation, mostly based on the four modes defined for DES [116]. A popular version of the 3DES algorithm is the Cipher Block Chaining (CBC) mode, which is a variant of DES-CBC. 3DES-CBC [57], as any CBC mode function, requires an Initialization Vector (IV). This has a length of eight octets. Unpredictability and non- repetition of this Initialization Vector are essential to support 3DES-CBC security against replay and cryptanalysis attacks.

It has been showed that DES is not a group. This means that composition of multiple DES rounds is not equivalent to simply using DES with a different key. 3DES is substantially stronger than DES, as it is more reliable against brute-force attacks. It also provides protection against flaws that had been found in 2DES. Finally, it deserves to be noted that if all three the keys used in 3DES are the same, 3DES is equivalent to DES. This property allows 3DES hardware implementations to easily operate in DES mode.

3DES is considered to be very sure and safe to be used. It has anyway one significant problem, being its performance 3 times slower than DES and being DES never been designed to be fast in software.

16.4.3 RC5 cipher

RC5 [6] is named after its author Rivest (Cipher). It is an encryption algorithm that provides fast symmetric block ciphering suitable both for hardware and software implementations. The RC5 algorithm consists of three components: a key expansion algorithm, an encryption algorithm, and a decryption algorithm. It has a variable word size, a variable number of rounds, and a variable- secret key length.

In RC5, the plaintext input is mapped into two w-bit words, which will be here denoted A and B. The algorithm assumes standard little-endian conventions for packing bytes into input/output blocks. The first byte occupies the low-order bit positions of register A, and so on, so that the fourth byte occupies the high-order bit positions in A, the fifth byte occupies the low-order bit positions in B, and the eighth (last) byte occupies the high-order bit positions in B.

RC5 uses an expanded key table, S [0...t-1], consisting of t=2(r+1) w-bit words. The key expansion algorithm uses two "magic constants" (binary constants) and consists of three simple algorithmic parts. The key-expansion function, determining K from S, is the core of the RC5 one-way feature. The three key expansion phases are as follows:

- In the first step of the key expansion algorithmic, the secret key K is copied into an array L. This operation is done in a natural manner, using consecutive key bytes of K to fill up successive words in L, low-order byte to high-order byte. Any unfilled positions of L are zeroed out.

- The second algorithmic step of key expansion is to initialize the array S (the expanded key table) to a particular fixed (key-independent) pseudo-random bit pattern, using an arithmetic progression modulo 2w determined by the two given "magic constants".

- The third step of key expansion is to mix the user's secret key in three passes over the arrays S and L. More precisely, due to the potentially different sizes of S and L, the larger array will be processed three times, and the other may be handled more times.

The final encryption algorithm assumes that the input has already been mapped into the two w-bit registers A and B. It also assumes that the key-expansion has already been performed, so that the array the expanded key table S has been computed. For a detailed description of RC5 encryption procedure refers to [6]. RC5 output is also contained in registers A and B.

16.4.4 Blowfish

The Blowfish cipher [106] is a 16-round, 64-bit block symmetric cipher with a key up to 448 bit, which may be useful in many cryptographic applications. Blowfish uses a variable-length key, from 32 bits to 448 bits, and this makes it ideal both for domestic and exportable use. Another important feature of Blowfish is that it is unpatented and unlicensed.

The cipher consists of a rather complex key initialization phase followed by an encryption phase. The Blowfish cipher is word oriented, operating on a block of 64 bits divided into two 32-bit words, with a key table of 18+1024 words. All data units are big-endian 32-bit words. Although the key initialization is complex and requires 4K for table allocation, the actual encryption of data is very efficient, especially on 32-bit processors.

The purpose of the key-expansion procedure is to expand the user's key K to fill two arrays of random binary key words. Each element of these key tables will be later used to compute the 16 Blowfish algorithm rounds. A table of given “magic number” is also used during the key expansion phase, which is divided in three main steps (array initialization, elaboration based on the secret key for the first array and elaboration based on the secret key for the second array) [106].

The core of Blowfish is the f-function, which is used to encrypt a 32-bit word plaintext data using the keys contained in the key-expanded tables. This function is used to compute a round function that is applied to all the 16 Blowfish encryption rounds [106].

Blowfish has been shown to be resistant to differential cryptanalysis, linear cryptanalysis, and related-key cryptanalysis (both conventional and differential). It has been designed neither for applications where key setup time is critical (for example large multiplexed systems) nor for implementations that lack a large table allocation capacity (e.g. smart cards). The main application area for Blowfish is bulk data encryption on powerful CPUs. In this case Blowfish can be considered as an excellent drop-in replacement for DES or IDEA.

16.4.5 AES proposals (MARS, RC6, Rijndael, Serpent, and Twofish)

The U.S. National Institute of Standards and Technology (NIST) has lately undertaken the Advanced Encryption Standard (AES) Project. This project aims in finding a new cipher as a replacement national (U.S.) standard for the Data Encryption Standard (DES).

The AES project set out several criteria for the new standard. Among these, it is of particular interest that the new AES standard will operate on 128 bit blocks of data instead of the 64 bit blocks of data operated on by the DES and similar ciphers. Other requirement on the future AES cipher are that there should be no requirement for licensing or other intellectual property constraints and the cryptographic processing should be as fast as possible.

3DES (section 16.4.2) has long been considered as one of the main candidates for AES, even if later it has been valued as too slow. It is likely that 3DES will remain for a while as the "safe but slow choice" but a different cipher will primarily be used as AES, especially in areas where performance is an issue [105]. So far five algorithms have been selected, as finalist to became the future AES standard. These are MARS, RC6, Rijndael, Serpent, and Twofish [79].

MARS [71] has been developed from the International Business Machines Corporation. It features a variety of operations, including the technique of rotating digits by a varying number of places that is determined by both the data and the secret key. Consequently, while MARS performs well in general, it performs particularly well on computer platforms that support its rotation and multiplication operations efficiently.

RC6 [98], developed by RSA Laboratories, is an evolution of RC5 (section 16.4.3). It is a very simple algorithm that should be easy and fast to implement both in software and hardware. RC6 does not use substitution tables. The main engine for its security is the technique of rotating digits by a varying number of places that is determined by the data. In general, RC6 is fast and it is particularly fast on platforms that support its rotation and multiplication operations efficiently. Key setup is also very fast in RC6.

Rijndael [100], created by Joan Daemen and Vincent Rijmen, performs excellently across many different platforms. Its key setup is fast and its memory requirements are low, so it also should perform well in hardware and in memory-constrained environments. The straightforward design and the conservative choice of operations should facilitate its further analysis, and the operations should be relatively easy to defend against certain attacks on physical implementations.

Serpent [107] is ultra-conservative in its security margin. Its designers choose to use twice as many iteration as they believed secure against currently known attacks. Consequently, Serpent's performance is relatively slow compared to the other four finalists. In some settings, however, this should be mitigated by the efficiency of optimized implementations using what the authors call the "bitslice" mode, for which the algorithm was specially designed. Serpent should fit well in hardware and in memory-constrained environments.

Twofish [115] exhibits fast and versatile performance across most platforms. It also should perform well both in hardware and in memory-constrained environments. It features variable substitution "tables" that depend on the secret key. Such tables generally offer greater security than tables with fixed values. The possibility of pre-computing these tables to varying degrees helps Twofish offer a wide variety of performance tradeoffs. Depending on the setting, Twofish can be optimized for speed, key setup, memory, code size in software, or space in hardware.

16.4.6 RSA

RSA [56] is named after its inventors, Rivest, Shamir, and Adleman. It is one of the most utilized public key cryptographic algorithms and it is often utilized to encrypt Message Authentication Code computed with other secret-key algorithms. It is also used to protect key exchange, as in ISAKMP or IKE.

One important feature of RSA is that the key length can be variable. Anyone using RSA can choose a long key for enhanced security, or a short key for efficiency. The block size in RSA is also variable. Block size is the size of the chunk of data to be encrypted. The plaintext block must be smaller than the key length. The ciphertext block will be the length of the key.

RSA, as most of public-key algorithms, is much slower to compute than other secret key algorithms, such as DES. As a result, RSA does not tend to get used for encrypting long messages Mostly it is used to encrypt a secret key, which is used to actually encrypt the message.

The basic algorithm is to generate two large primes p and q, multiply them together, and get the result n. The factors p and q will remain secret. A private key and a public key will be generated from p, q, and n.

RSA algorithm’s details are quite simple and may be briefly described as follows:

- Two large prime numbers p and q are chosen
- n = pq is computed
- Choose a random value e such that e and (p-1)(q-1) are relatively prime
- Compute d such that ed=1 mod (p-1)(q-1)
- e and n are public, while d is used as private key and it is kept secret

There are two ways authenticated RSA public keys can be used to provide authenticity and privacy for a datagram protocol, such as IP. The first one is to out-of-band establish an authenticated session key, using RSA in one of the several session key establishment protocols used prior to communication (like IKE (section 19.5) or ISAKMP (section 19.4). An alternative approach may be the use of RSA in a stateless key-management scheme. This might work through in-band signaling of the packet encryption key, where the packet encryption key is encrypted in the recipient's RSA public key. This is the way the SKIP protocol (section 19.3), Privacy Enhanced Mail (PEM) [68] and many other secure mail programs perform message encryption.

A lot of people have been trying to figure out how to break RSA, and they haven't come up with anything yet. The real premise behind RSA's security is the assumption that factoring a big number is hard. The best known factoring methods are really slow. To factor a 512-bit number with the best known techniques would take about a half million MIPS-years. However, it is not granted that factoring is the only way of breaking RSA. As with any encryption algorithm, there might be some other means of breaking RSA.

RSA is probably the most widely used public-key algorithm. Application areas for RSA are several. It is for instance used in the Secure Socket Layer protocol (SSL) to negotiate session keys between Web clients and a secure server, as well as in secure e-mail packages like S/MIME and Pretty Good Privacy (PGP).

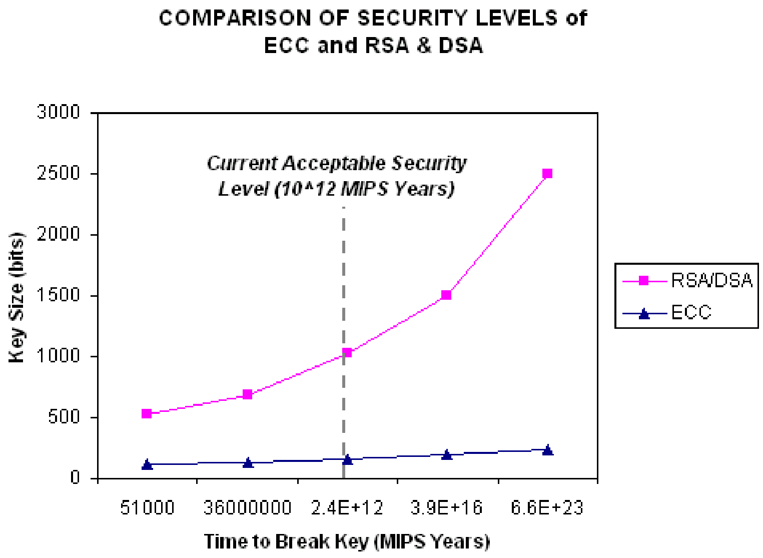
16.5 Elliptic curves cryptosystems

Public-key cryptosystems base their overall security on the computational difficulty of the underlying mathematical problem. The security level that a mathematical problem can guarantee is related to the time required solving it, given the use of a certain computational power. The three hard problem mostly used in public-key encryption are based on:

- The Integer Factorization scheme
- The Discrete Logarithmic problem
- Elliptic Curve Cryptosystem (ECC)

The integer factorization problem (e.g. used in RSA) and the discrete logarithm problem modulo p (e.g. used in DSA and in Diffie-Hellman key exchange) admit sub-exponential-time algorithms. This means that the problems are still considered very hard to be solved, but less than problems that can be solved only with exponential-time algorithms. The elliptic discrete logarithm problem, which is currently supported in some Diffie-Hellman scheme, is considered the only really exponential-time problem.

Following picture, taken from [24], compares the time required breaking Elliptic Curve Cryptosystem (ECC) with the one required to break RSA or DSA.



There have been several attacks against the elliptic curve discrete logarithm problem. The result of years of intensive study seems to be that ECCs are harder to be solved since they require computational exponential time. Moreover, the security gap between ECC and RSA/DSA grows with the key size. This means that the strength of ECC as key size increases grows much faster than for RSA or DSA algorithms.

17 Appendix C – User Authentication

Authentication is the process of reliably verifying the identity of someone (or something). Authentication mechanisms must be provided to ensure that communication at all times is performed between the correct entities, and this authentication mechanisms need to be protected against not authorized replays.

Message authentication, as opposed to encryption, has a "transient" effect. A published breaking of a message authentication scheme would lead to the replacement of that scheme, but would have no adversarial effect on information authenticated in the past. This is in sharp contrast with encryption (section 16.4), where information encrypted today may suffer from exposure in the future if, and when, the encryption algorithm should be broken.

17.1 Authentication schemes

Several schemes are possible to authenticate a remote user, depending on the security layer desired in authentication.

17.1.1 Clear text Password Authentication (PAP)

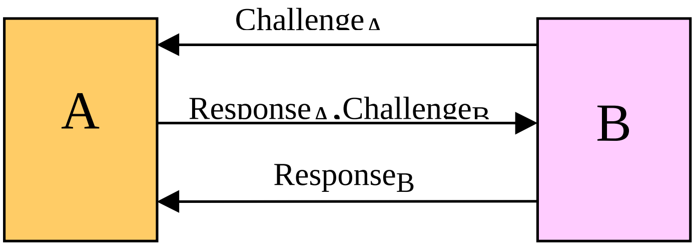
The easiest authentication scheme is to have users sending their passwords to a server who checks these passwords against his own records. This method is usually called Clear text Password Authentication (PAP).

Clear text Password Authentication is easy to implement and requires a very little number of messages (one or two, depending if the server wants to confirm authentication to the users). However, there are at least two major drawbacks with this method:

- Users can not know if they are actually talking with the correct server,
- It may be possible for a third entity to pick up the password when it is sent to the server and do a replay attack.

17.1.2 Challenge Handshake Authentication Protocol (CHAP)

An improved authentication scheme is to use a challenge protocol where the client may for instance encrypt a random number together with a time-stamp using its password. A simple value generated and challenged from the server may also be used without any time-stamps. The server then repeats the same process and verifies that the client actually did know the correct password.



Challenge Handshake Authentication exchange

This method is insensitive to replays from a third party and can be extended by having the server authorize himself against the client in a similar way. It does not send the secret key through the network, nevertheless showing the possession of this secret. This authentication mechanism may be easily built into small "smart-cards" containing a chip, which is able to perform the encryption of a random number.

A one-way function (section 17.3) is often applied to the random number, the time-stamp or nonce and the password to provide the challenged authenticated value (Message Authentication Code or Message Digest). This method is called Challenge Handshake Authentication Protocol (CHAP) [108].

17.1.3 Public key authentication

The above solutions have one common drawback, i.e. they do not scale well. A secret has to be established and shared in a secure way with each entity it is desired to communicate with. An alternative may be to share the same secret with several other parties. However, when a secret is known by many entities, the security in the system relies on the most insecure entity not to reveal his secrets.

A solution to this problem is the use of an asymmetric cryptosystem scheme, usually named public-key system. This cryptosystem has the property that an object is encrypted with one key and decrypted with another key. These two keys are named Private and Public, with obvious meaning.

If the encryption key is kept secret and the decryption key public, only one entity can perform the encryption but all entities can do the decryption. This makes it is possible for any entity to validate the contents and origin of a message. There is only one entity in possession of the encryption key and only this entity can have generated the authentication code. In this way it is possible for an entity to prove his identity to other entities without revealing any secrets and without the need of sharing a secret key with other parties.

17.2 Message Authentication Code (MAC)

To make it possible for the receiver of an object to verify the integrity of an object it is necessary to include some redundant information. This redundant information is usually named Message Authentication Code (MAC) (Encrypted Message Digest (MD) or encrypted Hash). A Message Authentication Code is generally appended to the message and can be verified only if the secret key is known. The receiver uses this information to authenticate the sender and the integrity of the message.

Redundant information used for authentication can be created by combining either a symmetric or asymmetric cryptosystem with a checksum that can be calculated from the contents of the object. The checksum mechanism has the advantage that the object itself does not have to be encrypted, and this saves a lot of time and effort in the communication process.

To select a checksum-generating algorithm is a rather complicated process. The statistical probability that the checksums for two different texts is equal should be zero or almost zero. Moreover, it is suggested that the checksum algorithm should not be a secret in itself, but rather use a secret key.

MACs are usually obtained from a one-way hash function, like MD5 or SHA. In particular, two relevant schemes are the Nested MAC (NMAC) and the Hash-based MAC (HMAC). It may also be possible to use a symmetric cryptosystem such as DES to generate the MAC directly. However, in this case, the effort of generating the check-sum would be as big as encrypting the whole message for data confidentiality.

17.2.1 Keyed-hash MAC (HMAC)

The Keyed-Hashing for Message Authentication Code (HMAC) [65] algorithm is a secret key authentication procedure, which may be considered as an improved version for Message Authentication Codes (MAC). It may be used both for data integrity and data origin authentication.

The HMAC algorithm only provides a framework for inserting specific hashing algorithms such as SHA-1, MD4 or MD5. The overall security will thereafter depend upon the dependability of these other algorithms. The HMAC computational cost is also very related with that of the underlying hash function.

HMAC can be used with any iterative cryptographic hash function, e.g. MD5, RIPEMD and SHA-1, in combination with a secret shared key. The cryptographic strength of HMAC depends on the properties of the underlying hash function. HMAC-MD5 and HMAC-SHA-1 are the two most widely used constructs of the HMAC message authentication function. Both constructs may be used by IPSec (AH and ESP) or other protocols to authenticate messages.

It is often common in HMAC to truncate the output resulting from the one-way hash function. There are several analytical advantages in doing this, even if the resulting security is still a debated question. A realization of HMAC that uses a hash function H with t bits of output is usually indicated as HMAC-H-t. HMAC-SHA-1-96 for instance would produce a 160-bit authenticator value but truncate it using only the first 96 bits. In this case, the receiver computes the entire 160-bit value and compares the first 96 bits to the value stored in the authenticator field.

HMAC specification [65] reports that keys length less than the authentication digest length decrease the overall security strength of the protocol, whereas keys longer than the authenticator length do not significantly increase security strength. Applications that use keys longer than the message digest length should first hash the key to the message digest length and then use the resultant byte string as the actual key for HMAC. HMAC-SHA-1 for instance uses a key of 160-bits, which is equal to its hash length [70].

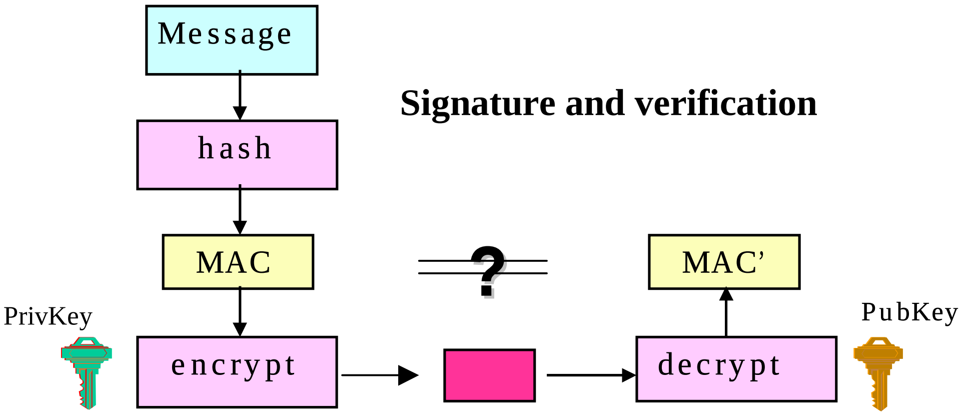
Being HMAC construction independent from the particular hash function used, the latter can be replaced by any other secure (iterative) cryptographic hash function without any severe problem. Test cases with relative results are available for HMAC and may be used as a conformance test for HMAC implementation [25].

17.2.2 Non-repudiation services and Digital Signature (RSA and DSS)

It is sometime desired to provide some not-repudiable proof that an entity has either transmitted (non-repudiation of origin) or received (non-repudiation of delivery) a message. The first of these two services (non-repudiation of origin) is usually provided through a special hash construction named Digital Signature and based on public-key cryptosystems.

An asymmetric cryptosystem can in fact be used as a digital signature to guarantee that a specific entity has created and sent an object. When an entity creates a digital signature, it cannot later deny to have created it. The signature consists in a message (for example plaintext data, a publicly known number or a Message Authentication Code) encrypted with the sender's private key. The receiver verifies the signature using the sender's public key.

Being the private key known only to the sender, this can not deny having created this object because no one else would have been able to perform the encryption. To verify a signature constructed with this type of scheme it is necessary to have the message itself. Signature schemes with appendix are in this way distinguished from signature schemes with message recovery, where it is not necessary to send a copy of the pre-encrypted message.



RSA (section 16.4.5) and DSS are two examples of asymmetric algorithms often used to construct digital signatures. RSA Data Security Inc. (RSADSI), for instance, has developed a set of standards (named Public Key Cryptography Standards-PKCS) for applying public-key cryptography while signing a message. The National Institute of Standards and Technology (NIST), instead, developed the Digital Signature Standard (DSS) [118] as a Federal Information Processing Standard (FIPS) publication, which uses the Digital Signature Algorithm (DSA) instead of the RSA algorithm.

DSA is used both to verify the sender's identity and the data integrity of the transmitted packets. As RSA, DSA is a public-key algorithm where the sender signs the data using his private key and the receiver verifies the signature using the sender's public key. The message to be sent and signed is first compressed using a hash function (section 17.3). Only the resulting message digest is used as input to DSA.

A weakness in DSS is that a secret random value has to be generated in each signature. The DSS private key is vulnerable if this random value is not well generated. It is hard to generate real random numbers in a deterministic machine like a computer. If an attacker guesses the random number, he can extract the private key and forge the signature.

Other services may be provided in the same area. The following is a list of them:

- Proof of origin of data (digital signature)
- Proof of original content (digital seal)
- Proof of delivery

- Proof of original content received

The first two services protect the receiver and the other two the sender. All four services are called non-repudiation services since neither the sender nor the receiver can deny having sent/received a message or deny the contents of a message.

17.3 One-way hash functions

A Hash, Message Digest (MD) or Message Authentication Code (MAC) is usually the result of a one-way function. One-way functions have the special property to take an input message and produces an output, while it is not practically possible to figure out what input resulted in what output.

A function used to compute hash values should also be collision resistant. This means that it must be computationally infeasible to produce two messages having the same message digest, or to produce any message having a given message digest as target. Also the function may produce a one-to-many mapping between the input and the output, hence given the output it should not be possible to determine the input.

A hash h (m) of a message m is considered one-way if it satisfies the following properties:

- For any message m, it is relatively easy to compute h (m).
- Given h (m), there is no way to find an m that hashes to h (m) in a way that is substantially easier than going through all possible values of m and computing h (m) for each one.
- Even though it's obvious that many different values of m will be transformed to the same value h (m) (there may even be infinite values of m), it is computationally infeasible to find two values that hash to the same h (m).

The strongest attack known against one-way functions is based on the frequency of collisions for the hash function. This is named "birthday attack" and is totally impractical for minimally reasonable hash functions.

In the rest of this section, some of the most popular algorithms used to compute message digests will be presented and compared.

17.3.1 MD4

The Message Digest 4 (MD4) [101] algorithm is one of the most studied one-way hash functions. It has been designed to be 32- bit word-oriented so that it can be computed faster on 32-bit CPUs. MD4 can handle messages with an arbitrary number of bits and can be computed in a single-pass over the data. This makes MD4 very fast to be elaborated. MD4 was indeed designed to be exceptionally fast. It is "at the edge" in terms of risking successful cryptanalytic attack.

The input message to be fed into the message digest computation must be a multiple of 512-bit (sixteen 32-bit words). If this is not the case for the original message, the latter is padded by adding a 1 bit, followed by enough 0 bits to leave the message 64 bits less than a multiple of 512 bits. Then a 64-bit quantity representing number of bits in the unpadded message, mod 264 is appended to the message.

The message digest computed in MD4 is a 128-bit quantity (32-bit words). The message is processed in 512-bit (sixteen words of 32-bits) blocks. The message digest is initialized to a fixed value, and then each stage of the message digest computation takes the current value of the message digest and modifies it using the next block of the message. The final result is the message digest for the entire message.

Each stage makes three passes over the message block. Each pass has a slightly different method of mangling the message digest. At the end of the stage, each word of the mangled message digest is added to its pre-stage value to produce the post-stage value (which becomes the pre-stage value for the next stage). Therefore, the current value of the message digest must be saved at the beginning of the stage so that it can be added in at the end of the stage.

17.3.2 MD5

Message Digest 5 (MD5) [102] is more conservative than MD4 in the sense that MD5 is more concerned with security than speed. This means that MD5 is slightly slower than MD4, at the same time providing a much greater security.

MD5 algorithm takes as input a message of arbitrary length and produces as output a 128-bit Message Digest or Message Authentication Code of that input. It, being very similar to MD4, has been designed to be quite fast on 32-bit machines and it does not require any large substitution tables.

MD5 may even be considered as an extension of the MD4 message-digest algorithm. The main differences between MD4 and MD5 are as follows:

- MD4 makes three passes over each 16-byte chunk of the message. In MD5 a fourth round has been added, i.e. MD5 makes 4 passes instead of 3 over the 16-byte chunk.
- The functions are slightly different and also the number of bit shifts is different.
- MD4 has one constant that is used for each message word in pass 2, and a different constant used for all the 16 message words in pass (no constant is used in pass 1). MD5 uses a different constant for each message word on each pass. Since there are 4 passes, each of which deals with 16 message words, there are 64 32-bit constants used in MD-5.

The padding in MD5 is the same as it is in MD4. Padding is always performed, even if the length of the message is already congruent to the needed length (448 modulo 512). The message digest processing is also done in the same way as it is done in MD4 and it is a 128-bit quantity as well. The message is processed in 512-bit (sixteen 32-bit words).

Each stage computes a function based on the 512-bit message chunk and the message digest that results in a value which is passed onto the next stage. The value resulting from the final stage is the final message digest. Each stage in MD5 makes 4 passes over the message block (compared with 3 for MD4). As in MD4, the value resulting from the function applied to the 512-bit chunk is added to the message digest value obtained from the last stage.

MD5 has been especially designed for digital signature applications, where a large file must be compressed in a secure manner before being encrypted with a private (secret) key with a public-key cryptosystem such as RSA (section 16.4.5).

17.3.3 RIPEMD-160

RIPEMD-160 [30] is a 160-bit cryptographic hash function, used as a secure replacement for the 128-bit hash functions MD4, MD5, and RIPEMD. These algorithms do not longer offer sufficient protection for high security applications with the increasing capacity of present computers, as described in several recent papers. RIPEMD-160 is a good candidate for use within IPSec AH and ESP [61]. It is the main alternative to SHA-1 (section 17.3.4).

RIPEMD-160 is a strengthened version of RIPEMD with a 160-bit hash result, and it is expected to be secure for the next ten years or more. Like its predecessors, RIPEMD-160 is tuned for 32-bit processors, which will remain important at least in the coming decade.

It consists of essentially two parallel versions of MD4, with some improvements to the shifts and the order of the message words. The two parallel instances differ only in the round constants. At the end of the compression function, the words of left and right halves are added to yield a 128-bit result. RIPEMD is believed to be stronger than extended MD4, which consisted of two parallel versions of MD4 with a 256-bit result.

The main part of the RIPEMD algorithm is known as the compression function: it computes the new state from the old state and the next 16-word block. The compression function consists of five parallel rounds, each containing 16 steps. The total number of steps is thus 5*16 * 2 = 160, compared to 3*16 = 48 for MD4 and 4*16 = 64 for MD5. First, two copies are made from the old state. Both halves are processed independently. Each step updates one of the registers based on the other four registers and one message word. At the end of the compression function, the new state is computed by adding to each word of the old state one register from the left half and one from the right half

RIPEMD-128 is also defined, as a plug-in substitute for RIPEMD (or MD4 and MD5) with a 128-bit result. However, 128-bit hash results do not offer sufficient protection for high level security. They may be still used only for low security applications.

17.3.4 SHA-1

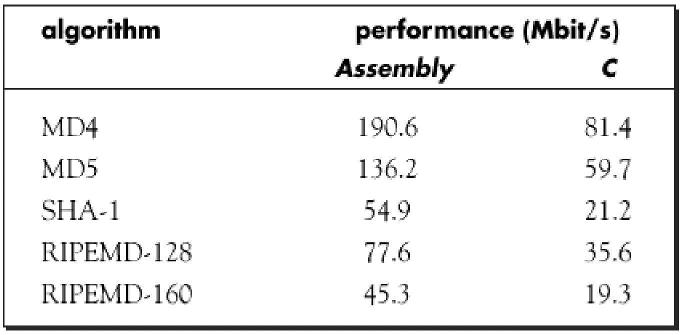
The Secure Hash Algorithm (SHA) has been derived from the MD4 algorithm and it may be considered as the present alternative to RIPEMD-160. A flaw was found in the original specification of this algorithm (SHA) [119], and a second version of SHA has been published to address this flaw, named SHA-1 [120]. SHA-1 appears to be cryptographically stronger than MD5. MD5 has better computational performance.

SHA's output (digest) length is 160 bits, usually 32-bit aligned. Key length in SHA-1 is not constrained to any particular size. Lengths of up to 160 bits are usually supported in most implementations, although key length may be shorter. Long keys are encouraged when high level security is desired.

There are no known flaws in the present version of the Secure Hash Algorithm. This means that there are no known attacks on SHA-1 or any of its components that are better than brute force attacks. The 160-bit hash output in SHA-1 is substantially more resistant to brute force attacks than the 128-bit hash size of MD4 and MD5. SHA is 62% as fast as MD5 and 80% as fast as DES hashing [75], i.e. MD5<DES<SHA (computational cost).

17.3.5 Performances of MD4-like one-way hash functions

The following table compares the performance of RIPEMD-160, RIPEMD-128, SHA-1, MD5, and MD4 [96]. Implementations were written in Assembly language optimized for the Pentium processor (90 MHz). Relative speeds coincide more or less with predictions based on a simple count of the number of operations. RIPEMD-160 is about 15% slower than SHA-1 and four times slower than MD4.



algorithm	**performance (Mbit/s)**	
	Assembly	***C***
MD4	190.6	81.4
MD5	136.2	59.7
SHA-1	54.9	21.2
RIPEMD-128	77.6	35.6
RIPEMD-160	45.3	19.3

Table 1: Performance of several MD4-based hash functions on a 90 MHz Pentium

18 Appendix D – IP Security protocol (IPSec)

IP is the most widely used network layer protocol. It underlies networks from medium-size corporate LANs up to the Internet itself. For this reason, the Internet Engineering Task Force (IETF) has created a working group (WG) to build a security protocol, named IP Security suit (IPSec), for securing IP networks.

IPSec [113] consists in a set of protocols that add connectionless security over the IP layer, so securing all the communications more transparently than it would have been using any other approach. IPSec is compatible both with current IP standard (IPv4) and with the upcoming next IP standard (IPv6). It is independent from specific encryption algorithms (sections 17.3 and 16.4), nevertheless supporting all of them.

Being algorithm independent, IPSec flexibly supports combinations of authentication, integrity, access control, and confidentiality. It also provides the instruments necessary to build up and manage secure Virtual Private Networks (VPN).

18.1 IPSec Framework

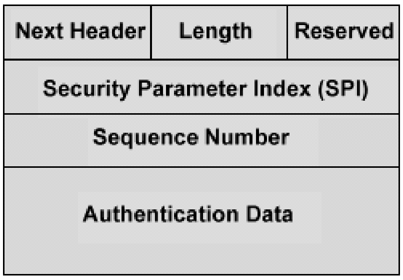
IPSec is an end-to-end protocol. Only the sender and the recipient have to be IPSec-compliant, while the rest of the network in between does not need any change to support IPSec. The Authentication Header (AH), used for authentication and integrity (section 18.2) and the Encapsulating Security Payload (ESP), used for data confidentiality (section 18.3), are the building blocks of the IPSec Protocol Suite [60].

In the IPSec structure, the Internet Security Association and Key Management Protocol (ISAKMP) also cover a critical role. This is a scalable secure key distribution and key management mechanism (section 19.4) used to upheld both the AH and ESP IPSec extensions. It provides an architecture to establish the cryptographic parameters and to exchange keys.

18.2 Authentication header (AH)

The IPSec Authentication header (AH) [58] is used to provide data origin authentication, integrity and anti-replay services for IP packets. It may also be used for non-repudiation of the IP header and payload. It is basically independent from the algorithm used to compute the hash value. This means that all of the one-way algorithms presented in section 17.3 may be used. The following is the structure of an AH header:

Authentication Header (AH)



The Next Header field is an 8-bit field that identifies the type of the higher-level protocol following the AH (e.g. ESP or TCP). The Security Parameter Index (SPI) is a 32-bit value used by the receiver to identify which set of security parameters is used, i.e. the Security Association (section 19.4).

The Sequence Number field contains a counter that increases each time a packet is sent to the same address with the same SPI. The receiver checks this sequence number against an anti-replay window. If the sequence number seems too old or duplicates any one in the window, this packet will be discarded as invalid. This allows the use of the same SA only for a certain number of connections and thwarts replay attacks.

The Authentication Data field contains a keyed hash of the data payload (e.g. HMAC-MD5 or HMAC-SHA-1). The authentication data field also contains the Integrity Check Value (ICV) for the packet. The AH hash value covers both the original IP packet and immutable fields in the outer header.

AH can be used in two different ways, named transport and tunnel mode. In transport mode, the AH header is inserted after the IP header and before the upper layer protocol (the encrypted original packet and possibly the ESP extensions). In tunnel mode, two IP headers are used. The inner IP header carries the original final source and destination addresses and is completely protected by the AH. The outer IP address may contain distinct IP addresses. Tunnel mode is usually implemented in security gateways, e.g. Firewall, or to provide some protection from flow analysis.

Original IP header	TCP/UDP	Data

Before applying AH

Original IP header	AH	TCP/UDP	Data

After applying AH in Transport Mode

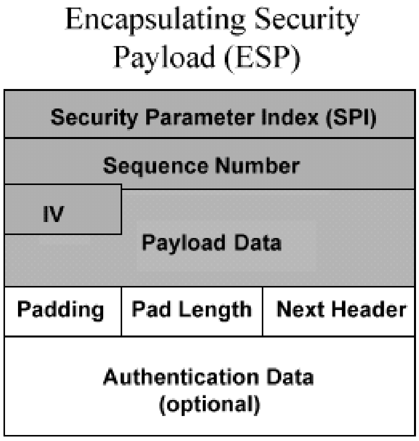
New IP header	AH	Original IP header	TCP/UDP	Data

After applying AH in Tunnel Mode

In both the cases, the SPI, in combination with the destination IP address and security protocol (AH), uniquely identifies the specific AH SA (authentication algorithm, authentication key, anti-replay enabled or not) which is used to protect the current IP packet.

18.3 Encapsulation Security Payload (ESP)

The IPSec Encapsulation Security Payload (ESP) [59] provides confidentiality service for IP packets. ESP can also provide data origin authentication, integrity and anti-reply services, but it only covers the original IP packet. None of the fields in the outer IP header are covered by ESP authentication. It encapsulates the upper-layer protocol data inside the ESP, encrypts most of the ESP payload, and then appends a new IP header to the resulting encrypted ESP.

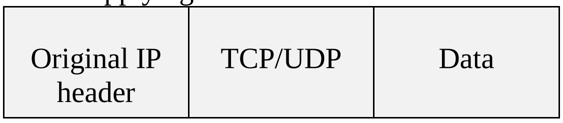

Encapsulating Security Payload (ESP)
Security Parameter Index (SPI)
Sequence Number
IV
Payload Data
Padding
Pad Length
Next Header
Authentication Data (optional)

The fields are used in the same way as described in AH. The Payload Data field contains the encrypted data. The Padding field is used mainly because same types of encryption algorithms require the data to be a multiple of a certain number of bytes (section 16.4).

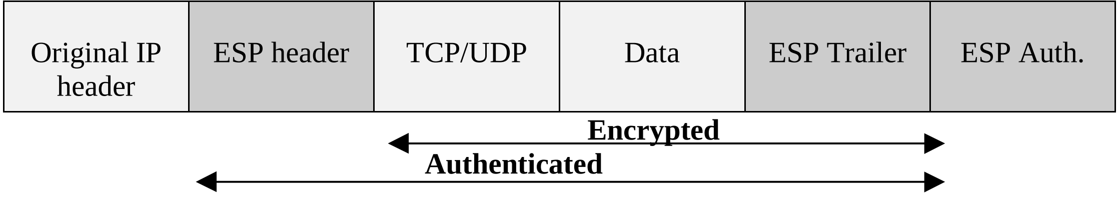
The Authentication Data field is optional and is used when authentication has been selected in the applied SA. If ESP is not used to provide authentication of origin, the sequence number and authentication data fields are not used. The SPI, in combination with the destination IP address and security protocol (ESP), uniquely identifies the specific ESP SA (encryption algorithm, algorithm mode, and encryption key) which is used to protect the current IP packet.

ESP may be used in tunnel or transport mode, as showed in the following pictures:

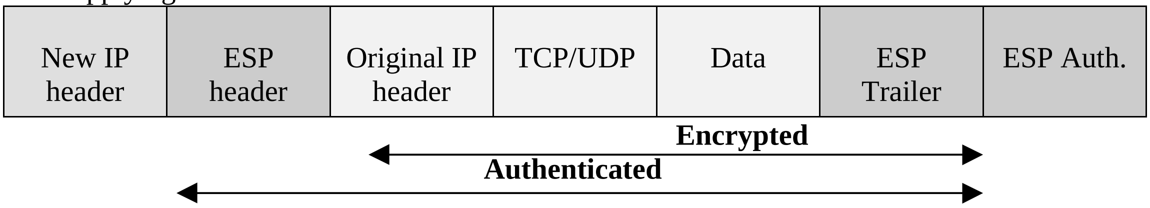
Before applying ESP



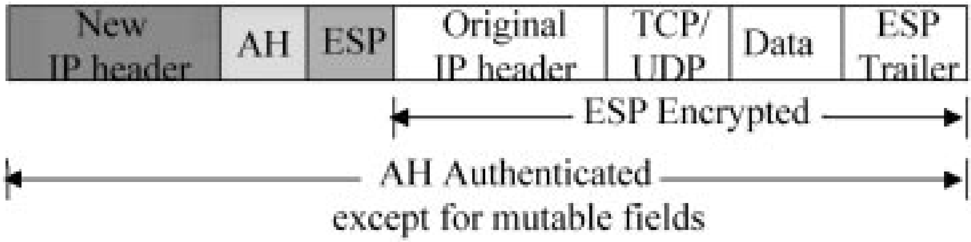
After applying ESP Transport Mode



After applying ESP Tunnel Mode



ESP and AH may also be combined as following:



19 Appendix E – Key Management Protocols

Key management and distribution is one of the most critical services for cryptography. Defining secure cryptographic algorithms and protocols is not easy, but generating, exchanging, and keeping secret keys is often an even harder goal. Many successful attacks break cryptosystems through flaws in their key management procedures.

Much of the overall security for a cryptosystem resides in the strength of key generation and distribution procedures. Not only distribution but also generation of real random-bit keys is not as simple as it may appear at a first glance.

It is possible to distinguish between two different types of keys:

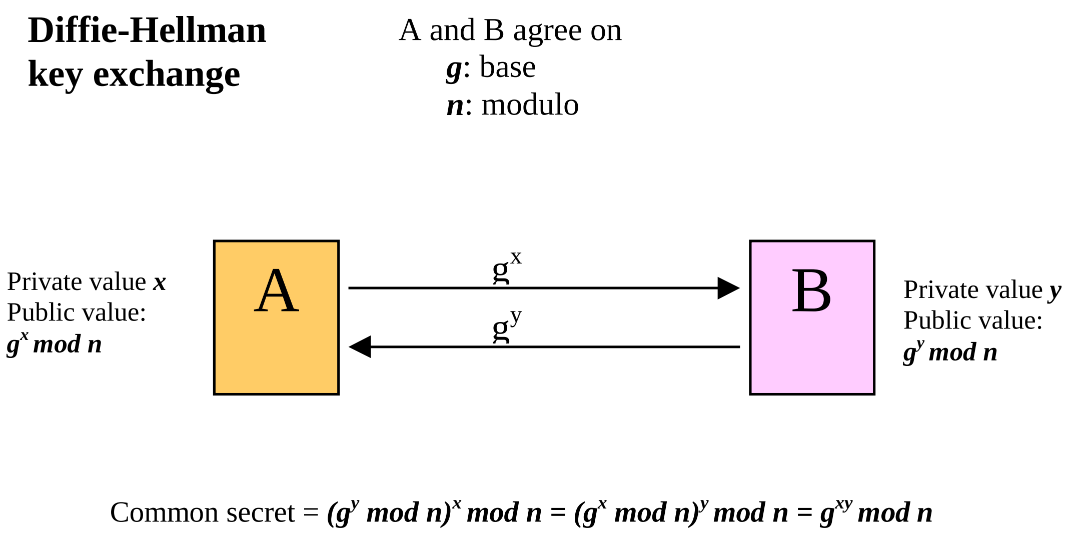
- *Master key-encryption keys, used to encrypt and established other keys*
- *Data encryption keys (session keys), used to encrypt and protect data*

19.1 Diffie-Hellman key exchange

The Diffie-Hellman key exchange is a well-known public-key technique, used to distribute keys but not to encrypt data. The Diffie-Hellman exchange allows two entities to agree on a shared secret key, even when they only can exchange messages in public.

To begin the Diffie-Hellman exchange it is necessary to define two numbers; p, a prime number, which is usually about 512-bits and g, which is a number less than p. p and g are known beforehand and can be publicly known. Once there is a mutually agreed upon p and g, both the entities choose a 512-bit number at random and keeps it secret

They elevate g to this randomly chosen value and send the result modulo p to the other entity (even through a public channel). Then, each entity uses the own generated value (exchanged only in a derived form) together with the other party's pre-computed value to generate a common secret. Even when the channel is not protected and enemies may eavesdrop the exchanged values, there would be no way to retrieve the final shared secret.



The Diffie-Hellman key exchange is patent-free and can be also used to achieve Perfect Forward Secrecy. Even if the two parties do not start out with any shared secret, yet after the exchange of two messages that may even be public they will share a secret key. This new-shared secret value may be then used inside a secret-key algorithm.

19.2 In-band and out-of-band Session Key exchanges

Any key-management scheme that needs to scale in a good way must be based on an authenticated public-key infrastructure. Public keys can be authenticated using a variety of mechanisms, such as public key certificates, secure directory servers (e.g. LDAP) and so on. ITU standard X.509 (section 19.6) provides an example of a public key infrastructure, built using certificates. Pre-shared master keys and other out-of-band key distribution techniques may also be used when the system must be upheld on a limited scale.

Private and public keys are usually not used to directly encrypt data traffic, e.g. IP packets. They are mostly used to authenticate some other key exchange mechanisms able to establish session keys. The latter will then be effectively used for data encryption. There are two different kind of key exchange where public keys can be used to establish session keys and eventually provide datagram protection.

The first alternative is to establish an authenticated session key, using some out-of-databand session key establishment protocols. This session key is then used to encrypt or authenticate IP data traffic. IKE/ISAKMP is the most used among these protocols and will be described in sections 19.4 and 19.5. Such a scheme has the disadvantage of having to establish and (securely) manage a pseudo-session layer underneath IP, which is a stateless protocol.

The IP source would need to first communicate with the IP destination to acquire this session key. Also, if and when the session key needs to be changed, the IP source and the IP destination need to communicate again to make this happen. Each such communication could involve the use of a computationally expensive public-key and Diffie-Hellman operation.

Secure management of pseudo-session states is further complicated by crash recovery considerations. If one side crashes, and looses all session state, then mechanisms need to exist to securely remove the half-opened session state on the side that did not crash. These mechanisms need to be secure because insecure (unauthenticated) removal of half-opened sessions opens the door to a trivial denial-of-service attack.

An alternative approach is to utilize public keys in a stateless key-management scheme, such as it is done in the Simple Key-Management for Internet Protocols (SKIP), described in section 19.3. This works through in-band signaling of the packet encryption key, where the packet encryption key is encrypted in the recipient's public key. This is also the way Privacy Enhanced Mail (PEM) [68] and other secure mail programs perform message encryption.

Although this avoids the session state establishment requirement and prior out-of-band communication to set up and change packet encryption keys, this scheme has the disadvantage of having to carry the packet encryption key encrypted in the recipient's public key in each IP packet. Since for instance an RSA encrypted key minimally needs to be 64 bytes, and can be 128 bytes, this scheme introduces the overhead of 64 to 128 bytes of keying information in every packet. Moreover, when the packet encryption key changes, a public key operation will need to be performed to recover the new packet encryption key. Thus, both the protocol and computational overhead of such a scheme is very high.

19.3 Simple Key-Management for Internet Protocols (SKIP)

The Simple Key-Management for Internet Protocols (SKIP) [5] is a stateless and sessionless key management protocol, used to upheld session keys protected with a public-key cryptosystem. In SKIP each IP-based source and destination has an authenticated Diffie-Hellman (DH) public value. This public value can be authenticated in numerous ways. Some possibilities for authenticating DH public values are the use of X.509 certificates (section 19.6), Secure DNS, or PGP certificates. These certificates can be signed using any signature algorithm, such as RSA (section 16.4.5) or DSA (section 17.2.2).

Each IP entity in SKIP has a secret value and a public value, as in any public-key cryptosystem. A shared secret is derived from these values and it is used as the basic key-encrypting key that provides IP packet-based authentication and encryption. This secret is called the long-term secret and it is used as key for a block symmetric cryptosystem like DES, RC2, IDEA, etc. (section 16.4).

This master key is used to encrypt a transient key, called session key, which is used for data packet encryption. This is done to limit the actual amount of data encrypted using the long-term key, since it is desirable to keep the latter for a relatively long period of time. Transient keys are only encrypted in this long-term key, and use the transient keys to encrypt or authenticate IP data traffic.

This limits the amount of data protected using the long-term key to a relatively small amount even over a long period of time, since session keys represent a relatively small amount of data.

If a node wants to change the session key, the receiver can discover this fact without having to perform a public key operation. It uses the cached long time value to decrypt the new session key. Thus, without requiring communication between transmitting and receiving ends, and without necessitating the use of a computationally expensive public key operation, the packet encrypting or authenticating keys can be changed by the transmitting side and discovered by the receiving side.

SKIP is simple to implement and provides straightforward and scalable solutions to permit dynamic rerouting of protected IP traffic through alternate encrypting intermediate nodes for crash-recovery, fail-over, and load-balancing scenarios. It is also especially suited for multicast and broadcast traffic. It may instead not be the best choice when bulk data transmission is done between two stabile and well-connected hosts. In the last case, session oriented protocols, such as IKE/ISAKMP (section 19.4), should be preferred.

19.4 ISAKMP

The Internet Security Association and Key Management Protocol (ISAKMP) [72] protocol provides a framework for authentication, security association management, and key generation and distribution. It does not specify any key exchange protocols because it is supposed to be key exchange protocol independent.

ISAKMP is a powerful and flexible protocol and supports the negotiation of the security policies under which the secure communication channel has to be established. However, ISAKMP's flexibility is provided at the cost of complexity, increased latency in association establishment, and increased packet traffic.

ISAKMP uses UDP as transport protocol, without relying on any UDP information (e.g., checksum) for its processing. ISAKMP can negotiate SAs for security protocols at any ISO/OSI layer.

19.4.1 Security Associations (SA)

The Internet Security Association and Key Management Protocol (ISAKMP) protocol is used to establish an agreement on the encryption algorithm, the authentication method, and the hash algorithm, which will have to be used in AH and ESP extensions. An instance of this set of parameters is named a Security Association (SA).

The ISAKMP protocol defines two different kinds of Security Associations:

- **ISAKMP SA**: It defines the policy used to establish a secure channel for further communication. This policy specifies the security parameters used, e.g., authentication method, encryption algorithm, Diffie-Hellman group, and hash algorithm.
- **Non-ISAKMP SA**: Once channel has been secured using an ISAKMP SA, a "protocol SA" or "non-ISAKMP SA" can be established. The non-ISAKMP SA is the set of the security parameters used for the safe communication of data, e.g., IPSec parameters for AH and ESP. A non-ISAKMP SA is also named session SA.

ISAKMP is also used to exchange keys in a secure way. Nevertheless, it provides this service, at the same time being independent from the actual key exchange protocol used. This is done in order to separate SAs and key management from the details of the key exchange itself. ISAKMP mainly uses the Internet Key Exchange (IKE) protocol as its actual key exchange protocol.

19.4.2 ISAKMP mean features

Being ISAKMP used to establish a secure channel for applications or network protocols, e.g. IPSec AH and ESP, it is important to provide authentication of the involved entities. The authentication method is part of the ISAKMP SA exchanged, even if the detailed implementation depends on the underlying key exchange protocol. ISAKMP/IKE, for instance, defines four possible authentication methods: signature, public-key encryption, encrypted nonces, and pre-shared key (section 19.5).

ISAKMP uses cookies to protect the protocol from Denial of Service (DoS), also named anti-clogging, attack. A cookie is a token card that provides a technique for partial protection against denial of service. Some basic secret material is used to generate cookies, which requires very little computational cost. In this way, the cookie generation is secure and unique. One way of generating cookies is to use Karn's method. A fast hash is performed over the IP Source and Destination Addresses, the UDP Source and Destination Ports, and a locally generated secret random value. Time and date are added, so the cookie is unique for each SA establishment.

ISAKMP exchanges are linked together to avoid insertion of fraudulent messages. The authentication, key, and SA exchanges are linked together to avoid an enemy jumping in and impersonating one of the peers. Time-variant information is inserted in the cookies against reflection back to the server and replay of old messages. Authentication is provided.

19.4.3 ISAKMP phases

The Internet Security Association and Key Management Protocol (ISAKMP) protocol negotiation consists of two separate phases. In the first phase (ISAKMP phase I) the peers establish a secure channel for further communication by negotiating ISAKMP SAs. In the second phase (ISAKMP phase II), the peers negotiate non-ISAKMP SAs, which will be used to protect real communication, e.g. AH and ESP for IPSec.

In ISAKMP phase I, the Initiator sends some proposals and the responder either chooses one of these proposals or rejects the communication. Each proposal contains all the security parameters used to establish an ISAKMP SA. ISAKMP phase II is done only if phase I has been successful. It is similar to phase I, even if it may be much simpler if no Perfect Forward Secrecy (PFS) is required. Phase II can indeed be very fast and simple if the correspondent Phase I guarantees a very secure channel.

If also the ISAKMP phase II has been successful, the peers can start to communicate using the security protocol (e.g., IPSec) with the non-ISAKMP SA they agreed on. ISAKMP SAs are bi-directional in nature: once the SA is established, each of the two entities can begin a Phase II negotiation.

This two-phase approach has a higher start-up cost than a single-phase approach, but the are also several advantages. The cost of Phase I can in fact be shared among several Phase II negotiations (multiple SAs can be established with a single Phase I).

ISAKMP defines a default set of exchange types for ISAKMP one and two phases, i.e. for establishment of both ISAKMP SA and non-ISAKMP SA. The payload format of these types is also defined. Exchanges types define the content and ordering of ISAKMP messages during communications between peers. The primary difference between exchange types is the ordering of the messages and the payload ordering within each message. These exchanges provide different security protection for the exchange itself and information exchanged.

Five default exchange types are currently defined for ISAKMP:

- Base Exchange: identity protection is not provided (identity identifier are exchanged before encryption)
- Identity Protection Exchange: separating the key exchange information from the identity and authentication information provides identity protection. The identities are encrypted before being sent.
- Authentication Only Exchange: the information is in plaintext. There is authentication without the computational expense of computing keys.
- Aggressive Exchange: faster since it has less round trips, but it does not provide identity protection.
- Informational Exchange: a one-way exchange with Error Notification or Deletion.

These exchange types can be used in either phase of negotiation. However, they may provide different security properties in each of the phases. With each of these exchanges, the combination of cookies and SPI fields identifies whether this exchange is being used in the first or second phase of a negotiation.

19.5 IKE

The Internet Key Exchange protocol (IKE) [47] combines ISAKMP, part of Oakley Key Determination Protocol [83], and part of SKEME Key Exchange Mechanism [64] to provide authenticated keying material for use with ISAKMP, IPSec AH/ESP and other security associations. Two separated phases, IKE phase 1 and IKE phase 2, compose IKE.

19.5.1 IKE phases and Security Associations

In IKE phase 1, the two entities negotiate security association (SA) for protection of further ISAKMP message exchange (ISAKMP SA). They negotiate both the authentication method (digital signature, two forms of authentication with public key encryption, or pre-shared key), the Diffie-Hellman group, hash algorithm, encryption algorithm, the ISAKMP SA lifetime, etc. With Diffie-Hellman key exchange, the peers agree on a common secret without exposing it to any third party.

This shared secret, together with other information (nonces, cookies, etc.), is then used to derive keying material for ISAKMP message encryption and authentication and for further derivation of keys for other services, i.e. IPSec, during IKE phase II exchanges.

In IKE phase 2, IPSec SAs are negotiated under the protection offered by ISAKMP SA. It is also possible to make an additional Diffie-Hellman key exchange to provide Perfect Forward Secrecy (PFS). Otherwise, the master keying material generated in phase 1 is used to derive keys for IP packet encryption and authentication.

As already pointed out, IKE is used to define ISAKMP SAs, i.e. a hash algorithm, an encryption algorithm, an authentication method, and a Diffie-Hellman group. Some possible values for the latter are defined as mandatory in IKE specification, others as optional. Any possible values may anyway be supported by IKE.

SA values as defined in present IKE specification:

- Hash algorithm: MD5 and SHA as mandatory, Tiger as optional
- Encryption algorithm: DES-CBC as mandatory, 3-DES as optional
- Authentication method: pre-shared key as mandatory, DSS and RSA as optional
- Diffie-Hellman group: MODP Oakley Group I as mandatory, MODP Oakley Group II as optional

IKE implementations may support any additional encryption algorithm (e.g. Blowfish, MD5, RIPEMD, etc.) and any other Diffie-Hellman groups (e.g. elliptic curve cryptosystems).

IKE also defined four different operation modes, based on the five ISAKMP exchange types:

- Main Mode**:** it is performed in IKE Phase I and it is an instantiation of ISAKMP Identity Protection Exchange
- Aggressive Mode**:** it is performed in IKE Phase I as an alternative to Main Mode and it is an instantiation of ISAKMP Aggressive Mode
- Quick Mode**:** it is used for IKE Phase II (i.e. for non-ISAKMP SA)
- New Group Mode**:** it is optional and it is used after IKE Phase I to define a new group for Diffie-Hellman key exchange

There is also a fifth mode, named informational exchange, which is directly derived from ISAKMP and it is used when anomalous situations have to be audited.

19.5.2 Main Mode and Aggressive Mode

Main Mode and Aggressive Mode are the two IKE Modes used during ISAKMP phase I exchange. They both establish a SA and provide an authenticated Diffie-Hellman key exchange to create a shared secret key to be used in IKE phase II.

Main Mode requires six one-way messages between the peers, whereas Aggressive Mode requires three one-way messages. This means that Aggressive Mode is a little faster than Main Mode. However, it does not provide identity protection because the peers' identities are transmitted before a secure channel has been established. With Aggressive Mode, an enemy can get information about the involved entities. Moreover, SA negotiation is limited in Aggressive Mode and the Diffie-Hellman group can not be negotiated. Main Mode provides the richest negotiation support.

More in detail, in Main Mode the first two messages are used for policy negotiation, the third and the fourth exchange Diffie-Hellman material and ancillary data (nonces), while the last two are used to authenticate previously exchanged data. In Aggressive Mode, the first two messages are used for negotiating policy, Diffie-Hellman exchange, and to provide ancillary data and identification. With the second message, the Responder also authenticates itself. The third message is instead used for the Initiator's authentication.

Keys used in IKE phase II or to generate IPSec keys are computed on some of the exchanged or derived values. In particular, the first key-seed that the entities have to calculate is SKEYID [47] and the peers may derive it in different ways according to the selected authentication method (section 19.5.3). Other keys are in various ways derived from SKEYD.

19.5.3 Authentication methods

The result of Phase I is the establishment of a secure channel through an authenticated key exchange. Four different authentication methods can be applied with either Main or Aggressive Mode. These four authentication methods are digital signature, two public-key encryption schemes, and the pre-shared key method.

Aside from the use of pre-shared key, a hash value is calculated to provide authentication. This value includes the two Diffie-Hellman public values, the two cookies, the SA payload originally sent by the Initiator, and the identity payload of the peer the hash is referred to. The hash algorithm is negotiated during the first IKE exchanges and SKEYID is used as key.

The hash value has to be signed and verified if authentication with digital signature is used. For the other three authentication methods (public-key, revised public-key, and pre-shared key), the hash directly authenticates the negotiation.

Some of the fields included in the definition of the hashes are exchanged under encryption. To be able to calculate the right hash value, each peer has to correctly decrypt each field, so giving a proof it possesses the decrypting keys. When the opposite peer receives the Hash, it calculates the Hash on its own and checks if the calculated value and the received one match. If so, the sender is authenticated being the possessor of the right secret keys.

If Authentication is done using Digital Signatures, after the Diffie-Hellman key exchange in phase 1, both the involved entities compute a hash value over the previously exchanged data. Each peer signs (encrypts) this hash with its private key, and sends the signature together with its identifier to the other peer. After receiving the signed data, a peer calculates the hash in the same way as the other entity has done. It also decrypts the signed data with the other peer's public key to recover the original hash value. Finally it compares the two values and, if they match, authentication is provided.

For a detailed description of the four IKE authentication schemes and exchange modes refers to [47].

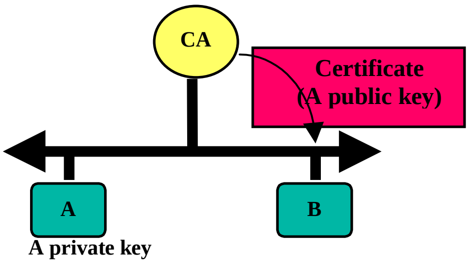
19.6 Public Key Infrastructure (PKI)

All entities that want to authenticate another party need to know either the secret key (if secret-key algorithms are used) or the public key (for public-key ciphers) for all clients in the system, which in a large system very soon becomes impractical from a management point of view.

Public-key systems partly solve this problem, allowing each entity to have only one secret key (encryption techniques based on public-key require a pair of keys but one of these is shared between all entities). While confidentiality is not a concern for public keys, integrity must be guaranteed.

The integrity of a public key is needed both to guarantee that the public key itself is authentic and that the key string has not been modified. It is important to assure that a public key really belongs to the owner and it is not someone else's key. If not, an attacker may manage to replace a user's public key (doing a man-in-the-middle attack), and in this way the user would be completely impersonated by the attacker. An attacker would be enabled to grab all the secret information intended for the user, and at the same time could do anything in the name of that valid user.

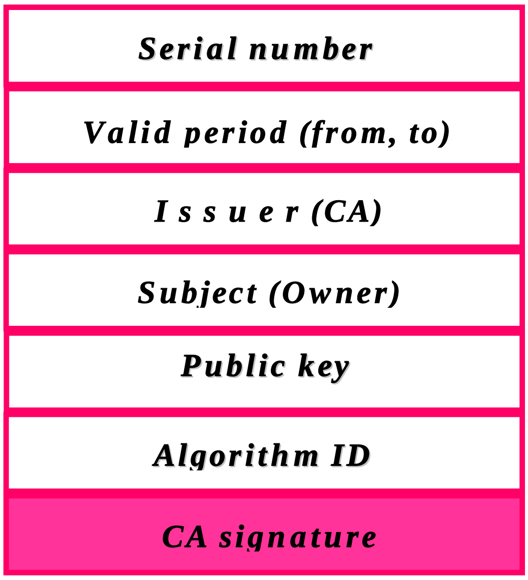
A possible solution is to have a special entity, a third party, performing the authentication process for all entities in the system. An effective way to provide this service is by creating a public-key infrastructure (PKI). This is a network structure where a trusted authority, called Certification Authority (CA), distributes authenticable certificates, containing among other things a public key, the identifier of the owner, and a digital signature of the Certification Authority.



Public Key scheme with Certificate Authority (CA)

At the present time, ITU-T X.509 [49] is the most widely used certificate format standard. It is currently at its third version (X.509 v3) and it has been for instance already defined for use in the Internet Privacy Enhanced Mail protocols, in several WWW applications, IPSec, and as a basis for the overall Internet security. The general structure of the X.509 certificates is as showed in the following picture:

X.509 Certificate



Certificates are securely transmitted encrypting the Message Authentication Code (MAC) of all the data fields (except the CA signature field itself) within the certificate with the private key for the CA, and in this way creating a CA's signature. Whenever a certificate is obtained, it should be verified by recalculating the MAC, decrypting the signature with CA's public key, and comparing the decryption result with the newly calculated MAC.

If they match the certificate is authentic, otherwise some field(s) must have been illegally modified. By this means a public-key certificate ensures the authentic binding between the public key and its owner as well as the integrity of the public key itself.

19.7 Kerberos

Kerberos is an example of trusted third-party authentication service. It is trusted in the sense that each of its clients believes Kerberos judgment as to the identity of each of its other clients to be accurate. Kerberos consists of a Key Distribution Center (KDC) that runs on a physically secure node somewhere on the network, and a library of subroutines that are used by distributed applications which want to authenticate their users.

Kerberos was designed so that once the user logs into a workstation by providing the name and password, then the workstation is able to obtain information from the KDC for any process to access remote resources on behalf of the user. Because Kerberos knows these private keys, it can create messages that convince one client that another is really who it claims to be.

Kerberos also generates temporary private keys, called session keys, which are given to two clients and no one else. A session key can be used to encrypt messages between two parties. Applications, including telnet and NFS, have been often modified to call subroutines in the Kerberos library as part of their startup.

20 Appendix F – IP Mobility

IP layer mobility refers to the ability of moving among different IP sub-networks without loosing network connectivity. It refers both to session and user mobility, even if session mobility is currently the designers' main goal. Session and user mobility will both be addressed in this paper.

Many requirements are applicable to macro mobility. Routing of IP datagrams to mobile nodes, for instance, should be transparent to applications and end-users. Mobile nodes should be able to communicate with other nodes that do not implement mobility functions. Protocol overhead should be as low as possible and only a small additional latency should be introduced in transmission of data, due to the requirements introduced by the mobility and wireless aspects.

Mobile Security is surely one of the hottest areas in the Internet community and this chapter will provide a broad overview of current state-of-the-art proposals.

20.1 Mobile IP

Mobile IP [89] is the IETF Mobile IP WG proposed protocol to support macro-mobility within the Internet. It has been adopted by most of the private and public networks and it is likely to become the standard protocol for future macro-mobility management. It is based on the idea that mobile nodes must be able to communicate with other nodes after changing link-layer and network-layer points of attachment, yet without changing public IP addresses.

In Mobile IP each *mobile node* (MN) is globally identified by a single IP address, regardless of its current network point of attachment. This address is called the *home address* and the mobile node obtained it during the initial phase of the Mobile IP protocol. When away from its home network, a mobile node has associated also a second address, called *care-of address*. However, this second address is only utilized within the Mobile IP protocol. It should not be used as mobile node public address for traffic from the correspondent node to the mobile node (even if it might be used for data sent from the mobile terminal). Indeed, Mobile IP has been designed with the goal of keeping the actual mobile node position unknown to any external users. This can be considered as something more than a simple transparency-based choice.

20.1.1 Mobility management

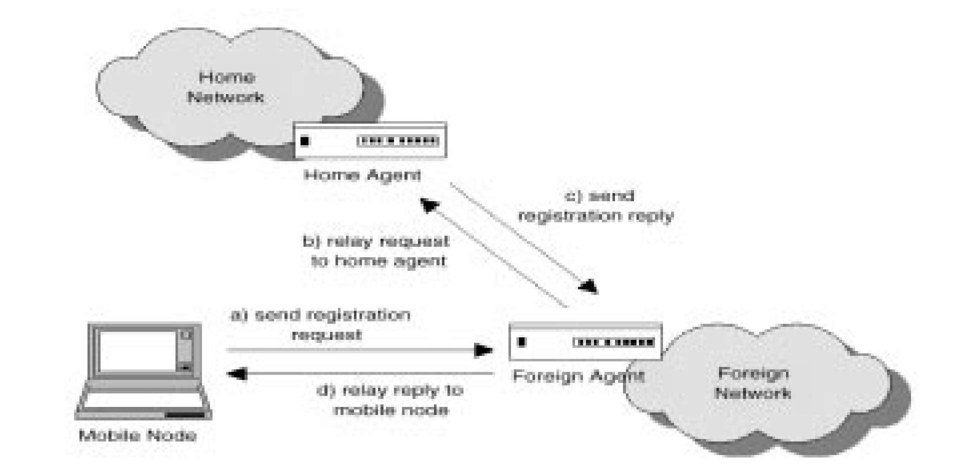
When mobility support is needed, Mobile IP associates a *Home Agent* with each Mobile Node. This agent is responsible for delivery of data to the mobile node's actual position. The home agent usually belongs to the network where the mobile node is most often connected, called *home network*, even though the home agent could be also dynamically assigned on any visited network (this method is not yet standardized).

When the mobile node is located on the same network as its related home agent, normal network procedures and protocols are utilized. When it is away from home, a temporary IP care-of-address needs to be assigned to identify the mobile node's actual position. This is usually obtained through an agent on the visited network, called the *foreign agent*.

The care-of address is sent to the home agent. Depending on the specific method of attachment, the mobile node could send this information either directly or through the foreign agent. In the latter case, the foreign agent simply forwards the registration request to the home agent.

As described here, a care-of address is the mechanism used to provide information about the mobile node’s current position to the *home agent*. Using the mobile node’s care-of-address, stored in a local binding table, the home agent is able to forward datagrams destined to the mobile node to its actual point of attachment. This is done through a tunnel that the home agent establishes between the home agent and the care-of address. Each datagram, after arriving at the end of the tunnel, is decapsulated and delivered to the mobile node.

When the mobile node detects that it has moved to a new IP (sub)network, it needs to obtain a new care-of address on the new foreign network. The movement is usually determined through foreign agent's advertisements, but link-layer support could also be provided. The new care-of address is obtained through the visited Foreign Agent. However, it could also be assigned by some external mechanism such as DHCP [31]. The mobile node will then update the binding table on the home agent, advising it of its newly assigned care-of address.


Home Network
Home Agent
c) send registration reply
b) relay request to home agent
a) send registration request
d) relay reply to mobile node
Foreign Agent
Foreign Network
Mobile Node

(Mobile IP) MN registration with the HA

Mobile nodes can use two main procedures to obtain care-of addresses. If they use *foreign agent care-of addresses*, they receive the temporary addresses through foreign agents. These foreign agents send Registration Replays messages containing the care of address values as answer to the mobile nodes’ Registration Request messages. The care-of address is an IP address belonging to the foreign agent. In this case, moreover, the foreign agent is the endpoint of the tunnel established by the home agent. Foreign agents receives the tunnelled datagrams and decapsulates them, delivering the inner datagram to the mobile node. This mode of acquisition is often preferred because it allows many mobile nodes to share the same care-of address. Therefore, it does not require allocation of many IP addresses for each mobile node.

If a *co-located care-of address* is instead used, the mobile node acquires the care-of address through some external mechanism. The mobile node could use a static address to be utilized only when it visits a specific foreign network. This creates the need of many IP addresses for each mobile node. A more effective solution may be to assign temporary address to the mobile nodes through the use of DHCP. When a co-located care-of address is used, the mobile node is the end-point of the tunnel. Therefore, it performs decapsulation of the datagrams tunnelled to it by the home agent.

In Mobile IP, registration requests and registration replays between home agents and mobile nodes are sent with UDP [93] using well-known port number 434. Agent Discovery messages, as agent advertisements or mobile node agent solicitations instead make use of the Router Advertisement and Router Solicitation messages, defined for the ICMP Router Discovery protocol [94].

20.1.2 ARP and Gratuitous ARP

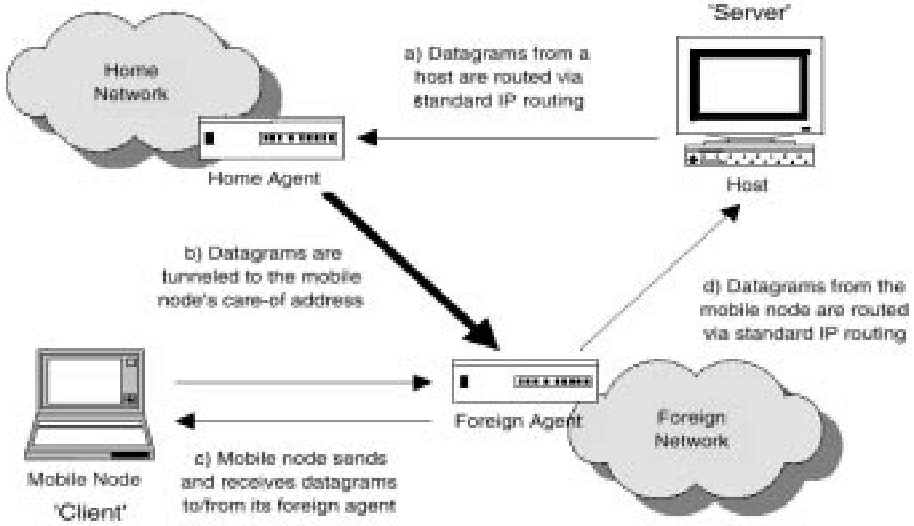
When mobile node is registered at a foreign network, the home agent needs to intercept any datagrams on the home network addressed to the mobile node [89]. All these packets will then be forwarded to the care-of-address that the mobile node is using on the visited network. Address Resolution Protocol (ARP) Messages [92] are usually used by home agents to intercept these datagrams.

Proxy Address Resolution Protocol mechanisms [95] are utilized by home agents to reply to any ARP Requests that asks for the mobile node's link-layer address. In this way, the home agent’s link-layer address is provided as destination for data sent to the mobile node. Moreover, when a mobile node leaves the home network, its home agent uses Gratuitous ARP [111] to update the ARP caches of the nodes on the home network. Also, when the mobile node returns to its home network, it or its home agent can use Gratuitous ARP to update the ARP caches of all the hosts on the home network, binding the mobile node link-layer address to its IP home address.

20.1.3 Triangular Routing

Triangular routing, i.e. asymmetric routing with respect to topology, is one of the main drawbacks in traditional Mobile IP. This leads both to an inefficient use of link resources, because of the use of tunnels that may span a large number of router hops, but also to problems with the TCP [111] protocol mechanisms. Finally, support for Quality of Service (QoS) is also problematic with triangular routing.

Triangular Routing in Mobile IP


Home Network
'Server'
a) Datagrams from a host are routed via standard IP routing
Home Agent
Host
b) Datagrams are tunneled to the mobile node's care-of address
d) Datagrams from the mobile node are routed via standard IP routing
Foreign Agent
Foreign Network
Mobile Node
'Client'
c) Mobile node sends and receives datagrams to/from its foreign agent

An alternative to triangular routing can be for the mobile node to tunnel its messages through the home agent, using *Reverse Tunnelling* [77]. In this way, all messages sent and received by the mobile nodes pass through the home agent and there is a symmetric, even if not optimized, routing path. This solution can also be useful when security mechanisms, i.e. Packet Filtering based on the source IP address, do not allow a mobile node to send data using its home address as data source address. This is especially the case for corporate networks protected by Firewalls.

20.1.4 Broadcast and multicast messages

A mobile node that wishes to receive multicast data must join the multicast group in one of the following ways. First, a mobile node may join the group via a local multicast router on the visited subnet. This option is usually referred as *Remote Subscription*. It assumes that there is a multicast router present on the visited network and that mobile nodes do not move too often. As an alternative choice, the mobile node may join the multicast groups via a bi-directional tunnel to its home agent. The mobile node tunnels IGMP messages to its home agent and the home agent forwards multicast datagrams down the tunnel to the mobile node.

If the mobile node is using a co-located care-of address, the home agent simply tunnels appropriate broadcast IP datagrams to the mobile node's care-of address. Otherwise, the home agent first encapsulates the broadcast datagram in a unicast datagram addressed to the mobile node's home address and then tunnels this encapsulated datagram to the foreign agent. This extra level of encapsulation is required so that the foreign agent can determine which mobile node should receive the datagram after it is decapsulated.

REFERENCES:

[1] Aboba, D. Lidyard, "The Accounting Data Interchange Format (ADIF)", Internet draft (expired), draft-roamops-acctng-06.txt, August 1999.

[2] Aboba, M. Beadles, “The Network Access Identifier”, RFC 2486, January 1999.

[3] Aravamudhan, M. O'Brien, B. Patil, “NAI Resolution for Wireless Networks”, Internet draft (work in progress), draft-aravamudhan-mobileip-nai-wn-00.txt, October 1999.

[4] Arkko et al, “DIAMETER Accounting Extension”, Internet draft (work in progress), draft-calhoun-diameter-accounting-03.txt, December 1999.

[5] Aziz, M. Patterson, “Simple Key-Management for Internet Protocols (SKIP)”, Proceedings of the Fifth Workshop on Enabling Technologies, IEEE Computer Society Press, 1996

[6] Baldwin, R. Rivest, “The RC5, RC5-CBC, RC5-CBC-Pad, and RC5-CTS Algorithms”, RFC 2040, October 1996.

[7] Beadles, “AAA Referral”, Internet draft (work in progress), draft-beadles-aaa-referral-00.txt, June 1999.

[8] Bellovin, “Report of the IAB Security Architecture Workshop”, RFC 2316, April 1998.

[9] Bennett, B. Volz, A. Westerinen, “DHCP Schema for LDAP”, Internet draft (work in progress), draft-ietf-dhc-schema-01.txt, October 1999

[10] Blunk L., J. Vollbrecht, “PPP Extensible Authentication Protocol (EAP)”, RFC 2284, March 1998.

[11] Blunk, J. Vollbrecht, “PPP Extensible Authentication Protocol (EAP)”, RFC 2284, March 1998.

[12] Boström T., L. Hellström, O. Lindroos, “Simple Mobile IP”, MSc Thesis, Royal Institute of Technology & Ericsson Radio Systems AB, Stockholm, November 1999.

[13] Brownlee, E. Guttman, “Expectations for Computer Security Incident Response”, RFC 2350, June 1998

[14] C. Rigney et al, “Remote Authentication Dial in User Service (RADIUS)”, RFC 2058, January 1997.

[15] Calhoun R. et al, “DIAMETER Framework Document”, Internet draft (work in progress), draft-calhoun-diameter-framework-05.txt, December 1999.

[16] Calhoun R. et al, “DIAMETER Framework Document”, Internet draft (work in progress), draft-calhoun-diameter-framework-04.txt, October 1999.

[17] Calhoun R., C. Rubens, H. Akhtar, “DIAMETER Base Protocol”, Internet draft (work in progress), draft-calhoun-diameter-10.txt, October 1999.

[18] Calhoun R., C. Perkins, “DIAMETER Mobile IP Extensions”, Internet draft (work in progress), draft-calhoun-diameter-mobileip-05.txt, December 1999.

[19] Calhoun R., C. Rubens, H. Akhtar, “DIAMETER Base Protocol”, Internet draft (work in progress), draft-calhoun-diameter-12.txt, December 1999.

[20] Calhoun R., E. Perkins, “Mobile IP Network Access Identifier Extension”, Mobile IP Working Group, Internet draft (work in progress), draft-ietf-mobileip-mn-nai-07.txt, January 2000.

[21] Carrara E., “Wireless adaptation of a security management protocol suite”, Universita’ degli studi di Genova, March 1999.

[22] Castelluccia C., L. Bellier, “A Hierarchical Mobile Management Framework for the Internet”, MoMuc’99, November 1999

[23] Castelluccia C., L. Bellier, “Toward a Unified Hierarchical Mobility Management Framework”, Mobile IP Working Group, Internet draft (work in progress), draft-castelluccia-uhmm-framework-00.txt, June 1999.

[24] Certicom Whitepaper, “Current public-key cryptographic systems”, ECC Whitepapers, 1998

[25] Cheng, R. Glenn, “Test Cases for HMAC-MD5 and HMAC-SHA-1”, RFC 2202, September 1997

[26] CSL - Computer Systems Laboratory Bulletin, January 1994

[27] CSL - Computer Systems Laboratory Bulletin, March 1994

[28] Deering, “ICMP Router Discovery Messages”, RFC 1256, September 1991.

[29] Distributed Computing Group, “Security in a Public World: A Survey”, Stanford University, February 1996.

[30] Dobbertin, A. Bosselaers, B. Preneel, ”RIPEMD-160, a strengthened version of RIPEMD” Fast Software Encryption, Springer-Verlag, 1996

[31] Droms R., “Dynamic Host Configuration Protocol”, RFC 2131, March 1997.

[32] Droms R., W. Arbaugh, “Authentication for DHCP Messages”, Network Working Group, Internet draft (work in progress), draft-ietf-dhc-authentication-12.txt, October 1999.

[33] Eastlake, S. Crocker, J. Schiller, “Randomness Recommendations for Security”, RFC 1750, December 1994.

[34] Egevang, P. Francis, “The IP Network Address Translator (NAT)”, RFC 1631, May 1994.

[35] European Telecommunications Standards Institute - ETSI, “Broadband Radio Access Networks (BRAN); HIgh PErformance Radio Local Area Network (HIPERLAN) Type 2”, RTR/BRAN-0022001

[36] Farrell S. et al, “AAA Authorization Requirements”, Internet draft (work in progress), draft-ietf-aaa-authorization-reqs-01.txt, October 1999.

[37] Farrell S. et al, “AAA Authorization Requirements”, Internet draft (work in progress), draft-ietf-aaa-authorization-reqs-01.txt, October 1999.

[38] Forsberg et al, “Distributing Mobility Agents Hierarchically under Frequent Location Updates”, MoMuc’99, November 1999

[39] Fraser, “Site Security Handbook”, RFC 2196, September 1997

[40] Glass S., S. Jacobs, C. Perkins, “Mobile IP Authentication, Authorization, and Accounting Requirements”, Mobile IP Working Group, Internet draft (work in progress), draft-ietf-mobileip-aaa-reqs-01.txt, January 2000.

[41] Gupta V., “Flexible Authentication for DHCP Messages”, Internet draft (work in progress), draft-gupta-dhcp-auth-01.txt, October 1999

[42] Gupta V., “Inline Security Parameter Payload for Mobile IP”, Mobile IP Working Group, Internet draft (expired), draft-gupta-mobileip-inline-secparams-00.txt, June 1999.

[43] Gustafsson et al, “Requirements on Mobile IP from a Cellular Perspective”, Mobile IP Working Group, Internet draft (work in progress), draft-ietf-mobileip-cellular-requirements-02.txt, June 1999.

[44] Guttman, L. Leong, G. Malkin, “Users' Security Handbook”, RFC 2504, February 1999

[45] Haller N. et al, “A One-Time Password System”, RFC 2289, February 1998.

[46] Hamzeh et al, “Point-to-Point Tunneling Protocol (PPTP)”, RFC 2637, July 1999

[47] Harkins, D. Carrel, “The Internet Key Exchange (IKE)”, RFC 2409, November 1998.

[48] Hiller et al, “3G Wireless Data Provider Architecture Using Mobile IP and AAA”, Internet draft (work in progress), draft-hiller-3gwireless-00.txt, March 1999.

[49] Housley R. et al, “Internet X.509 Public Key Infrastructure Certificate and CRL Profile”, RFC 2459, January 1999

[50] IEEE Std 802.11-1997, “Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications”, ISBN 1-55937-935-9, New York, Institute of Electrical and Electronics Engineers, Inc.

[51] Irwin S., T. Bakey, “Tips for staying alive in a competitive industry”, Insurance Software Review, March 1989

[52] Jacobs S., “Mobile IP Public Key Based Authentication”, Internet draft (work in progress), draft-jacobs-mobileip-pki-auth-02.txt, March 1999.

[53] Johnsson M., “HiperLAN/2 – The Broadband Radio Transmission Technology Operating in the 5 GHz Frequency Band”, HiperLAN/2 Global Forum White Paper, 1999

[54] Johnsson M., “Simple Mobile IP”, Internet Engineering Task Force, Internet draft (work in progress), draft-ietf-mobileip-simple-01.txt, March 1999.

[55] Jokela P., “Wireless Internet Access Using Anonymous Access Methods”, Mobile Multimedia Communications (MoMuc`99) conference, November 1999.

[56] Kaliski, J. Staddon, “PKCS #1: RSA Cryptography Specifications”, RFC 2437, RSA Laboratories, October 1998

[57] Karn, P. Metzger, W. Simpson, “The ESP Triple DES Transform”, RFC 1851, September 1995

[58] Kent S., R. Atkinson, “IP Authentication Header”, RFC 2402, November 1998.

[59] Kent S., R. Atkinson, “IP Encapsulating Security Payload (ESP)”, RFC 2406, November 1998.

[60] Kent S., R. Atkinson, “Security Architecture for the Internet Protocol”, RFC 2401, November 1998.

[61] Keromytis, “The Use of HMAC-RIPEMD-160-96 within ESP and AH”, Internet draft (work in progress), draft-ietf-IPSec-auth-hmac-ripemd-160-96-04.txt”, September 1999

[62] KPMG UK – Information Risk Management, Information Security Survey, 1998

[63] KPMG UK – National Computer Security Survey, 1996

[64] Krawczyk, "SKEME: A Versatile Secure Key Exchange Mechanism for Internet", IEEE Proceedings of the 1996 Symposium on Network and Distributed Systems Security, 1996

[65] Krawczyk, M. Bellare, R. Canetti, “HMAC: Keyed-Hashing for Message Authentication”, RFC 2104, February 1997

[66] Laprie J.C., “Dependability: Basic Concepts and Terminology”, Springer-Verlang, October 1991

[67] Li Gong, N. Shacham, “Multicast security and its extension to a mobile environment”, Wireless Networks 1, 1995.

[68] Linn et al, “Privacy Enhanced Mail”, RFCs 1421-1424, February 1993

[69] Litvin M., R. Shamir, T. Zegman, “A Hybrid Authentication Mode for IKE”, Internet draft (work in progress), draft-ietf-IPSec-isakmp-hybrid-auth-03.txt, December 1999.

[70] Madson, R. Glenn, “The Use of HMAC-SHA-1-96 within ESP and AH”, RFC 2404, November 1998

[71] MARS Information Site: http://www.research.ibm.com/security/mars.html

[72] Maughan et al, “Internet Security Association and Key Management Protocol (ISAKMP)”, RFC 2408, November 1998.

[73] McAuley et al, “Dynamic Registration and Configuration Protocol (DRCP)”, Internet draft (work in progress), draft-itsumo-drcp-00.txt, October 1999

[74] McCann P., T. Hiller, “IP Transform Policy Distribution using Mobile IP/DIAMETER”, Internet draft (work in progress), draft-mccann-transform-00.txt, June 1999.

[75] Metzger P., W. Simpson, “IP Authentication using Keyed SHA”, RFC 1852, September 1995

[76] Mogul J.C., J. Postel, “Internet Standard Subnetting Procedure”, RFC 950, August 1985.

[77] Montenegro, “Reverse Tunneling for Mobile IP”, RFC 2344, May 1998.

[78] Neumann P., “Security Risks in Key Recovery”, Computer Science Laboratory (CSL), August 1997.

[79] NIST bulletin, “NIST Announces the AES Finalist Candidates for Round 2”, Federal Information Processing Standard (work in progress), August 1999.

[80] Olovsson T., “A Structured Approach to Computer Security”, Chalmers University of Technology, Gothenburg, 1992

[81] Opplinger R., “Authentication Systems for Secure Networks”, Artech House Publishers, London, 1996.

[82] Opplinger R., “Internet and Intranet Security”, Artech House, London, 1997

[83] Orman, “The OAKLEY Key Determination Protocol”, RFC 2412, November 1998

[84] Patel et al, “DHCP Configuration of IPSec Tunnel Mode”, Internet draft (work in progress), draft-ietf-IPSec-dhcp-04.txt, December 1999.

[85] Patil, R. Narayanan, E. Qaddoura, “Security Requirements /Implementation Guidelines for Mobile IP using IP Security”, Internet Engineering Task Force, Internet draft (work in progress), draft-bpatil-mobileip-sec-guide-00.txt, June 1999.

[86] Patrick M., “DHCP Relay Agent Information Option”, DHC Working Group, Internet draft (work in progress), draft-ietf-dhc-agent-options-09.txt, March 2000.

[87] Pereira R., S. Anand, B. Patel, “The ISAKMP Configuration Method”, Internet draft (work in progress), draft-ietf-IPSec-isakmp-mode-cfg-05.txt, August 1999.

[88] Pereira R., S. Beaulieu, “Extended Authentication within ISAKMP/Oakley”, Internet draft (work in progress), draft-ietf-IPSec-isakmp-xauth-06.txt, December 1999.

[89] Perkins C., ”IP Mobility Support”, RFC 2002, October 1996.

[90] Perkins C., P. Calhoun, “AAA Registration Keys for Mobile IP”, Internet draft (work in progress), draft-ietf-mobileip-aaa-key-01.txt, January 2000.

[91] Perkins C., R. Calhoun, “Mobile IP Challenge/Response Extensions”, Mobile IP Working Group, Internet draft (work in progress), draft-ietf-mobileip-challenge-09.txt, February 2000.

[92] Plummer C., “An Ethernet Address Resolution Protocol”, RFC 826, November 1982.

[93] Postel J., “User Datagram Protocol”, RFC 768, August 1980.

[94] Postel, “Internet Control Message Protocol”, RFC 792, September 1981.

[95] Postel, “Multi-LAN Address Resolution”, RFC 925, October 1984.

[96] Preneel, A. Bosselaers, H. Dobbertin, “The cryptographic hash function RIPEMD-160”, CryptoBytes Vol. 3 No. 2, 1997

[97] Ramjee R., T. La Porta, S. Thuel, K. Varadhan, “IP micro-support using HAWAII”, Internet draft (expired), draft-ramjee-micro-mobility-hawaii-00.txt, February 1999.

[98] RC6 Information Site: http://www.rsa.com/rsalabs/aes

[99] Rigney et al, “Remote Authentication Dial in User Service (RADIUS)”, RFC 2058, January 1997.

[100] Rijndael Information Site: http://www.esat.kuleuven.ac.be/~rijmen/rijndael/

[101] Rivest R., “The MD4 Message-Digest Algorithm”, RFC 1320, April 1992.

[102] Rivest R., “The MD5 Message-Digest Algorithm”, RFC 1321, April 1992.

[103] Rivest R.," Chaffing and Winnowing; Confidentiality without Encryption“, MIT Lab for Computer Science, April 1998.

[104] Sanchez L., N. Condell, "Security Policy System", Internet draft (expired), draft-ietf-IPSec-sps-00.txt, November 1998

[105] Schiller, "Cryptographic Algorithms for the IETF", Internet draft (work in progress), draft-ietf-saag-aes-ciph-00.txt, August 1999

[106] Schneier, P. Gutmann, "Description of the Blowfish Cipher", Internet draft (work in progress), draft-schneier-blowfish-00.txt, August 1999

[107] Serpent Information Site: http://www.cl.cam.ac.uk/~rja14/serpent.html

[108] Simpson W., "PPP Challenge Handshake Authentication Protocol (CHAP)", RFC 1994, August 1996.

[109] Simpson, "The Point-to-Point Protocol (PPP)", RFC 1661, July 1994

[110] Solomon J., S. Glass, "Mobile-IPv4 Configuration Option for PPP IPCP", RFC 2290, February 1998.

[111] Stevens R., "TCP/IP illustrated, Volume 1: The protocols", Addison-Wesley, Reading, Massachusetts, 1994.

[112] Sufatrio, Kwok-Yan Lam, "Scalable Authentication Framework for Mobile IP", Mobile IP Working Group, Internet draft (work in progress), draft-rio-mobileip-safe-mip-00.txt, October 1999.

[113] Thayer R., N. Doraswamy, R. Glenn, "IP Security Document Roadmap", RFC 2411, November 1998.

[114] Townsley et al, "Layer Two Tunneling Protocol - L2TP", RFC 2661, August 1999

[115] Twofish Information Site: http://www.counterpane.com/twofish.html

[116] U.S. Department Of Commerce, "Data Encryption Standard (DES)", Federal Information Processing Standard (FIPS) Publication 46-1, January 1988

[117] U.S. Department Of Commerce, "Data Encryption Standard (DES); specifies the use of Triple DES (TDES)", Federal Information Processing Standard (FIPS) Publication 46-3, November 1999

[118] U.S. Department Of Commerce, "Digital Signature Standard (DSS)", Federal Information Processing Standard (FIPS) Publication 186-1, December 1998

[119] U.S. Department Of Commerce, "Secure Hash Standard", Federal Information Processing Standard (FIPS) Publication 180, May 1993

[120] U.S. Department Of Commerce, "Secure Hash Standard", Federal Information Processing Standard (FIPS) Publication 180-1, April 1995

[121] US Department of Defence, ”Trusted Network Interpretation of the Trusted Computer System Evaluation Criteria”, July 1987

[122] Valencia, T. Kolar, “Cisco Layer Two Forwarding (Protocol) - L2F”, RFC 2341, May 1998

[123] Valko, A. Campbell, J. Gomez, “Cellular IP”, Internet draft (work in progress), draft-valko-cellularip-01.txt, October 1999.

[124] Wing P., B. O’Higgins, “Using Public-Key Infrastructures for Security and Risk Management”, IEEE Communications Magazine, Vol. 37 No. 9, September 1999

[125] Xu et al, “Mobile IP Based Micro Mobility Management Protocol in the Third Generation Wireless Network”, Mobile IP Working Group, Internet draft (work in progress), draft-ietf-mobileip-3gwireless-ext-02.txt, January 2000.

[126] Yi Cheng, “Security Solutions in Wireless LAN Systems”, Ericsson White Paper, January 1999.

[127] Yingchun Xu, K. Peirce, E. Campbell, “Mechanism to Support CHAP Mobile Node Authentication for RADIUS/DIAMETER Hybrid AAA Networks”, Internet draft (work in progress), draft-peirce-radius-challenge-00.txt, June 1999.

21 Glossary

Accounting	the process of measuring resource usage of a pricipal
(IP) Roaming	the ability of a customer to use any one of multiple networks or Internet service providers (ISPs), while maintaining a formal, customer-vendor relationship with only one
Address spoofing	a type of attack in which the attacker steals a legitimate network (e.g. IP) address of a system
Birthday Attack	an attack in which it is necessary to obtain two identical values from a large population
Brute-force attack	every possible key is tried until the recovered plaintext is meaningful
Certification Authority	a trusted agent that issues digital certificates to entities.
Chosen-plaintext attack	an attack in which the enemy has access to ciphertext and associated plaintext for several messages
CRL version 2	a format and content for certificate revocation lists. Sites need to have a way to interchange revocation information
Data integrity	the reasonable assurance that data has not been changed while transmitted from a sender to its intended recipient
Denial of service	an attack where an attacker floods the server with bogus requests, or tampers with legitimate requests
Digital certificate	a structure for binding a principal's identity to its public key. A certification authority (CA) issues and digitally signs it
Digital signature	a method for verifying that a message originated from an entity and that it has not changed en route
Domain of Interpretation	collects IANA’s assigned parameters for a particular protocol
DSA	this algorithm uses a private key to sign a message and a public key to verify the signature
IP (Session) Mobility	applications and TCP sessions are not affected by the fact that the terminal is moving within and between subnets, except for some packet loss which may result out from a handover between radio base stations

IPSec a protocol suit that provides encryption and uthentication for network sessions using the Internet Protocol (IP)

Man-in-the-middle attack an attack in which an attacker inserts itself between two parties and pretends to be one of the parties

MD5 a message digest (MD) algorithm that digests a message of arbitrary size to 128 bits

Message digest the fixed-length output of a one-way function

Non-repudiation the reasonable assurance that a principal can not deny being the sender of a message

PKCS family a protocol family that defines the format and behavior for public-key exchange and distribution architectures. It allows different vendors’ implementations to request and move certificates in a way that all understand

PKINIT emerging standard for using public keys to log on to networks that uses the Kerberos authentication protocol

PKIX an emerging PKI standard that many major vendors and enterprises are adopting in place of the PKCS standards

Replay attack an attack in which an attacker captures a message and at a later time communicates that message to an entity

RSA a public-key cryptosystem invented by Ron Rivest, Adi Shamir, and Leonard Adleman

Session key a temporary symmetric key that is only valid for a short period. It is generally exchanged through a public-key protocol

SHA a message digest (MD) algorithm that digests a message of arbitrary size to 160 bits

SSL version 3 is the best known and most widely used security protocol at ISO layer 3 on the Internet, but it's subject to export controls

X.509 version 3 a format for digital certificates (without a standard for certificate formats, there's no way to exchange certificates between vendors)

22 Abbreviations and Acronyms

AAA	Authentication, Authorization and Accounting
ACH	Access feedback CHannel
ACK	Acknowledgment packet
ACL	Access Control List
AES	Advanced Encryption Standard
AH	Authentication Header
AP	Access Point
ARP	Address Resolution Protocol
ARQ	Automatic Repeat Request
ASCH	ASsociation control Channel
ATM	Asynchronous Transfer Mode
AVP	Attribute-Value-Pairs
BCH	Broadcast Channel
BSS	Basic Service Set
CA	Certification Authority
CBC	Cipher Block Chaining
CFB	Cipher Feedback
CHAP	Challenge Handshake Authentication Protocol
CL	Convergence Layer
CRL	Certificate Revocation Lists
CSMA/CA	Carrier-Sense Multiple Access, Collision Avoidance
CTS	Clear-To-Send
DBS	DataBase Server
DCCH	Dedicated Control CHannel
DES	Data Encryption Standard
DF	Diffie-Hellman key exchange
DFS	Dynamic Frequency Selection
DHCP	Dynamic Host Configuration Protocol
DL	DownLink
DLC	Data Link Control

DLCC	DLC Connection
DNS	Domain Name System
DOI	Domain of Interpretation
DRCP	Dynamic Registration and Configuration Protocol
DS	Distribution System
DSA	Digital Signature Algorithm
DSS	Digital Signature Standard
DSSS	Direct Sequence Spread Spectrum
EAP	Extensible Authentication Protocol
EC	Error Control
ECB	Electronic Codebook
ECC	Elliptic Curve Cryptosystem
ESP	Encapsulating Security Payload
ESS	Extended Service Set
ETSI	European Telecommunications Standards Instutute
FA	Foreign Agent
FCH	Frame CHannel
FHSS	Frequency Hopping Spread Spectrum
GRE	Generic Routing Encapsulation
HA	Home Agent
HiperLAN	High Performance Radio Local Area Network
HTTP	HyperText Transfer Protocol
IANA	Internet Assigned Numbers Authority
ICMP	Internet Control Message Protocol
IDEA	International Data Encryption Algorithm
IEEE	Institute of Electrical and Electronics Engineers
IETF	Internet Engineering Task Force
IKE	Internet Key Exchange protocol
IMT	International Mobile Telecommunications
IPSec	IP Security protocols
IPX	Internet Packet eXchange
IR	Infrared
ISAKMP	Internet Security Association and Key Management Protocol
ISM	Industrial, Scientific, and Medical

ISO	International Organization for Standardization
ISP	Internet Service Provider
ITU	International Telecommunication Union
KDC	Key Distribution Center
KEK	Key Encryption Key
L2F	Layer 2 Forwarding
L2TP	Layer 2 Tunneling Protocol
LAN	Local Area Network
LCCH	Link Control Channel
LCH	Long transport Channel
LCP	(PPP) Link Control Phase
LDAP	Lightweight Directory Access Protocol
LLAN	Local LAN
LLC	Logical Link Control
MAC	Message Authentication Code
MAC-ID	MAC IDentifier
MD	Message Digest
MER	Mobility Enabled Router
MIB	Management Information Base
MIP	Mobile IP
MLAN	Mobile LAN
MN	Mobile Node
MT	Mobile Terminal
NAI	Network Access Identifier
NAS	Network Access Servers
NAT	Network Address Translator
NAV	Network Allocation Vector
NCP	Network Control Protocols
NetBEUI	Network Basic input/output Extended User Interface
NSP	Network Security Policy
NTP	Network Time Protocol
OAM	Operation and Maintenance
OFB	Output Feedback
OFDM	Orthogonal Frequency Digital Multiplexing

OTP	One-Time Password
PAN	Personal Area Network
PAP	Password Authentication Protocol
PEM	Privacy Enhanced Mail
PFS	Perfect Forward Secrecy
PGP	Pretty Good Privacy
PHY	Physical Layer
PKI	Public Key Infrastructure
PPP	Point-to-Point Protocol
PPTP	Point-to-Point Tunneling Protocol
PRNG	Pseudo Random Number Generator
PVC	Permanent Virtual Circuits
QoS	Quality of Service
RADIUS	Remote Authentication Dial-In User Service
RC	Rivest Cipher
RCH	Random CHannel
RF	Radio Frequency
RFC	Request for Comments
RLC	Radio Link Control
RR	Resource Request
RRC	Radio Resource Control
RSA	Rivest, Shamir, and Adelman
RTS	Ready-To-Send
SA	Security Association
SAD	Security Association Database
SAP	Service Access Policy
SBCH	Slow Broadcast CHannel
SCH	Short transport Channel
SDU	Service Data Unit
SEA	Spokesman Election Algorithm
SHA	Secure Hash Algorithm
SKIP	Simple Key-management for Internet Protocols
SMIP	Simple Mobile IP
SNMP	Simple Network Management Protocol

SNR	Signal to Noise Ratio
SPD	Security Policy Database
SPI	Security Parameters Index
SR	Selective Repeat
SSK	Session Secret Key
SSL	Secure Sockets Layer
TCP	Transmission Control Protocol
TDD	Time-Division Duplex
TDMA	Time-Division Multiple Access
UDCH	User Data Channel
UDP	User Datagram Protocol
UMTS	Universal Mobile Telecommunications Service
UP	UpLink
U-PDU	User Protocol Data Units
VACM	View-Based Access Control Model
WCDMA	Wideband Code Division Multiple Access
WEP	Wire Equivalent Privacy
WG	(IETF) Working Group
WLAN	Wireless Local Area Network
WLG	WLAN Guards
VoIP	Voice-over-IP
VPN	Virtual Private Network
X-AUTH	Extended AUTHentication ISAKMP method
X-HA	Extended Home Agent