Alulírott Szalay Máté, a Budapesti Műszaki és Gazdaságtudományi Egyetem hallgatója kijelentem, hogy ezt a diplomatervet meg nem engedett segítség nélkül, saját magam készítettem, és a diplomatervben csak a megadott forrásokat használtam fel. Minden olyan részt, melyet szó szerint, vagy azonos értelemben de átfogalmazva más forrásból átvettem, egyértelműen, a forrás megadásával megjelöltem.

Szalay Máté

Safety and Security Considerations of IP Micro Mobility

- Examine the basic safety and security issues of the IP micro mobility environment.
- Safety: Analyse and compare various network topologies from a reliability point of view. Examine how network errors are handled.
- Security: Examine possible attacks and design an authentication method for micro mobility environments.

Contents

1 Introduction

1.1 Why IP?

IP is the protocol of the Internet. Now it is only the protocol of the Internet, but it might become more.

The Internet, telecommunications, and advertising seem to converge nowadays. The resulted so-called ‘infocom’ network of this convergence will probably be based on IP. There will be an IP backbone, but IP can also be used in the access networks. The goal is to take the IP as far as to the terminal equipments.

The terminal equipments of this network will not necessarily be PCs, but for example mobile phones or personal digital assistants (PDA).

IP is much more widespread than anyone would have expected it at the time of its birth, but it will be even more commonly used in the near future.

1.2 Why IPv6?

Even without the new ‘infocom’ network a new version of the protocol would be needed. Considering the Internet, we are running out of IP addresses, and the new applications of the IP will raise new demands.

As the circumstances and requirements were different at the time of the design of IPv4, the new version of the protocol, IPv6 has several improvements and additions.

1.3 Why Mobility?

One of the requirements that are hard and difficult to fulfil using IPv4 is the support of mobility. However, Mobile IP will be an integral part of IPv6. Powerful mobility support is an essential part of an IP based telecommunication network.

1.4 What is Micro Mobility?

Infocom equipments have to be mobile, but not in a way that is provided by Mobile IP, the standard IP mobility solution. Mobile IP [1] requires a lot of communication between the mobile node and its home agent, it provides a large-scale but slow mobility. Below this Mobile IP mobility a small-scale but fast mobility protocol is needed. This small-scale mobility is often called micro mobility, referring to Mobile IP as macro mobility.

1.5 Why Security?

The more wide-spread use of the Internet Protocol raises a strong need of security solutions. As the provided services may include for example trading, banking, shopping, ticketing, security may be at least as important as mobility.

1.6 About the Tasks

In my thesis the “micro mobility network” can be a LAN connected to the Internet or an access network of the ‘infocom’ network. Thus the phrases ‘connection to the Internet’ and ‘connection to the IP backbone’ are synonyms here.

The three tasks are the following:

1. Examine the basic safety and security issues and problems of the IP micro mobility environment.
2. Analyse and compare various network topologies from a reliability (safety) point of view. Examine how network errors are handled.
3. Security: Examine possible attacks and design an authentication method for micro mobility environments.

Task 1 is a summary of tasks 2 and 3. In task 2 the most commonly used topologies are examined. Task 3 can be divided into two separate tasks. One is to examine the various micro mobility specific attack situations. The other one is to design an authentication method.

The examination of the attacks can not be exhaustive of course, and a security architecture that I can not break should not be considered safe. I just try to give an overview of the security of micro mobility networks, as it is an extremely important point.

When designing the authentication protocol I will not build a new protocol up to the last bit, but rather try to show how a general security principle (authentication) can be used in micro mobility environments, what are the special requirements.

1.7 Structure of the Thesis

Chapter 2 gives the foundations on which this thesis is built. The first section is about IPv6, then it gives a short explanation of IPsec and MobileIP. The mobile IP section also touches micro mobility.

Chapter 3 is about reliability. It is the answer to task 2. After examining several network topologies, in Section 3.6. I recommend a new network topology that suits the micro mobility requirement very well and has very good safety qualities.

Chapter 4 is about the security of micro mobility networks. It is the answer to first part of task 3. It gives an overview of micro mobility security through the examination of various attack situations.

Chapter 5 is the answer to the second part of task 3. It describes how authentication should be solved in micro mobility environments.

2 Basics of IP

2.1 IPv6

2.1.1 Introduction

The success of the Internet shows, that IPv4 is a very well designed protocol without any critical errors, but a new protocol is needed. It is mainly caused by a shortage of IP addresses. At the time the IPv4 was designed, in 1978, the 32 bits seemed to be enough. Nobody thought that the Internet would connect as many networks and computers as it connects today. And it is still growing very quickly. It could be enough just to increase the address space, and leave everything else unchanged. The designers today are not any smarter, but they have an advantage: they could examine the Internet in practice for several years. And it taught some lessons.

Now, not only the address space has to be extended, but because of the wider application areas some enhancements are also needed.

The IPv6 is a new version of the Internet Protocol, it is not a new protocol. It has improvements, but based on IPv4.

In this chapter the word “Internet” will be used, and not the phrase “IP backbone”.

2.1.2 Addressing

Address Space

There has been a lot of debate about the length of the addresses in the new version of the protocol. Some said that with choosing any fixed length we will have to face the same problems again, so they advised a variable length address as the only solution to the problem. It is true that using fixed address length sets up a limit, but on the other hand fixed addresses are much easier to handle, and that means faster processing (e.g. routing), and that is a crucial question. Some of the proposals recommended a 128-bit fixed length, some others said 64-bit length was enough.

Actually, if all the addresses were used, 64 bits would be more than enough, but experience shows, that this is not the case. So after a log debate, the final decision came, the 128-bit fixed length addresses were chosen.

Unicasting, Multicasting, Anycasting

IPv4 addressing had unicast and multicast addresses. Anycasting is a new feature of IPv6. Different routing is required for different applications.

Unicasting means that the packet should be routed to the specified host (to be more exact, to the specified interface of the specified host).

Multicasting makes it possible to route a packet to a group of hosts (to be more exact, to a group of interfaces). This group is called a multicast group. The multicast packet should be routed to all of the interfaces in the group. The host should be able to join or leave any of the multicast groups any time. IPv4 allowed multicasting using class D addresses.

Anycasting is similar to multicasting. There are anycast groups, which the hosts should be able to join or leave any time. The difference is that anycast packets are not routed to all of the members, but to one (the nearest) of them. There were some attempts to enable anycasting in IPv4, and the IPv6 version is still an object of research.

2.1.3 The Header

The Basic Header

The IPv6 header has 6 fixed length header fields with a total length of 64 bits followed by two 128-bit addresses. Thus, the length of the IPv6 header without extensions is 40 bytes.

Actually, the basic IPv6 header is much simpler than the IPv4 header. The options have been moved from the IP header to the extension headers. Table 2.1. and Table 2.2. show the details of the IPv6 and the IPv4 header.

Table 2.1. The IPv6 header

bits	field name
4	Version
8	Class
20	Flow Label
16	Payload Length
8	Next Header
8	Hop Limit
128	Source Address
128	Destination Address

Table 2.2. The IPv4 header:

bits	field name
4	Version
4	IHL
8	Type of Service
16	Length
16	Identification
3	Flags
13	Fragment Offset
8	Time-To-Live
8	Protocol
16	Header Checksum
32	Source Address
32	Destination Address
?	Options
?	Padding

The *version* field of the IPv6 header is 6 of course (0110 binary). The *class* and *flow label* fields are defined to enable QoS or to differentiate the traffic, and are objects of research.

The IPv6 header contains a *payload length* field instead of IPv4's *total length* field. The *next header* field contains the type of the next header in the daisy-chain. If its value is TCP, then there are no extension headers.

The *hop limit* field is instead of the *time-to-live* field of IPv4. The value of the TTL was a time interval in milliseconds. If the packet cannot reach its destination before that time expires, it should be discarded. In practice, most of the routers decremented this value by 1 at each relay. It was already used as a hop count. The *hop-count* field is meant to be decreased by 1 at each relay. This field is encoded on 8 bits, so the maximum is 255. This means that the maximum distance between two hosts on the Internet should not be more than this.

The *header checksum* field is omitted. It makes routing faster, because much less processing is needed. The disadvantage is obvious. But the field was suppressed, because the risk was thought to be minimal. It is minimal, because most of the lower layer protocols have a checksum, and even if the packet is delivered to the wrong destination or corrupted some other way, it will probably be discarded at higher levels.

Extension Headers

The IPv4 header had a variable length field, the options field. This field has been suppressed, and a daisy-chain of extension headers has been introduced instead. There can be several types of extension headers, and only a few types are defined up to now. The extension headers are inserted between the IP header and the TCP header, and each of the headers contains information about the type of the next header. Figure 2.1. shows the structure of a packet with extension headers.

IP header	AH header	TCP header	payload data
next=AH	next=TCP		

Figure 2.1. An IPv6 packet example

2.1.4 Autoconfiguration

Autoconfiguration means that a host can automatically determine all the parameters (e.g. own IP address) it needs to connect to the Internet. It is extremely important when one has to manage hundreds of computers, so its importance grew as the Internet grew. It is unpractical (if not impossible) to set up the parameters manually for all the computers one-by-one. IPv4 had the DHCP (Dynamic Host Configuration Protocol) for this purpose. DHCP has an IPv6 version, but IPv6 offers more ways of autoconfiguration.

2.1.5 Security

History

IPv4, when designed, contained no security features. As the Internet grew, there was an increasing demand for security. There were some application level solutions, but security features had to be added to the IP level. It was added, but unfortunately it had to be added after the design of the protocol, and for example IPsec contained some security weaknesses (or even) holes, which were fixed later. (See [Sza00]).

Now, when designing and standardizing IPv6, it is clear, that security is a crucial question, and IPv6 will be designed with plenty of security features from the beginning.

Security Payloads

The IPv6 (as the IPv4) contains specifications of two types of security payloads: *Authentication Header (AH)* and *Encapsulating Security Payload (ESP)*. At IPv6, they both use extension headers, and some cryptographic protocols.

The *Authentication Header* contains some authentication data, so the receiver can be sure that the origin of the packet is really the host mentioned in the *Source Address* field, and that the packet arrived untouched. It provides absolutely no confidentiality, anybody eavesdropping somewhere on the Internet can "read" the payload. There are several explicit algorithms for computing the authentication data, *Message Digest 5 (MD5)* is suggested.

The *Encapsulating Security Payload* should be used to provide confidentiality. It means that a third party somewhere on the Internet cannot listen to the communication. Actually ESP also adds some authentication data to the packet, but its authentication is much weaker than that of AH, because it does not authenticate the entire packet. The authentication feature of the ESP can be treated as a part of the encryption algorithm. When authentication is needed, the AH should be used.

When using ESP, the entire packet is encrypted except the IPv6 header, the extension headers inserted after the ESP header, and the authentication data. Encrypting the IPv6 header would make routing impossible. The strength of ESP depends on the strength of the cryptographic algorithm used. Various algorithms can be used, DES-CBC is recommended.

Security Associations

AH and ESP are end-to-end security services. The two parties have to agree on a set of parameters (e.g. the algorithms, the parameters of the algorithms, the keys). This set of parameters that belongs to one end-to-end connection is called a *Security Association*.

Key Management

There are several cryptographic algorithms that are believed to be strong. It is not easy to design a strong protocol using these strong algorithms, but there are a lot of already published algorithms that are treated strong by professionals. Actually they are so strong that offering $1.000.000 for breaking them was not enough...

If it is relatively easy to get the key somehow, then the strength of the algorithm or the protocol does not really matter.

Key management is the toughest question of today's cryptography. Today, when really strong security is needed, manual key distribution is used. The Internet Security Association And Key Management Protocol (ISAKMP) defines some generic, algorithm-independent protocols, and there are some well-known key exchange protocols, such as Diffie-Hellman or COMSET ([Sch96] Chapter 22), which can be used with ISAKMP.

2.2 IPsec

2.2.1 Introduction

IPsec is designed to add security services to the Internet architecture. As security is added at the IP level, the IP level and all upper levels are protected. IPsec has several security services:

- authentication
- integrity
- confidentiality
- ...

With IPsec security can be added to IPv4 networks, but its main scope is IPv6. Actually, IPsec will be a part of the IPv6 protocol.

2.2.2 The parts of IPsec

IPsec is made up of four different parts:

1. traffic security protocols
2. Security Associations
3. SA and key management
4. various algorithms

There are two traffic security protocols: Authentication Header (AH) and Encapsulating Security Payload (ESP). The objective of AH is to provide source authentication, so the receiver of a packet can be sure about the sender, and that the packet arrived untouched, it was not altered by an active attacker. ESP provides confidentiality to prevent third parties from listening to the communication. Two extension headers are defined in IPv6: one for AH and one for ESP. AH and ESP can be used separately or simultaneously.

A Security Association contains the security parameters for the communication between two specific hosts. The SA describes which algorithms to use with which parameters with which keys when sending packets to a host, and how to verify the authentication or how to decrypt, when receiving packets from a host.

SA and key management protocols can be manual or automatic.

Most of the protocols are algorithm independent. So IPsec also describes exactly how some cryptographic algorithms should be used for authentication and encryption.

2.2.3 Security Associations

When using authentication or encryption, the sender and receiver must agree on the specific algorithm to be used, the parameters that the algorithm should be used with, and the keys. A Security Association contains these data for a unidirectional link. For two-way communication, two SAs are needed, one for each direction.

The negotiation of the Security Association is usually a part of the key exchange procedure.

A Security Association can be uniquely identified by the following triplet:

- Destination IP address
- Security Parameter Index (SPI)
- Security protocol (AH or ESP)

When a packet is received, it can only be verified and decrypted if the receiver knows which SA it belongs to. The IP address is in the basic IPv6 header, the security protocol and the SPI are in the extension header.

There are two types of Security Associations:

- transport mode SA
- tunnel mode SA

The transport mode SA is an SA for the communication between two hosts. The sender encrypts the packet or adds the authentication information, and the receiver decrypts or checks. Note that the IP header cannot be encrypted, because it would make routing impossible.

The tunnel mode SA is an SA applied to an IP tunnel. If either of the communicating hosts is a security gateway, tunnel mode SA has to be used. When using a tunnel, there is an outer IP header, and an inner one. The inner IP header can be encrypted, and routing can be based on the outer one.

The IPsec standard defines two databases: the *Security Association Database* (SAD) and the *Security Policy Database* (SPD). The *Security Association Database* contains the parameters of the Security Associations. Each SA has one entry in the SAD. The *Security Policy Database* is consulted when processing inbound or outbound traffic. Then a decision is made for each packet whether

IPsec has to be applied on it, or it can pass untouched. Basically the SAD tells what to do, the SPD tells when to do it.

2.2.4 Traffic security protocols

Authentication Header

Authentication Header is an extension header of IPv6. Figure 2.2. shows the structure of an authenticated packet.

IP header	Authentication Header	TCP header + packet data

Figure 2.2. Structure of authenticated packet

The AH itself is very simple, because it is independent of the used authentication protocol. The AH is made up of a 96-bit header and the authentication data, that is encoded as a variable number of 32-bit words. Table 2.3. shows the structure of the AH.

Table 2.3. Authentication Header

bits	field name
8	Next Header
8	Payload Length
16	Reserved
32	SPI
32	Sequence Number
?	Authentication Data

The *Next Header* field contains the type of the next header in the daisy chain. The *Payload Length* field contains the number of 32 bits following the *SPI* field.

SPI stands for Security Parameters Index. This field contains the Security Parameter Index of the Security Association.

The Sequence Number field is there to make replay attacks impossible. The sender numbers all the packets, so the receiver can detect and discard old packets. Sequence numbers are not that easy to implement, because the Internet does not guarantee ordered delivery. Some window-technique should be used. If the communication is long enough, and more than 2^32 packets are transmitted, the sequence number cycles, and it does not protect against replay attacks any more. Before the sequence number cycles, new keys should be negotiated.

The authentication data is a cryptographical checksum. This is the checksum of the payload data, some fields of the IPv6 header, the extension headers and a shared secret. There are various algorithms for this purpose, the MD5 (Message Digest 5) is specified as a default. All implementations must contain MD5, but the actual algorithm is negotiated as a part of the SA.

A one way hash function can be used as a cryptographic checksum, if a secret is concatenated to the message before the checksum is computed.

Some of the features of a cryptographic checksum:

- It should be algorithmically difficult to compute the checksum of a plaintext without the key (even if the plaintext is very similar to one with a known checksum)
- It should be algorithmically difficult to find a message with a specific checksum
- It should be algorithmically difficult to find two messages with the same checksum

The receiver computes the checksum again, and checks whether it matches the *Authentication Data*. If it does not match, the packet should be discarded.

It is evident that the checksum has to be independent of the transformations of the routing. When computing the checksum, some of the fields have to be excluded. Actually they are not excluded, but the packed is copied, and these fields are filled with zeroes before computing. The hop count is excluded of course, and options where the C (change en route) flag is set.

Encapsulating Security Payload

When using the AH, the payload is not encrypted, eavesdropping is still possible. If confidentiality is required, ESP should be used. Table 2.4. shows the general, algorithm-independent structure of ESP.

Table 2.4. Encapsulating Security Payload

bits	field name
32	SPI
32	Sequence Number
?	Parameters (Encrypted)
?	Encrypted Payload
?	Authentication Data

The *SPI* is the Security Parameter Index of the SA. The *SPI* and *Sequence Number* are the same as that of AH. The variable *Parameters* field contains the parameters of the encryption algorithm used. The *Authentication Data* is at the end of the entire packet. This authentication is used to authenticate the Sequence Number field and the encrypted data, not the IP header itself. Actually it can be considered as a part of the encryption, its purpose is to prevent attacks that truncate and modify encrypted messages. If origin or packet authentication is needed, the AH should be used.

The ESP header should be the innermost one, because all headers that were inserted before the ESP would be encrypted, thus become useless.

In case of transport mode ESP between two security gateways, the inner IP header is also encrypted, so the eavesdropper cannot even find out the true origin and destination of the packet. The packet is routed to its "destination", a security gateway, where it is decrypted, and then routed to its real destination.

Several algorithms can be used with ESP, DES-CBC is specified as a default. As in the case of AH, every implementation should contain DES-CBC, but the actual algorithm is negotiated.

2.2.5 Key management

Here, in the case of IPsec, the key management can be considered as a part of the SA management, as most of the keys are parts of SAs.

The traffic security protocols (AH and ESP) are independent of the SA management.

There are to two types of key/SA management protocols: manual and automated.

Manual SA/key management techniques

Manual management is much simpler than automated. It basically means that someone configures all the hosts manually. Keys and all relevant data are "taken" there by that person. It works well if the system does not change too quickly. Host cannot be reconfigured manually in every five minutes. A problem with manual SA/key management is scalability. If the network contains a few hundred (or thousand) nodes, it is practically impossible to configure them manually. At the same time manual management is "more secure", than automated. When security is really crucial, keys are always distributed manually by a trusted agent.

Automated SA/key management techniques

If IPsec is widely used (and it shall be), the SA/key management has to be standardized and automated. Of course a scalable solution is needed.

In the case of automated key management, there is a sort of key hierarchy. The top-level secret ("master secret") cannot be distributed automatically. When sharing a secret, the parties can generate keys (session keys). If a session key becomes compromised, it should not compromise the master secret. No master secret is needed, if one knows the public key of a *Certificate Authority* (CA) he trusts.

The standard for automated key and SA management is called ISAKMP (Internet Security Associations and Key Management Protocol). (RFC2408).

Authentication

When building an SA, the question of authentication is crucial. Symmetric key authentication requires manual key distribution (at the top level). This is an inscalable method, so public key authentication is needed. In the case of public key authentication, the most important question is the following:

How can A be sure, that K_B (the public key of B that he received via the Internet) is really the public key of B?

He can be sure, if he did not receive it via the Internet, but it was manually configured. But then we face manual key distribution again. The other solution is a hierarchy of certificates and Certificate Authorities (CA). Certificates require an infrastructure for generation, distribution, verification and management. There are several cryptographic algorithms for this purpose that are believed to be strong. Certificates bind public keys to entities. If a specific CA gives me a certificate, it means that it signs a document that says that my public key belongs to me. These certificates have expiration dates of course. If I send this certificate along with my public key, and the other party trusts that CA, he can be sure, that the public key I sent really belongs to me.

ISAKMP

ISAKMP is a collection of procedures, protocols and packet formats for SA and key management. It is a generic framework that is independent of the algorithms used for authentication, encryption and key generation.

SA management mechanisms are currently defined for unicast addresses only. Different security services are required for different environments. Sometimes no security needed at all, sometimes authentication is required, but no confidentiality at all, and sometimes really strong authentication and encryption are required. When using ISAKMP all the security parameters are negotiated by the communicating parties.

While negotiating the SA parameters, the communication has to be secure, otherwise it would be all useless. ISAKMP has protection mechanisms against connection hijacking denial-of-service attacks and man-in-the-middle attack. The most important part (and the most vulnerable in practice) is authentication.

A Security Association is a set of information that has to be shared between the parties of communication. First the entities agree upon a basic security parameter set (there are default algorithms that every implementation must contain). This basic set is used to protect the following steps of the negotiation. When the parties agree upon this basic set, the identities have to be authenticated to avoid impersonating and man-in-the-middle attacks. Some public key authentication algorithm has to be used, but the ISAKMP does not specify the exact protocol or the CA. When the identities are authenticated, key exchange is the last step of the SA negotiation.

2.2.6 Algorithms

SAs, SA management (ISAKMP) and even AH and ESP are independent of the algorithm. Without explicit algorithms, IPsec would be just an empty shell. Some RFCs tell how to use specific algorithms for different purposes.

2.3 Mobile IP

2.3.1 What is MobileIP?

In the Internet architecture the address of a node id built up of the network part and the local part. So whenever the node changes its point of attachment to the Internet (changes its position), it has to change its IP address too. The goal of MobileIP is to overcome this problem, and let the nodes keep their IP addresses even if their change their positions. [Per96]

Mobile IPv4 is an extension of the IPv4 protocol. Only special nodes have to implement it, it is transparent for normal nodes.

Mobile IPv6 is an integral part of the IPv6 protocol itself, it has to be implemented at every node.

2.3.2 Architecture

The mobile node (MN) has its home IP address in its home network (HN). There have to be at least one Home Agent (HA) in the home network. If the MN stays in its home network it does not use mobility at all. When the MN moves to another network, called the foreign network (FN), it obtains a care-of-address there. The care-of-address can be the address of a foreign agent (FA) of the FN or a local IP address in the FN. The later is called “co-located care-of-address”. When the MN has its care-of-address, it has to tell it to its home agent. This is called registration.

A node that the MN communicates with is called a corresponding node (CN). The CN sends IP packets to the MN and receives packets from it.

Packets sent to the home IP address of the MN are intercepted by the HA and forwarded to the care-of-address through an IP tunnel. In the IPv4 case the FA then gets the original packet from the tunnel and passes it to the MN. This is shown in Figure 2.3.

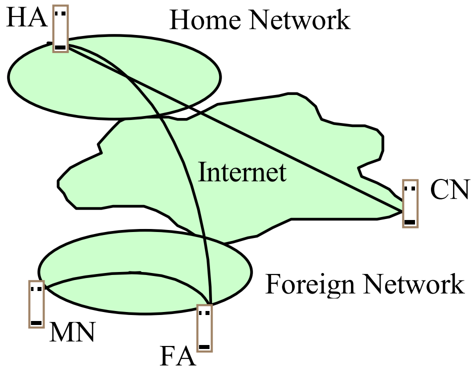

Figure 2.3. IPv4 mobility architecture

In the IPv6 case there is no FA at all. The MN itself is the end of the IP tunnel, it has to get the original IP packets. The IPv6 case without the FA is shown in Figure 2.4.

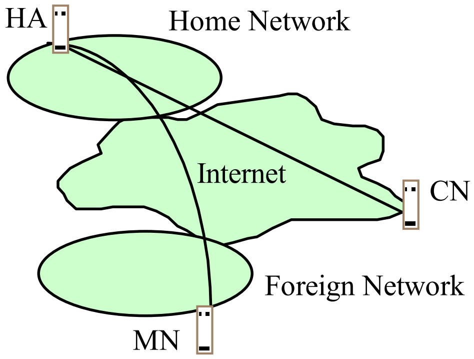

Figure 2.4. IPv6 mobility architecture

When the MN sends packets to the CN there are two possibilities. It can either send it directly to the CN or tunnel it back to its HA. In the later case the HA forwards the packets to the CN.

In the IPv4 case sending packets directly to the CN should be avoided, because routers implementing ingress filtering may drop these packets. When the MN sends a packet directly to the CN the *source address* field of the header contains the home address, and routers may receive the packet from a different interface than where they should have expected it, and may discard the packets considering them false packets. This is why in the IPv4 case packets originating from the MN sent to the CN should be tunnelled back to the HA first.

When MobileIPv6 is used there is a special extension header for the care-of-address, so the IPv6 header contains both addresses, and routers will not discard the packets.

In an IPv6 environment it is also possible for the CN to learn the care-of-address of the MN and send packets directly to that address and not to the home address. This way the triangle routing can be completely avoided.

IPv6 provides a much powerful mobility, as in an IPv6 environment all routers and hosts has to implement features that help mobility, as it is an integral part of the protocol itself.

2.3.3 Hierarchical Mobility, Micro Mobility

The mobility provided by MobileIPv6 is global, but slow. Slow means that it might take a few seconds to register to the home agent. For example, the home network of my laptop is at the Budapest University of Technology and Economics. After flying to Helsinki and plugging into the connector at the hotel it takes a few seconds before I can use the Internet. Here the few seconds are not to much.

But if I have a cellular phone using MobileIP over a wireless interface, than I may have a few handovers in a minute, and my connection will be interrupted for a few seconds at every handover.

The solution is hierarchical mobility shown in Figure 2.5.

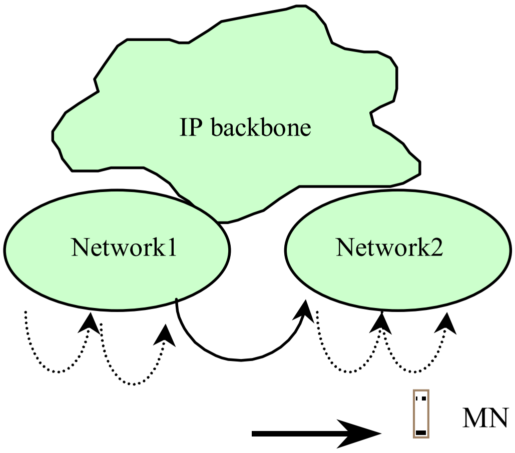

Figure 2.5. Hierarchical mobility

While the MN stays in the same network, handovers are handled locally. This is called micro mobility. Handovers are fast but mobility is limited to a domain.

When the MN moves from one micro mobility domain to another, a macro mobility handover takes place. This handover is much slower, but a global mobility is provided.

More levels of mobility can be introduced under micro mobility. These are usually called pico mobility solutions.

3 Reliability, Network Topology

In this chapter I examine various micro mobility topologies concentrating on reliability. The classical micro mobility network topology is the tree. After analysing several topologies I show in detail how two of them can be used as a micro mobility network. These two are the *ring* and the *scalable hierarchy of rings*, see [Sza01]. Then I compare the two newly introduced topologies with the classical tree and with each other.

3.1 Reliability in General

Reliability means how the system can tolerate failures. Reliability deals with accidental and mot intentional malfunctioning. Links of the network may break or host may malfunction. A reliable system keeps functioning although some of the components are broken.

I will use a graph model for the network. The hosts are the vertices of the graph, and the connections between the hosts are the edges. All the hosts and links have two states: working/broken, and they are stochastic variables. The breakdowns of different nodes or links are independent events. This is a good assumption at least until the first breakdown.

My reliability analysis is not very deep, but one important point has to be noted. The probability of a node or link staying in the broken down state is very low, usually below 10^{-6}. So the system made up of the component is very likely to have no errors, and much more likely to have one error than more than one.

So a system that can surely tolerate one error is much more reliable than a system that can tolerate no errors. This is because the system is very unlikely to have more errors simultaneously. At the same time systems that can surely tolerate one error are much more likely to tolerate more.

3.2 General Micro Mobility Network Architecture

IP micro mobility access networks are connected to the IP backbone via gateways. Because of the wireless access, service access points (SAP) are called base stations (BS). The traffic shape of a micro mobility network is characteristic. Most of the traffic flows between a gateway and a SAP. Downlink traffic (that is sent form a gateway to a SAP) is usually much more than uplink traffic (mobile nodes get long answers to short questions), see [Sza01b].

As mobile nodes (MN) are wandering around within the micro mobility network, dynamic routing is needed. This makes routing an important question of a micro mobility network. The actual positions of the MNs have to be stored in a (possibly shared) database.

3.2.1 The "Classical" Tree

Most micro mobility protocols define one gateway, and a tree topology network of routers with the gateway as the root. Every node has one uplink neighbour (parent towards the gateway) and may have some downlink neighbours (children towards the MN). In Figure 3.1 D is a downlink neighbour of C and A is an uplink one respectively. The nodes that do not have any children are called leaves. The leaves are base stations in the micro mobility network, the nodes with children are routers.

The root node or gateway is connected to the IP backbone, and all the traffic of the mobile nodes flows through it. All the routers maintain a routing cache [Glo00], where data is stored about the MNs that are in the subtree under the router. The routers know which child packets have to be passed to. As we go higher and higher in the tree, more and more link capacity is needed.

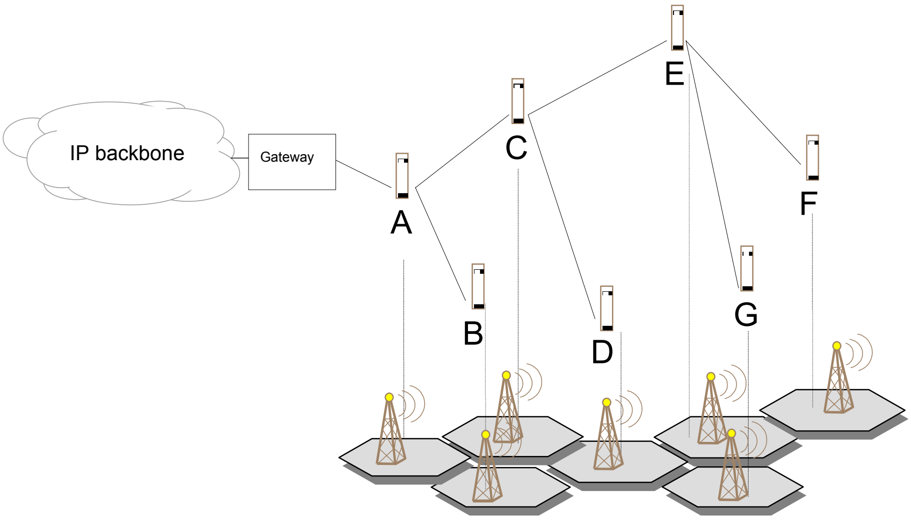

Figure 3.1. General architecture of micro mobility networks

3.2.2 The Reliability of the Tree

Tree topology means that there is exactly one path between any two nodes. So there is exactly one path between a base station and the gateway.

This is a rather vulnerable network architecture. Consider our graph-model, where links and nodes have two states: working/broken. If we suppose that all the traffic flows between a base station and the gateway, a link failure is equivalent to the failure of the node that is at the bottom end of that link. The result is the same. A subtree is separated from the network.

It is even more severe, when the gateway router or the link between the gateway and the backbone breaks down. Then the whole micro mobility network is separated from the backbone, and no communication is possible between an MN in the network and another host on the Internet.

If this topology is so vulnerable, then why are almost all the micro mobility solutions based on a tree topology network? It is because the tree suits the routing requirements of the micro mobility network and signalling requirements of the micro mobility protocol very well. Both uplink and downlink routing are simple so simple and relatively cheap routers can be used. And at the same time the tree is a very scalable solution. The problem that we are concentrating on is the weak reliability of the tree.

There are two basic solutions for the reliability problem. One is to use a completely different network topology, the other is to try to make the tree topology more reliable somehow. If another topology is used in the micro mobility network instead of the tree, it has to be chosen carefully. There are several aspects. The network should be less vulnerable than the tree, of course, but too complex routing or too complex signalling should be avoided, and scalability is also very important for micro mobility networks.

If the tree topology is kept but improved, the aspects are similar. Links and nodes have to be duplicated and physically separated for safety reasons. This new network inherits a lot of the attributes of the tree, for example it probably suits the signalling requirements, and remains scalable. But routing and signalling becomes much more complex. The micro mobility protocol and the routing have to be redesigned.

3.3 Examination of Various Topologies

In this section, various network topologies are investigated, which are suitable for micro-mobility networks. The special features of a micro-mobility network make some of the otherwise not that important aspects really crucial, and at the same time raise some new problems. The most important problems related to micro-mobility networks are:

- reliability, vulnerability,
- scalability,
- connection to other networks (Internet),
- wandering MN, complexity of routing,
- special traffic.

3.3.1 Tree

The tree is the "classical" micro mobility network topology. Both Cellular IP [Cel00] and HAWAII [Haw99] use tree network topology. Almost all requirements are met, the major weakness is vulnerability.

3.3.2 Broadcast Medium, Bus

A bus topology network can be connected to the Internet via gateways. If multiple gateways are used, the reliability is probably satisfactory. There are no routing problems, an access protocol is used instead. (ALOHA, CSMA). The serious problem with the broadcast medium is inscalability. If it is used in a micro mobility network, the size of the network is strongly limited.

3.3.3 Star

Star is a centralised network topology. All the nodes are connected to the central node. The central node can be used as a gateway to the Internet. All routing capabilities can be concentrated to the central node, other nodes are very simple, thus very cheap. Routing at the central node is not very complex, and there is no routing at the other nodes. This network topology really suits the traffic shape of a micro mobility network, where most of the traffic flows between the gateway and a base station. Vulnerability is a weakness, as a central node breakdown is critical. This is one of the reasons why a double star is often used. In a double star, the central node is duplicated, and the two central nodes are probably connected to each other. Packets then can be sent to both gateways.

Another weakness is inscalability. As the number of base stations increases, routing at the central node becomes resource time consuming.

3.3.4 Ring

In a ring there are exactly two paths between two nodes. If a link or node breaks down, there is still one path left, so it is much more robust than the tree. In a micro mobility ring multiple gateways should be used of course. Routing in a ring is simple. The ring does not expressly suit the traffic requirements, and inscalability is another problem. As the number of BSs increases, routing does not get more complex, but links may get overloaded.

An important ring type is the self-healing ring. In a self-healing ring only one half of the capacity is used for normal operation, the other half is reserved for critical situations. It is like the MSSP (Multiplex Section Shared Protection) ring in an SDH environment. If a link breaks down, the two neighbouring nodes realise the breakdown and the spared capacity of all other links is used to replace the broken link, see Figure 3.2. Thus, one error can be corrected below the micro-mobility level, and a reliable communication network is provided for the micro mobility protocol.

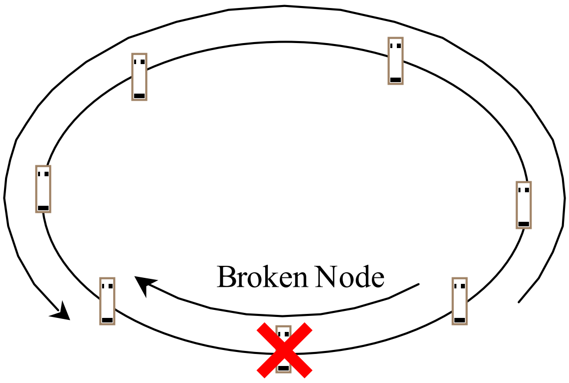

Figure 3.2. Self-healing ring with broken node

3.3.5 Mesh

A full mesh is nonsense of course, because it is extremely inscalable, and does not suit the traffic shape anyway. A partial mesh can be scalable and multiple gateways can be used for safety reasons. It is robust, if there are several paths between any two nodes. The only problem is that routing becomes difficult. The packets have to be routed correctly even when some of the links are broken. So a complex routing protocol has to be used, and all of the nodes have to function as routers, so unless a very sophisticated routing is used, it is an inefficient and expensive solution.

3.4 Possible Alternatives

It is the tree that suits the micro mobility requirements best. The only weakness of the tree is safety. The mesh looks quite attractive as an alternative, but when designing a routing protocol for meshes, a lot of things should be reinvented, and the result would probably be an extremely complex protocol.

The double star and the ring seem to be the two topologies that suit the requirements beside the tree.

In the next sections the ring topology is examined in detail concentrating on the micro mobility aspects and then a new reliable micro mobility network topology will be introduced based on the results. That new topology is the *hierarchy of rings*. It is a combination of the tree and the ring.

3.5 Elaborating the Ring

The ring consists of several nodes and two-directional links. This is of course not an autonomous system, but an IP micro mobility domain, so it is connected to the Internet (or IP backbone) by a router or possibly routers.

In the following subsections I present how a micro mobility domain can be based upon ring network topology. Of course there might be some other micro mobility architectures over a ring than the one described here, however this solution improves not only the reliability of the system but the routing capabilities too.

3.5.1 Multiple Gateways, Multiple Connections

The reliability of the ring is satisfactory, but how should we connect the network to the IP backbone? Some of the nodes in the ring have to function as micro mobility gateways. At least one gateway is needed, but having only one might make the connection between the backbone and the access network too vulnerable. A gateway failure or a link failure at the gateway can separate the whole micro mobility network from the backbone. On the other hand having more than one gateways obviously makes the micro mobility protocol more complex. But as we will see, in ring network topology, it is easy to handle multiple gateways.

How many gateways should we have then? A larger micro mobility network does not necessarily require more gateways. The main reason of using multiple gateways is the increased reliability, and having more than three gateways does not make notable improvements to reliability. So it is rational to assume that the network does not have more than ten gateways.

It is important, because it makes it possible to store some data (e.g. how far is it? which direction is it closer?) about all the gateways at each node.

Besides increasing reliability, having more gateways has another advantage. If a gateway is closer to a node, a packet sent by the node can be routed out from the micro mobility network earlier (i.e. on a shorter path) and a received packet can reach the node earlier. Thus, having more gateways allows the link capacities to be better utilised.

3.5.2 Node Types

Now we summarise what kind of nodes should be used in a ring topology micro mobility network. The self-healing mechanism has to be implemented at each node. This is what the nodes have in common, but the functionality of the nodes can be different. At the micro mobility level, the node types in the ring are:

- gateway + router: The router routes the packets between the two neighbouring nodes and the gateway. The gateway sends packets out to the IP backbone and receives packets from there. There are probably about two or three nodes of this type in the access network.
- BS + router (SAP): The router routes packets between the two neighbours and the BS. The BS sends packets to the MNs and receives packets from them. This is probably the most common node.
- router + special function: There can be nodes that neither function as gateways nor as BSs, but have some other function such as packet authentication or traffic analysis.

- combined: A combined node for example is a node that functions as a BS and as a gateway at the same time. It is better to avoid these combined nodes and separate the functions.

Figure 3.3 shows the architecture of a ring topology micro mobility network.

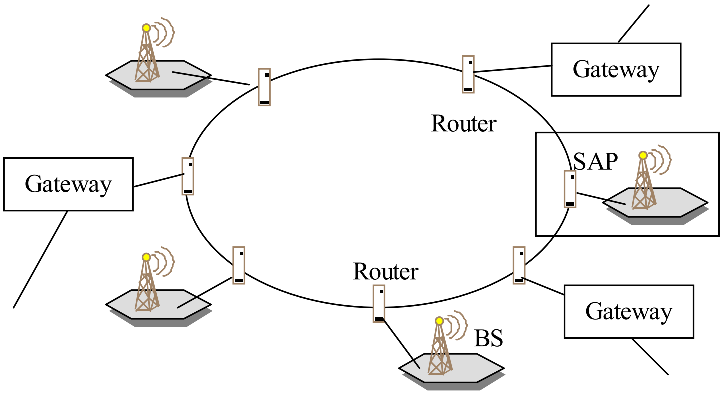

Figure 3.3. Ring topology micro mobility network

3.5.3 Uplink and Downlink Routing

Once we have a reliable network with reliable connections to the IP backbone we want our micro mobility protocol to work over this network. Micro mobility routing questions over a ring topology network are addressed in this subsection.

As already mentioned the architecture presented here is not the only solution, some even more powerful solutions may exist, but this one can illustrate how the features of a ring can be profited when building a micro mobility domain.

Most of the nodes of the ring are the service access points (SAP) of the micro mobility network. Some of the nodes are gateways to the Internet.

The Basic Structure

As it was mentioned in Section 3.2, most of the data flows between a BS and a gateway. The two directions of this kind of traffic will be handled separately. *Uplink* traffic means data originating from a SAP and being delivered to a gateway, *downlink* traffic means data flowing from a gateway to a SAP. Traffic between two SAPs will be called *internal* traffic. In a micro mobility environment the internal traffic is insignificant compared to uplink and downlink traffic.

The basic idea is that we have a ring with two-directional links and use one direction for the normal traffic. The full capacity of the other direction is reserved for self-healing, see Section 3.3.4. Thus we have a safe ring with one-directional links. This makes routing much more simple, because nodes on the ring do not have to decide which direction to send packets. On the other hand packet routes are not always optimal in the micro mobility network.

Routing in the micro mobility domain is based on some kind of ID of the MNs. This ID can be the home IP address or the care-of-address of the MN or some other ID that identifies the MNs uniquely. The care-of-address is probably the best choice.

SAPs

The SAP consists of one BS and one router. The function of the BS is independent of the network topology, a BS in a ring topology network operates the same way as a BS of a tree topology network. Time Division Multiple Access (TDMA), Frequency Division Multiple Access (FDMA), Code Division Multiple Access (CDMA) or a combination of these can be used to access the common medium.

In our model the router of the SAP has to maintain a database about the MNs that are connected to that SAP. A SAP does not have to have any information about other MNs, BSs or gateways, it just has to know which MNs are connected to the BS next to it. These database entries are soft state, which means that if the MN leaves without notification, they are torn down after a time.

Downlink Traffic

IP routing at the Internet is beyond the scope of this paper. Micro mobility downlink routing begins when one of the gateways receives a packet from the Internet. The packet can be originated from a Mobile IP HA (Home Agent) or a CN (Corresponding Node) somewhere on the Internet. The packet is passed to the router, and as we have a one directional ring, the packet is launched onto the ring.

All the nodes except the SAP that can find the recipient MN in its database forward the packet along the ring. If the router of a SAP identifies the destination address as an address of an MN connected to that SAP, it passes the packet to the BS instead of forwarding it. If the recipient is in none of the databases, the packet travels around the ring, and the gateway that sent it off receives it again. From receiving it, it can know that the MN cannot be found in this network, and can act accordingly.

Uplink Traffic

If the BS of a SAP receives a packet from a MN, it passes it to the router. The router launches it onto the ring. All the nodes forward the packet along the ring until it reaches the first gateway. As the gateway router recognises that it is an uplink packet, instead of forwarding it along the ring it passes it to the gateway. If a router of a SAP receives a packet that it has sent out, it can know that the packet has travelled around the ring. It can deduce that probably there were more than one link or node errors in the ring, the ring has fallen apart, and there is no gateway in that part.

It is of course possible to allow the BS to specify which gateway it wants to use for sending out the packet to the backbone instead of the first one the packet reaches.

3.5.4 Registration

When the MN enters the micro mobility network or it is switched on there, it has to register both to its Mobile IP home agent, and to the micro mobility network. Mobile IP registration is beyond the scope of this paper.

The micro mobility registration is very simple. Unlike in the case of a tree topology network, the position of the MN has to be stored only at the SAP where it accesses the network. When the BS sends a registration message to the router, the router adds a new record to the database. Later when it receives a packet sent to the MN, it will pass it to the appropriate BS.

Because database entries are soft state, the MN has to send registration update messages from time to time to prevent the entries from timing out, like in Cellular IP [Cel00]. These registration update messages are almost the same as the registration messages. When receiving a registration update message, the router resets a timer instead of adding a new entry to the database.

3.5.5 Handover

In a micro mobility environment handover is initiated by the MN. The BS which it moves away from is called the *old BS*, the BS it moves to is referred to as the *new BS*. We consider soft handovers because all the members of the 3rd generation

IMT2000 family use wideband CDMA. Therefore the MN is able to communicate with more than one BSs at the same time.

When soft handover is used, the MN informs the new BS that it wants to access the network there. It is similar to a registration at the new SAP. The router of the new SAP adds a new record to the database. From that time the router passes packets to the BS and does not forward them. Then the MN informs the old BS about its leaving. The old SAP’s router sets up a timer for the database entry, and when the timer expires, it deletes the entry. This timer is needed to avoid the loss of the packets that were already on the way between the old and the new BSs when the new BS was notified. There is a fixed amount of time between the addition of the new entry at the new SAP's router and the deletion of the old entry at the old SAP's router. During that time the MN can send packets to both BSs. Depending on the order of the old and the new BSs in the ring, during the time of the handover, packets may arrive only from the old BS or from both the old and the new ones This is a soft handover type called *simple* handover. The simple handover is always soft.

There is another handover mechanism that can be used in a ring topology micro mobility network, called *advertised* handover. The advertised handover has a soft and a hard variation. MNs that are unable to maintain connections to more than one BSs can also use the advertised handover. The disadvantage is that it requires communication on the ring, and hence it is slower and more resource consuming than the simple handover. In the advertised handover it is not the MN but the new SAP that notifies the old SAP about the handover. As the new BS receives an advertised handover message, it passes it to the router. The router launches the message on the ring, where it travels to the old BS. The old BS deletes the database entry when it receives this message. In the hard variation, the MN is only connected to the new BS, thus packets sent to it by the old BS are lost.

3.5.6 Standby Mode, Paging

MNs that are not transmitting or receiving any data may switch to "idle" state to prolong battery life. When the MN is in idle state, it does not have to notify the network about each handover, hence the network does not know its exact location. In order to locate the MN in case of an incoming call, Paging Areas (PAs) are defined, and the MN has to notify the network only when it moves from one paging area to another. In case of a ring topology network it is obvious to define the whole ring as one paging area.

When the MN switches to idle state, it sends a control message to the BS. The BS router sends the message around the whole ring, and all the routers put the ID of the MN in their database as an idle state MN. These entries have to be refreshed, because they are also soft state, but they timeout period is much longer than that of the routing entries.

When a packet destined to an idle state MN arrives to the network, it travels around the ring, and all the BSs send a paging message to the MN, and delete the paging entry from the database. When the MN receives the paging message, it has to switch to active state, and register to the micro mobility network. At registration, the appropriate SAP adds a routing entry to its database, and packets can be sent to the MN.

3.5.7 Further Considerations

Breakdown Considerations

If a BS breaks down in the ring, the ring heals itself, and both uplink and downlink traffic routes remain the same. What happens when a gateway breaks down? The ring heals itself and the new network topology will be a ring without the broken down gateway node. Figure 3.4 shows how the uplink traffic that the broken gateway used to send out to the Internet is sent out by the next gateway in the ring.

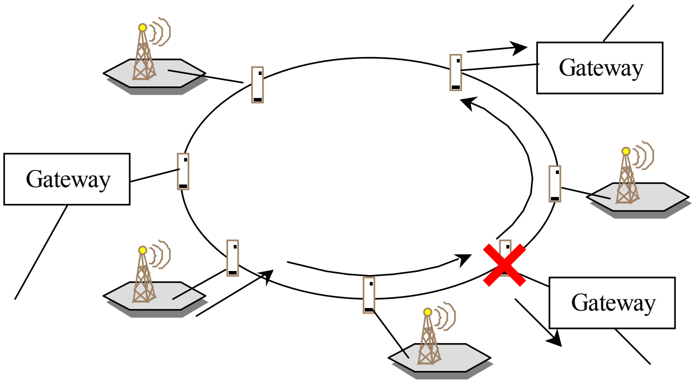

Figure 3.4. Broken gateway in a ring topology micro mobility network

Gateways with Enhanced Databases

In a micro mobility network there are a lot more BSs than gateways. Using more complex gateways with huge databases does not significantly raise the costs. To build a cost effective network the SAPs should be as simple and cheap as possible. It might be a good idea to store some more data at the gateways. In Section 3.6., at the hierarchy of rings topology it will turn out that it is useful to have a list of the MNs staying in the network at each of the gateways. If we would like to maintain these lists in the gateways, the registration has to be more complex. When the MN registers at the SAP, after updating the database a message has to be sent around the ring informing all the gateways about the new MN in the network.

It is also possible to store the ID of the idle state MNs at the gateways instead of the routers. When a packet arrives to the idle state MN, a paging message is sent around the ring by the gateway. This paging message makes the MN switch to active state.

Network Size

It is hard to estimate the size of the micro mobility domain that a ring topology network can serve. We can have probably a few hundred SAPs and a few (not more than ten) gateways.

Serving 50 MNs should not be too much for a SAP, so the range of a ring topology micro mobility network can be a building for example.

In order to extend the size of the network multiple ring topology networks can be used, and when the MN moves from one network to another, a macro mobility (Mobile IP) handover takes place. In the next section it is presented, how to build large size scalable network from the above introduced basic ring topology, where these Mobile IP handovers can be avoided.

3.6 Scalable Hierarchy of Rings

The most serious weakness of the ring is that it does not scale with the size of the network. To solve this problem I recommend a hierarchical interconnection of ring topology networks.

3.6.1 The Topology

In a tree topology micro-mobility network every node has a parent node (except for the root), and every node may have some child nodes under it. A hierarchy of rings is similar to this topology. It is a tree with rings instead of the nodes. Every ring has a parent ring (except for the root ring), and every ring may have some child rings under it. In our network there is exactly one root ring, and all the rings are self-healing rings like described in Section 3.3.4.

By using this topology the advantages of the tree and the ring topologies can be combined.

All the rings are self-healing rings, so one error in a ring is corrected below the IP level. To build a robust network that can handle failure of the interconnection links that connect the rings of the access network, every ring should have multiple connections to its parent ring.

3.6.2 Node Types

In a hierarchy-of-rings topology micro mobility network there are more node types than in the basic ring topology. These are the following:

- gateway + router: This is the same type of node that we had in a single ring network. Our hierarchical network has gateways only in the root ring.
- BS + router (SAP): This is the same as what we had.
- Interconnection node: these nodes connect the parent ring with a child. There can be two functions separated in an interconnection node. It acts as a base-station-like router in the parent network, and acts as a gateway-like-router in the child network. These two functions can even be separated physically.
- router + special function: The hierarchy can have special nodes too.
- combined: The hierarchy can have combination of the above listed node types too.

3.6.3 Ring types

In our micro mobility network a ring has either BS routers or child rings, but not both. Actually both of them can be allowed, but this restriction makes the network much more structured. So there are two types of rings: *access rings* (with BSs but no children) and *transport rings* (with children but no BSs).

Access rings are actually ring topology micro mobility networks with enhanced database gateways like the one described in Subsection 3.5.7., but with a bit more functionality. The interconnection nodes that connect the access ring to its parent act like the gateways in the simple ring structure. As they are extended database gateways, they have information about all the MNs connected to this access ring. Figure 3.5 shows a micro mobility network with this topology. This network has three access rings, two transport rings and the root ring has two gateways.

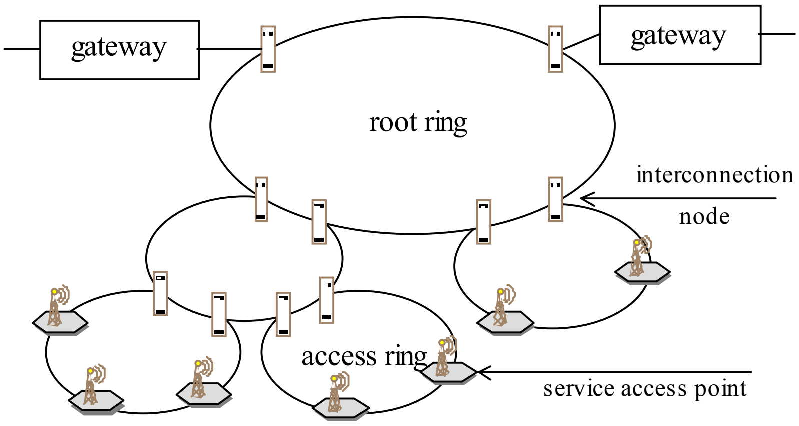

Figure 3.5. Hierarchy of rings topology micro mobility network

A transport ring is also similar to a ring topology micro mobility network. The interconnection nodes that connect the ring to its parent act like gateways, the interconnection nodes that connect the ring to its children act like the SAPs. They have information about all the MNs that are directly or indirectly connected to the child ring.

Thus all the interconnection nodes have a database of all the MNs that are connected to the micro mobility network "under" them.

3.6.4 Uplink and Downlink Routing

Uplink Traffic

MNs send packets to a BS of an access ring over the air interface. The BS passes them to the BS router. The router launches them on the ring, and they travel along the ring until they reach an interconnection node. As the interconnection node identifies these packet as uplink packets, it passes them "up", thus launches them on the parent ring instead of forwarding them. Packets are passed up ring by ring to the root ring by the interconnection nodes. From the root ring the first gateway sends them out to the IP backbone.

Downlink Traffic

When a packet arrives from the Internet, a gateway router launches it on the root ring. As all the interconnection nodes know which MNs are connected under them, the first interconnection node that has the recipient address in its database launches the packet on the child ring instead of passing it on. The packet travels down from transport ring to transport ring until it reaches the proper access ring, where the appropriate BS passes it to the MN.

It is easy to route packets correctly if all the interconnection nodes have the information about the MNs under them. It will be elaborated how these databases can be maintained.

3.6.5 Registration

Now let me present what happens when a MN registers to the micro mobility network and how the databases should be built up. The micro mobility registration message travels from the access ring where the MN registers, up to the root ring and all the databases are set up. The registration message travels around the whole access ring, the whole transport rings on the way, and the whole root ring, so all the interconnection nodes and gateways can set up their databases correctly. The registration message can be passed up by all the interconnection nodes or just the first one it reaches. If they are passed up by all of them, multiple registration messages have to be handled in the parent ring. If only the first interconnection node passes them up, then the message has to be altered by the first interconnection node so that other interconnection nodes can know that it has already been passed up. It is safer if all the interconnection nodes pass up the registration messages.

3.6.6 Handover

If the MN cannot be connected to more than one base station at a time a Cellular IP-like [Cel00] hard handover mechanism can be used. Here soft handover is explained in detail.

Consider the two paths: the one to the old BS and the one to the new BS. Going uplink the first ring that is part of both paths is called the crossover ring. (Like the crossover router in the Cellular IP tree.) The soft handover is similar to a registration. A control message travels up to the crossover ring. There the interconnection nodes that connect the crossover ring to its parent know that the MN is under them, they do not change the entries in their databases. A registration update message can be sent up to the root ring to prevent database entries from timing out. The interconnection nodes that connect the crossover ring to its child ring that is on the path towards the old SAP delete the database entries, so they do not route the packets down any more. A release signal may be sent down to the old BS that tears down the database entries, but they will time out anyway. During the time of the handover the MN may send packets to both BSs, and packets may arrive from both BSs. If a release signal is sent down, it can be passed to the MN, so the MN knows that it will not receive any more packets from the old BS.

3.6.7 Standby Mode, Paging

It is obvious to define the paging areas as the access rings. When an MN switches to idle state, a message is sent around the ring, all the BS routers delete the MN from their database, and all the interconnection nodes put a paging entry in theirs. To prevent the database entries from timing out at higher levels, the interconnection nodes have to send route update messages up to the root ring while the MN stays in idle state.

When a packet arrives to the MN, it is routed down to the appropriate access ring, where the interconnection node sends a paging message around which makes the MN register at one of the SAPs and it switches to active state.

When the MN moves from one access ring to another in idle state, it has to notify the network. This case is very similar to a handover in active state. The MN sends an idle-state registration message to the new BS in the new access ring. This idle-state registration message travels around the new access ring. All the interconnection nodes put the paging entry in their databases, and from this point exactly the same procedure happens as in case of normal active-state handover. The message travels up to the crossover router, and the path to the old access ring may be cleared explicitly or they can just left to be timed out.

3.7 Comparison

The "classical" topology of a micro mobility domain is a tree. The two alternatives presented here are the single ring and the hierarchy of rings.

Both of the new topologies provide better reliability than the tree. Both of them can tolerate at least one node or link error, while the tree is very vulnerable.

Routing in the ring is much more simple than in the tree. In the hierarchy of rings routing is more complex than in the tree, but not significantly.

A lot of the solutions of Cellular IP [Cel00] and HAWAII [Haw99] can be used in both the ring and the hierarchy.

The scalability of the hierarchy of rings is as good as that of the tree, the single ring is weaker in scalability.

The tree and the hierarchy of rings have to be connected to the Internet at the root or the root ring, while a single ring can have connections anywhere.

An advantage of the ring is that uplink and downlink traffic use the same links, so we do not have to decide beforehand how much capacity to allocate for uplink and downlink traffic.

In the ring and the hierarchy of rings the handling of idle state MNs is much more "local", than in a tree. In a hierarchy of rings, only one access ring has to know about an idle state MN.

Authentication is a crucial point. Similar security considerations are needed in case of all the three network topologies.

As the *hierarchy of rings* topology is a combination of the tree and the ring, it inherits several qualities from them. It inherits its good reliability from the ring, routing methods and its good scalability from the tree.

4 Security Considerations

4.1 Introduction

In case of IPv4 a *home agent* (HA) and a *foreign agent* (FA) are needed for Mobile IP. The packets sent to the *mobile node* (MN) are forwarded to the FA by the HA, the FA forwards them to the MN, and the MN receives them as if it were staying in its home domain. So the mobility is limited to networks, which have FAs. In case of IPv6 mobility no FA is needed. Mobile IPv6 is implemented in the *home agent* and in the *mobile node*. Actually Mobile IP is a part of the IPv6 protocol.

We consider domain-based micro mobility protocols. The domain where the MN is wandering using micro mobility is called the *foreign domain*(FD) or *foreign network*(FN).

The micro mobility is transparent for the *home agent*. When the MN is wandering around in the *foreign domain* using micro mobility, the HA is not even notified about the handovers. Signalling only takes place within the *foreign domain*. If the MN moves to another domain, the HA is notified, and normal MobileIP handover (macro mobility handover) takes place.

Of course micro mobility has to be transparent for the *corresponding node* (CN). Any host communicating with the MN should not even know that the MN is not staying in its home network, or the CN learns the care-of-address of the MN, it should not know that the *foreign domain* is a micro mobility domain. The CN sends the packets to the home IP address of the MN or to the care-of-address to avoid triangle routing.

While the MN stays in the micro mobility domain, the HA and all the CNs send the packets to the same care-of-address. The packets sent to this IP address get to a special host (router) of the *foreign domain*, called the *gateway router*. The *gateway router* is responsible for forwarding the packets to the MN.

Mobility is usually used in wireless environments. The MN is connected to one of the wireless access points called the *base stations* (BS).

Micro mobility is also transparent for the MN in the sense of IP. It has a fix address, packets are routed to it, and it can send IP packets to any IP addresses. But the MN has to communicate with the foreign domain hosts to make it possible for them to route the packets correctly. The MN and hosts of the *foreign domain* have to exchange micro mobility control messages. When the MN enters the *foreign domain* it has to register. When it moves to a new BS (or new *paging area*), it also has to notify the network. The macro mobility (MobileIP) is either handled by the *gateway router* or by the MN. If the macro mobility is handled by the gateway, it is similar to the MoblieIPv4 case. The gateway router acts like a foreign agent. The MN can handle MobileIP in an IPv6 environment.

There has to be two different protocols implemented in the MN: IPv6 (with MobileIP and IPsec) and a micro mobility protocol (for example Cellular IP [Cel00]).

4.2 Security of Mobile IP (macro mobility)

When some micro mobility protocol is used within a foreign domain, standard Mobile IP is used as the macro mobility solution. A Mobile IP handover takes place every time, the MN moves to another foreign domain (another micro mobility network). IPsec is the standard answer to security questions in a Mobile IP environment. [Sec00]

When the IP tunnel ends at the *gateway router*, the case is similar to IPv4 Mobile IP with a *foreign agent*. The gateway router also acts as a foreign agent. There should be security associations (SA) established:

- between the MN and the HA (required by Mobile IP)
- between the MN and the CN (to provide end to end security)
- between the *gateway router* of the foreign domain and the HA of the home domain

If the MN is the end of the tunnel, the 3rd SA can be omitted.

Instead of the SA between the *gateway router* and the *home agent* there can be a tunnel mode SA between the *security gateways* of the two domains. This SA can be easily set up if there is a service level agreement (SLA) between the home domain and the foreign domain, and in this case only one SA is needed even if there are more than one MNs registered in the *foreign domain* from the same *home domain*. This tunnel prevents active and passive attacks on the Internet, and also hides all the information about the MN in the *foreign domain*. This means that no attacker somewhere on the Internet can get any information about the MN in the *foreign domain*. This way even the care of address can be hidden (of course the HA has to know it). The domain where the MN is attached to the Internet is known, but the IP address in the network is completely hidden.

The SAs between the MN-HA and the MN-CN prevent attacks from the home and foreign networks. The SLA between the two domains does not mean that it should be possible for even the *foreign domain* administration to listen to the communication of the MN.

4.3 Security Within the Micro Mobility domain

Standard IPsec ESP and AH can be used for end-to-end message integrity, sender authentication and confidentiality. A tunnel mode SA between the *home domain* and the *foreign domain* can make attacks from the Internet even harder to carry out.

In this chapter security of the communication within the foreign domain is discussed, attacks from the *foreign domain* are examined.

4.3.1 Micro Mobility Protocol

There are several BSs in the micro mobility domain. In most of the approaches there is a tree hierarchy with the *gateway router* as the root of the tree. Signalling messages has to be sent between the BSs and routers, and between the BSs and the MN. Some of these links can be wireless, what makes attacks easier.

The control messages that flow between the hosts of the FN (BSs and routers) do not go "out" to the Internet. The source and destination of these messages are in the same administrative domain, the *foreign domain,* where any asymmetric or symmetric key authentication and encryption algorithm can be used. Asymmetric algorithms are probably better, because then if a successful attack is carried out against a BS, only a part of the network is corrupted, and what is more important: the *gateway router* and other important hosts are not.

All the hosts of the foreign domain have to have a shared secret, a key. This key is not given to any MNs in the domain or any other hosts outside of the network. No matter what kind of algorithms are used, key distribution in the foreign domain is not difficult, so the gateway router and all the BSs of a domain can have the shared secret. This shared secret can not only be used to encrypt and authenticate control messages within the domain, but certificates can be issued by hosts of the domain, that other hosts can verify without contacting the issuer. This method can also be used for MN authentication, see Chapter 5.

4.3.2 Wireless Links

In the micro mobility environment wireless links will be used. What are the special features of a wireless interface from a security point of view?

The wireless link is extremely simple to eavesdrop. The “medium” is easy to access and the hardware requirements for eavesdropping are cheap.

Sending false messages on the air interface is simple too.

On the other hand substitution of messages (or parts of messages) are almost impossible on the air interface. It is difficult for the malicious attacker to prevent the receiver from receiving the message.

Thus the man-in-the-middle type attacks described in the following sections are not as trivial to launch as they might seem.

4.3.3 Registration

When the MN is switched on in the *foreign domain* or it enters the *foreign domain*, it has to register to the network. It has to authorize itself to the Micro Mobility network. The authorization is based on a database, and is done by authentication of identity. The gateway router, or a dedicated host can handle this authentication. As this authentication is the most important part of micro mobility, it is better to have a dedicated authentication server in the domain. This authentication can be based on a SLA and may require communication between the *foreign domain* and the *home domain* of the MN, or it can be based on certificates. This authentication is elaborated in Chapter 5.

Another issue is that the foreign domain has to "authenticate" itself to the MN too, to prevent impersonations. An attacker can try to impersonate the whole micro-mobility network. It simulates all the protocol steps, and can try to eavesdrop or do some other nasty things. This authentication can be based on the SLA too. This means that the MN knows the public key of the network, and the network can for example authenticate itself in a challenge and response way. Certificate based authentication can be used here too.

After registering to the foreign domain, a *binding update* message has to be sent to the HA. It has to be sent by the MN, because IPsec AH must be used for authenticating *binding update* messages. [Mip00] An IP address has to be given as the care of address of the MN. Using the security gateway as the end of tunnel is not weaker in security than using the MN itself, because the care-of-address was given to the MN by the gateway router itself or some other host of the foreign network, and all the packets form or to the MN can be intercepted by the *gateway router* anyway. ESP and AH have to be used between the HA and the MN for end-to-end integrity and encryption.

Personal Identifier

The mobile host will have to authenticate all the control messages of the Micro Mobility protocol. To avoid using complex authentication algorithms for every control message a Personal Identifier (PID) can be used [Cel00]. This PID is generated at the time of registration, and it is a shared secret known by (or easily calculated by) all the BSs and routers in the foreign domain, see Chapter 5, and is known by the MN. This PID is given to the MN at registration time, and can be used for message authentication or encryption.

Note that all the BSs and routers know or (can calculate) this secret, they all can verify the integrity of control messages, they all can sign, encrypt or decrypt control messages.

Stealing the PID

If an attacker can get the PID of a MN, it can send false control messages, and can steal or drop a connection. Confidentiality is not compromised, but the attacker can impersonate the MN to the micro mobility domain.

If the PID is not sent to the BSs and routers, but calculated by them, the attacker cannot get it there. The attacker can try to get this PID when it is sent to the MN.

It can either eavesdrop or try to carry out a man-in-the-middle attack at the registration. To avoid this, the PID must be sent encrypted. This public key of the MN should be acquired from a trusted party, for example a host in the *home domain* using IPsec of course. Or this PID can be on a chip card in the MN (like at the GSM system).

Denial-of-Service Attack

Let us consider that an attacker is sending a false registration control message to a BS of a micro mobility domain. The BS sends the message to the host that handles authentication. During this authentication procedure packets may be sent to the *home domain*, and it may require a lot of computing even if the authentication fails. The attacker may try to flood the network with false authentication requests, and while the network is struggling with these messages, authorized hosts cannot register. This is a denial-of-service attack. These types of attacks are very hard to defeat. One idea can be not to process the authentication requests in a FCFS order, but to have separate queues for every BS, then only registration at some BSs can be made impossible by the attacker, but the whole micro mobility domain can not be paralysed.

4.3.4 Data Transmission

Let us suppose, that the MN has registered successfully, it has its PID, and communicates with various hosts. For confidentiality and authentication the IPsec ESP and AH can be used. The MN can establish an SA with every host it communicates with. This confidentiality and authentication means that not even the *gateway router* can impersonate the MN or eavesdrop the communication. Note that the gateway router is an especially good place for eavesdropping as all the traffic of the MN passes through.

All the data packets sent by the MN should be authenticated using the PID, so false packets sent by an attacker can be discarded immediately by the *base station*. If they were not authenticated, the attacker could send false data packets that would be sent to the HA (where they would be discarded), but resources of the FN would be used.

4.3.5 Handover

Handover is initiated by the MN, and can be based on measuring various parameters of control channels of different BSs (like in the GSM environment).

The handover control messages (request and reply) have to be authenticated. If they were not, a malicious attacker could send false handover requests, and could make the packets sent to a MN routed somewhere else, so the MN should not get the packets sent to it. This handover control message authentication can be based on the shared secret (PID), so the receiving BS can verify the authentication, and forged handover requests can be ignored immediately, so the network cannot be flooded by forged messages (denial-of-service attack).

Another denial-of-service attack can be to try to make the handover impossible. In the micro mobility environment handovers are frequent and fast. If the attacker can prevent the handover request from reaching the BS (by emitting some noise in a wireless environment), the routers will keep routing the packets to the old BS, these packets will be lost. But as soon as the MN can communicate with any of the BSs, it can notify the micro mobility network of its location, so only a few packets will be lost. Note that by this attack confidentiality or authentication is not compromised at all.

4.3.6 Idle State - Paging

If the MN is in the idle state, and a packet arrives to it, the micro mobility network pages it. If a paging message is received, the MN has to switch to active state. If the paging message is not authenticated, security is not compromised, the attacker can only activate MNs by sending false paging messages. If the authentication is done by using a challenge-response algorithm, the MN has to transmit some data (the challenges) to verify the authentication of the paging message. Security is not compromised again, but the attacker can seriously reduce the battery life of the MNs by sending false paging messages. So it is important to authenticate the paging messages in a way that the MN can verify the authentication without transmitting. For example digital signatures can be used.

4.3.7 Leaving the Network

It is possible that the MN leaves a Micro Mobility network without sending any notification messages. This can happen for example because of a link loss.

If the MN is alive, it has to send some control messages to the network. If the network does not receive any messages from a MN, all routing and other information for that MN is lost. If the attacker can prevent the MN from sending these control messages for a time, the network will tear down all information, and the MN has to register again.

5 Authentication

5.1 The importance of authentication

The conclusion of Chapter 4 is that security of a micro mobility environment strongly depends on authentication. The micro mobility network has to be authenticated too, but the authentication of the MN is even more important. Impersonating a MN is much easier than impersonating the whole micro mobility network.

A shared secret between the hosts of the network and the MN can be used to encrypt or sign messages or to generate session keys. This way symmetric key algorithms can be used. There are many algorithms these purposes.

If the MN has the public key of the FN and the FN has the public key of the MN, asymmetric key cryptographic protocols can be used or they can agree upon session keys for symmetric key algorithms.

The importance of the authentication is that the MN can be sure that it is communicating to the real network, and the network can be sure that the MN is really the MN it claims to be.

5.2 Cellular IP PID

In the Cellular IP protocol a *personal identifier* (PID) is used. This PID is the shared secret between the network and the MN.

Here is Section 3.5 “Security” from the Cellular IP internet draft [Cel00]:

> Each Cellular IP Network has a secret network key of arbitrary length known to all Cellular IP nodes. The network key is kept secret from mobile hosts and other nodes outside the Cellular IP Network, however. Upon initial registration the Gateway must authenticate and possibly authorize the mobile host. This initial authentication and authorization can be based on any known symmetric or asymmetric method. After authentication the Gateway concatenates the key of the network and the IP address of the mobile host and calculates the PID of the mobile host by an MD5 Hash similarly as in [4]:

> PID := MD5(network key, IP address of MH)
>
> Then it acquires the public key of the mobile host from a trusted party, encrypts the PID and sends it to the mobile host. This way the mobile host and the Cellular IP network have a shared secret. The PID remains the same during handoff and can be easily computed by each Base Station.
>
> The PID can be used to authenticate (and optionally to encrypt) IP packets over the air interface. Authentication is performed by creating a short hash from the (PID, timestamp, packet content) triple that is placed into the transmitted packets. The validity of each packet can be easily checked by any Base Station even immediately after a handoff and without prior communication with the mobile host or with the old Base Station.
>
> In addition to authenticating control packets, PID can optionally also be used to provide security for data packets transmitted over the wireless link. To this avail, any known shared secret based security mechanism can be used where PID serve as the shared secret.

Fist of all it does not completely solve the authentication, it just says “it acquires the public key of the mobile host from a trusted party”. This is just an explanation why we only have to concentrate on the authentication at registration. The network has to authenticate the MN only once at the time of registration. After that, the PID can be used for authentication encryption and message integrity purposes.

The clever idea is that all the hosts can calculate the PID of a MN, it does not have to be stored.

A Weakness of the Cellular IP PID

The PID in the way it is recommended by the Cellular IP draft has a serious weakness. If the PID belonging to a specific IP address gets corrupted somehow and the corruption is even known, that IP address can not be used without the change of the network key. The PID is a one way hash function of the IP address concatenated to the network key.

This means that the network key should be changed every time a PID gets stolen, which is nonsense. The users in the network should not be able to force the change of the network key.

Enhanced PID

This problem is eliminated, if the PID contains a timestamp. Instead of a hash function of the network key and the IP address only, the PID can be the hash function of the network key, the IP address and an integer. This integer represents time, it is sent to the hosts encrypted by the network key once a day or once an hour. For example if it is sent once a day, and a PID gets stolen, the IP address cannot be used for a day. The following day the PID of that IP address is changed. IP addresses with corrupted PIDs must not be given to MNs at registration. The node that handles registration has to maintain a list of the corrupted PIDs. This list is not indefinitely growing now, it is reset every day.

A weakness of this system is that successful attacks can be launched if the attacker is able to manipulate system time. If the clock of a node is changed, an old PID can be used.

5.3 Another Symmetric Key System

Here is another solution for secret sharing. The shared secret can be used for symmetric key cryptographic algorithms, but for the secret sharing asymmetric cryptography is used as in the previous example.

We would like to use symmetric key cryptography. The key will be called PID as in the Cellular IP example. In this solution the hosts neither know the PID nor can they calculate it, but the system has all the advantages of the previous one.

At the time of registration, a random PID is generated for the MN. Then this random PID and a “certificate” are sent to the MN encrypted with its public key.

The structure of the certificate is the following:

$$C=E_K(IP,T,V,PID)$$

Here IP is the IP address of the MN, T is the time of issue and V is validity. E_K means that the quadruple is encrypted with the network key.

Note that this is not a true certificate, because only hosts of the network can verify it. Without the network key, it can not be even read. This “certificate” certifies that a PID belongs to an IP address. The validity of certificates can be limited in time in any way. The MN uses the PID as the key for symmetric key encryption and message authentication, and always sends its certificate along with the message. The BS receiving the message has the network key, so can extract the PID from C.

BSs can maintain a cache of the most recently used certificates along with the IP addresses. When a BS wants to talk to the MN it can look up the PID in the cache, if it is not found, it has to send a PID request message. As a reply the MN sends its certificate to the BS.

An advantage of this method, that although a symmetric key algorithm is used, the key itself is not stored in the network it plaintext form, only encrypted with the network key.

A disadvantage is that it causes a communication overhead, as the certificate is sent several times.

5.4 Using Asymmetric Key Cryptography

Asymmetric key cryptography can provide user authentication too. The FN and the MN both have a private key / public key pair. This solution assumes a hierarchy of certificate authorities (CA). A CA certifies that the public key of the FN really belongs to the FN. This certificate will be limited in time of course. So the FN can be easily authenticated. All it has to do is to send this certificate to the MN. The MN sends all the messages encrypted with the public key of the FN, only hosts of the FN will be able to decrypt it anyway. The FN will sign all the messages with its private key, the MN verifies the signatures.

How can the MN authenticate itself to the FN? The problem is that the MN does not only have to be authenticated (it has to prove its identity), but also have to be authorized (it has to prove it has the right to use services offered by the FN).

The solution is that at the time of registration, the FN gives the MN a certificate similarly as in the previous solution.

When the MN registers, the host that handles the registration gets the public key of the MN from a trusted party (for example the HA). The certificate can be one of the following:

$$C=E_K(IP,T,V,M_{pub})$$

$$C=S_n(IP,T,V,M_{pub})$$

IP is the IP address of the MN, T and V are the time of issue and validity as in the previous example, M_{pub} is the public key of the MN.

E_K means encryption with the network key, so that only hosts of the FN can verify the validity of the certificate, S_n means signing with the networks private key, so that everyone can verify the validity of the certificate.

The MN signs all the messages with its private key and sends its certificate along with the message.

The hosts of the FN maintain a cache of the most recently used certificates like in the previous example. All messages sent to the MN has to be encrypted with the public key of the MN.

5.5 The Public Keys

All these solutions, even the symmetric key ones are based on public key cryptography. In the Cellular IP PID solution the FN sends the PID to the MN encrypted with its public key. How can the FN and the MN be sure that the public keys really belong to the other party. There are basically two methods. One is to get the public key from a trusted party, the other is the use of certificates.

Obtaining the Public Key

The FN can broadcast its public key. The MN cannot be sure that it is really the public key of the FN, but as IPsec is used for end to end encryption and authentication, confidentiality can not be compromised anyway.

If there is an SLA between the FN and the HN, the MN can get the public key from a trusted host of the HN, and the FN can acquire the public key of the MN from the HA for example.

CAs can issue certificates certifying that a public key belongs to a network. The FN can have a certificate for its public key. If the HN has a certificate too, it can issue certificates for all its MNs.

In practice the FN and the HN have certificates, and the public key of the MN will be obtained from the HN.

Another possibility is similar to the GSM system. The MN can have the public key of the FN, its own private key, and certificates in untamperable hardware (i.e. chipcards as the SIM card of GSM).

6 Conclusions

In micro mobility environments both reliability and security are important issues. There are several ways to improve the reliability of a network. In Chapter 3 I examined the reliability of various network topologies. As an alternative of the widely used tree topology I proposed the ring, and the combination of the tree and the ring: the scalable hierarchy of rings.

In Chapter 4 I gave a survey on various security considerations and attack situations in micro mobility networks. Chapter 5 deals with authentication, ss authentication appears to be the most important point of the security of micro mobility environments. After examining the authentication system of Cellular IP, I recommend some more symmetric and asymmetric key authentication protocols to be used in micro mobility environments.

Abbreviations

AH	Authentication Header
BS	Base Station
CA	Certificate Authority
CDMA	Code Division Multiple Access
CN	Corresponding Node
ESP	Encapsulating Security Payload
FA	Foreign Agent
FCFS	First Come First Served
FD	Foreign Domain
FDMA	Frequency Division Multiple Access
FN	Foreign Network
HA	Home Agent
HD	Home Domain
HN	Home Network
IP	Internet Protocol
LAN	Local Area Network
MD5	Message Digest 5
MN	Mobile Node
MSSP	Multiplex Section Shared Protection
PC	Personal Computer
PDA	Personal Digital Assistant
PID	Personal IDentification
SA	Security Association
SAP	Service Access Point
SDH	Synchronous Digital Hierarchy
SIM	Subscriber Identity Module
SLA	Service Level Agreement
TDMA	Time Division Multiple Access

References

[Per96]	C. Perkins, "IP Mobility Support", IETF RFC 2002, http://www.ietf.org/rfc/rfc2002.txt, 1996
[Cel00]	A. T. Campbell, J. Gomez, C. Y. Wan, S. Kim, Z. Turanyi, A. Valko, "Cellular IP", draft-ietf-mobileip-cellularip-00.txt, IETF Internet Draft, 1999
[Haw99]	R. Ramjee, T. La Porta, S. Thuel, K. Varadhan, L. Salgarelli, "IP micro-mobility support using HAWAII", draft-ietf-mobileip-hawaii-01.txt, IETF Internet Draft, 1999
[Glo00]	B. Gloss, C. Hauser, "The IP Micro Mobility Approach", EUNICE 2000, September 2000, Eschende pp. 195-202.
[Hui97]	C. Huitema:IPv6 – The New Internet Protocol, Prentice Hall, 1999.
[Per98]	C. E. Perkins: Mobile IP –Design Principles and Practices, Addison-Wesley, May 1998.
[Ste00]	CA. Stephane, A. Mihailovic, A. Aghvami: “Mechanism and Hierarchical Topology for Fast Handover in Wireless IP Networks”, IEEE Communications Magazine, November 2000, pp. 112-115.
[Sza01]	S. Imre, M. Szalay: “Application of Ring Topology in Wireless IP networks”, CONTEL2001, Zagrab, Croatia, 2001.
[Sza01b]	S. Imre, M. Szalay: “Reliability Considerations of Micro Mobility Networks”, submitted to DRNC2001, Budapest, Hungary, 2001.
[Lui00]	L. Barriga, R. Blom, C. Gehrmann, M. Näslund: "Communications Security in an all-IP World", Ericsson Review No. 2, 2000, pp. 96-107.
[Sea00]	IETW Working Group, Context and Micro-mobility Routing (seamoby), http://www.ietf.org/html.charters/seamoby-charter.html