Corporate Wireless IP Telephony

RAÚL GARCÍA HIJES

KTH Information and
Communication Technology

Master of Science Thesis
Stockholm, Sweden 2005

IMIT/LCN 2005-20

Corporate Wireless IP Telephony

Raúl García Hijes

Stockholm, Sweden
29 July 2005

Supervisor and Examiner: Professor Gerald Q. Maguire Jr.

Abstract

IP telephony is defined as the transport of telephony calls over an IP network. IP telephony exploits the integration of voice and data networks. However, enterprises are still reluctant to deploy IP telephony despite the potential increase in productivity and reduction of costs. The principal concerns are: can IP telephony provide the same level of performance in terms of security, reliability, and scalability as traditional telephony?. If so, are its proclaimed benefits such as flexibility and mobility cost-effective?.

The aim of this thesis is to analyze how to deploy IP telephony in large corporations - while providing the necessary security and facilitating mobility. Through the different parts of this thesis, we will analyze the applicable technologies, along with their integration and management. We will focus on the essential requirements for an enterprise of scalability, reliability, flexibility, high-availability, and cost-effectiveness.

The massive changes brought about due to the deregulation of telecommunications in nearly all countries, the increasingly global nature of business, and the progressively affordable and power technology underlying information and communication technologies have lead to increasing adoption of IP telephony by residential and commercial users. This thesis will examine these technologies in the context of a very large distributed corporation.

Sammanfattning

IP telefoni är definierat som transporten av telefon samtal genom ett IP nätverk. IP telefoni utnyttjar integrationen av tal och data nätverk. Dock är affärsföretag fortfarande motsträviga till att införa IP telefoni trots potentiell ökning i produktivitet och minskade kostnader. Huvud bekymren är: kan IP telefoni tillhandahålla samma nivå av prestanda med avseende på säkerhet, tillförlitlighet, och skalbarhet som traditionell telefoni? Och i så fall, är dom proklamerade fördelarna flexibilitet och rörlighet kostnadseffektiva?

Målet för detta examensarbete är att analysera hur IP telefoni kan införas i stora affärsföretag - medan samtidigt tillhandahålla nödvändig säkerhet och främja rörlighet. Genom olika delar av detta examensarbete, analyserar vi tillämpliga teknologier, inklusive deras integrering och skötsel. Vi kommer att fokusera på de grundläggande kraven för ett affärsföretag gällande skalbarhet, tillförlitlighet, flexibilitet, hög tillgänglighet, och kostnadseffektivitet.

Dom massiva förändringarna frambringade i och med avregleringen av telekommunikation i stort sett alla länder, affärsverksamhetens alltmer globala natur, och de progressivt kostnadseffektiva och kraftfulla underliggande teknologier bakom informations och kommunikations system har lett till ökande adoptering av IP telefoni av både privata och kommersiella användare. Detta examensarbete undersöker relevanta teknologier i samband med mycket stora utbredda affärsföretag.

Acknowledgements

First of all I would like to express my most sincere gratitude to my project advisor, Professor Gerald Q. Maguire Jr., for his time, his encouragements, and for his inestimable comments.

Of course, special thanks to my parents, Jose Luis & Elena, for their love, for giving me so many opportunities, and for their unconditional support. This wouldn't have been possible without you.

Thanks also to Tero Rautiainen for his opposition and for his swedish translation.

Finally, I want to thank to my family and friends, those in Spain and those here in Sweden, for supporting and encouraging me in the difficult moments. Special thanks to Adrián, David, and Juansa, my "swedish" family.

Table of contents

List of Figures and Tables

Figures

Tables

1. Introduction

1.1. General Overview

Business success of large corporations relies, more and more, on their communication infrastructure. Information exchanges between enterprise branches, inter-business transactions, and relationships with suppliers and consumers are directly dependent on the enterprise's communication infrastructure. Traditionally, this infrastructure consisted on two separated networks: an IP data network and a circuit-switched voice network. This network division complicated management and maintenance –increasing associated costs. However, the solution to this problem, integration of both networks into a single voice *and* data IP network, has become feasible with the appearance of IP telephony and Voice over IP (VoIP) technologies.

Although both terms (IP telephony and VoIP) are often used as synonyms, we could difference them by defining IP telephony as the complete solution (including servers and clients) that makes use of VoIP technologies to transport telephony calls over an IP network. With the deployment of IP telephony, corporations can take advantage of their own data infrastructure and Internet to route calls between employees, thus decreasing costs. This cost reduction comes not only from using the corporate data network and Internet to route calls between far away enterprise branches at a much lower cost than using the traditional Public Switched Telephone Network (PSTN) infrastructure, but also by offering simplified maintenance (as there is only a single infrastructure to maintain - rather than two). In order to provide these VoIP capabilities we will focus our study on the use of the Session Initiation Protocol (**SIP**).

The use of a public shared infrastructure such as Internet has, on the other hand, one main drawback, security. **Security** has always been a basic requirement of a corporate network due to the characteristics of the information which an organization must utilize. However, without appropriate measures the use of a shared communication infrastructure could compromise the security of this communication. This has become a more pervasive concern due to telecommunications deregulation (as there are more parties involved and the trust relations are no longer as simple as they were) and due to the extension of services to wireless communication links (with their shared transmission medium). That's why we will analyze how to make our corporate network secure by the use of Virtual Private Network (VPN) technologies. We will focus not only on securing voice and data communications, but also on securing the corporate network from 'compromised' endpoints and unauthorized users.

Another main part of this thesis is **mobility**. Deploying mobility solutions allows corporations to extend their 'office' to mobile workers, which is essential for large enterprises because a large portion of business activities are handled outside the corporation's physical boundaries. Additionally, wireless users need constant access to corporate resources while roaming from different networks.

In addition to security and mobility, there are other aspects which might not be so important in other environments, but are *essential* for any corporate network. In fact, they are always carefully considered by corporations when deciding whether or not to implement a solution. These requirements are scalability & flexibility, reliability & high-availability, simplified management, and cost-effectiveness.

- **Scalability and flexibility** means that it must be possible to adapt the network to the growth of the company, especially regarding the number of employees, thus supporting changing from one network configuration to another as required without interrupting the on-going use of the network, except perhaps for those nodes directly involved in a change (i.e., configuration changes should be as invisible to users and processes as possible). Besides it must be possible to adapt to future applications and operations, as change **will** happen.

- **Reliability and High-Availability** are also essential aspects of an enterprise network. It is very important to maximize them since increasingly the operations of the business depend upon network connectivity. One of the main goals in the deployment of VoIP in a corporate network is to provide a similar grade of availability as traditional telephony solutions. This is important because business users are used to fairly high availability, thus they *expect* it.

- **Management** can be used to increase security and reliability, as well as, help to define, modify, or enforce the corporate policies of use and meet legal requirements. As corporations grow, their infrastructure becomes larger and more complicated to manage and the deployment of an automatic, reliable, and efficient network management system becomes essential for business success.

- The network and its operations must be **cost effective**. This requires a balance between features and cost. It also requires that the network facilitate the business operations. Network deployment must be carefully studied according to real business needs and future benefits derived from migration have to be clearly stated.

1.2. Problem statement

The aim of this thesis is to analyze the technical feasibility of deploying IP telephony in large corporations. The solution has to provide security and enable mobility. Besides, essential requirements of corporate environments, such as scalability and flexibility, reliability and high-availability, easiness of management, and cost effectiveness have to be carefully considered.

We will analyze and compare different technologies and their integration as a means to achieve these goals. This analysis will be qualitative in general but also quantitative when required. In this latter case, we will analyze data from a large corporation in order to study server and bandwidth requirements, cost savings and scalability of the proposed solutions.

1.3. Report Outline

This thesis report in divided in 9 different sections:

- *Section 1* contains the introduction to the thesis
- *Section 2* is related to security. It covers VPN technologies, endpoint security, and admission control.
- *Section 3* studies the introduction of mobility in the network by the use of Mobile IP.
- *Section 4* describes the SIP protocol - which is one of the protocols that we will use to provide VoIP capabilities.
- *Section 5* explores different solutions to provide high availability, reliability, and scalability using the technologies explained in prior sections.
- *Section 6* analyzes integration between the different technologies (VPNs, SIP, and MIP) in our network, as well as studies the integration between the corporate VoIP network and the PSTN infrastructure.
- *Section 7* concerns management of the network, covering the different architectures and services required in an enterprise network.
- *Section 8* is a case study based on data of two large corporations. The use of concrete numbers in this section helps to clarify, as well as, to study missing aspects from the prior sections.
- *Section 9* states from final conclusions and describes possible future work.

2. Securing the network

One of the main concerns when planning a wireless IP telephony network in a corporate environment is security. Security has always been a basic need of a corporate network due to the characteristics of the information which an organization must utilize. With the deployment of wireless technologies, and consequently the use of a shared transmission medium, security becomes even more important. Considering the wireless network as part of the corporation intranet may entail major security risks due to the inherent insecure characteristics of the transmission medium, especially as some supposedly 'secure' layer 2 protocols have been compromised [64]. On the other hand, if wireless users are considered insecure and the wireless network is situated outside the intranet we can take advantage of Virtual Private Network (VPN) technologies, which have already demonstrated their security when applied to remote access from public fixed network infrastructure.

2.1 Virtual Private Network (VPN)

A Virtual Private Network (VPN) is defined by the VPN Consortium [1] as a private data network that makes use of the public telecommunication infrastructure, maintaining privacy through the use of a *tunneling protocol* and *security procedures*.

VPNs were originally motivated by corporations' need to provide secure remote access to enterprise resources to its employees at a reasonable cost. The method of decreasing costs was to use the existing public telecommunication infrastructure, but its 'public' nature demanded security methods to protect the sensitive information while transmitted over an insecure medium. This necessity for security when transmitting information over an insecure medium is one of the principal reasons that also lead to the deployment of secure VPN technologies in wireless environments. The other reason is that the use of the same technology to provide secure access both to remote and wireless users facilitates management and control of **all** users.

A conclusion is that when planning a network we should consider the wireless part of the network as *external and insecure*, situating it outside the corporation's intranet; thus we can use VPN technology to provide the necessary security. (See Figure 1)

There are two main secure VPN protocols that can be applied to the design of a corporate wireless network: IPsec [2] and SSL/TLS [3]. Both technologies provide confidentiality, authenticity, and integrity; however, they are implemented at different layers so the differences between them are noticeable and significant.

First, we will explain their fundamentals, and then we will compare their respective advantages and disadvantages when deployed in a corporate wireless environment.

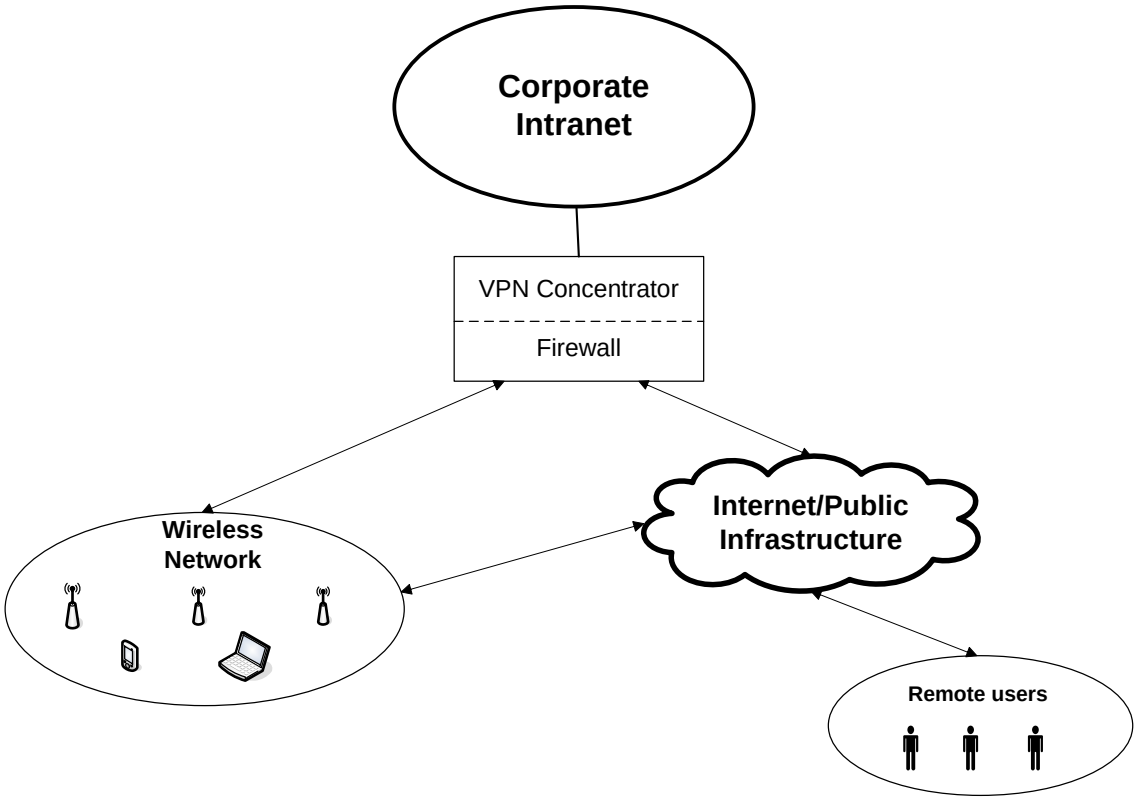

Figure 1: Wireless network and remote users situated outside corporate intranet

2.1.1. IPsec

IPsec is a set of open standards developed by the Internet Engineering Task Force (IETF) [4] to provide secure real-time communications over unprotected networks. It is implemented at the *network* layer, often inside the operating system, so all applications are protected without needing to modify the individual applications. IPsec provides confidentiality by bulk encryption algorithms such as DES, 3DES, and AES; and integrity and authenticity by hashing algorithms such as MD5 and SHA1; and identity verification by means of using Digital certificates. [5]

To provide stateful security, despite being based on a stateless protocol (IP), IPsec defines Security Associations (SAs) which are cryptographically protected connections. In order to create these SAs, the Internet Key Exchange (IKE) [6] protocol was defined. IKE provides both mutual authentication and key establishment.

IPsec Header Formats

There are two different security protocols in IPsec: Authentication protocol (AH) [7] and Encapsulating Security Payload (ESP) [8]. The former provides integrity protection only, while the latter provides encryption and/or integrity protection. A

single SA can use either of them, but not both in the same connection. To support both requires the use of two SAs.

IPsec Modes of Operation

IPsec provides two different modes of operation: tunnel and transport. Tunnel mode keeps the original IP packet intact while adding an additional IP header and IPsec information (ESP or AH) before it, while Transport mode adds an IPsec header between the IP header and the rest of the IP packet.

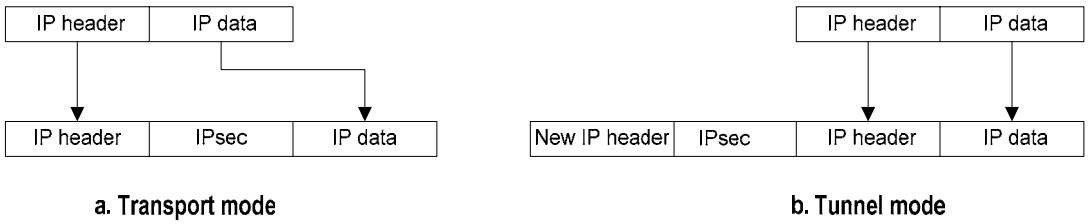

Figure 2: (a) Transport mode and (b) Tunnel mode packets

Transport mode is usually used when communication is end-to-end and tunnel model when the data is only protected along some part of the path, for example from firewall to firewall or from endpoint to firewall. (Figure 3)

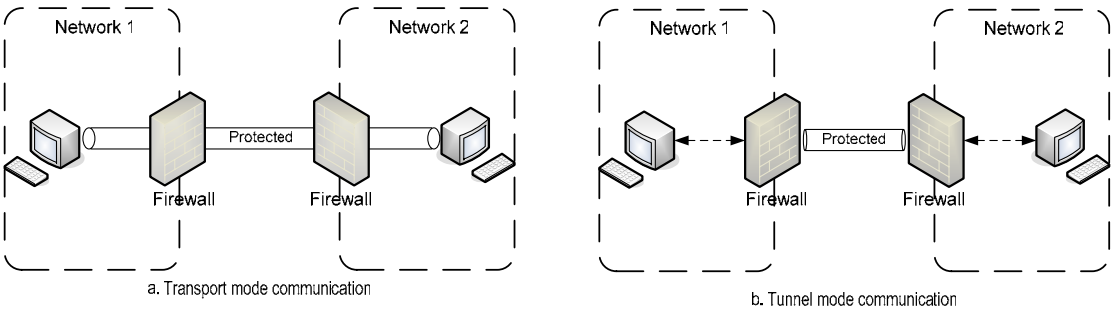

Figure 3: (a) Transport mode and (b) Tunnel mode communications

2.1.2. IKE

Internet Key Exchange (IKE) [6] is a protocol for mutual authentication and key establishment in order to create an IPsec Secure Association (SA). IKE defines two phases: During phase 1 the mutual authentication and the key establishment is done, while in phase 2 an IPsec SA is created.

IKE Phase 1

There are two types of IKE phase 1 exchanges: Aggressive mode and Main mode. The principal difference between them is the number of messages each needs to perform mutual authentication and key establishment. Aggressive mode uses three messages while **Main** mode requires six different messages. These 3 'extra' messages of the Main mode add flexibility when negotiating cryptographic algorithms. Such algorithms can be for example encryption algorithms (DES, 3DES, and AES), hash algorithms (MD5, SHA1), or authentication methods (RSA, PKI, pre-shared keys).

IKE Phase 2

Phase 2, also called quick mode, establishes an ESP or AH Secure Association using the keys established during phase 1.

2.1.2 Secure Sockets Layer/Transport Layer Security (SSL/TLS)

Secure Sockets Layer (SSL) [3] is a family of protocols which includes SSL v2, SSL v3, and Transport Layer Security (TLS). SSL was designed to provide secure and reliable communications and it runs on top of layer 4, specifically Transmission Control Protocol (TCP). SSL uses the octet stream provided by TCP and divide it into **records**, adding a header and cryptographic protection. The protocol is composed of two layers: The lowest is the SSL Record Protocol used for encapsulating higher level protocols, and the higher is the SSL Handshake protocol which manages client authentication and communication encryption. The advantage of being designed to run on top of layer 4 is that is doesn't require operating system changes and allows deployment at the user-level. SSL provides encryption (via DES, RC4,…), authentication (via PKI, RSA,DSS,…), and integrity (via SHA, MD5, …).

2.1.3 Comparison of IPsec vs. SSL/TLS

Having presented the main characteristics of both technologies, we have to analyze the advantages and disadvantages of using them in a corporate wireless network. This analysis must be done in terms of the flexibility, scalability, mobility, security, ease of management, costs, and end user complexity associated with them.

Flexibility and mobility

SSL can be used in any client with an SSL enabled Web browser installed whereas IPsec needs a specific client or kernel software. This is a very important advantage of SSL, because almost every client (desktop, laptop, PDA, or smart phone) supports one or more web-browsers. However, many SSL solutions need an Active X or Java applet to run, and remote public machines (i.e., those in airports and

hotels) may not allow installing them, nor do these machines allow the installation of IPsec clients.

Another disadvantage of IPsec vs. SSL is that the former ties the user to a machine while the latter identifies the user and not the machine, permitting greater flexibility. IPsec doesn't support dynamic and changing IP addresses so firewalls and NAT may interfere with remote access. On the other hand IPsec provides complete application access while SSL mainly focuses on Web-enabled applications.

Scalability

SSL is more scalable than IPsec; given that the latter requires configuration not only for every new user, but for every new machine. Moreover this configuration is not trivial and depends on the specifications of the user's device.

Security

In terms of security, the specific client software of IPsec provides strong device authentication while SSL identifies the user and not the machine, so users can come from untrusted machines (i.e. a public terminal situated in an airport). In this case, there is no guarantee that the client's machine has an antivirus or firewall installed, since SSL doesn't require the installation of specific client software to check the system in advance. That is why, if we want to use SSL, we need to deploy some admission control system as well as strong user authentication add-ons, in order to ensure that only endpoints complying with network security policies can access the corporate network.

Another security risk when using SSL is that sensitive information can be left in public machines after their use due to the caching system deployed in many browsers. Also an application can remain running if the user 'forgets' to log out, and could be accessed by other untrusted users. An example of a commercial solution to solve these problems is Cisco Secure Desktop (CSD) Security Suite [66]. CSD Security Suite forces logging out after a variable inactivity time, performs endpoint check before allowing remote access, and includes a component which erases all data downloaded and cached by the browser. All these security issues when deploying a SSL VPN based system are further explained in section 2.2.

Easy of Management

Easy of management is one of the key advantages of SSL vs. IPsec systems. With IPsec every new machine associated with a user needs configuration both at the core network level and at the end-user level. At this later level, IPsec client software needs configuration, upgrades, and frequent user support. These aspects will complicate management as the number of users grows, as well as, increasing the cost associated with maintenance.

End-user complexity

Almost every user has gotten used to utilizing Web based applications; consequently SSL doesn't require additional investments in user training as users will assimilate it quickly. On the contrary, as already mentioned, IPsec needs specific software which

is usually complicated and requires training and support. This maintenance will involve additional costs.

2.2. Endpoint Security

Increasing mobility means that portable devices are increasingly used to access corporate resources, and infected devices can compromise the security of the entire corporate network. Therefore, it is necessary that a system checks the integrity of the endpoints before allowing them to access the corporate intranet, i.e., only devices complying with network security policies are granted access. Examples of such security policies could be that the device had: the latest operating system patch, updated anti-virus software, and a firewall solution running. To achieve these goals, we have to adopt measures both at the endpoint level, using an application that checks the device, and at the network level using an intelligent system that applies the corresponding policies.

The deployment of such systems is easier when IPsec VPN technology is used, because as already mentioned, IPsec VPNs require the installation of specific software to run. Configuration utilities can be added to this software, so only properly secured devices can establish a VPN tunnel. On the other hand, if an SSL VPN solution is used, the endpoint device doesn't need any specific software. In this case, there are two major solutions: (a) installing checking software or (b) doing a 'remote' check.

The adoption of the first solution (a) implies that some of the greatest benefits of SSL are lost, due to the requirement for local installation of local software. Complexity of use would increase, the number of potential access devices (PDAs, Smart phones, public terminals,…) would diminish, and maintenance costs would grow (due to increased user support, upgrades, installation costs,…).

Solution (b) would require network systems that after an initial connection perform a remote scan of the device looking for any vulnerability and apply the corresponding policies. These could include: granting access, denying access, granting a restricted access, remotely updating the device,… . The advantage of this solution is that SSL's benefits remain.

2.3. Admission Control System

An admission control system is a network system that checks endpoint security, providing access to corporate network resources only to those devices compliant with security policies. It also identifies non-compliant devices, taking different measures according to security policies.

The basic function of an Admission Control System is shown in Figure 4. First the user has to log into the system. Once the user has been authenticated, the admission system checks if the user's device is compliant with corporate security policies. This scan process can be implemented in several ways depending on which VPN technology is used. With *IP sec*urity (IPsec), its specific client software can perform a local check and send the result to the admission control system, which will process and validate it.

Another possibility is to perform the scan remotely; this is more suitable for SSL based systems. If the check result is satisfactory, the device is granted access to corporate resources. Otherwise, different measures can be taken, according to security policies. Examples of theses measures are: denying access to corporate network, granting restricted access (to certain network elements or services), or allowing only connection to an specific server from which necessary upgrades and patches can be downloaded.

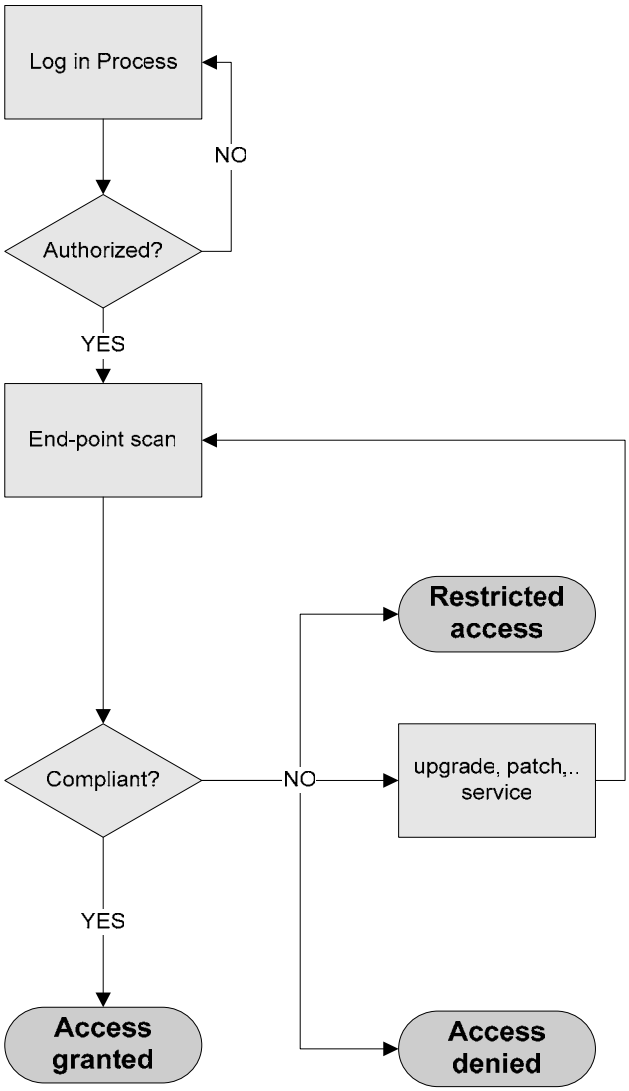

Figure 4: Conceptual function of an Admission Control System

A commercial example of this system is Network Admission Control (NAC) [9] an industry-wide collaboration led by Cisco Systems and integrated by other companies such as IBM, Symantec, Trend Micro, and Network Associates.

3. Enabling Mobility

Today, a large portion of business activities are handled outside the corporation's physical boundaries. An increasing number of employees need to access corporate resources from an airport, a bus, a hotel, or in a meeting in another city or country. Consequently, mobility is becoming one of the most important requirements of corporate networks.

Traditional network layer protocols, such as IP [10] or IPsec [11], or session layer protocols, like TLS [12] were developed without concern for mobility. Thus, to take advantage of all the benefits that mobility can provide to enterprises, the development, and adoption of new protocols, designed to provide mobile capabilities, is needed. Supporting mobility can be approached mainly in two different ways: providing mobility at the network layer (IP mobility) or at the session layer (session mobility). In the following pages we will describe both alternatives.

3.1. IP Mobility

Here, the strategy is to make IP address changes transparent to the transport layer. That means that transport layer connections can be maintained despite a mobile device roaming from one network to another.

3.1.1. Mobile IP (MIP)

Mobile IPv4 is a standard defined by the Internet Engineering Task Force (IETF) in RFC 3344 [13]. This standard defines a mechanism that allows a mobile node to keep its IP address, and thus to maintain higher layer connections, despite roaming between different IP networks. It is based on three components: the Mobile Node (MN), the Home Agent (HA), and the Foreign Agent (FA) as shown in figure 5.

Mobile node (MN)
A MN is any device (i.e., laptop, PDA, routers…) that changes its point of attachment from one network to another.

Home agent (HA)
The HA is a router on the mobile node's home network that controls communications with the mobile node. It keeps track of the MN's current location and tunnels all the packets to the MN or to a Foreign Agent in the foreign (visited) network.

Foreign agent (FA)
The FA is a router on the mobile node's foreign network. It delivers packets received from the home agent to the mobile node.

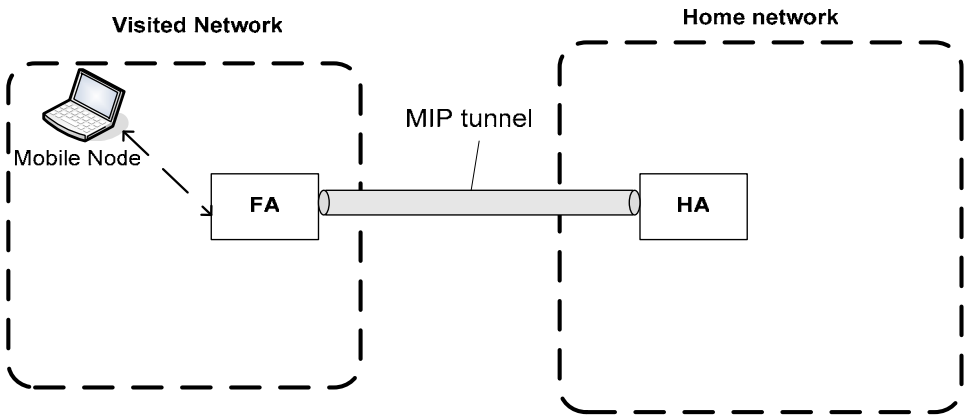

Figure 5: Basic MIP network architecture

A basic overview of Mobile IP is that a mobile node has a 'permanent' IP address regardless of its location. This address is given by its home network. Every time the MN roams to another network, a new address in this visited network, called a *care-of address*, is assigned to this MN. The association between the 'permanent' and 'tempory' addresses is stored at the home agent, thus when a packet is sent from any node to the MN the home agent intercepts it and resends it via a tunnel to the mobile node's care-of address.

To further explain Mobile IP´s operation we must describe its three main phases: (a) Agent discovery, (b) registration, and (c) tunneling.

a) Agent Discovery

During the Agent Discovery phase the home agent (HA) and the foreign agent (FA) advertise their presence on the network via Agent Advertisement messages. Optionally a mobile node can solicit these messages via an Agent Solicitation message. Based on these Agent advertisement messages, the MN determines whether it is (1) connected to its home network or (2) connected to a foreign network.

1) If the MN is connected to its home network it uses its normal IP address, without mobility features.

2) If the MN is connected to a foreign network it obtains a care-of address from this network. There are two types of care-of addresses:

   i) "Foreign agent care-of address" which is provided by the FA via its agent Advertisement messages. This address is **this** FA's IP address, so all the packets destined to the MN will be send by the HA via a tunnel to the FA. The FA will *locally* deliver them to the MN using the mobile node's 'permanent address' as the destination address (Figure 6). This approach has the advantage that many MN's can share the same foreign agent care-of address.

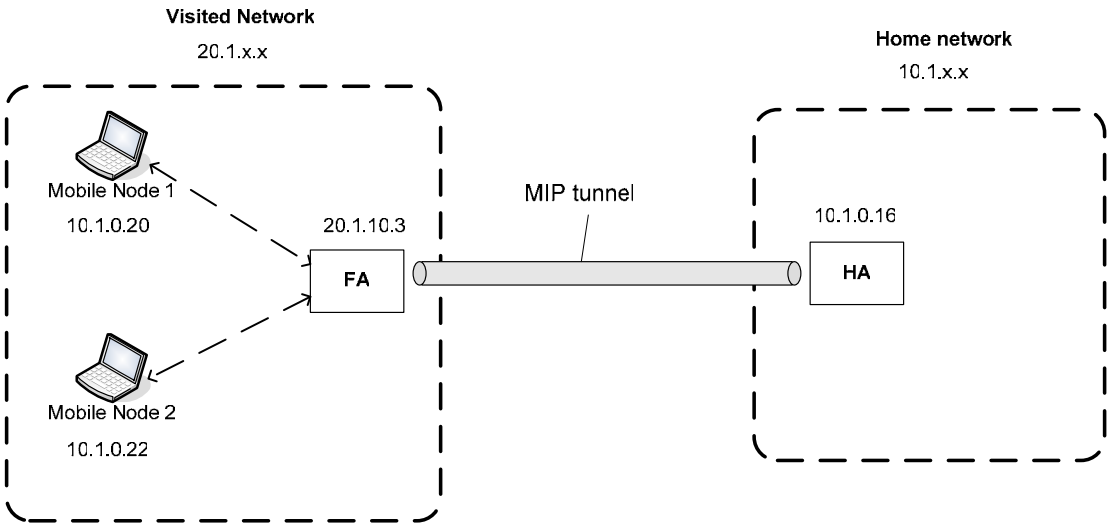

Figure 6: Mobile nodes using Foreign agent care-of addresses

ii) "*Co-located care-of address*" is an address in the foreign network obtained by the mobile node which it associates with one of its network interfaces. This approach increases the number of IP addresses needed on the foreign network (one per mobile node), but eliminates the necessity for a separate Foreign Agent. (This is sometimes referred to as a co-located foreign agent, as the MN performs the FA functions) (See Figure 7).

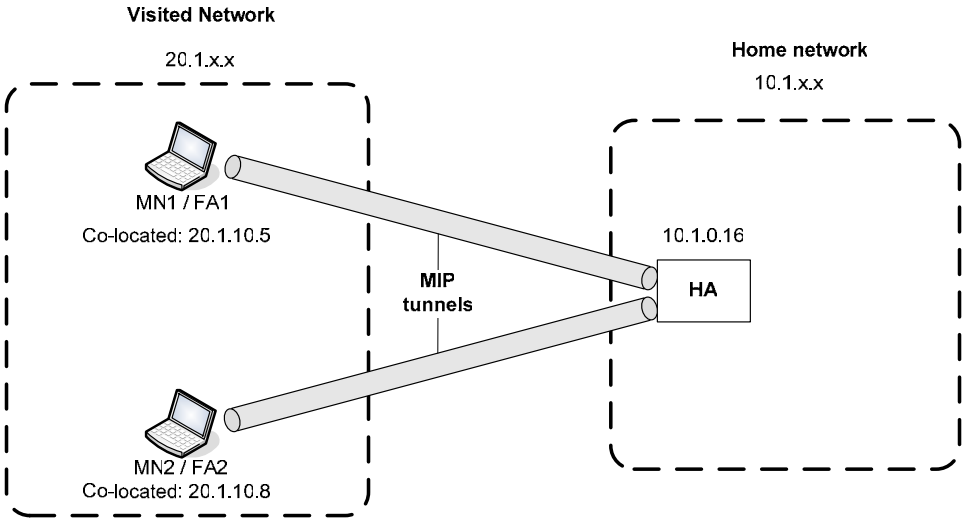

Figure 7: Mobile nodes using co-located care-of addresses

b) Registration

Once the mobile node has obtained a care-of address in the agent discovery phase, it must register this address with its Home Agent. This registration can be done directly with the HA or through the Foreign Agent (which forwards it to the home agent).
The registration process is shown in Figure 8:

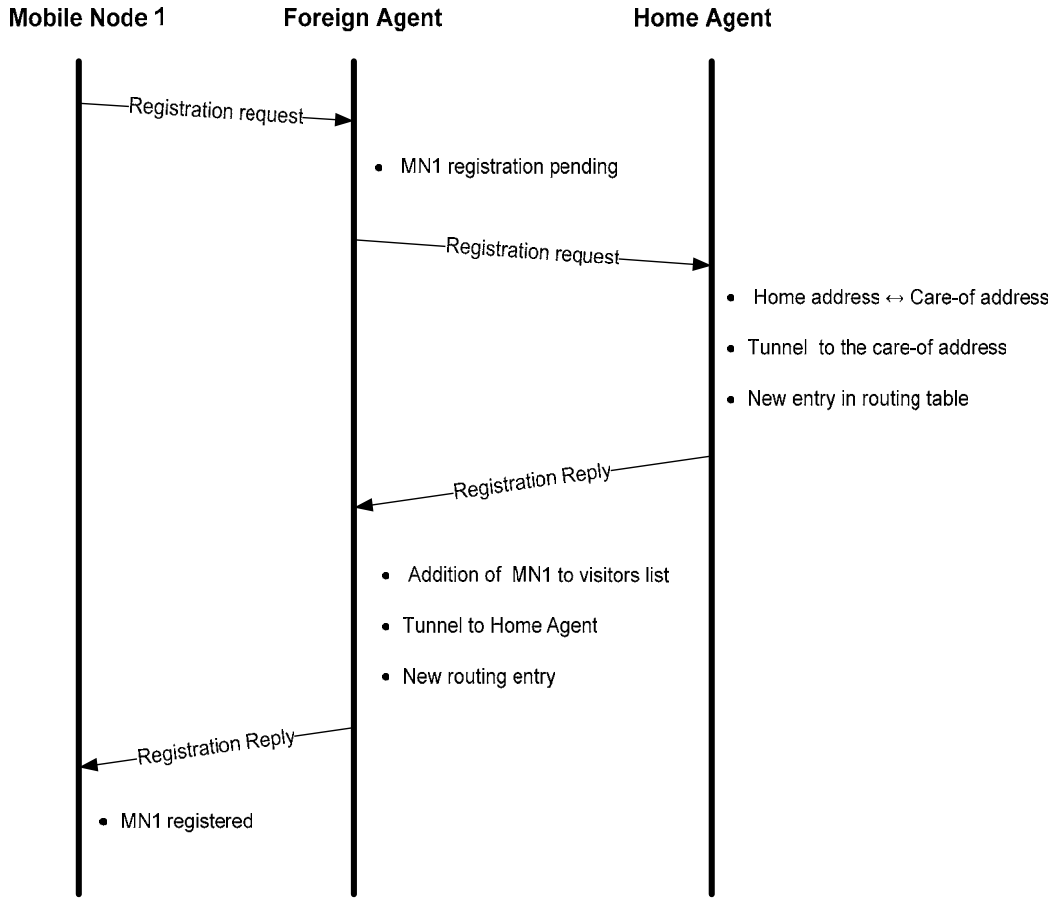

Figure 8: Registration process when a FA is used

1. The mobile node sends a registration request message directly to its home agent or to through its foreign agent (depending on the type of care-of address it is being used).
2. When a foreign agent is used, the FA verifies the registration request and if the request is valid it adds the request to a list and sends it to the home agent.
3. The Home agent verifies the request and if it is valid:
   - i. Creates an association between the MN's 'permanent' address and the MN's care of address.
   - ii. Establishes a tunnel to the care-of address
   - iii. Adds an entry to its routing table, to forward all the packets to the MN through the tunnel.

4. The HA sends a registration reply message to the MN either directly or through the FA. In this later case, the FA adds the MN to its visitor list, establishes a tunnel to the HA, and creates a local routing entry.
5. The MN receives the registration reply and if valid, considers itself to be registered.

c) Tunneling

There are two cases to consider: To or From the mobile node.

*1. Datagrams sent **towards** the mobile node*

When datagrams are sent to the mobile node's 'permanent' address (home address). These packets are intercepted by the MN's Home Agent which sends them through a tunnel (to hide the home address from intermediate routers) to the MN's care-of address. This care-of address can be the Mobile node itself, or the MN's foreign agent which will deliver the packet to the mobile node. (See Figure 9)

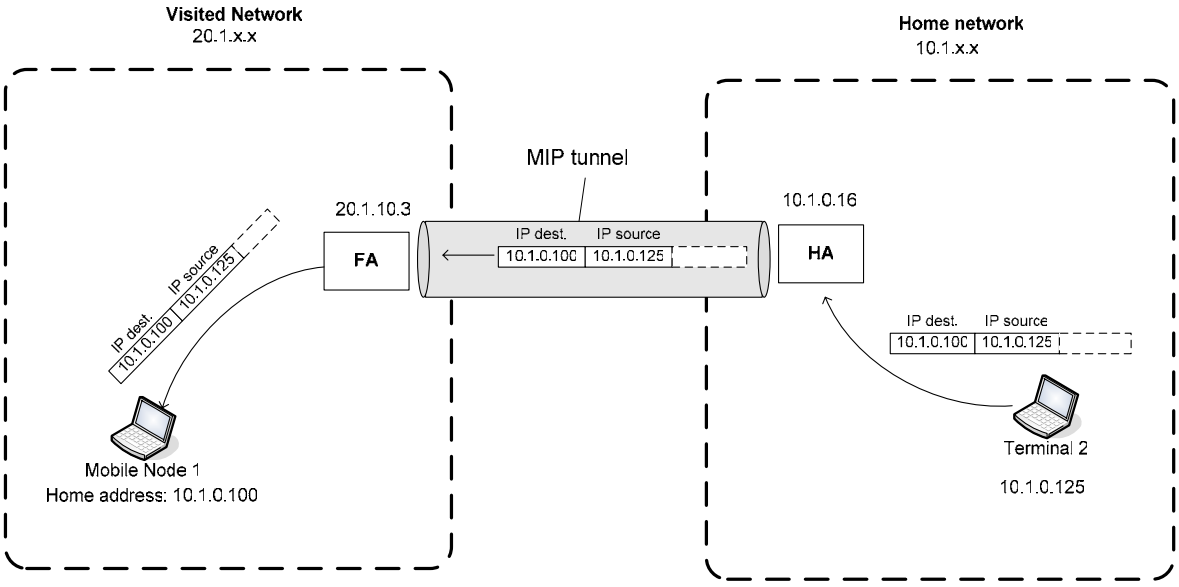

Figure 9: Example of a path for datagrams sent towards the MN when a FA is used

*2. Datagrams sent **from** the mobile node*

The mobile node uses its home address as the source address in datagrams. If a foreign agent is being used, the mobile node uses it as its default router and sends all packets through it. Then the FA either forwards the packets directly to their destination (normal IP routing) or to avoid ingress filtering problems (because the source address is not topologically from the current network) it can send the packets through a tunnel to the MN's home agent (*reverse tunneling*).

3.1.2 Host Identity Protocol (HIP)

The Host Identity Protocol (HIP) [14] is a new protocol defined by the IETF in order to provide rapid authentication and continuous communication between two hosts independently of the networking layer.

It is based on assigning to each host a new cryptographic identity called a Host Identity. This Host identity can be represented in two ways: the full Host Identifier (HI) and the Host Identity Tag (HIT). The HI is a public key that will identify the host, and the HIT is a 128 bits long hash of this public key.

By introducing this new namespace, every host will have a constant identifier *independent* of its IP address, and consequently of its point of attachment in a network. Applied to mobility, this host identifier is similar to the care-of address of Mobile IP. A host can continue to use its actual IP address as its source address, but it will include its Host identity Tag in the HIP payload of every packet. Consequently, the receiver will not only be able to identify the sender of a packet, but also knows its current point of attachment in the network. Moreover, in HIP the payload of each packet is signed with the sender's private key providing communication integrity.

However, the HIP protocol is still very 'new', and thus, commercial products based on HIP barely existent, this makes its deployment in a corporate environment very difficult at the moment.

3.2. Session Mobility

Session mobility solutions are implemented, as its name suggest, at the session layer. Thus, they don't try to maintain transport connections while roaming (as network layer solutions do). The goal is to provide recovery mechanisms at the session layer. Thanks to these mechanisms, the re-establishment of transport layer connections will be faster.

Wireless Transport Layer Security (WTLS)

Wireless Transport Layer Security (WTLS) [15] is a wireless adaptation of the Transport Layer Security protocol [12]. This adaptation is mainly focused on two characteristics typically associated with wireless devices: higher mobility and lower power/processing capabilities. To enhance mobility, WTLS utilizes optimized handshaking which provides faster re-establishment of lost connections. In wireless environments connection instability is common, and a complete session setup would consume a lot of power and processing capabilities of devices such as PDAs. Thus fast re-connection without heavy computational requirements is important. WTLS also adds datagram support, allowing the use of transaction recovery mechanisms to deal with lost packets. On the other hand, security in WTLS is weaker than in TLS. WTLS uses shorter security parameters (shorter shared key, session ID's, Client and server randoms, and a truncated version of SHA-1). This is because encryption algorithms consume power and processing resources, so at the cost of reduced security, mobility and performance are enhanced. Due to this weakness we will not consider this method further.

4. Providing VoIP capabilities with SIP

The Session Initiation Protocol (SIP) is one of the most important protocols to manage and create VoIP sessions i.e. allowing voice transmission over IP networks. Many VoIP product manufacturers are focusing their research and development on implementing SIP-based products. Additionally, SIP is considered to be simpler than the other principal VoIP standard H.323 and consequently SIP is expected to become the dominant VoIP standard. Therefore, we will focus our analysis only on SIP-based solutions. In this section we will describe SIP basics, and in section 6 we will analyze how to integrate SIP network elements in a corporate network.

4.1. SIP overview

The Session Initiation Protocol (SIP) [16] is an application-layer signaling protocol for setting up and modifying multimedia sessions (i.e. Internet telephony calls), that works independently of underlying transfer protocols. Using SIP, internet endpoints (called User Agents) can find one another and agree upon communication parameters. SIP can also invite users to join a session and, by supporting name mapping and redirection services, allows personal mobility (a user can maintain a single identifier regardless of the terminal he is using), terminal mobility (a terminal can roam between subnets), and session mobility (a session is maintained even while the terminal being used changes).

SIP is not a general purpose protocol. It only covers the signaling part of a media session establishment, and thus, it has to be used in conjunction with a protocol which carries the real-time multimedia data, such as the Real-Time Transfer Protocol (RTP) or its secure version (SRTP). SIP also makes use of the Session Description Protocol (SDP). By carrying SDP messages inside an INVITE payload, to describe the media content of the session.

To identify communication resources SIP utilizes a type of Uniform Resource Identifier (URI) [17]. A SIP URI has the form *sip:user@host* where *host* is a domain or IP address and *user* is a specific resource at this host or in this domain. This user field can be either numeric or non-numeric (i.e., sip:54321@kth.se ; sip:raul@kth.se) which provides flexibility when deciding upon a “numbering” plan. Moreover, a SIP URI can identify a user, a specific device, or an instance of a user at a given UA [18].

4.2. SIP network elements

The essential components of a SIP-based communication system are SIP User Agents. In the simplest SIP configuration two SIP endpoints (User Agents) can establish a communication session by means of exchanging SIP messages between them. Nevertheless, a typical SIP network is composed of four other basic components: proxy, registrar, redirect, and location servers. This division into ‘servers’ is purely logical, as

some of these logical entities can be located in the same physical machine. However, in the case of large corporations, with high capacity requirements, usually each server has high processing and memory demands, and thus runs on dedicated hardware.

4.2.1. SIP user agents

SIP user agents (SIP UAs) are the endpoints (i.e. IP phone sets, soft-phones) that negotiate a session's parameters by sending SIP requests and receiving SIP responses.
SIP UAs are composed of two logical entities: (1) the User Agent Client (UAC) and (2) the User Agent Server (UAS). The User Agent Client initiates a request and the User Agent Server generates a response (to accept, redirect, or reject a request).

4.2.2. SIP servers

SIP Proxy Server
A SIP Proxy Server (also referred simply as a SIP Proxy) is an intermediary that acts both as a client and as a server by making requests on behalf of other clients. A SIP proxy interprets a request, then either serves or forwards that request to another server closer to the targeted user (re-writing, if necessary, specific parts of the request message).

There are two different types of Proxy Servers: stateless and stateful.

Stateless Proxy server
Stateless proxies are servers that don't maintain a record of transactions, thus acting as simple message forwarders.

Stateful Proxy Server
Stateful proxies maintain a record of transactions by remembering information about requests they receive and send, and they use that information to process future messages associated with that request.

Redirect Server
A redirect server receives SIP requests and generates responses directing to the requesting client to contact an alternate URI or URIs.

Registrar Server
A registrar server receives registration messages (i.e., REGISTER requests) from user agents, extracts information about their location, and stores that information in a database (to implement a Location Service).

Location Service
The Location Service (or Location Server) stores information about the location of the users, and provides that information to proxy and redirect servers when requested.

SIP entities can use the Domain Name System (DNS) to locate SIP servers (See section 4.4.)

An example of the basic architecture of a SIP network (the SIP Trapezoid) is shown in Figure 10.

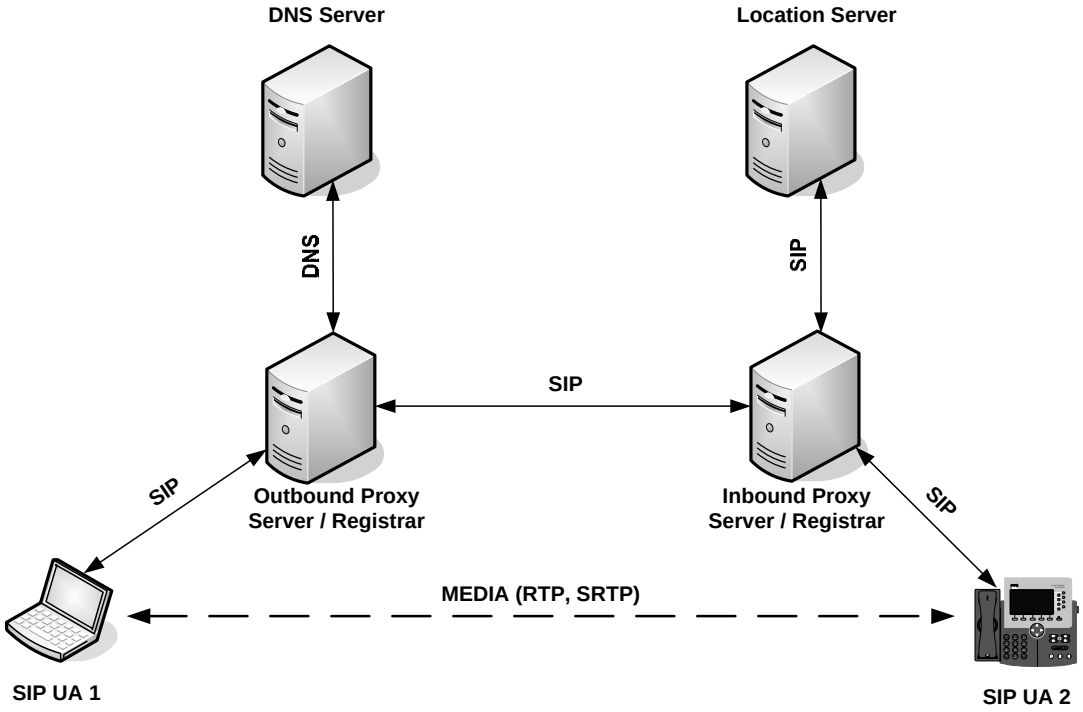

Figure 10: SIP architecture (SIP Trapezoid) [19]

4.3. SIP messages

The SIP specification [16] defines two main types of messages: REQUEST messages and RESPONSE messages. There are six main REQUEST messages: REGISTER for registering clients, INVITE, ACK, and CANCEL for setting up sessions, BYE for terminating sessions, and OPTIONS for requesting clients' capabilities. RESPONSE messages are, as usual, numerous, but they are grouped into six subtypes: 1xx Information, 2xx Success, 3xx Redirect, 4xx Request Failure, 5xx Server Failure, and 6xx Global Failure.

4.3.1. Sample Message Flows

In order to clarify SIP messaging, we are going to explain in greater detail, two important SIP processes: (a) registration and (b) a basic call.

a) Registration

Registration is the first operation that a SIP UA performs when connected to a SIP system. During Registration a User Agent sends information about its current location to its Registrar Server by means of a REGISTER message.

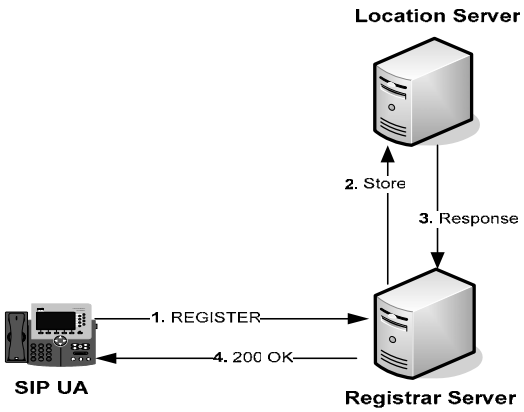

Figure 11: Example of Registration process

This registration process is very important since without this location information the system wouldn't be able to route subsequent SIP messages towards the User Agent, and thus a UA wouldn't be able to receive incoming calls. The servers (registrar server and the location server) and messages involved in a simple registration process are shown in Figure 11.

b) Basic call

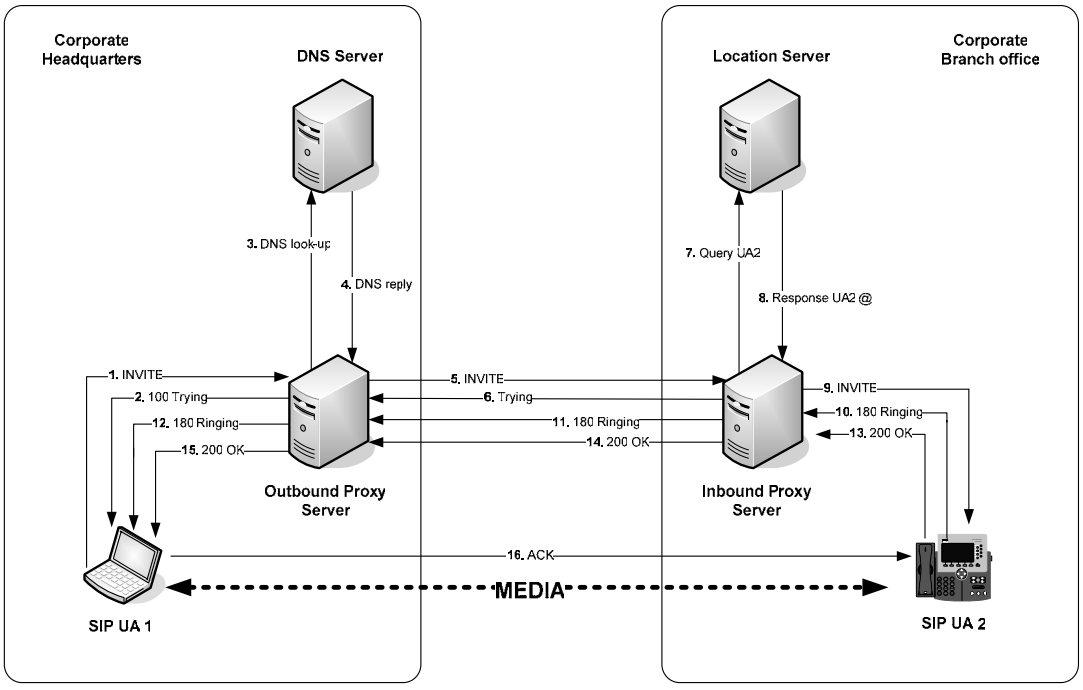

Figure 12: Basic Call message flow

To analyze the basic call message flows, we will use an example with two user agents, situated in different domains (i.e., the corporate headquarters domain and a branch office domain) as shown in Figure 12. Two users want to establish a media session.

1. INVITE
The INVITE message is sent by the caller (SIP UA1) to invite the callee (SIP UA2) to establish a media session. The details of the session the caller wants to establish (i.e., type, supported codecs, ports, and media protocol) are defined in a SDP attachment inside the SIP INVITE message body. As the caller doesn't know the current location of the callee, UA1 sends this INVITE message to the Outbound Proxy.

2. 100 Trying
When the INVITE message reaches the Outbound Proxy server, this server sends back a Trying message to the caller. This message indicates the correct reception of the INVITE message and that the Outbound Proxy is processing this request.

Messages 3 and 4
These two non-SIP messages are used by the outbound proxy to resolve the SIP URI into an IP address and port of the callee domain's proxy server (Inbound Proxy).

5. INVITE
The Outbound proxy forwards the INVITE towards the Inbound Proxy associated with the domain of the URI.

6. 100 Trying
This message indicates correct reception of the INVITE message.

Messages 7 and 8
These two non-SIP messages are used by the Inbound Proxy Server to locate the callee, based upon an earlier registration.

9. INVITE the callee
The Inbound Proxy forwards the INVITE message to the callee (SIP UA2), if the INVITE meets the requirements of UA2.

Messages 10-12. 180 Ringing
The callee's UA starts ringing and sends this message through the network to the caller, so that the caller knows the callee's device is ringing.

Messages 13-15. 200 OK
When the user of SIP UA2 accepts the call, generally by 'picking-up the phone', a 200 OK message is sent through the network to the caller. This OK message also contains an SDP part indicating the media session's parameters as selected by the callee (from those 'offered' from the caller in the INVITE).

16. ACK
With this message the caller confirms to the callee the reception of the OK message (thus completing a three-way handshake INVITE, OK, ACK). This message can be sent through the proxies or as in this example sent directly from the caller to the

callee, bypassing the two proxies. This is possible because the endpoints now know each others address from the INVITE and OK messages.

Once the media session has been established, the media data exchange can begin. This media session (i.e., carried by RTP or SRTP) is routed directly between endpoints, as they already know each others address. This aspect **must** be emphasized: data in a media session established by means of SIP does **not** need to follow the same path as the SIP signaling did.

4.3.1.2 SIP Forking

Forking is an interesting characteristic of SIP for corporate users. It is based on sending INVITE messages to more than one destination. This allows users with more than one terminal to receive calls automatically (simultaneously or sequentially) in all terminals, thus enhancing their mobility. For example, one user may have a phone in an office, a cellular phone, and a phone at home. With SIP forking, this user can program the system to forward a call to all locations simultaneously (parallel forking), or first to the office, then to the cellular phone, and finally to his/her home phone (sequential forking).

4.4. Using DNS to locate SIP servers

The use of DNS procedures in SIP allows clients to resolve SIP URIs into IP addresses, ports, and transport protocols. Additionally, it allows servers to send responses to a back-up client when the primary client fails.

We will explain the necessity of DNS to locate SIP servers by means of a simple example based on Figure 10. In this figure a SIP User Agent (SIP UA1) wants to establish a session, with another User Agent (SIP UA2) situated in a different domain. To do so, SIP UA1 communicates with a proxy situated in its domain (the Outbound proxy) which needs to forward the request to a proxy situated in the destination domain (the Inbound proxy). Using DNS procedures (i.e., DNS SRV [20]) the outbound proxy determines the IP address, ports, and transport protocol for the inbound proxy. The necessity of also determining the transport protocol occurs because SIP can run over different transport protocols, such as: TCP, UDP and SCTP. Therefore, the outbound proxy needs to choose a transport protocol that is supported by the inbound proxy. Note that in this example we assume that SIP UA1 already ‘knows’ its outbound proxy. If not, then SIP UA1 would also need to use DNS procedures.

Scalability and availability are essential in a corporate environment. There also benefit by the use of DNS procedures. Usually, a SIP proxy is not a single ‘machine’ but a cluster of proxies to provide redundancy and availability. Using DNS SRV we can associate a priority and weight with every server, thus, providing redundancy and scalability. We discuss this aspect in more detail in next section.

5. Providing Reliability, Availability, and Scalability

5.1. General requirements

A foundation for obtaining high availability and reliability is correct planning of the corporate policies and procedures for the network devices and users. Firstly there must be adequate resources, both for current needs and for growth, over estimation will provide greater scalability, at the expense of increased costs so there should be a compromise between desires and costs.

Redundancy is important because critical services, information and people must remain available even in the event of failure of network and other equipment. Thus automatic recovery methods are useful. For some settings and needs the existence of a secondary network based on other technology (such as wired or GSM), may be needed in the event of a general failure of the wireless network in other to avoid a complete absence of communication channels.

To obtain scalability and flexibility when planning the use and deployment of wireless networks, it is important to analyze the present and future capacity needs, and intelligently over estimating them considering scalability and the time to procure & install additional capacity. It should be possible to design facilitates so that increasing the number of servers or changing their configuration has minimal negative impact on network performance or operations. Furthermore, the reliability of subsystems must be analyzed from three different aspects: hardware, software, and power supply.

Hardware reliability

Today the mean time to repair (MTTR) is much more important since individual subsystems are low in cost so spares are not expensive. Hence relatively few devices require optimization of a low the mean time between failures (MTBF),

Software reliability

There are some important parameters related to the reliability of the software, such as frequency of crashes or time to reboot. Carefully study of the compatibility between platforms and protocols in the network is also important.

Power Supply

Continuous power is indispensable for the correct operation of the corporate network and computing systems since without it all the equipment is useless. Thus a fundamental requirement is Uninterruptible Power Supply (UPS) systems, generator backup, auto-restart capability, and of course UPS system monitoring.

Additionally, back-up sites don’t have common power sources, whether by situating them far away geographically or by powering them by independent generators. In this way, if the main network elements suffer a power failure, back-up sites will remain

available. This is an area where the corporate environment of distributed sites is a major advantage.

5.2. MIP Home Agent redundancy

In Mobile IP [21], as explained in section 3, when a mobile node is away from its home network, the Home Agent creates a binding between the mobile node's permanent address and the mobile node's current care-of address in the visited network. Using this binding table the Home Agent can forward all the packets destined to the mobile node to its current care-of address. However, if the Home Agent fails, then all the mobile nodes registered with it will lose connectivity. That's why a Home Agent could be a single point of failure, thus it is essential to introduce redundancy to prevent such a failure. Two alternatives to realize this redundancy have been studied: the Virtual Home Agent and Distributed Home Agents.

5.2.1. Virtual Home Agent

This alternative is based on the existence of a secondary (redundant) Home Agent (HA). This Home Agent remains in stand-by status while the primary Home Agent is working. If the main HA fails, then the secondary HA assumes the role of active HA (see Figure 13).

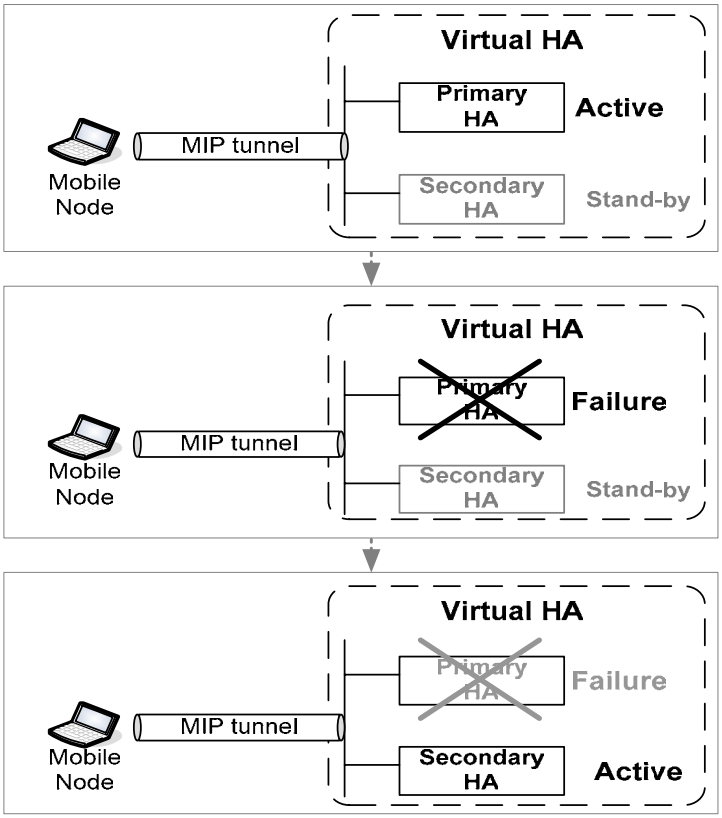

Figure 13: Failure in case of redundant HA

In this configuration, both (or all Has, if there are more than two) Home agents share an IP address, thus forming a virtual Home Agent, so configuration changes are transparent to the mobile nodes. These agents also share the binding information, thus if the active HA fails, the secondary home agent's binding tables are updated and it is able to assume immediately the active role thus providing continuous service. The way in which this binding information is exchanged has been studied and some protocols have already been developed. Examples are: the *Home Agent Redundancy Protocol* (HARP) [22] which is basically an extension of Mobile IP based on the addition of three new messages: (1) Harp tcp dump, (2) Harp udp forward, and (3) Harp udp ping; and the *Mobile IP Home agent redundancy feature* [23] developed by Cisco which runs on top of the Hot Standby Router Protocol (HSRP) [24].

However, this configuration, doesn't take advantage of the existence of multiple Home Agents, in terms of load balancing, because secondary HAs only operate when the primary one fails. Nor does this scenario capitalize on the geographically distributed presence of most large corporations. In this later case, it would be desirable to have the multiple home agents distributed geographically around the corporate network so that it would be possible for the mobile nodes to choose among Home agents depending, for example, on proximity or traffic load.

5.2.2. Distributed Home Agents

As noted above, one of the main characteristics of actual enterprises, especially in the case of large corporations, is their geographically distributed presence. As well as the main office, there are corporate branches situated in different cities, countries, or even continents, and the corporate network links all these locations. Thus, when introducing Mobile IP technology, the enterprise can take advantage of its existing distributed infrastructure. Thus we can distribute the different Home Agents across the corporate network (see Figure 14).

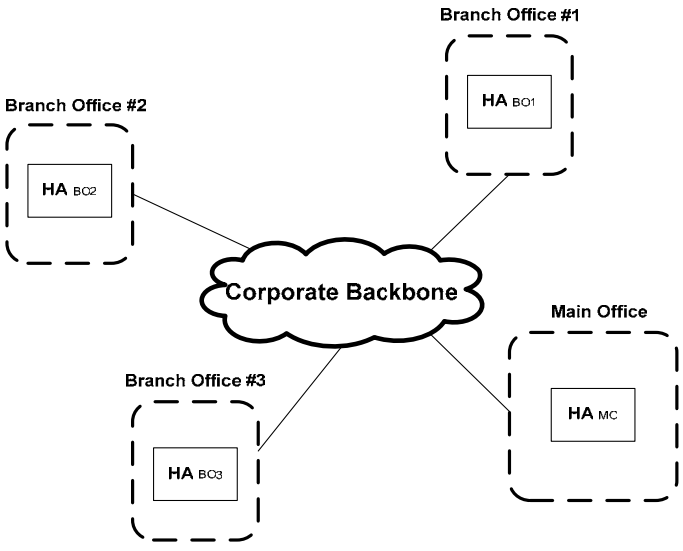

Figure 14: Distributing Home agents across the corporate network

The introduction of several Home Agents in different locations provides redundancy as well as enables load sharing and regional registration policies. Examples of these policies could be that a mobile mode should register with its nearest home agent, thus reducing network latency; or if the closest Home Agent is overloaded, then register with the next closest or the least loaded Home Agent.

In the virtual HA solution, where the HAs are physically co-located, a local (power, network, …) failure could affect the entire network, as these HAs will likely share fate, but in the distributed solution the Home Agents are far apart, thus if a part of the network fails, only the Home agents situated in some locations will be affected. Similarly, fires and power failures are unlikely to affect all HAs at the same time. Thus, the distributed solution is more robust than the virtual solution.

5.2.2.1. Dynamic Home Agent Assignment

Mobile IPv4 Dynamic Home Agent Assignment [25] is a mechanism to dynamically assign an optimal HA to a mobile node for a given mobile IP session. This mechanism is based on extensions to the mobile IP messages, such us the registration request and registration reply messages. Mobile nodes must also obtain a dynamically assigned home address in order to be assigned a dynamic home agent. Thus, the use of the Network Access identifier (NAI) extensions of IPv4 [26] is mandatory (see section 5.2.2.2). There are two alternatives for this dynamic assignment: Dynamic HA assignment and HA redirection.

a) Dynamic HA assignment

We will use a specific example (a mobile node using a Foreign agent care-of address) to explain the message exchange (see Figure 15).

1. The MN sends the Registration Request to the FA. In this request the Home Address field can be set either to all ones to indicates preference of a HA in the home domain or to all zeros to indicate no preference about HA. However, if the MN knows the IP address of its desired HA it can add that address in the Requested HA extension.

2. The FA receives and forwards the registration to a HA (Requested HA). If the Requested HA extension is present the registration request is sent to this HA address. If the extension is not present the FA determines the Requested HA.

3. The HA processes the Registration request and if it accepts the request creates a mobility binding and becomes the Assigned HA for that MN. Then, the Assigned HA send a Registration reply to the FA containing its IP address in the field HA address.

4. The FA forwards the Registration Reply to the MN. The MN extract the Assigned HA address from the HA Address field, and uses that address for the remainder of the session.

5. The MN sends later Re-Registration and De-Registration directly to the Assigned HA.

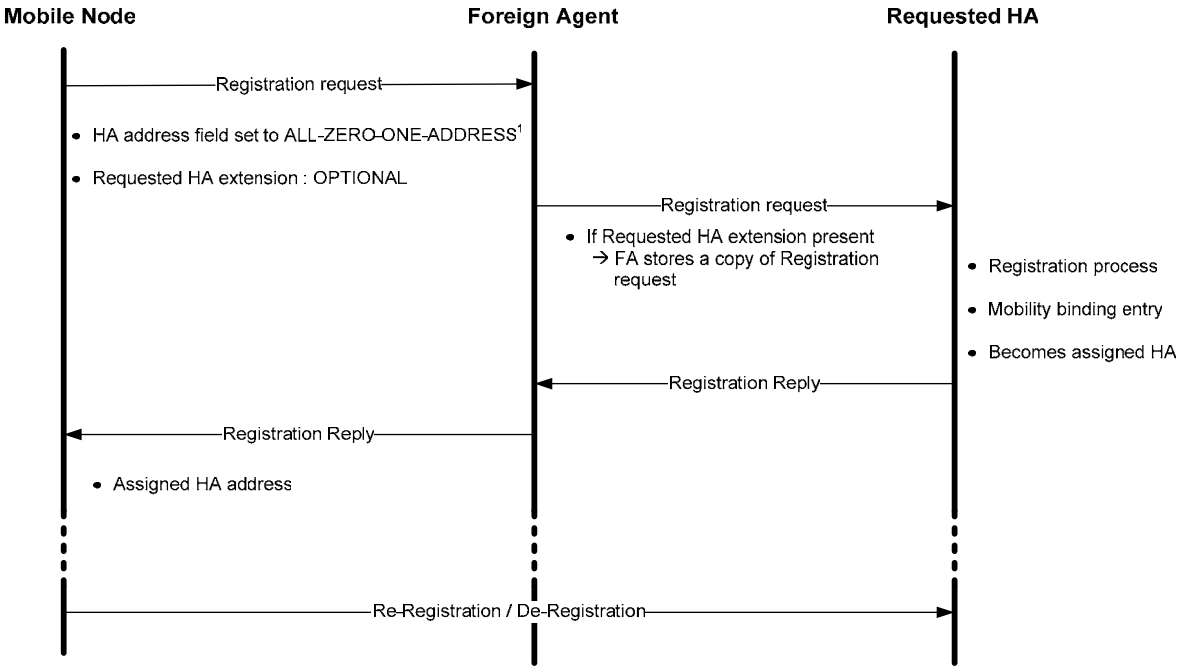

Figure 15: Dynamic HA assignment message exchange example

b) HA redirection

This redirection occurs when the requested Home agent doesn't accept the registration request, but redirects the mobile node to another home agent (Redirected HA).

We use the same example as the previous section (MN using FA care-of address) to explain the message exchange (see Figure 16).

1, **2**. The first two messages are similar as before.

3. When the Registration Request arrives at the HA it rejects it. This can be because local configuration or administrative policy directs the HA to refer the MN to another HA. Consequently, the HA sends a Registration Reply reject to the FA adding an extension to this message where indicates the address of the Redirected HA.

4. The FA forwards the Registration Reply to the MN.

[1] ALL-ZERO-ONE-ADDRESS: IP address 0.0.0.0 or 255.255.255.255. An address of 255.255.255.255 indicates a preference for an HA in the home domain. An address of 0.0.0.0 indicates no preference.

5. When the MN receives this Registration Reply reject it authenticate it and extracts the HA address from the redirected HA Extension. Then the MN sends a Registration Request to the Redirected HA.

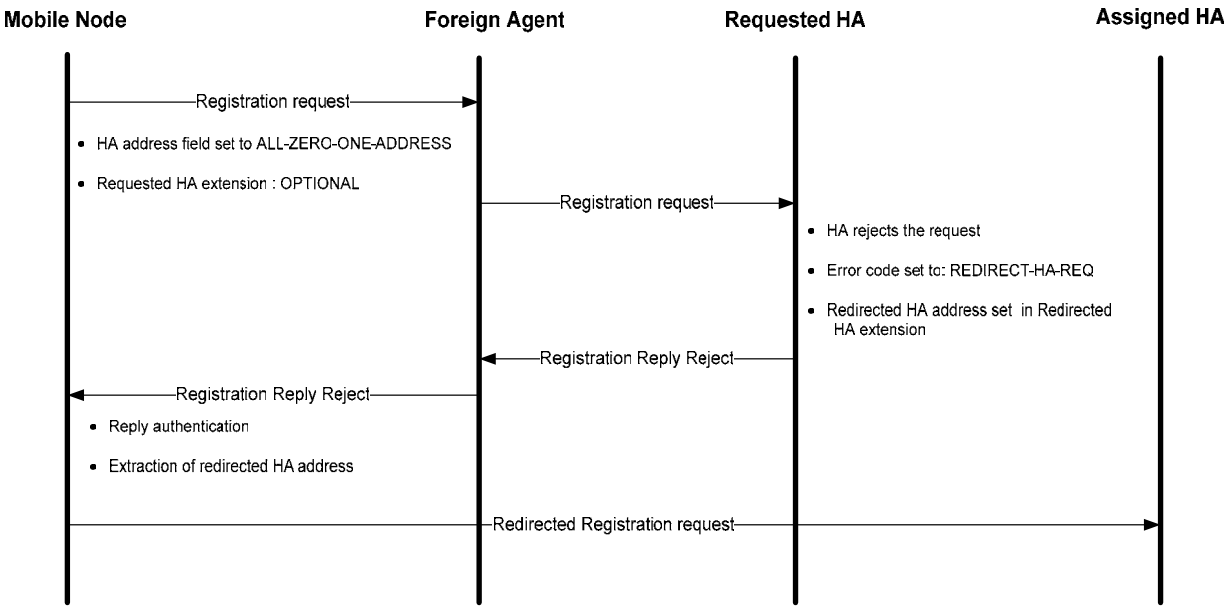

Figure 16: Example of redirected HA message exchange

5.2.2.2. Mobile IP Network Access Identifier (NAI) Extension for IPv4

The NAI Extension for MIPv4 is based on the addition of a Network Access Identifier (NAI) field in the Mobile IP registration request message. A Network Access Identifier [27] is a unique userID submitted by the client during a PPP authentication. It was developed in order to enhance interoperability of different roaming and tunneling services. By using the NAI, the mobile node is uniquely identified without need for a fixed home IP address. Thus, home IP addresses can be *dynamically* assigned to mobile nodes. To request that a home IP address is assigned, when the NAI extension is used, the home address field in the registration request is set to 0. The registration reply will contain the assigned home address. This also requires the use of dynamic DNS (DDNS), so that the binding between the host name and the dynamically assigned IP address can be made.

Some benefits of implementing a dynamically assigned home agent are:

- ***Decreases the latency between the home agent and the mobile node***
  When dynamic assignment is not used and a MN is in a visited network far away from its home network, then the signaling delay for registrations can be long. Additionally, all the traffic between the home agent and the node will travel over long distances, thus increasing delay and generating unnecessary network traffic. If a closer home agent is assigned to the mobile node, then the distance between

the HA and MN will be shorter and delays will be reduced. Of course the critical issue is the delay on the path between the MN and CN.

- ***Load balancing mechanisms between home agents can be implemented***
  In some situations, such as when there are many mobile nodes in the area associated with a home agent, it may be desirable to assign a distant but less loaded home agent, thus trading network delay for processing delay.

- ***Possibility to implement various network policies***
  The dynamic assignment of a home agent allows for different administrative policies, what in the case of corporations can be very useful. Examples include: to allow only 'high priority' registrations (i.e., those mobiles nodes belonging to high ranking employees) at a specific HA when it is overloaded, or giving preference to local users (those belonging to that branch office) vs. visiting users.

Other mechanisms by which a Mobile node can choose an optimal Home agent have already been commercialized, for example the *Priority HA assignment feature* developed by Cisco [28]. Where a Mobile node can be configured to use multiple home agents with different priorities, based on availability or proximity. The selection is based on:

- Each HA having an access list with the foreign care-of addresses of its region.
- The MN sends a registration request to the best HA which subsequently will accept or deny the request depending on whether the MN's care-of address is in its access list or not.
- If the MN is unable to register with its first choice for HA, it sends a registration request to the second best HA, and so on.
- If for any reason the MN is not able to register with any home agent, then the MN will wait for an agent advertisement message, then it will again start the registration process.

5.3. Gateway redundancy

Gateways are also single points of failure. Since reliability and availability are essential needs of corporate networks, and large corporate networks are usually geographically distributed and the number of users is high (and can grow quickly), the use of a single gateway for the entire corporate network is not appropriate. Thus, large corporate networks require, in a similar manner to distributed Mobile IP Home Agents deployments, the existence of redundant gateways fits together with load balancing.

5.3.1. Hot Standby Router Protocol

The Cisco Hot Standby Router Protocol (HSRP) [24] allows hosts to maintain connectivity when the first hop router fails. Using HSRP, hosts see a group of routers, known as *HSRP group* or *standby group*, as a single virtual router. A HSRP group can be formed by many routers, but only two have a simultaneously defined role. These are the a*ctive* router and the *standby* router. The active router is responsible for forwarding all the packets sent by hosts to the virtual router. If the active router fails, then the standby router assumes the packet forwarding duties and becomes the active router. Thus, another router of the HSRP group needs to be elected the standby router (see Figure 17). To minimize the signaling traffic, once the election process has ended, only the active and the standby routers periodic send HSRP messages.

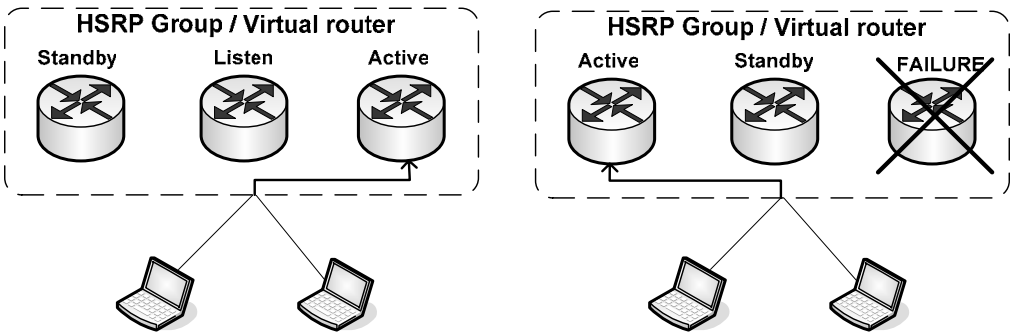

Figure 17: Failure of a router in a HSRP group

The Hot Standby Router Protocol increases availability by providing gateway redundancy, but doesn't take advantage of multiple gateways when the network is functioning correctly and there are no failures, that is, HSRP (in normal use) doesn't provide load balancing or optimized routing capabilities. To obtain load balancing with HSRP one can use multiple HSRP groups in a common subnet [29]. This is known as Multi-Group HSRP.

5.3.2. Multi-Group HSRP

Multi-Group HSRP is based on the use of different HSRP groups simultaneously. Routers are configured to belong to various HSRP groups, but can only be active in one group at a time. That is, when a router is 'active' for one group is 'standby' for the other groups (see Figure 18).

This solution, while providing load sharing and redundancy capabilities to the network, needs manual configuration and design. These tasks can become complicated and costly in the case of large networks. Additionally, corporations need to focus on their business processes rather than on the configuration and maintenance of their networks. Consequently, it would be preferable to use an automatic system that provides similar capabilities, but with lower maintenance costs.

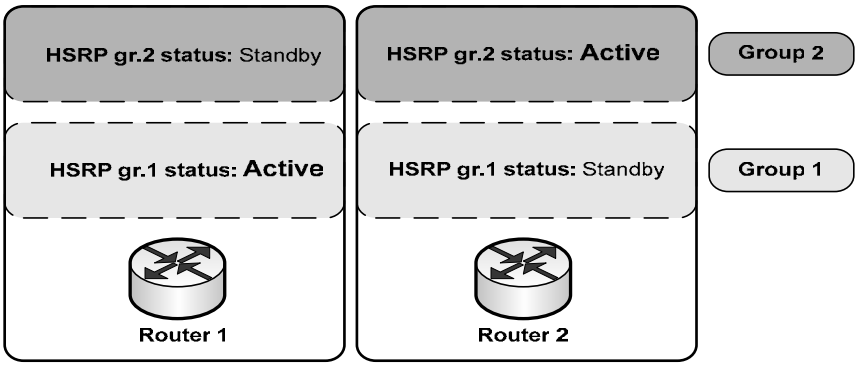

Figure 18: Example of Multi-Group HSRP

5.3.3. Gateway Load Balancing Protocol

Cisco's Gateway Load Balancing Protocol (GPLB) [29] was developed for their IOS Software as an extension to their Hot Standby Router Protocol (HSRP) that dynamically assigns responsibility for a virtual IP address and distributes multiple virtual MAC addresses to members of a GLBP group [29]. In an IP internet, a host can discover a gateway by sending an Address Resolution Protocol (ARP) request. The gateway that receives the request, answers with an ARP reply that includes its MAC address. The host will cache then that address and will now send the packets through that gateway (while it remains the best gateway). One means to obtain load balancing capabilities is to utilize the different gateways in a group by having them reply in a given order. Thus, the traffic will be shared over all these gateways. To control this, there is a Gateway Load Balancing Protocol master controller (called Active Virtual Gateway) that assigns virtual MAC addresses to the routers in a GLBP group and handles all ARP requests destined to the virtual IP address of the group. The use of virtual MAC addresses provides failure transparence. That is, when the primary gateway (A) fails, the secondary gateway (B) will assume its role, and will forward the packets destined to A's virtual MAC address. Consequently, the gateway failure will be unnoticed by the end nodes.

Virtual MAC addresses

Cisco's Gateway Load Balancing Protocol defines, in order to avoid addresses' conflicts, virtual MAC addresses in the form: *0007.b4yy.yyyy* where *yy.yyyy* equals the 24 less significant bits. The first 6 bits are set to 0, the next 10 bits specify the group, and the last 8 bits correspond to the virtual forwarder number. These 8 last bits allow the existence of 255 different virtual forwarders in each group; however, currently configuration is capped at 4.

Note that in intra-group communications, routers use their real IP and MAC addresses to avoid address conflicts.

ARP considerations

The use of ARP techniques, such as proxy ARP or gratuitous ARP, in conjunction with GPLB (which uses both real and virtual IP and MAC addresses) can involve some addressing conflicts.

Proxy ARP
When a host sends ARP requests for the IP address of the destination host to discover its MAC address, any router with a direct route to that remote host will reply to this request, on behalf of the remote host by sending an ARP reply with its own MAC address. When GPLB is used, different routers can reply, so there can be several ARP replies with different source MAC addresses. As a result the endpoint will probably use as the destination the MAC address of the last ARP reply received. This means that load balancing provided by GPLB will be lost. Therefore, GPLB is not recommended when using Proxy ARP.

Gratuitous ARP
Gratuitous ARP is an ARP request sent by a host for its own IP address to check if there is another host with the same IP address. It is used to detect duplications and thus conflicts. In GPLB all routers in the same group use the same virtual IP address, so if a router uses gratuitous ARP to resolve its virtual address all the other routers in the group will reply. In order to avoid this, when GPLB is used all gratuitous ARP mechanisms for the virtual IP address of the GPLB group members have to be disabled.

5.4. High-Availability and Scalability with SIP

A corporate IP telephony solution, in order to successfully replace 'traditional' telephony, has to satisfy essential requirements of any enterprise communication system. Thus, when designing a SIP based telephony system, high availability, reliability, and scalability have to be carefully analyzed.

High-Availability relies on two principal factors: Capacity and Redundancy. Capacity measures the volume of traffic a network can handle. In the case of voice networks capacity is usually measured in calls per second or Busy Hour Call Attempts (BHCA) -which are the number of call attempts the system supports during the busiest hour-. Redundancy is related both to capacity, since extra network capacity is needed in the event of equipment failure, and to reliability, since redundancy increases system reliability. The other essential requirement of our SIP based system is scalability. It must be possible to adapt the network to the growth of the company, especially regarding the number of employees. This growth will be seen as the BHCA increases, indicating increased capacity is needed. Thus new SIP servers will have to be added. Finally, our design must include and exploit the distributed nature of large corporations, since geographically distributing SIP servers will enhance availability (i.e., avoiding common power and local failures) and scalability.

In the following subsections we will analyze different solutions to provide such a SIP system, while considering the required characteristics.

5.4.1. First approach

The simplest solution is to configure redundant server per main server (see Figure 19). The redundant server is in a stand-by status until the active server fails. At that moment, the redundant server takes the role of active server, and the failure remains unnoticed for users. Examples of protocols used to achieve this solution are the Virtual Router Redundancy Protocol (VRRP) [30] and the Cisco Hot-Standby Router Protocol (HSRP) [31]. (See sections 5.2.1 and 5.3.1).

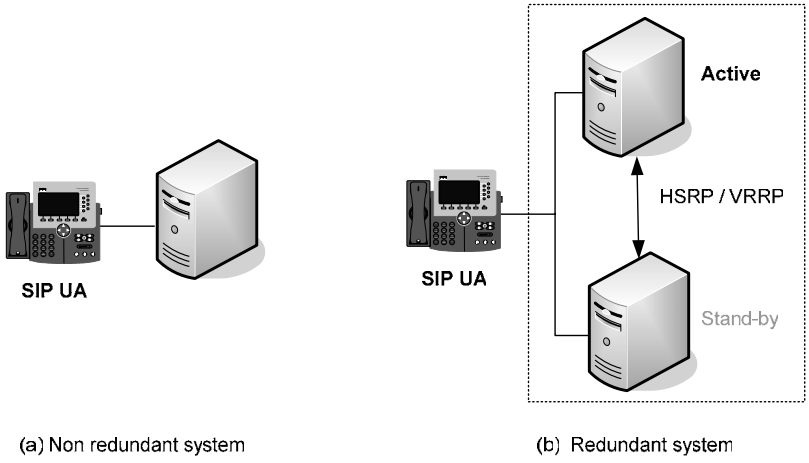

Figure 19: (a) Non redundant system and (b) Redundant system

This solution increases global reliability, but does not increase capacity since redundant equipment only becomes active when the main equipment fails. With this configuration, when capacity is reached and we want to add an extra server, we have to add yet another server for redundancy and load sharing is still impossible (configuration of two active servers and one redundant is not supported by VRRP and HSRP). Thus with four severs we still have only simple redundancy. Therefore this solution is neither efficient nor scalable.

A better solution would be the use of groups of servers. In such a group, there can be more than one active server and one or more redundant servers. With this solution capacity can be easily increased by adding new active servers to the group or adding new groups, what also increases scalability. In addition it would be desirable to use load balancing techniques to distribute the load across the active servers. This can be done by using suitable DNS procedures, specifically the DNS SRV mechanisms.

5.4.2. DNS based redundancy and load sharing

The use of DNS SRV (a DNS resource record for specifying the location of services) [20] mechanisms by SIP clients to locate SIP servers can provide the load-sharing and scalability capabilities desired for our system. A DNS SRV record lists a group of servers ordered by priority and with a weight associated with each service and domain. For example:

```
corporation.se

_sip._udp   0     50    server1.corporation.se
            0     50    server2.corporation.se
            1     0     redundant1.corporation.se
```

The numbers in the first column indicate the priority of each server. A lower value is higher priority. And the second column is the weight associated with each server (this value is used for load sharing). In this example there are three servers, two of them are preferred (priority 0) and the other is redundant (priority 1). This redundant server will only receive traffic when both of the active servers become unavailable. The traffic will be distributed among the active servers in accordance with the weight (in this case 50% for server1 and 50 % for server 2).

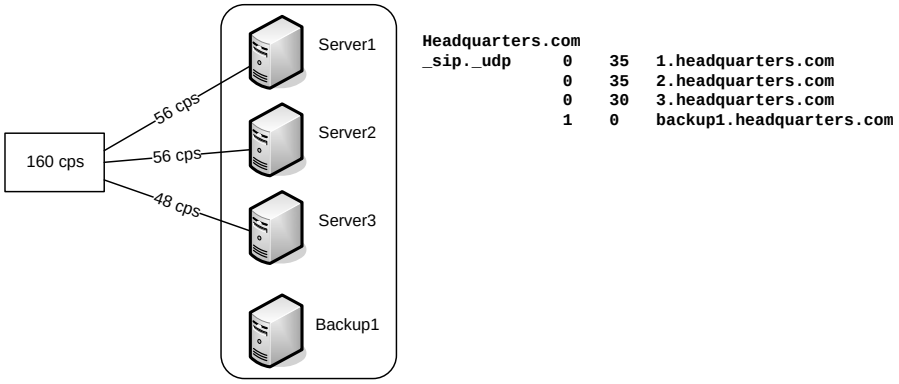

Figure 20: Example of load sharing with DNS SRV

In Figure 20 we can see the distribution of the traffic load in an example. Here each SIP server has a capacity of 100 calls per second. The system is designed to be able to process 300 calls per second with one redundant server (with 100 calls per second capacity). The incoming traffic load is 160 calls per second.

Some of the advantages of using DNS SRV with SIP servers are:

Servers don't need to be co-located
One of the main advantages of using these DNS procedures is that the servers don't need to be co-located. Thus, corporations can take advantage of their distributed nature and situate back-up servers in a different location that main servers (see Figure 21). Thus, in case of a localized failure, back-up servers will remain available.

Flexibility
It is possible to change the behavior of the entire system by modifying the DNS SRV records. These changes could: redistribute the load assigned to each server, change the group distribution of the servers, change or add back-up or main servers, etc.

Scalability
This configuration enhances scalability. Simply by adding new servers and modifying the DNS SRV records, the capacity of the system can be easily increased.

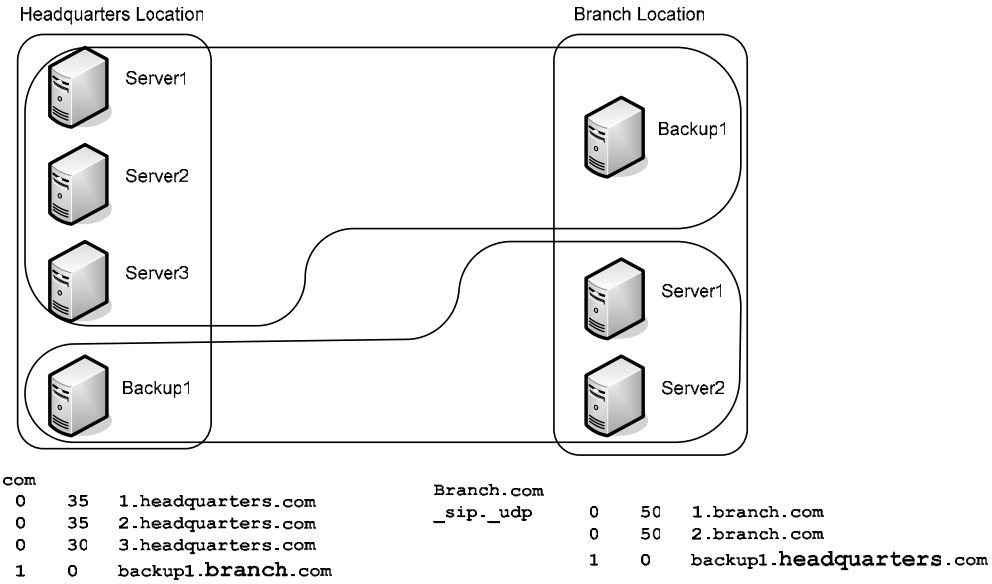

Figure 21: Example of SIP servers distributed geographically

Reliability

The existence of more than one main server and one or more back-up severs increases reliability.

For example in a configuration of 3 servers, if each server has only 99 % reliability (R_{server}) we obtain a collective reliability of:

$$R_{group} = 1-(1- R_{server})^3 = 1-(1-0.99)^3 = \mathbf{99.999}9\ \%$$

This reliability is even higher than the “Five Nines” reliability goal of traditional telephony systems.

When applying this to a SIP system we have to realize that all the servers in a group have to **share** the same REGISTER information. The two main solutions for this are shown in figure 22.

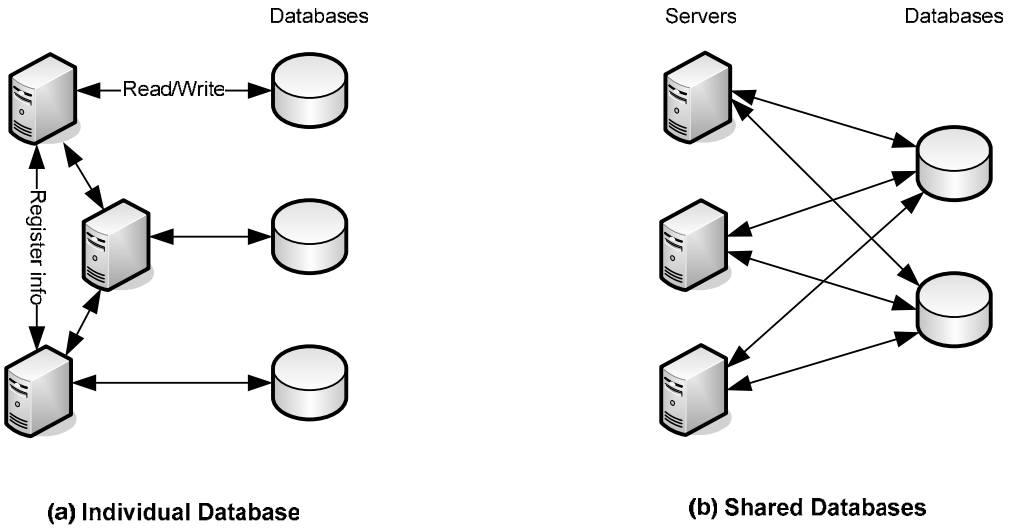

Figure 22: Example of (a) Individual database and (b) Shared database configurations

a) Each server has its own Database (Location Server). In this case every time a server receives a registration message it has to replicate it and send it to the other servers in the group, so all servers can store information about this client in its own database. The problem of this solution is that there must be as many databases as servers and every database contains the same information. Additionally, replication of REGISTER messages increases system traffic load.

b) The servers in a group share the same databases (It is highly advisable to have more than one database per group, as a redundancy measure). In this case, the server that receives a REGISTER message has to write the information in every database. But when a server has to read information (because of an INVITE message) from a database, it only has to read one of them. The bottleneck of this configuration is the time that takes to write the information in the databases, especially since all the servers of the group have access to them, this can affect scalability.

If we analyze again reliability, but now including the database reliability, we obtain:

$$R_{solution\_a} = (1-(1-R_{server} \cdot R_{database}))^{\#servers}$$

$$R_{solution\_b} = (1-(1-R_{server})^{\#servers}) \cdot (1-(1-R_{database})^{\#databases})$$

If the case of Figure 4, with $R_{server} = R_{database} = 99.5\ \%$

$$R_{solution\_a} = (1-(1-0.995 \cdot 0.995))^3 = 97.037\ \%$$

$$R_{solution\_b} = (1-(1-0.995)^3) \cdot (1-(1-0.995)^2) = 99.997\ \%$$

We see that the first solution (a) decreases reliability even though there is an additional redundant database. In general, in solution (a) reliability decreases as the number of

servers increases, while in solution (b) reliability increases if the number of servers or databases increases, therefore (b) solution is preferred.

As a REGISTER adds a binding to the database (DB), the entry for a given node must be locked during this update –while for a read the entry in the DB does not need to be locked- hence the read can (independently) come from a copy of the database. Thus the probability of failure is not multiplicative.

6. Integrating network elements

6.1. Integrating Mobile IP and IPsec VPNs

The use of Virtual Private Network (VPN) technologies provides corporate users situated outside the corporate intranet with a secure method to access enterprise resources. The biggest problem of IPsec solutions, as said in section 2, is that they don't support mobility and every time the tunnel's endpoint IP address changes a new IPsec secure association has to be renegotiated, which can be a costly process (in both time and energy). This can be a major burden for mobile users which will need to rapidly roam to different external networks during a communication session and even for other corporate wireless network users, as this network is situated outside the intranet. Thus a mobility solution like Mobile IP (MIP) has to be used in conjunction with IPsec. However, integration of both technologies hasn't been standardized yet, presenting challenges and thus, results will differ depending the enterprise needs.

One of the principal decisions is whether to run MIP above or below IPsec, i.e., to run MIP inside and IPsec tunnel or IPsec inside a MIP tunnel (Figure 23).

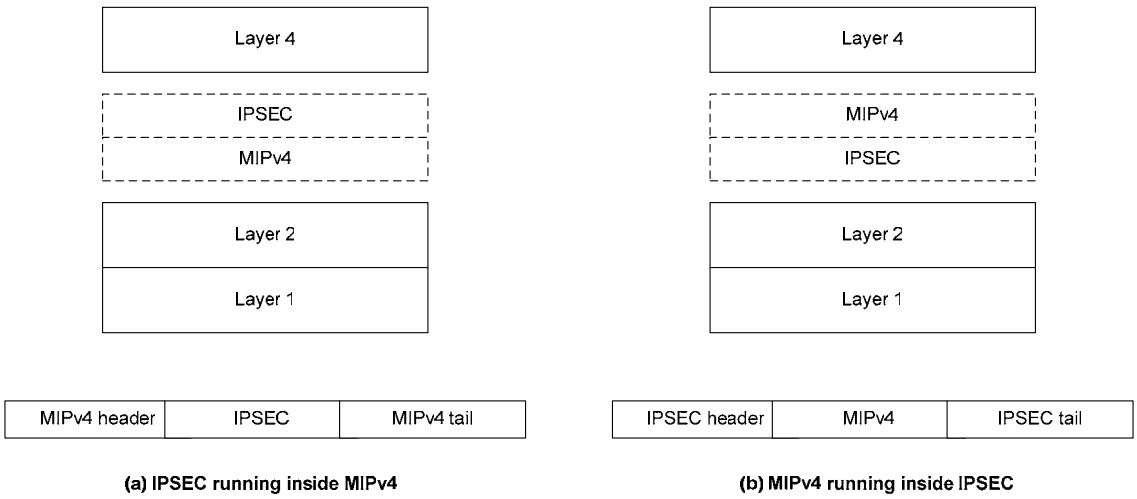

Figure 23: (a) IPSEC inside MIPv4 and (b) MIPv4 inside IPSEC

The selection of one solution or the other will depend on where the MIP agents are situated (inside, outside, or at the boundary of the corporate intranet). Therefore we will analyze these scenarios where MIP agents and VPN gateways have to coexists, presenting the problems of such integration [32].

6.1.1. MIP HA situated inside the corporate intranet

In this scenario the mobile users can only access the home agent(s) through the VPN gateway, because the home agent is situated inside the corporate intranet and all traffic between external users and the intranet must be IPsec protected. Thus, it is mandatory to use MIP inside IPSec which leads to two main problems:

a. If a foreign agent in the external network is used, then registration becomes impossible because all the traffic between the mobile node and the VPN gateway will be IPSec protected, and the foreign agent will not be able to read the MIP headers (see figure 24).

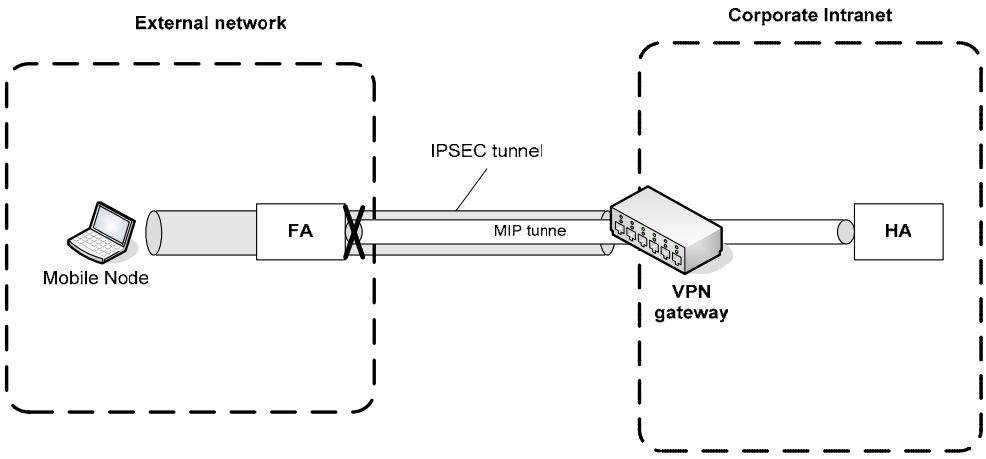

Figure 24: MIP inside IPsec with external FA

b. If the registration of the mobile node is done without a foreign agent, that is, using a co-located care-of address, the problem of above is solved, but on the other hand, the node will have to re-establish the VPN tunnel every time it changes its IP address (see figure 25), consuming more processing and adding delay.

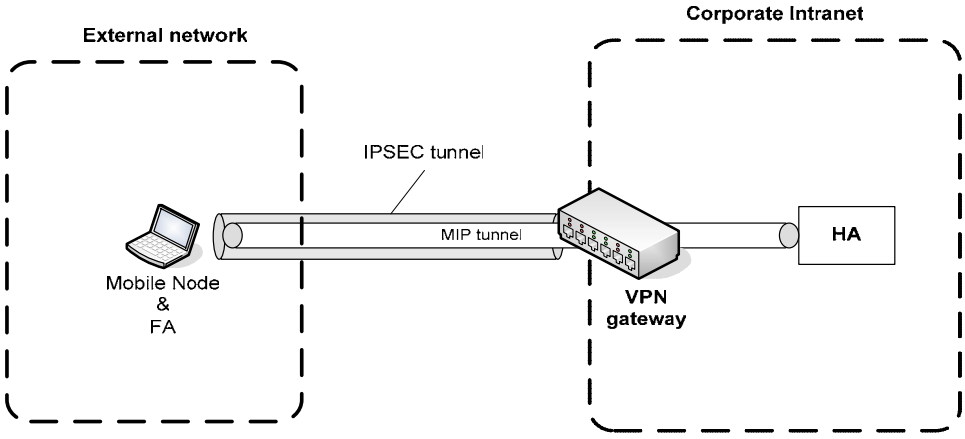

Figure 25: MIP inside IPsec without external FA

6.1.2. MIP home agent situated at the border of the corporate intranet

In this scenario MIP's home agent and the VPN gateway are both situated at the border of the intranet. The HA has a public IP address, hence it is reachable for all the mobile users.

Once again, if MIP inside IPsec is used the problems are the same as when the HA is situated inside the intranet. In contrast, when using IPSec inside MIP the VPN tunnel can take advantage of the mobility features of MIP. Thus, the IPSec connection won't be affected by the mobile node's roaming, as the IP address changes will be transparent to IPSec. However, there will be some routing problems to solve because packets from the mobile node have to go through **both** the VPN gateway and the home agent. This can be solved by running the VPN gateway and the home agent on the same physical machine (Figure 26).

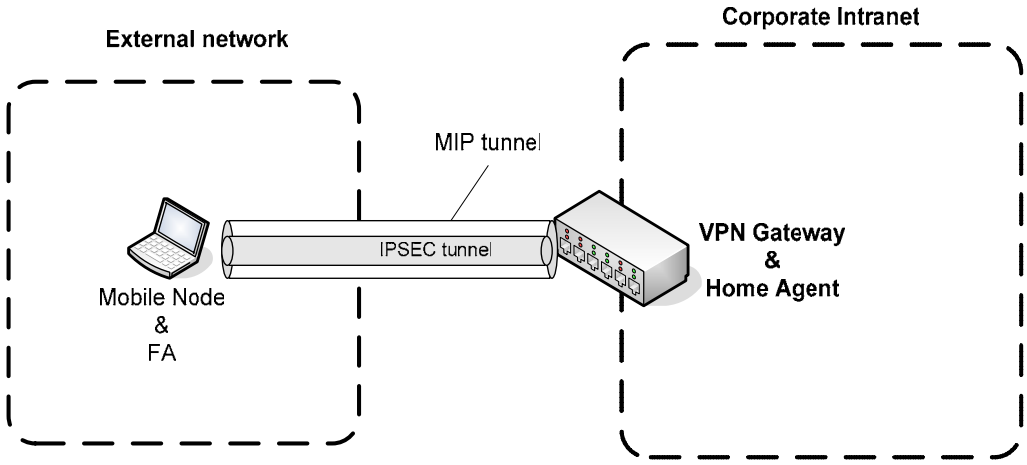

Figure 26: IPsec inside MIP when VPN gateway and HA on the same physical machine

Another problem in this scenario occurs if multiple HAs are needed for scaling because as we have pointed, if we want to avoid the routing problems the VPN gateway and the HAs have to be running on the same machine, what is not feasible as the number of HAs grows. Consequently, in case of large corporations adopting this solution is not the best option. Additionally, this approach does not facilitate multi-vendor equipment integration.

6.1.3. MIP home agent situated outside the corporate intranet

In this scenario both (a) IPsec inside MIP or (b) MIP inside IPSec can be used. However, as we mentioned in section 6.1.1 it is always preferred, if possible, to use IPsec inside MIP, so that IPsec can take advantage of Mobile IP and there is no need to re-establish an IPSec tunnel every time a mobile node changes its IP-address. Since the Home agent is situated outside the intranet there will be no problem during the

registration phase and the VPN tunnel between the MN and the VPN gateway will be established (see figure 27).

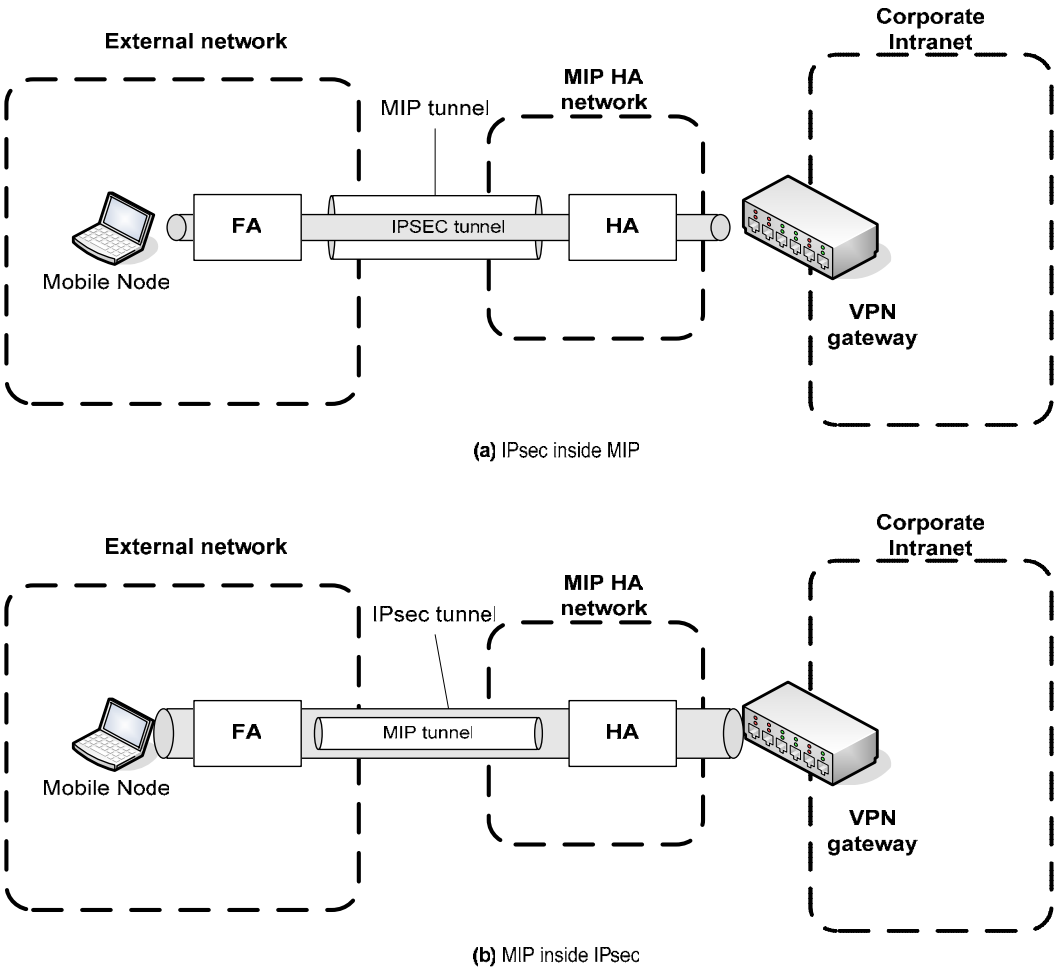

Figure 27: (a) IPsec inside MIP and (b) MIP inside IPsec when HA situated outside intranet

The conclusion reached based on all these integration scenarios is that if we want IPsec to take advantage of Mobile IP capabilities (which implies that IPsec's Secure Associations remain valid after a mobile node changes its network point of attachment), then the only possible solution is to run **IPsec inside a Mobile IP** tunnel. This choice of running IPsec inside a MIP tunnel implies that we should situate the **HA outside the corporate intranet** -- in order to avoid the registration, routing, scalability, and interoperability problems shown in the other scenarios.

6.2. Situating SIP elements in the corporate network

There are two main questions to answer when deciding where to situate SIP servers: (a) Should they be outside or inside the corporate intranet and (b) which location is better to situate each type of server? We will answer both questions in the following sections.

6.2.1. Placing SIP servers outside or inside the intranet

An important decision when designing a corporate SIP network is where to situate the SIP servers. The two main possibilities are: (1) to situate SIP servers inside the corporate intranet and thus behind the VPN gateway or (2) to situate them outside the corporate intranet and thus making them directly reachable for all users. Each possibility has some advantages and disadvantages which we will analyze.

6.2.1.1. SIP *inside* the corporate intranet

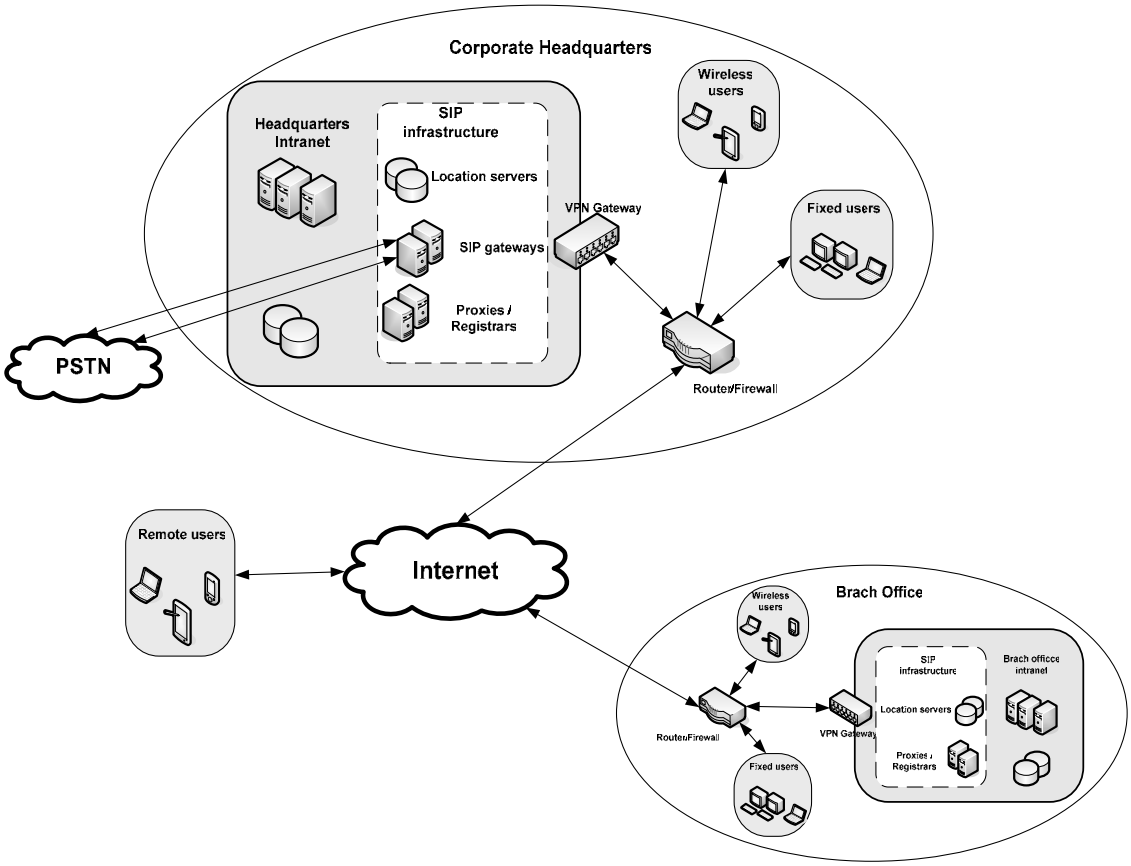

Figure 28: Example of SIP elements situated inside corporate intranet

In this configuration (see figure 28) SIP servers are situated inside the corporate intranet and thus behind the VPN gateway. Consequently, all IP-based users (wireless, remote, and fixed) need first to establish a VPN tunnel to reach the SIP servers. Note in the figure that there are *no users situated inside* the intranet, this is because we consider all users as untrusted. Therefore, every user who wants to access the corporate intranet resources must authenticate and establish a secure tunnel first. The advantages of this include identical configurations for all user computers, no need to reconfigure when moving, all users must authenticate for any access, logging of data movement is easier, etc.

The different steps in establishing communication depend on both where these users are located (in the same or different locations) and if the PSTN network is involved or not. Therefore, we will analyze the different combinations in order to extract conclusions about this configuration.

a) Both users situated in the same location

First, when any user wants to register their SIP user agent utilize a tunnel through the VPN concentrator to reach the SIP Registrar server. Assuming that both users have already registered and want to establish a media communication, the steps are as follows:

- UA1 wants to send an INVITE message to UA2. It must send that message through its Outbound Proxy server which is situated inside the intranet, so it must utilize a tunnel.

- The Outbound Proxy server forwards the INVITE to the Inbound proxy of UA2. In this case, as both users are situated in the same location, usually both servers will be inside the same intranet so no tunnel is necessary. (If users were in the same location but they were 'associated' with servers located in different intranets the process would be the explained in situation (b)).

- The UA2 inbound proxy forwards the INVITE to UA2, and for this must also utilize a tunnel to UA2.

- The following messages can use the same tunnels already created/used.

- Once the session is established, the media data will flow directly between UA1 and UA2 without traversing the intranet, and therefore no tunnel is needed. This is because, as we already mentioned, both endpoints already know each others location, thanks to the earlier SIP messages.

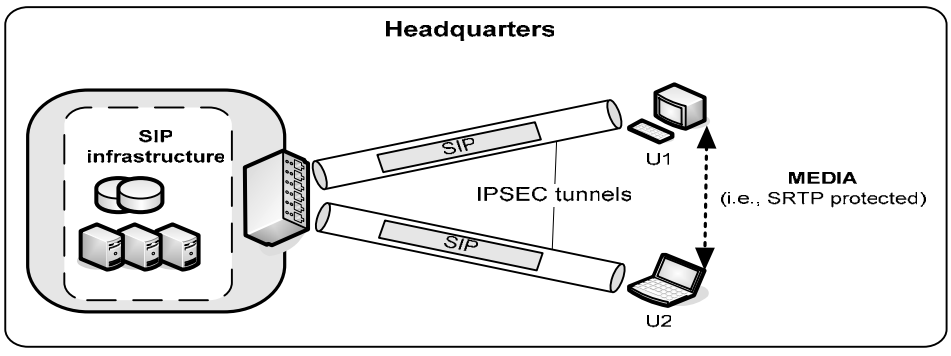

Figure 29: Example of tunneling when both users are in the same location

So in this case, as show in Figure 29, it was necessary to utilize two tunnels to the intranet (one for each mobile user) for registration and session establishment. The subsequent and media communication between endpoints is not tunneled.

b) Users situated in different locations

In this situation, assuming again that both users are already registered, the steps are:

- UA1 wants to send an INVITE message to UA2. It must send that message through its Outbound Proxy server which is situated inside the intranet, so it must utilize a tunnel.

- The Outbound Proxy server forwards the INVITE to the Inbound proxy of UA2. Because the users are situated in different locations, these proxies are likely to be in different locations, and thus a secure tunnel must be establish between the proxies.

- The UA2 inbound proxy forwards the INVITE to UA2, and for this it must also utilize a tunnel to UA2.

- As in case (a) subsequent SIP messages can use the tunnels already created/utilized.

- Again when the session is established, the media data will flow directly between UA1 and UA2 without traversing the intranet, and therefore is not tunneled.

In this case the session establishment required the use of three tunnels and again the media communication is not tunneled (see Figure 30).

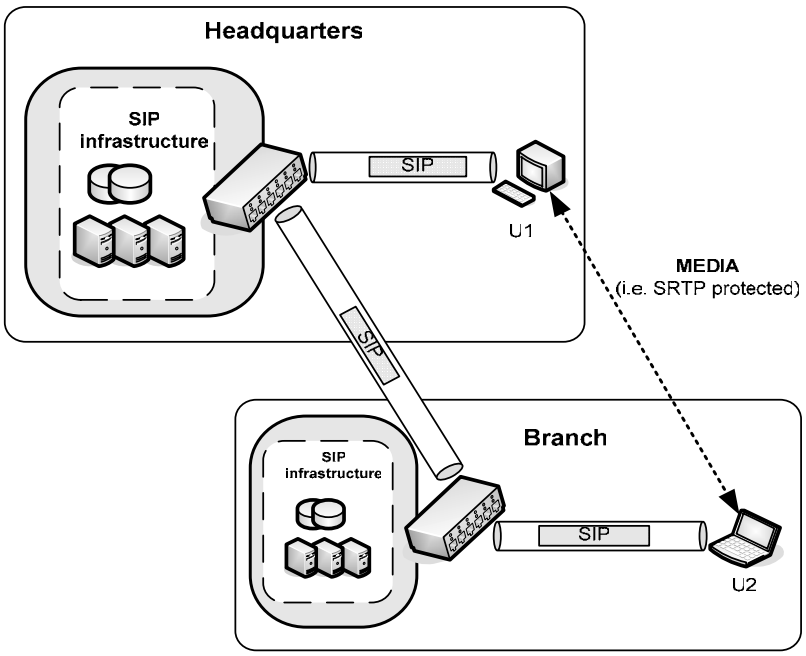

Figure 30: Example of tunneling when both users are in different same location

Note that the utilization of tunnels doesn't imply that these tunnels have to be created for each session. Since users will regularity want to access resources within the intranet, it is likely that a tunnel for each endpoint always exists while a user is using this

endpoint. In a similar manner, it is likely a full mesh between all instances of the intranet always exist.

c) Communication to PSTN

When a user wants to communicate to the PSTN, they must use the PSTN gateway which is situated inside the intranet. Hence, they would need to utilize a tunnel both for SIP signaling and media communication. If the location where the user is situated has a local PSTN gateway, there only is 1 tunnel needed. If not, there will be two tunnels, one between the user and the SIP servers and another between its intranet and the intranet with PSTN gateway. In both cases the media data **is** tunneled.

6.2.1.2. SIP *outside* the corporate intranet

When SIP servers are situated outside the corporate intranet (see Figure 31) they are directly 'reachable' by users as they don't need to go through the VPN gateway in order to communicate with these servers.

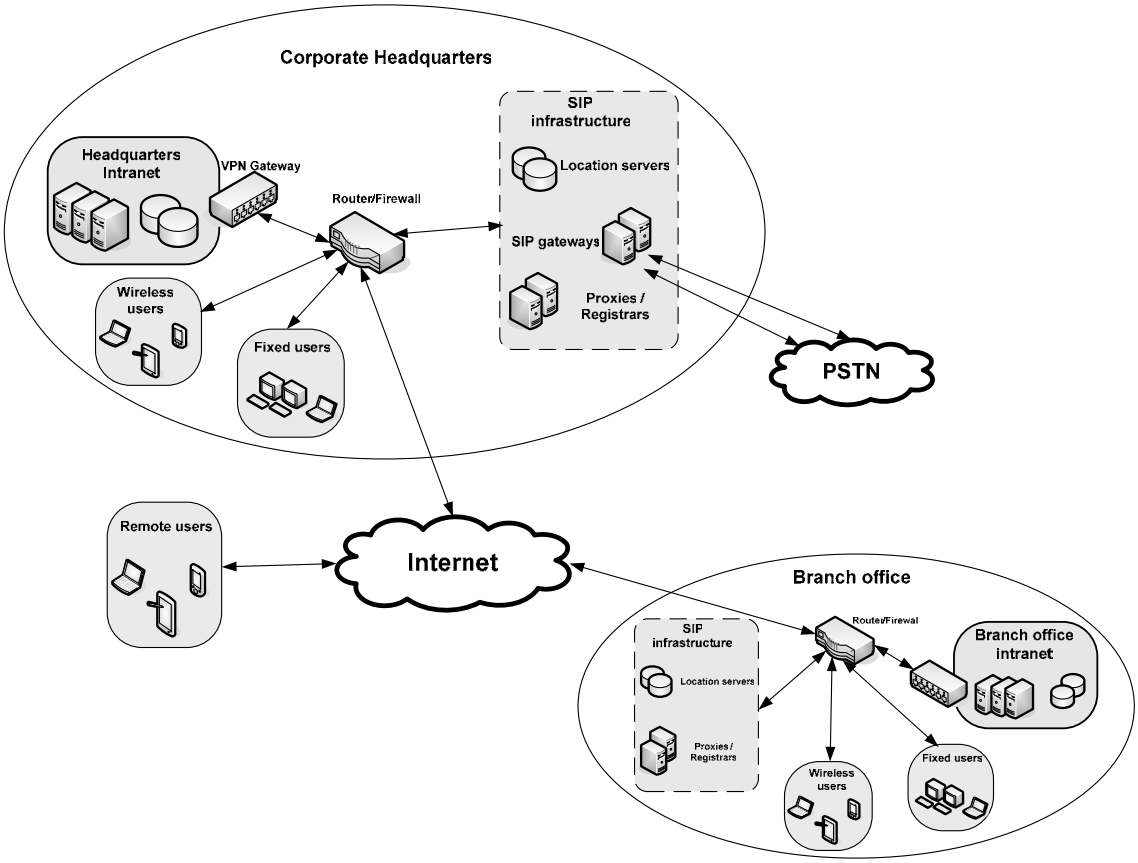

Figure 31: Example of SIP elements situated outside corporate intranet

In this case there is **no** need to utilize VPN tunnels, since communication between clients and servers will not traverse the intranet. However, the UA could use TLS to connect to the registrar and Outbound & Inbound proxies to protect the SIP signaling (see section 6.2.3 and Table 2). Since a given UA will have a specific registrar, Inbound, and Outbound proxy in a given location it can keep the TLS connected.

The VPN tunnels necessary for IPsec security related to the situation of SIP servers (inside or outside the corporate intranet) is resumed in the following table:

Table 1: VPN Tunnels and IPsec security related to SIP servers location

	SIP servers INSIDE			**SIP servers OUTSIDE**
Number of IPsec tunnels	2	3	1 or 2	**0**
Signaling Security	**YES**	**YES**	**YES**	NO
Media Security	NO	NO	**YES**	NO
	Users Same Location	**Users different Location**	**To PSTN**	

As a conclusion, by situating the SIP infrastructure outside the corporate intranet users can access SIP resources without accessing corporate intranet, and therefore without utilizing VPN tunnels through the VPN gateway. This reduces the amount traffic traversing the corporate intranet as well as favors SIP scalability, contrasting with the solution of situating SIP infrastructure inside the corporate intranet. However, unless TLS tunnels are used the SIP messages will not be protected. Similarity without using SRTP the media data is also not protected. Therefore, additional security measures have to be deployed to protect SIP signaling and media communications. We explain these security measures in greater detail in section 6.2.3.

6.2.2. Where to locate each type of SIP server

a) Proxy servers

The main function of proxy servers is to act as intermediaries between two SIP clients or a SIP client and a SIP server. Sometimes they also authenticate users before forwarding their registration requests to the registrar servers. Thus, it is reasonable to think that they should be as near as possible to users. This is possible in most large corporations, due to the distributed nature of the organization. We can situate a group of proxy servers in each location (or in locations with a large number of users). The number of proxies in each group will depend on the number of users in that location. Besides, as we already mentioned, we can take advantage of DNS SRV mechanisms and situate some proxy servers to act as back-up servers in different locations to increase availability. This solution is flexible and scalable since we can easily modify the proxy groups, and the load share if necessary, by modifying the DNS SRV records.

b) Registrar servers

Registrar servers are responsible for storing users' location in the location servers or databases. Registration messages are not very frequent compared with other SIP signaling messages, since they are mainly sent when a user first uses the system, when the user agent changes its IP address (thus if we are using a network mobility protocol such as Mobile IP, there would be no such changes), and after a refresh time (which is usually set to one hour). Therefore, if the number of register servers needed is small, most of them can be situated in the same location (i.e., the headquarters location). Of course, some of them can be placed as back-ups in different locations as we explained earlier for proxy servers.

c) Location Servers

The location servers are databases where the registrar servers store information about the users' current location. This information is also accessed (but only read) by proxy servers for routing. The first solution would be to locate these servers next to the registrar servers, that is, in a centralized location, since registrar servers must write and read from them. But this is not optimal for two reasons. The first reason is that proxy servers need to access (read) that information more often than registrar servers, so it would be better to place these location servers next to the proxy servers. The second reason is that it is better for redundancy and availability to have location servers situated in different locations, so that local failures don't affect all of them. Thus, we can geographically distribute the location servers near and co-located the proxy servers.

6.2.3. Security considerations

When designing a corporate SIP network, security is a key element. It is not only necessary to protect all SIP messages (i.e. signaling), but also media communications established by means of SIP; or, at least, to give users and/or administrators the possibility of securing them. As it has been explain before, situating SIP servers outside the corporate intranet, while reducing the traffic going through the VPN gateway and being more scalable than situating them inside the intranet, doesn't secure the SIP based signaling. Moreover, neither configuration secures the communications between users. As a result, additional security measures have to be deployed.

Several mechanisms to provide secure VoIP communications with SIP have already been studied in previous theses. Israel M. Abad Caballero in his thesis "*Secure Mobile Voice over IP"* [33] studied the use of Transport Layer Security (TLS) for securing SIP, Secure Real Time Protocol (SRTP) for protecting the media stream, and the Multimedia Internet KEYing (MIKEY) as the Key management protocol. This work was continued by Johan Bilien in his thesis "*Key Agreement for Secure Voice over IP"* [34] focusing especially on MIKEY. As a result of both theses, security was added to the SIP User Agent *miniSIP* [35] to provide secure VoIP. This implementation was tested in order to analyze how this added security affected call establishment delay. Results were published in the paper "*Call establishment delay for secure VoIP"* [36] and showed that the additional computational requirements were insignificant (NOTE: in these tests the

User Agents were running on 1.4 GHz. Pentium 4 laptops and the add call setup delay was less than 100 ms.).

Another possibility is to use IPsec VPNs to secure both SIP signaling and the media stream(s). This solution would also mean that all network communications would be protected by the same technology, as a VPN gateway is already present to protect corporate intranet. The use of IPsec in minisip for securing SIP was studied and implemented by Joachim Orrblad in his thesis *"Alternatives to MIKEY/SRTP to secure VoIP"* [37]. However, experimental results weren't as good as it might be expected as the answering delay turned to be 0.7 seconds, what is 'too much'. He explained that this delay is because of the specific IPsec implementation used and that by modifying minisip's MIKEY exchange messages the delay can be reduced. This matter has been studied by minisip's developers.

6.2.4. Integration of SIP with Firewalls and NATs

A firewall is defined as a security system that prevents unauthorized access to or from a private network by filtering messages according to specific security policies. Network Address Translation (NAT) is based on changing the source or destination IP addresses of packets. In Network Address and Port Translation (NATP) port numbers are also changed in order to hide internal IP address and to allow several endpoints in a private network to share the same external IP address.

Firewalls and NAT/NATP provide essential functions, such as security and flexibility, to corporate networks; but also represent a considerable challenge when deploying a SIP network. The main problem of such integration is that SIP carries the session description information (including IP address and port) inside its payload, and firewalls can't read it. Consequently, a firewall won't know that it has to allow communication using that (IP, port) and will drop these packets. One solution is to make the firewall capable of reading and understanding SIP messages. However, problems increase if NAT is used, because firewalls have not only to be able to understand SIP packets, but also to understand and to be able to translate private addresses into public ones.

There are different solutions to these problems, such as the use of Application Level Gateways (ALG), as studied in the previous thesis *"SIP, NAT, and Firewalls"* [38] by Fredrik Thernelius, the Realm Specific IP (RSIP) protocol [39], or Middlebox solutions [40]. Other solutions to the NAT problem include Simple Traversal of UDP through NATs (STUN) or an RTP relay also studied in the thesis of Amos Muhunda Nungu "*VoIP Service Provider"* [41].

The use of IPsec tunneling also solves some of the previous firewall problems as packets are tunneled straight through the firewall. However, NAT problems continue, due to nature of IPsec encapsulation. Nevertheless, there are specific solutions to this problem such as UDP encapsulation [42] based on encapsulating IPsec ESP packets inside UDP packets and using the same ports as those used by IKE traffic for traversing Network Address Translators.

To clarify this table 2 shows with the most important characteristics of both solutions for securing SIP signaling and media sessions.

Table 2: Solutions for securing SIP signaling and media traffic

	TLS + MIKEY/SRTP	IPSEC+MIKEY/IPSEC
Advantages	**Low delay Low hardware requirements**	**Secures all the traffic**
Disadvantages	**Only secures VoIP traffic**	**High Delay High Hardware requirements**
NAT/Firewall solutions	**STUN RTP relay ALG**	**UDP encapsulation**

6.3. Connecting the corporate VoIP network to the PSTN

By means of SIP, as has already been explained in previous sections, corporations can take advantage of their existing IP infrastructure and the Internet to deploy an IP telephony system capable of handling all their internal communications. This IP telephony system also allows communications with external IP based systems. However, an important part of the external communications of the enterprise are links to the Public Switched Telephone Network (PSTN) infrastructure. This PSTN infrastructure is not an IP-based infrastructure, thus corporations need an element acting as a gateway between both 'worlds'. The function of this gateway is to convert calls and call signaling between the internal SIP format and the external PSTN format; this is called a SIP/PSTN gateway (see figure 32).

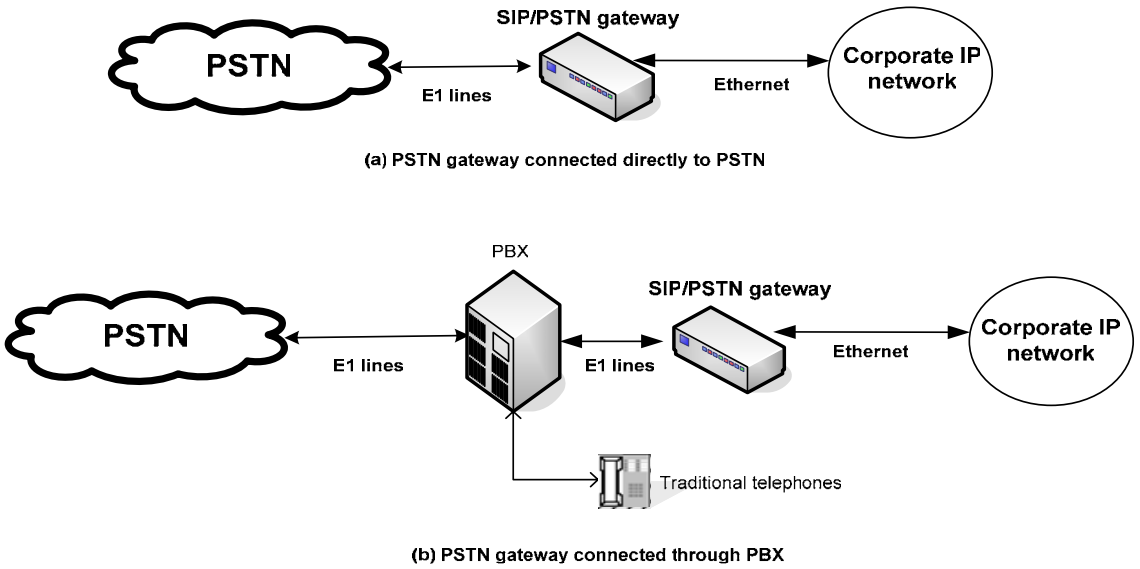

Figure 32: Examples of connections between a SIP/PSTN gateway and the PSTN

Figure 32(a) shows the SIP/PSTN gateway's external interface connected directly to the PSTN. This is possible because we are assuming that there are no remaining analogic telephones inside the corporation and all internal communications are SIP based. If this isn't the case, the PSTN will be connected to the corporate PBX. Figure 32(b) shows this PBX connected to the PSTN gateway. From now on we will assume the first case (a), as this is the most demanding.

The capacity of the PSTN gateway is an important parameter of the system. This capacity is measured by the number of simultaneous calls that the gateway is able to handle. These simultaneous calls are usually restricted by the number of E1 or T1 interfaces on the gateway. We will clarify this by means of an example:

> Consider a PSTN gateway with 4 E1 interfaces, all of which are connected to E1 lines. Each E1 line can carry 30 simultaneous calls, so in this example the PSTN gateway can handle 120 simultaneous calls to and from the PSTN infrastructure. If we consider that the mean time of a telephony conversation is 3 minutes, then this gateway can serve almost *0.667* calls per second (cps). If there are 2000 users in the office, 1/4 of them are calling during the busy hour, and among these calls the 30% are external (towards the PSTN), then the gateway has to serve 3000 BHCA or almost *0.84* cps. So in this case, we would need *2* gateways to be able to serve all the calls.

7. Managing the Network

Business operations of enterprises rely, more and more, on their communication infrastructure. Information exchanges between enterprise branches, inter-business transactions, relationships with suppliers and consumers, and mobile workers are directly dependent on the enterprise's communication infrastructure. As corporations grow, this infrastructure becomes larger and more complicated to manage; and manual management and configuration of the network is no longer efficient. Therefore, deployment of an automatic, reliable, and efficient network management system becomes essential for business success. This management system must allow the corporation to control every part of their communication infrastructure in a simple, efficient, reliable, secure, and flexible manner.

7.1. Management Architecture

The first step in the design of a management system is the selection of a suitable management architecture. The two main possibilities are: Centralized and Decentralized architectures (see Figure 33). Each of them has its advantages and disadvantages, and according to a report by the META Group [43] 60% of the 500 largest enterprises will deploy centralized management architecture by 2006. Thus both alternatives have to be considered.

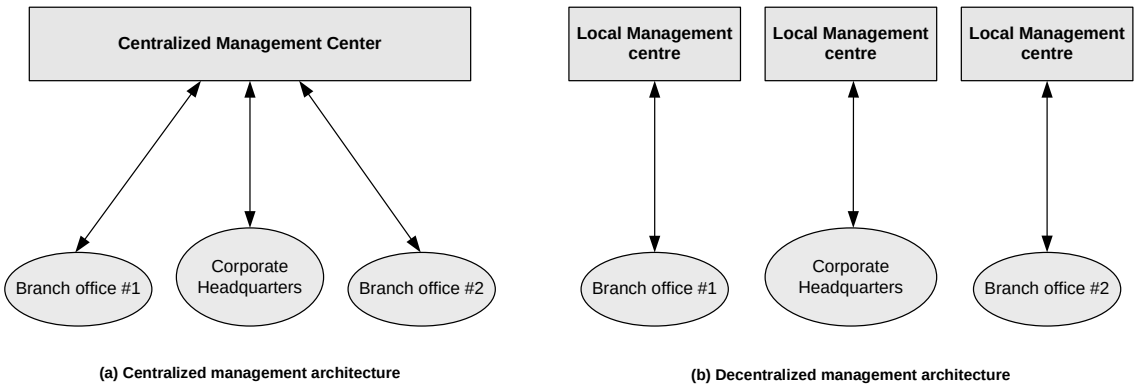

Figure 33: Example of (a) Centralized management architecture and (b) Decentralized management architecture

7.1.1. Centralized management architecture

In this architecture the entire management system can be controlled from a central location. This architecture allows corporations to group management resources and staff in highly specialized centers, also referred as *command centers of excellence (C-COE).*

Some of the main advantages and disadvantages of deploying this type of architecture are explained below.

Advantages

- ***Decision making consistency***
  With a centralized architecture corporate decisions and policy enforcement are consistent across the entire network. Changes can be quickly and consistently applied.

- ***Global view***
  This architecture provides a single global view of the entire corporate communications infrastructure.

- ***Cost reduction***
  A centralized architecture allows enterprise to focus management resources on specialized operational centers. Concentration of management equipment and qualified staff in a single location can reduce operating costs.

Disadvantages

- ***Interoperability problems***
  Network services and equipment are usually from many different manufacturers and its integration in centralized management architecture can be difficult.

- ***Central Coordination requirements***
  Any local problem or failure, without consideration of its importance, needs to be handled by the central management center, decreasing the timeliness of response.

- ***Lack of resilience***
  A single failure of the corporate management facility or its connectivity compromises the entire corporate communications infrastructure, potentially stopping business operations.

7.1.2. Distributed Management architecture

In a distributed management architecture each business unit manages its own services and processes (i.e. those within their boundaries). Consequently, the management of every unit, branch, or business area is ‘independent’ of others. This independence can be interpreted either as an advantage, because, for example, it provides flexibility and faster response to local problems, or as a disadvantage given that it hinders a global vision and limits integration between business units. However, there is some loose correlation since each management unit must follow general corporate policy. Some of the main advantages and disadvantages of implementing this architecture are:

Advantages

- ***Business structure accuracy***
  This architecture reflects the business structure of the corporation.

- ***Local management flexibility***
  Every business unit is responsible for its own management; hence it can adapt to the specific characteristics and needs of each environment.

- ***Resilience***
  A single failure will not compromise management of the entire corporation, as may happen with centralized solutions.

- ***Management tools independence***
  A management unit can locally use specific management tools, without worrying about interoperability with solutions adopted in other units. However, this can be also seen as a disadvantage since using independent tools obstructs data sharing between different business units. This possibility of sharing some data between different business units can be useful even when a distributed architecture is used.

Disadvantages

- ***Partial vision***
  Distributed management doesn't facilitate a global vision of the corporate communications infrastructure and different or independent monitoring reports are difficult to integrate.

- ***Inconsistent policies***
  It is more difficult to apply general corporate policies than in a centralized solution, this can lead to conflicts between business units.

- ***Maintenance costs***
  The number of different management tools is greater and there is a need for specialized management staff located in each management center. Therefore, the associated maintenance and upgrading costs may be higher (depending on the license terms for these tools).

7.1.3. Additional design considerations

There are additional important design considerations when deploying corporate communications infrastructure management systems. These considerations reflect general requirements for any corporate infrastructure: security, scalability, availability, and performance. Along with requirements for flexibility and cost- effectiveness that have already been considered; since they depend to a large extent on the architecture selected.

Security

Security considerations are essential when implementing a management architecture as management tools have privileged access to all the software and hardware of the network. Thus, authentication, authorization, and data encryption should be used for all management procedures. However, these security processes can be difficult to centralize. Consequently, both security and manageability must be considered when deploying management architectures.

Scalability

When designing the management system future capacity requirements derived from the expected growth of the corporation must be considered. The first step is to plan for additional capacity. As the network requirements grow new capacity must be added to the system, so that sufficient capacity is *always present*. An important concept is that not only it is necessary to scale as requirements grow, but also to maintain a network performance. Of course, management tools scalability must also be considered, as they are responsible for management functionality.

Availability

Availability requirements for every management service depend on their importance. That is, if a service provides critical information, it is more valuable and, consequently, its availability must be higher. However, if a service is non-essential, then providing high availability may not be cost-effective (as high availability measures usually include additional hardware, software, and additional skilled management staff). So, when a cost/benefit analysis is done in order to decide if a specific availability measure is necessary, the importance of that service for management and operations must be carefully considered.

Performance

Because management systems usually share the communications infrastructure with other corporate systems, the load caused by management activities must be considered. For example, when management agents transfer management data over the corporate network they consume network bandwidth that may adversely affect the whole network's performance.

7.2. Management Services

Regardless of the architecture selected, there are key services that any corporate management system has to provide. These services are monitoring, control, and software distribution. The first two, monitoring and control, are the base on which

management relies because it is not possible to efficient manage a system without continuously knowing its status, and being able to apply necessary changes. The last, software distribution, is vital in a corporate environment where users (1) may not have the technical knowledge needed to configure and maintain their own equipment and/or (2) they can't waste time doing it.

7.2.1. Monitoring and Control

Service monitoring and control are key to a management system. This system must observe the overall function of the network, network attached services, …; helping to prevent and detect network and service failures. Thus failure detection and prevention are essential for business operations since more and more of the corporate operations **depend** on this infrastructure.

The basic functionality necessary is:

- Collect information from network and other communication elements about their functionality, performance, or faults.
- Analyze this information (especially alerts) try to generate an automatic response or an incident report.
- Allow management staff to solve problems by changing configuration, installing new software releases, … remotely.

7.2.2. Software distribution

In corporate networks the amount of equipment and the number of software packages can be very high. Routers, gateways, servers, mobility agents, …, each representing a very important part of the infrastructure, but are still small in number compared to the very large number of user devices. Today, many corporate users utilize more than one machine (i.e. a laptop ,a PDA, and/or a smart phone) to access enterprise resources, and it is very important for the correct function of the enterprise to manage the software installed in these devices. For example, from a security point of view, it is very important that every machine has the latest operating system patches installed and that the antivirus system is up to date. It is also essential that every configuration change can be easily and quickly distributed to all network attached equipment. These tasks shouldn't be performed manually by system administrators, as this would be too time-consuming, costly, and inefficient, nor should users be required to explicitly perform all the changes as this would cause configuration problems, increase costs for support, and waste of the users' time. One solution to all these problems is the use of an automatic software distribution system (see Figure 34).

The purpose of an automatic software distribution system is to manage and distribute approved corporate software to enterprise's equipment. This software can include operating system patches, program upgrades, antivirus updates, specific applications, or any software necessary to support business functions. This distribution should be

selective (devices running different operating systems need different upgrades), automatic, and potentially scheduled to minimally affect business operations.

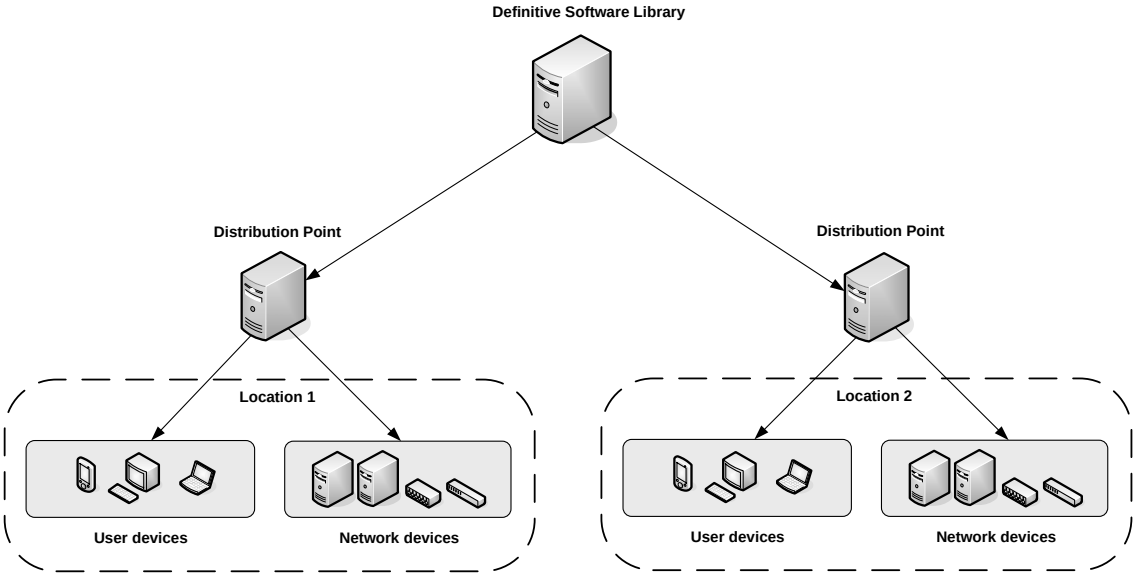

Figure 34: Example of a software distribution system architecture

The basic function of such a software distribution system can be divided into three different parts:

- ***Definition of installation policies***
  Which devices should receive which software and when, has to be defined according to corporate policies. This process generally requires manual configuration, but need not be very time-consuming if a correct classification of users and devices exists.

- ***Software distribution***
  Software distribution is usually based on client/server architecture. Enterprise approved software is stored in a central repository sometimes referred as *Definitive Software Library (DSL)* and distributed to servers, also called *distribution points*, which will in turn distribute the software to the clients. While more than one distribution point can be connected to the DSL, they are generally distributed geographically (see figure 34). Thus, distribution, although it is centrally managed, can be decentralized in practice and clients should be assigned an optimal distribution point based on parameters such as link capacity, delay, cost, capacity, … .

- ***Distribution reports***
  Once a distribution process has finished a report is generated with information about its success or failure. The gathering of this information can provide statistical data valuable to detect problems or to plan system upgrades.

Some examples of commercial software distribution systems are *Microsoft's System Management Server (SMS) 2003* [44] and *Novell's ZENworks* [45].

7.3. Policy-based management

Efficient management of corporate resources becomes more and more important as network complexity and heterogeneity increases. The number of different devices, network technologies, and applications present in actual corporate networks makes manual management infeasible or inefficient. Besides, trying to manage every network element independently moves the enterprise network further away from its main goal, to support business operations instead of hindering them. The solution is to manage the network as a single entity (even if done in a distributed fashion), which applies corporate business policies to the entire network by specifying a collection of conditions and their associated actions, rather than continuously analyzing what measures should be applied to each network component. This is the main concept of Policy-based management.

7.3.1. Policy and Policy rules

The IETF Policy Framework Working Group [46] defines policy as a collection of policy rules, and policy rules as a set of conditions and a corresponding set of actions. A policy rule can be expressed in the form: If <*condition*> then <*actions*>. *Condition* defines when a policy rule has to be applied; and when that activation takes place, one or more *actions* associated with that policy rule are applied. Thus, policy rules define, according to corporate guidelines, how network resources and applications can be used, by whom, when, and in what situations. Once these policy rules have been defined, the policy-based system will apply them automatically in each situation, modifying network parameters as necessary and configuring appropriate network elements without human intervention.

7.3.2. Policy-based management system architecture

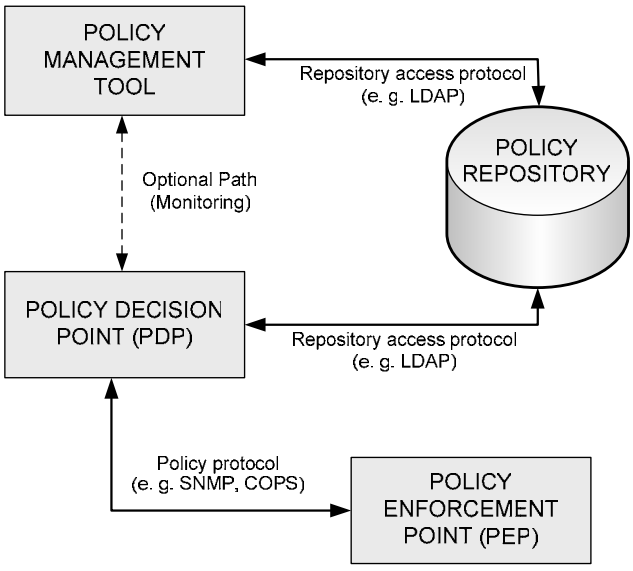

Figure 35: Policy-based management system architecture [52]

An example of the logical policy based management architecture is shown in Figure 35. The main elements of this architecture are:

Policy Management Tool
The Policy Management Tool is used by the administrator to interact with the policy-based system, allowing him or her to define and edit policies, as well as to monitor the deployment of policy rules. To facilitate management, it allows definition of these rules in a human understandable syntax that the Policy Management Tool later translates to the policy model syntax of the policy repository.

Policy repository
The policy repository stores and retrieves all the policy rules. It uses a specific syntax defined by the IEFT Policy Framework Working Group to ensure interoperability between products of different manufacturers.

Policy Decision Point (PDP)
The PDP is the entity which makes the decisions based on policy rules. It interprets the policies stored in the Policy Repository, and when applicable translates these policies into a form usable for the Policy Enforcement Point (PEP).

Policy Enforcement Point (PEP)
The PEP enforces a policy decision communicated by the PDP by taking the necessary actions. For policy-aware devices, the PEP is a component that resides within them.

Policy communication Protocols
Different protocols are used for communication between the elements of the architecture. The communication between PDP and PEP can be carried for by SNMP [47] or COPS [48] [49], while the Policy Repository could be a directory server accessed with LDAP [50].

The adoption of policy-based management capabilities has already been studied by Konstantinos Avgeropoulos in his thesis “Service Policy Management for User-Centric Services in Heterogeneous Mobile Networks” [51]. He focused his study on the user-side service policy management in SIP networks, and his findings show that policy-based management systems are feasible, although they can present some scalability challenges.

To sum up, a corporate network management system should:

- Be flexible, reliable, and secure;
- Provide high availability while remaining cost-effective;
- Be effective without notably decreasing network performance;
- Include monitoring and control as well as a software distribution system; and
- Have policy-based management capabilities

8. Case study

In previous sections we have analyzed different aspects and solutions for deploying IP telephony in a large corporate environment. Some aspects such as the use of VPN technologies to protect the corporate intranet, access by wireless users, integration of technologies, and redundancy measures have already been explained. However, there are some questions such as the number of SIP severs, PSTN gateways, and LAN or WAN requirements which need a deeper study focused on 'actual' numbers.

In order to do this we have created a fictitious large corporation based on data of two actual large corporations belonging to telecommunications and manufacturing sectors. Note that the numbers given are not the actual numbers of a specific corporation, but are meant to be representative of such a corporation. First in section 8.1 we present the company structure as well as its calling patterns, and in following sections we analyze each of the following aspects: SIP servers requirements, PSTN requirements, LAN and WAN requirements, Mobility requirements, and cost savings associated with the deployment of VoIP.

8.1 Company data

8.1.1. Company structure

Our company is a large enterprise with more than 66.000 employees distributed all over the world with a presence in more than 100 countries. The company is divided into three different units: Production (which is related to manufacturing processes), Engineering (focused principally on research and development), and Administration (comprising internal administration, Sales, Marketing, Aftersales, and Support functions). A detailed distribution of employees per country and business unit is shown in Table 3.

Table 3: Employees per country and business unit

	Production	Engineering	Administration	Total
Sweden	20.000	5000	6000	31.000
China	8000	1300	1200	10.500
Brazil	3000	200	600	3.800
India	0	1500	2000	3.500
UK	0	1500	1800	3.300
USA	0	0	3000	3.000
Japan	0	800	1500	2.300
Spain	0	800	800	1.600
Germany	0	0	1300	1.300
Canada	0	400	700	1.100
France	0	0	1000	1.000
Italy	0	0	1000	1.000
Australia	0	0	700	700
Others (90):	0	0	2000	2.000
TOTAL	**31.000**	**11.500**	**23.600**	**66.100**

There are some conclusions we can extract from data shown in Table 3:

a) There are 13 main locations combining more than 96% of the workforce. The remaining 4% - all belonging to the administration unit- are distributed among 90 different countries. This diversity reflects the distributed nature of large corporations aiming for expand to their business market.

b) Employees are divided into 3 different business units: Production, Engineering, and Administration. If we examine the employees' distribution shown in percentage in figure 36 we see that Production is the largest unit (47 %), but Administration is close with 36%. Some years ago Production was over 60%, but outsourcing greatly reduced the number of employees in production in order to reduce costs. While this caused a reduction in the number of production employees it lead to an increase in the Administration workforce.

c) Production is concentrated in only 3 countries: Sweden, China, and Brazil. Sweden is the only western country, and production remains in it because of the Swedish origin of the company. The remaining production is located in countries (China and Brazil) where workforce is cheap compared to western countries and where entry into the local market was facilitated by domestic manufacturing.

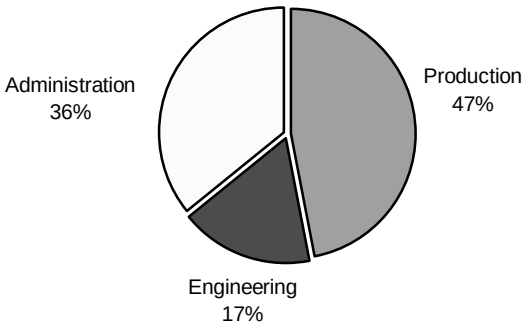

Figure 36: Distribution of employees in business units

8.1.2. Calling Patterns

There are three different parameters of calling patterns needed for the study:

- The percentage of users calling during the Busy Hour
- The Mean Holding Time of a voice call
- The percentage of calls that are originated or destined to the PSTN infrastructure

These parameters depend more on the type of business unit than on the location, since employees belonging to the same business unit have similar tasks wherever they are

located, thus the job tasks depend completely on their type of work and consequently their business unit. The data shown in table 4 shows these parameters for each business unit.

Table 4: Calling patterns of business units

	Users calling BH (%)	Mean Holding time (sec)	PSTN calls (%)
Production	20	180	20
Engineering	30	180	30
Administration	60	180	70

If we analyze this table for the 3 patterns:

- The number of users calling simultaneously during the busy hour varies from 20% for Production employees (whose main calls are for production control, requesting materials, and communicating incidents), and 60% for Administration employees (who have to keep in contact with users, providers, and coordinate all internal and sales processes).
- The mean holding time is 180 sec for all units. Usually business' mean holding times vary from 120 to 240 seconds, so we decided to use the mean value. Although the holding time could vary for different units this variation was not clear so we decided to use the same value for each unit.
- The number of external (to/from PSTN) calls follows a similar pattern to the percentage of users calling during the Busy Hour. Production calls are mainly internal, while administration calls are mainly external.

8.2 SIP servers study

8.2.1 First approach

The first step to decide upon is the number and location of the SIP servers the company needs in order to estimate the load these SIP servers must support. This load is usually measured in calls per second for proxies and registrations per second for Registrar servers.

From Table 4 we know that the mean hold time of a call is 180 seconds, so during the busy hour the number of calls can be up to 1 hr/180 sec = 20 calls per active user. We can also calculate the number of active users during the busy hour by multiplying the total number of users (Table 3) and the % of active users (Table 4). Considering also that a call needs two users which will both be employees when the call is internal, and only one when the call is external (to PSTN), we can obtain the number of calls during the busy hour and thus the number of calls per second during the busy hour (see Table 5).

Table 5: Calls per second in busy hour

	Production	Engineering	Administration	**Total**
Sweden	20,00	7,08	14,17	41,25
China	8,00	1,84	2,83	12,68
Brazil	3,00	0,28	1,42	4,70
India	0	2,13	4,72	6,85
UK	0	2,13	4,25	6,38
USA	0	0,00	7,08	7,08
Japan	0	1,13	3,54	4,68
Spain	0	1,13	1,89	3,02
Germany	0	0,00	3,07	3,07
Canada	0	0,57	1,65	2,22
France	0	0	2,36	2,36
Italy	0	0	2,36	2,36
Australia	0	0	1,65	1,65
Rest(90):	0	0	4,72	4,72
Total	31,00	16,29	55,72	**103,01**

And if we consider that every endpoint must send a re-registration message every 60 minutes (the typical value) and we assume each user has one endpoint we obtain the number of registrations per second (RPS) shown in table 6.

Table 6: Registrations per second in busy hour

	Production	Engineering	Administration	**Total**
Sweden	5,56	1,39	1,67	8,61
China	2,22	0,36	0,33	2,92
Brazil	0,83	0,06	0,17	1,06
India	0	0,42	0,56	0,97
UK	0	0,42	0,50	0,92
USA	0	0	0,83	0,83
Japan	0	0,22	0,42	0,64
Spain	0	0,22	0,22	0,44
Germany	0	0	0,36	0,36
Canada	0	0,11	0,19	0,31
France	0	0	0,28	0,28
Italy	0	0	0,28	0,28
Australia	0	0	0,19	0,19
Rest(90):	0	0	0,56	0,56
Total	8,61	3,19	6,56	**18,36**

The other important parameter is the number of simultaneous calls and simultaneously registered users the servers must support. We can estimate the maximum number of simultaneous calls by simply multiplying the number of calls per second and the mean holding time. For the number of simultaneous registered users, we are going to assume in this first analysis that all users are registered all the time (we will refine this later).

Table 7: Simultaneous calls and registered users requirements of our system

	Simultaneous calls	Registered users
Sweden	7.425	31.000
China	2.282	10.500
Brazil	846	3.800
India	1.233	3.500
UK	1.148	3.300
USA	1.275	3.000
Japan	842	2.300
Spain	544	1.600
Germany	553	1.300
Canada	400	1.100
France	425	1.000
Italy	425	1.000
Australia	298	700
Rest(90):	850	2.000
Total	**18.543**	**66.100**

8.2.2. Commercial solutions

During the last several years, due to the increasing popularity of VoIP and especially the SIP protocol, the number of available solutions has grown considerably. Among these solutions, we find open-source based SIP servers such as Asterisk [53], SER [54], or Minisip [35] and proprietary servers from Cisco [55], Avaya [56], or Alcatel [57].
As usual, performance and capabilities are different for each solution. For our study, we have chosen the Cisco SIP Proxy Server which presents some of the best characteristics [58] and comes from a well recognized manufacturer.

Performance values of the Cisco Proxy Server for three different hardware configurations suggested by Cisco [55] are show in Table 8.

Table 8: Cisco SIP proxy Server Performance[2]

Hardware	Processor	RAM	CPS	RPS	Sim. Calls	Registered Users
Sun Fire V120	1 x 550 Mhz.	≥512 MB	**90**	**55**	**-**	**-**
Sun Netra 2.0	2 x 900 Mhz.	≥1 GB	**375**	**67**	**20.000**	**20.000**
IBM x335	2 x 2.0 GHz. Intel Xeon[3]	≥1 GB	**1000**	**67**	**20.000**	**20.000**

[2] This performance table supposes a redundant configuration of **two** SIP servers using load balancing mechanisms (DNS SRV).

[3] Note that although IBM server x335 can use processors up to 3.2 GHz. Xeon, Cisco tests were conducted on servers running 2.0 GHz. Xeon processors.

As we can see in the table above, the maximum number of simultaneous calls and registered users in this system is not influenced by the processor capabilities unlike the number of calls per second which depends to a large extent on the processor speed. This relationship between calls per second and processor speed is similar for every solution and can be seen in Figure 37.

Note that in Figure 37, to establish a clearer relationship between processor speed and calls or registrations per second we have considered that a dual processor performance is about 1.8 times the performance of a single processor [59] [60] [61]. For example, for the IBM x335 series server performance results for the SPECint_rate_2000 are 18.6 for a single 3.2 GHz. processor and 33.8 for a dual 3.2 GHz. processor what is equivalent to a factor of 1.817 [61].

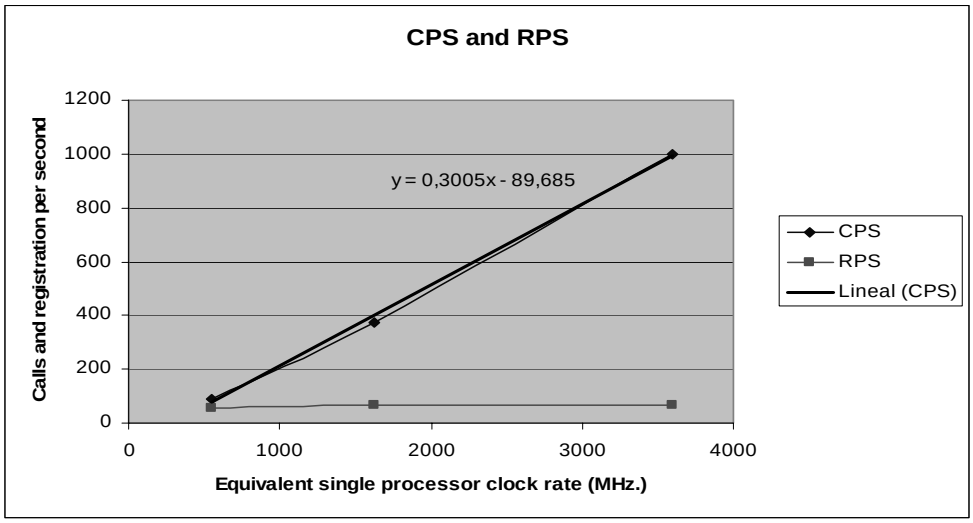

Figure 37: Calls and registrations per second tendency

Fitting a curve to this data we see that an estimation of the performance of a SIP server for actual processors is as shown in figure 37 and summarized in table 9.

Table 9: Approximated performance of SIP servers

GHz	CPS	RPS
2	511	67
3	**811**	**67**

If we compare server performance (from now on we will use data for a 3 GHz. processor) with the company's needs, we observe that the calls per second and registrations per second of a single configuration of two redundant servers (remember these values are for a configuration of two redundant servers with load sharing) is sufficient. However, while the simultaneous calls performance is very close to the

requirement, the number of registered users is clearly not sufficient using a single cluster of two servers (see Table 10).

Table 10: First approach: Requirements and Capacity

	Proxy Servers		**Registrar Servers**	
	Requirements	Capacity	Requirements	Capacity
CPS/RPS	103,1	811	18,36	67
Simultaneous	18.543	20.000	**66.100**	**20.000**

So in this first approach we will need 2 proxy servers (a clusters of two servers), although the use of **3 or 4** proxy servers would be better since the simultaneous call requirements are close to capacity and for redundancy. Additionally we will need **8** registrar servers (4 clusters of 2 servers).

8.2.3. Considering Local Times

For all our previous calculations we have assumed that all locations share the same local time and thus their busy hours coincide. However, this is not true. The busy hour, as it has been already explained, is the hour of the day of maximum call load, business calls during a day according to each local time usually follow the pattern shown in figure 38.

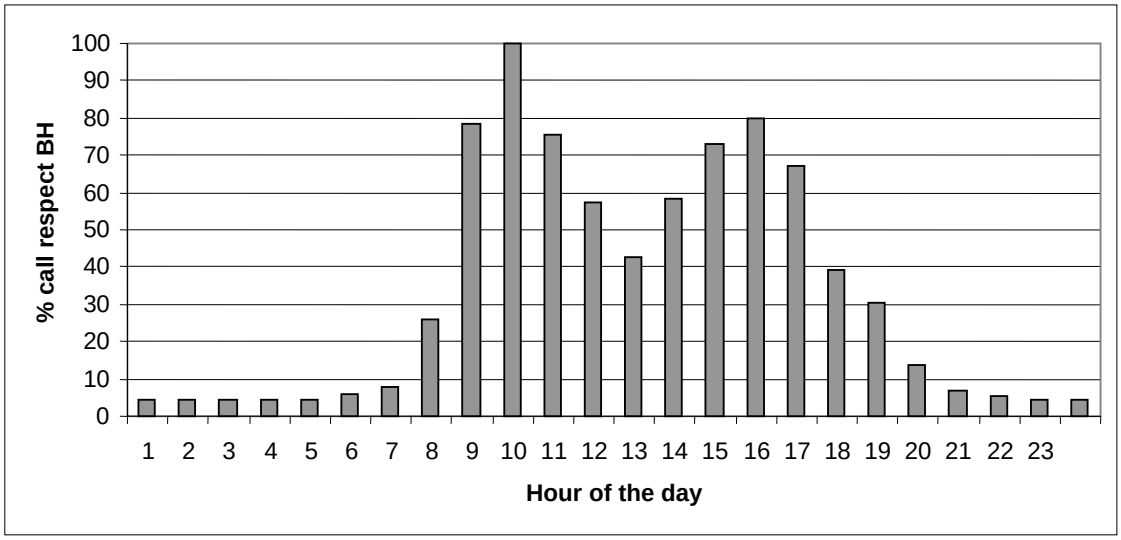

Figure 38: Typical daily business calls distribution [Adapted from 65]

Thus, the distributed nature of a large corporation implies that the load will depend on the local time, and considering time zone difference between countries the different busy hours will not coincide. This will decrease our requirements, but the importance of this decrease has to be calculated to realize its relevance. We analyze this situation for

our case, considering the time zones for our main corporate locations as showed in Table 11. In addition to this local time shift we will also consider the fact that the number of registered users will vary with the working schedule, thus not all endpoints will be registered at the same time. For this measurement we assume that during the local working day for each location (8 hours) all endpoints are registered and during the rest of the day registrations decrease accordingly.

Table 11: Country time shift respect to Swedish local time[4]

USA Canada	**Brazil**	**UK**	**SE. ES. DE. IT. FR.**	**India**	**China**	**Japan**	**Australia**
-6	-4	-1	**0**	+4	+7	+8	+9

Repeating each calculation adding this new information, and considering also the percentage of calls for each hour related to the busy hour traffic extracted from figure 38, we obtain the following values for the global busy hour (see Table 12):

Table 12: SIP requirements considering local time

CPS	RPS	Simultaneous calls	Registered users
72,55	16,48	13.059	58.000

As Table 12 shows, all the requirements have decreased, as was expected. The number of calls per second and registrations per second were not very important since needs were already covered with a single server configuration. However, the decrease in the number of simultaneous calls and registered users can affect the final design (Table 13).

Table 13: Needs and Capabilities considering local time

	Proxy server		**Registrar Server**	
	Requirements	Capacity (2)	Requirements	Capacity (2)
CPS/RPS	72,55	811	16,48	67
Simultaneous	13.059	20.000	58.000	20.000

Examining these new values, we only need **2** Proxy servers and 6 Registrar servers instead of 3 Proxies (as 2 was very close to the requirements) and 8 registrars from the first estimate Nevertheless, in the case of Registrar servers with 6 servers, the requirements is tight, and as we have already mentioned, these calculations are approximate so we still recommend the use of **7** or **8** Registrars.

[4] Note this is a simplification as several of these countries have multiple time zones

8.2.3.1. Location and load sharing

Proxy servers

The number of proxy servers needed is only two, so we only have two possibilities: (a) locate both proxies together in a single location and (b) locate each server in a different location.

(a) Same location

This location, of course, would be Sweden which has the greatest number of users (and thus the largest facilities), and most simultaneous calls. The use of DNS SRV load balancing techniques (see section 5.4.2.) is very simple since for all locations the weight and priority of the two servers would be the same. In this case, we would strongly recommend the use of at least another server for redundancy, which *should* be located far from the other 2.

(b) Different locations

This solution would increase reliability as explained in section 5, since both servers won't be co-located. In our case these locations would be Sweden, for the same reason as before, and China which is the second most important location, with big facilities, and it is 'close' to other important locations like those in India and Japan. The use of DNS SRV load balancing mechanisms in this case can provide greater benefits than in case (a). We should configure DNS records for locations closer Sweden to use mainly the proxy in Sweden, and locations closer to China to use the server in China. This configuration also takes advantage of the time swift between these locations (8 hours) so when Swedish server is very loaded and call messages are directed to the server in China, this server will have a low local load and vice versa. Finally, as for case (a), we would recommend the use of another redundant server located for example in Sweden.

Registrar Servers

The number of registrar servers is 8 (as we said is 6 is very close to the requirements), we would chose to distribute them among main locations according with the number of 'close' users and their distance. That would be for example: 4 in Sweden, 2 in China, and another 2 in USA. Although this distribution increases maintenance costs (as it is cheaper to locate all equipment in the same location) it increases availability and reduces network load due to registrations from remote locations.

Location Servers

The Cisco SIP Registrar servers used in out study utilize one location server (database) per registrar, and thus the number of these databases is the same as the number of registrars (8). Consequently, registrar servers have to share dynamic registration data as explained in section 5.4.2.

If we estimate the overhead due to these replicated Registration messages shared between registrars for our case (8 registrar servers and assuming an average registration message size of 400 bytes) we obtain an overhead of less than 400 Kbps in the entire network.

8.2.3. Increasing needs

As we have seen, the processing capabilities of existing processors can fulfill even the need of a large corporation needs in terms of new calls and registrations per second. It is, in fact, the number of simultaneous calls and registered users what set the number of servers. If we calculate the number of servers needed for the same corporate structure and calling patterns but for a larger number of employees we obtain figure 39.

As we can see in figure 39 the number of registrar servers increases faster than the number of proxies. This is mainly because while simultaneous calls decrease notably when considering different time shift between location, registered users decrease less (all users in one location are assumed to remain registered during 8 hours while calls in different hours than the busy hour are significant lower).

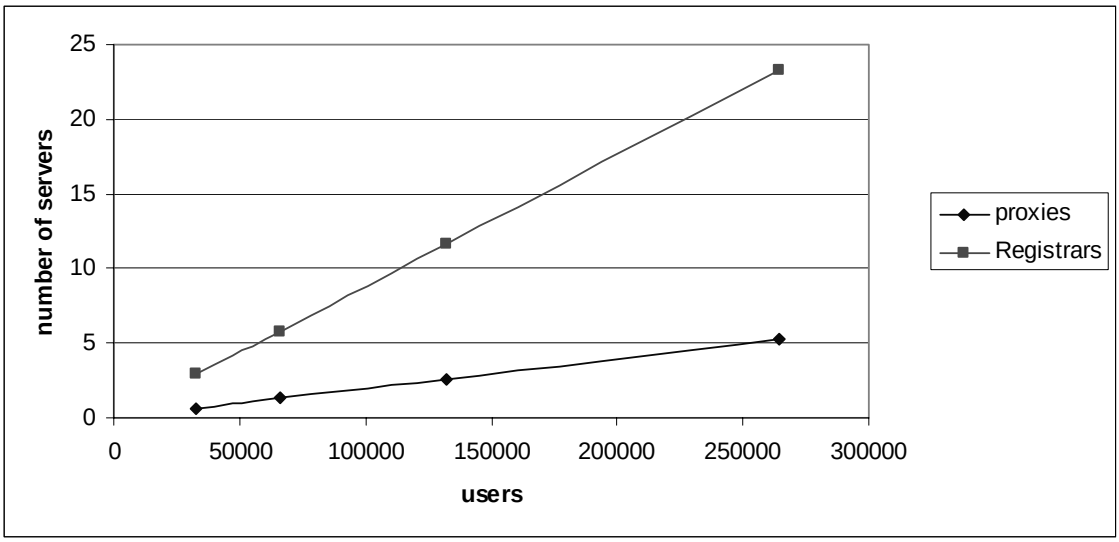

Figure 39: Proxies and registrars need for different numbers of users

The location of these servers would be mainly in the location with the greatest number of users, in our case Sweden. The rest of the servers can be distributed in 1 or 2 other important locations. This number of additional locations can be increased when the number of servers needed is very high (for example 36 for 400.000 users), but not to broadly spread considering that locating servers in many different sites will increase maintenance costs. As before, DNS SRV procedures should be used to achieve load sharing and increase availability.

8.3. PSTN Gateways

Another important part of the study is to calculate the number of lines needed to interconnect the corporate VoIP network with the PSTN infrastructure and the number of gateways associated with these lines (see section 6.3.). Using the company structure data (Table 3) and the associated call patterns (Table 4) we can calculate in a similar way as before the number of external calls per second (to or from the PSTN) and the load in Erlangs associated with these calls. Note that we are going to use the unit Erlang because when studying PSTN lines requirements we can make use of the Erlang-B formula to calculate the number of required lines for a desired grade of service. If we use these tables for a grade of service of *0.01%* (i.e. the blocking probability), this implies designing the network in such a way that only 1 every 10.000 calls won't be served, we obtain the values shown in Table 14.

Table 14: Number of required lines for a grade of service of 0.01%

	External calls per second	Erlangs	Number of required lines
Sweden	16,04	2887,5	2919
China	4,14	744,45	774
Brazil	1,68	301,8	326
India	3,94	709,75	738
UK	3,61	650,25	679
USA	4,96	892,5	922
Japan	2,82	507,45	535
Spain	1,66	299,2	324
Germany	2,15	386,75	412
Canada	1,33	238,85	261
France	1,65	297,5	322
Italy	1,65	297,5	322
Australia	1,16	208,25	231
Rest(90):	3,31	595	623
Total	50,09	9016,75	9388

These lines can be grouped into a number of either E1 trunks, E3 trunks, or any combination[5]. One E1 trunk corresponds with 30 lines and an E3 line transports 530 lines. So for example our company needs, can be grouped as shown in Table 15:

[5] We ignore the fact that in some countries T1, T3, or some combination of them would be used.

Table 15: E1s and E3s needs to interconnect the corporate network and the PSTN

	Number of required lines	Approx. E1 trunks	Approx. Combination
Sweden	2919	98	5 E3 + 9 E1
China	774	26	1 E3 + 9 E1
Brazil	326	11	11 E1
India	738	25	1 E3 + 7 E1
UK	679	23	1 E3 + 5 E1
USA	922	31	1 E3 + 14 E1
Japan	535	18	1 E3 + 1 E1
Spain	324	11	11 E1
Germany	412	14	14 E1
Canada	261	9	9 E1
France	322	11	11 E1
Italy	322	11	11 E1
Australia	231	8	8 E1
Rest(90):	623	Depends upon location	Depends upon location
Total	9388	317[6]	10 E3 + 120 E1[3]

If we compare the number of lines needed with the number of users in each location, we realize that the percentage of employees is not simply proportional to the percentage of lines. This is because as we already said, the call pattern of each business unit is independent of its location. For example (see table 4), Administration users call more than any other type of employee and their percentage of external calls is also higher. Consequently, Administrative employees will mainly determine the number of PSTN lines (see Figure 40).

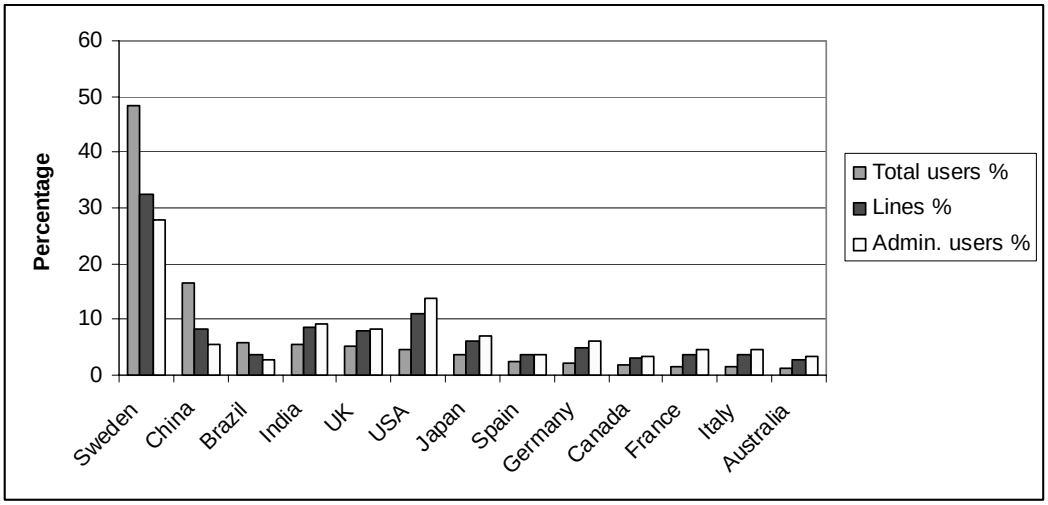

Figure 40: Relationship between required lines (%) and users (%) for each location

[6] We have not considered "Rest" since we don't have data about traffic at each of these locations.

If we plot the influence of each type of user on the number of PSTN line we obtain figure 41:

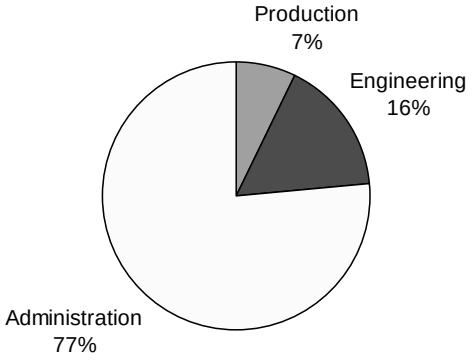

Figure 41: Influence of each type of user with regard to external lines

Regarding the number of PSTN gateways required, as happened with proxy servers the number of calls per second they have to handle is **not** the important factor. What limits a gateway's capacity is the number of lines it supports. The number of these interfaces varies a lot depending on the specific gateway. Supposing we are using two types of gateways: gateway 1 (up to 2 E3 trunks) and gateway 2 (up to 16 E1 trunks), and distributing gateways (thinking about scalability) we would need for example:

Table 16: number of PSTN gateways per location

	Gateway 1	Gateway 2
Sweden	3	0
China	2	0
Brazil	0	1
India	1	0
UK	1	0
USA	1	0
Japan	1	0
Spain	0	1
Germany	0	1
Canada	0	1
France	0	1
Italy	0	1
Australia	0	1
Rest(90):	0	90
Total	9	97

Note that the cause why the number of type 2 gateways is so high is because for every country grouped in rest (90) we need at least one gateway to connect our VoIP network to the PSTN infrastructure of that country. Thus, the enterprise will be able to take advantage of its own network to route the traffic from these countries' PSTN

infrastructures to other PSTNs or our internal network at the price of a national call instead of the price of an international one. However, this may not be cost effective since equipment and maintenance costs could exceed the call cost (remember the number of employees and thus calls in that locations is low). In this later case, one option would be **not** deploy VoIP telephony gateways in these locations.

Another feasible option would be to connect these 'Rest' locations (or most of them) with **other** types of Gateways of smaller capacity (this capacity should be adapted to each location requirements). This solution would allow us to take advantage of VoIP in the entire corporation while reducing acquisition costs. (For example for AudioCodes Mediant2000 gateways the dual (2E1s) solution costs US$8400 while the 16E1s solution rises up to US$32.700)[67].

If we establish a relationship between the number of gateways needed and the number of users, we can estimate the number of users with our call patterns supported per gateway (table 17). Note that in table 17 we only consider type 1 and 2 gateways for simplicity and significance.

Table 17: Employees per gateway type

	Whole network with VoIP	Only main sites with VoIP
Employees per Line	7	7,31
Employees per Gateway 1	7345	7122
Employees per Gateway 2	681	9157

We observe that the number of users per gateway is between 7000 and 9000 if only the main sites deploy VoIP (or locations with low requirements use the other types of gateways) and this number is much lower for gateway type 2 when only two types of gateways are used for providing VoIP capabilities in the whole network.

If we extend these values (second column) for different number of users we obtain figure 42. In this figure we observe that the number of gateways grows, as expected, linearly with the number of employees.

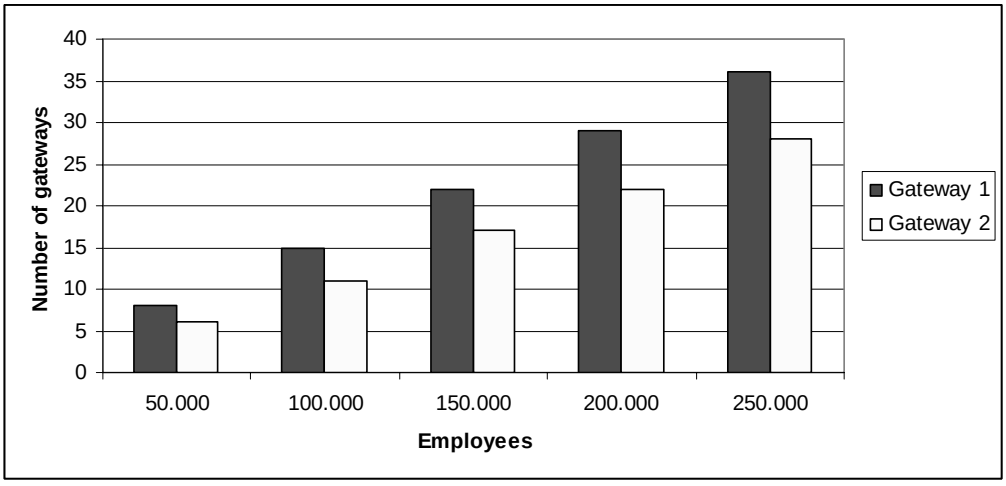

Figure 42: Number of gateways

8.4. IP bandwidth to support VoIP

Once we have calculated the number of servers, we need to estimate the bandwidth necessary in our network to support our call volume. Note that this bandwidth can be high because, in contrast with SIP servers which only take care of signaling packets, the network has to carry also the media streams, which is much more bandwidth per 'consumer'. We are going to divide the analysis into WAN and LAN bandwidth. WAN bandwidth refers to the bandwidth each site needs for connecting to other sites, while LAN bandwidth refers to the specific bandwidth of each location's LAN. The main difference between them is that LAN bandwidth requirements will be higher than WAN ones. This is because each LAN has to support all its own VoIP traffic (intra-site, inter-sites) while the WAN connection only needs to support the inter-site traffic.

8.4.1. WAN bandwidth

In order to estimate the WAN bandwidth requirements of each site we need to know the amount of inter-site calls per site. These data in practice would be different for each location and business unit, but since all our calculations are approximation we will only consider the difference between business units (see Table 18), we estimate these as:

Table 18: Inter-site call percentage per business unit

	Production	Engineering	Administration
Inter-site calls	10 %	20 %	30 %

As we can see in Table 18, most calls are intra-site. However, differences between units are again important and while only 10 % of Production calls are inter-site, Administration inter-site calls raise up to 30 %.

Applying this new data we can easily obtain the number of inter-site simultaneous calls for each location, but in order to estimate the bandwidth in a typical unit (i.e. Kbps) we need to know how many Kbps each call requires. To answer this question we need to look at the packet structure.

An IP packet consists of the payload and some overhead. The overhead of a media packet due to IP, UDP, and RTP is 40 bytes. Using the PPP overhead of 7 bytes (Note that this value would be different if, for example, Frame Relay (9 bytes) were used, but this wouldn't affect very much final results) the total overhead is 47 bytes. However, since WAN links are 'expensive' we can also use compressed RTP (cRTP) [62] which reduces the IP,UDP,RTP overhead to 2 or 4 bytes (we will use the 4 byte value) so the total overhead for each packet would be only 11 bytes.

The VoIP payload is the sampled voice and its size depends on the voice codec used. The two most common codecs are: G.711 with an associated transmission rate of 64 Kbps and the compressed G.729 at 8 Kbps. The packet size is usually determined by the packetization in milliseconds per packet. By multiplying these two values: transmission

rate and milliseconds per packet we can calculate the number of bits of payload per packet.

Putting all values together we can calculate the bandwidth needed for each unidirectional media stream (see Table 19).

Table 19: Bandwidth per unidirectional stream

	G.711 (Kbps)		G.729 (Kbps)	
Packet size	RTP	cRTP	RTP	cRTP
20 ms	80,86	66,80	26,17	12,11
30 ms	74,74	65,36	20,05	10,68
40 ms	71,68	64,35	16,99	9,96

As we can see in the previous table the bandwidth associated with an unidirectional media stream is much lower when the G.729 codec is used. Our intention is to reduce the bandwidth needed as much as possible, thus we will use the G.729 codec with silent suppression (this furthers reduces the bandwidth about 45 %), we obtain our WAN bandwidth requirements (see Table 20). We have also taken into account that each conversation involves a bidirectional media stream (we have not considered multiparty conferencing).

Table 20: WAN bandwidth requirements in Kbps

	G.729 (20 ms, RTP)	G.729 (40 ms, RTP)	G.729 (20 ms, cRTP + suppression)	G.729 (40 ms, cRTP + suppression)
Sweden	72.234,38	46.898,44	**18.382**	**15.120**
China	19.016,48	12.346,52	**4.839**	**3.980**
Brazil	7.364,77	4.781,60	**1.874**	**1.541**
India	17.351,95	11.265,82	**4.416**	**3.632**
UK	16.017,19	10.399,22	**4.076**	**3.352**
USA	20.021,48	12.999,02	**5.095**	**4.191**
Japan	12.146,37	7.886,07	**3.091**	**2.542**
Spain	7.474,69	4.852,97	**1.902**	**1.564**
Germany	8.675,98	5.632,91	**2.208**	**1.816**
Canada	5.739,49	3.726,39	**1.461**	**1.201**
France	6.673,83	4.333,01	**1.698**	**1.397**
Italy	6.673,83	4.333,01	**1.698**	**1.397**
Australia	4.671,68	3.033,11	**1.189**	**977**

As we can see in this table the greatest WAN bandwidth needed is in Sweden - approximately 16 Mbps. In the other locations the requirements are much lower, being around 4 Mbps in 5 other locations and only 2 Mbps in the rest. This bandwidth is not very high since the use of VoIP allows integration of Voice and Data networks and

prices for SDSL connections are cheaper per Mbps than traditional telephony lines (see section 8.6).

8.4.2. LAN bandwidth

To calculate LAN bandwidth we have to perform similar calculations as before but remembering that now each call destined or initiated in that location has to be considered. Now the overhead is due to IP + UDP + RTP + Ethernet or cRTP + Ethernet headers (as there is not PPP inter-site link overhead).

Table 21: LAN bandwidth requirements in Kbps

	G.729 (20 ms, RTP)	G.729 (40 ms, RTP)	G.729 (20 ms, cRTP + suppression)	G.729 (40 ms, cRTP + suppression)
Sweden	452.460,94	284.238,28	**133.998**	**98.903**
China	139.028,91	87.338,67	**41.173**	**30.390**
Brazil	51.553,13	32.385,94	**15.267**	**11.268**
India	75.105,47	47.181,64	**22.242**	**16.417**
UK	69.925,78	43.927,73	**20.708**	**15.285**
USA	77.695,31	48.808,59	**23.009**	**16.983**
Japan	51.278,91	32.213,67	**15.186**	**11.209**
Spain	33.150,00	20.825,00	**9.817**	**7.246**
Germany	33.667,97	21.150,39	**9.970**	**7.359**
Canada	24.344,53	15.293,36	**7.209**	**5.321**
France	25.898,44	16.269,53	**7.669**	**5.661**
Italy	25.898,44	16.269,53	**7.669**	**5.661**
Australia	18.128,91	11.388,67	**5.368**	**3.962**

Again, bandwidth requirements are much higher for Sweden (130 Mbps) than for any other location. This 130 Mbps are important, but for example with current gigabit Ethernet capabilities (1 Gbps), even when data traffic load is added, this shouldn't create a problem.

8.5. Mobile IP

Another important part of the study is the analysis of how mobile users affect the design. Thus, we are going to estimate what implies if we introduce mobility into our network, specifically Mobile IP (see section 3). We are going to estimate both the number of MIP agents needed and the increase in LAN and WAN bandwidth due to MIP overhead.

The first data we need is the number of users who are mobile. This number may differ for each business unit, but in order to make calculations simpler and more general, we are going to simply consider the total fraction of users mobile. This percentage of mobile users is about 10 % of total employees. However we will vary this percentage from 10% to 40 % to see how this would affect the network requirements.

8.5.1. MIP agents

MIP agents' capabilities are determined by two parameters: the number of simultaneous bindings and the number of bindings (registrations) per second they support. If we assume that every endpoint must re-register every 180 seconds and that typically there will be 2 handoffs per user per hour, then we obtain the requirements shown in Table 22.

Table 22: Bindings requirements for different percentages of mobile users

Mobile users %	10%	20%	30%	40%
Simultaneous bindings	6.610	13.220	19.830	26.440
Bindings per second	40,39	80,78	121,18	161,57

If we use data from Cisco Home Agent which when running on a Cisco 7200 router supports up to 262.000 simultaneous bindings and 100 bindings per second [63], we realize that the number of simultaneous bindings is not the dominant factor (as with a single HA we could serve up to 262.000 mobile endpoints). What really determines the number of HAs is the number of bindings per second. However, this number is not very high and with 2 HAs we could serve up to 33.000 mobile users (up to 50% of our company).

If we extrapolate these calculations relative to the number of mobile users instead of the percentage we obtain figure 43.

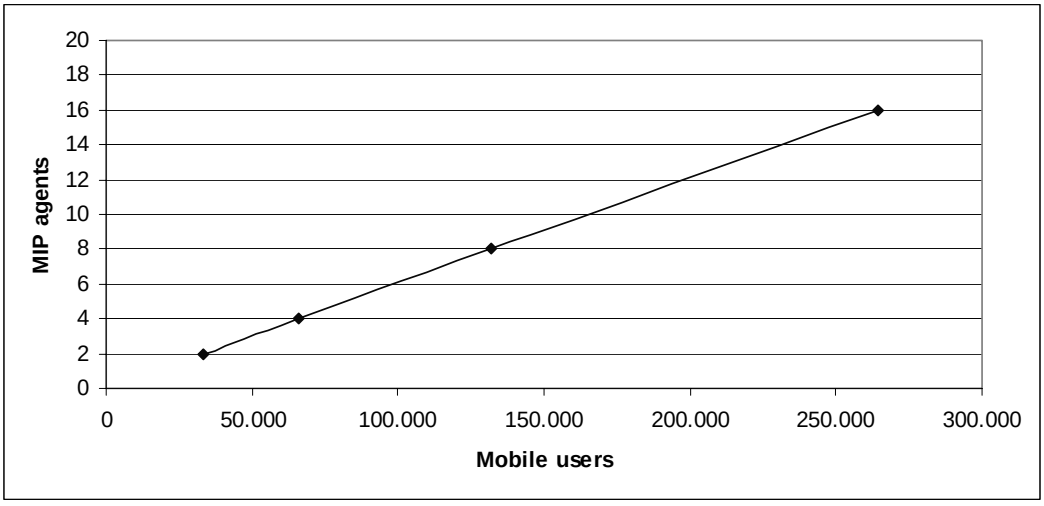

Figure 43: MIP agents' requirements

8.5.2. MIP overhead

The overhead due to Mobile IP can be divided in two main overheads: the overhead due to registration messages and the overhead in call packets.

Registration messages

When a mobile node registers with its Home Agent it sends a packet of about 100 bytes. If we take the same patterns as before (re-registration every 180 seconds and 2 handoffs per user and hour) we see that for 6610 (10%) endpoints using Mobile IP simultaneously the registration overhead is less then 32 Kbps. This value is much smaller than the Registration overhead of SIP (around 500 Kbps). The reasons of this difference are the number of registering users (MIP users are 10% of SIP users), the different size of registration messages, and the registration messages share of SIP.

An option to reduce this overhead would be to share registration information between SIP and MIP servers. Of course, this wouldn't be by replicating messages, but by sharing databases. However, this solution wouldn't reduce overhead significantly in our case because the more frequent messages (MIP re-registration) only correspond to a small percentage (10%) of SIP registered users.

Another option is to increment the time between SIP re-registrations since the use of Mobile IP make them unnecessary. This is because SIP re-registrations will contain the same information (IP address's changes remain unnoticed for upper layers thanks to Mobile IP).

Call messages

When a mobile node is in a foreign network all packets destined to the mobile are intercepted by its home agent which re-sends them through a tunnel to that mobile node or to its foreign agent (see section 3.1.1). In order to tunnel these packets the HA adds a header of (usually) 20 bytes (IP in IP encapsulation). Thus, the overhead for packets (going from the HA to MN/FA) is 20 additional bytes per packet. From the other endpoint to the MN, as packets from the MN to the other endpoint use normal IP routing, there is not additional per packet overhead (unless we must do reverse tunneling). Calculating this overhead is more difficult than before because we would need the number of simultaneous calls in which mobile nodes are outside the home network. However, to give a rough idea of this overhead, if we suppose that 20% of employees are mobile, and 10 % are calling simultaneously during the busy hour, with 40 % outside their home network, and they are using G.729 with 20 ms payload packet size and cRTP, we can obtain the following figure for different number of total employees (Figure 44).

As we can see in Figure 44 the effect of MIP overhead is not negligible, especially as number of users grows. For example for 100.000 users the overhead is more than 20 Mbps. However, this overhead is distributed between locations so the effect in each LAN or WAN network will be less than 20 Mbps. Moreover, the benefits of mobility (see section 3) can easily support the cost of this bandwidth. It should be noted that this is less total bandwidth than 1041 additional calls in the entire corporate network!

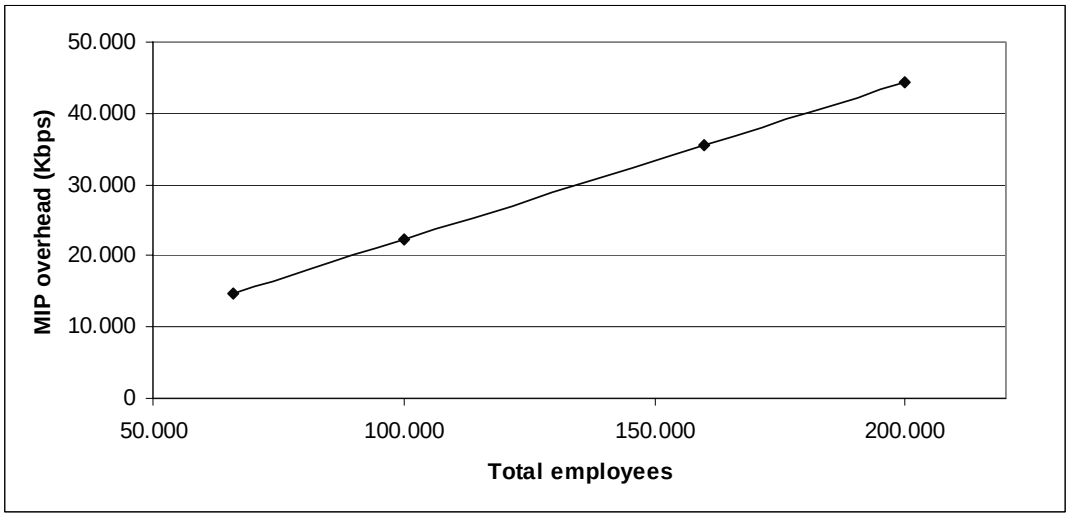

Figure 44: MIP overhead

8.6. Delay considerations

8.6.1. Voice packets delay

Other important parameter when transporting voice packets over an IP network is the delay these packets experiment. The International Telecommunication Union (ITU) recommends maximum one-way delays of 150 ms for voice communications. In practice, however, delay limits can vary for different applications and delays up to 250 ms can be acceptable for some business communications.

This delay can not be estimated accurately in theory since it will mainly rely on network topology and our corporate network uses Internet to route some calls. There are however some delay sources with predictable values:

Coding delay
Compression algorithms such as G.729 reduce the LAN and WAN bandwidth as shown in section 8.4 but, on the other hand, introduce delay. The coding delay of G.729 codec is approximately 15 ms.

Packetization delay
Packetization delay is the time needed to fill a packet payload with the compressed samples. This delay depends on the payload size (number of compressed speech blocks in each packet). In the worst case this delay equals payload size. That is, if we are using 20 ms or 40 ms packets, packetization delay will be 20 ms and 40 ms respectively. That is why in order to decrease this delay it is better to use 20 ms packets rather than 40 ms. This choice increases LAN and WAN bandwidth requirements, but as we have seen in the previous section, the bandwidth difference is not very important.

Serialization Delay
Serialization delay is the time needed to put a packet onto a network interface. This delay depends on the interface speed and on the packet size. For relatively high speed lines (> 512 Kbps) and small packets (< 64 bytes), this delay is smaller than 1 ms.

De-jitter delay
Voice is a real time communication and the relative difference between packets' arrivals (jitter) has to be avoided. This can be achieved by the use of a jitter buffer. However, the use of this buffer introduces an additional delay, usually referred as de-jitter delay. This delay is the time needed to fill the jitter buffer, and thus depends on its size. A typical size is 1.5 times the packet size. In our case using 20 ms packets, delay would be 30 ms.

Decoding delay
De-coding delay for G.729 codec is approximately half of the coding delay. That is 7.5 ms.

The total delay for these sources is around **73.5** ms which means that our voice packets can support additional LAN / WAN one-way delays up to (76,5 ms-176,5 ms).

In the case of the corporate LAN it is not very difficult to achieve this goal when the network is well designed. On the other hand, WAN delays are outside the corporate control, but actual carrier backbone networks usually keep within these limits even in case of transcontinental links [69]

8.6.2. Additional delay considerations

- Call establishment delay, as already commented in section 6.2.3, is small when TLS/MIKEY+ SRTP is used for securing conversations, and it is important when IPsec is used. Thus, the first solution (TLS/MIKEY+SRTP) is preferred.

- Mobile IP routing can introduce significant delay if the direct path between the Mobile Node and the Correspondent Node is much shorter than the path involving the Home Agent. This delay can be important in transoceanic communications. Consequently, in our example, even when only 2 HAs are needed for 33.000 mobile users according to requirements of section 8.5, the best option would be to use at least 3 Home Agents (in Europe, America, and Asia) and to use Dynamic Home Agent assignment methods to reduce this delay.

- Hand-off delays associated with mobile users (VoWLANs and MIP) are high and can entail problems. These delays can be greater than 500 ms what would be noticed by users. Hence, there are many on-going studies in this area with different solutions to reduce this latency, claiming to reduce it down to 100 ms. [70] [71]

8.7. Cost savings

One of the most important benefits of deploying VoIP in a corporate environment is cost savings. These cost saving come from the reduction of cost associated with telephony bills. The two main benefits in this area are:

- **Integration of data and voice networks**
  VoIP calls are routed through the corporate IP network what allows integration of voice and data in one network. With such integration all internal calls make use of the corporate infrastructure (LAN and WAN) instead of needing a separate and more expensive network. For example an E1 voice line (which is able to carry up to 30 simultaneous conversations) costs about 900€ per month. On the other hand using VoIP 30 conversations need less than 598 Kbps (30 x 2 x 9,96 Kbps) using G.729 codec and cRTP (see table 19). These 598 Kbps (half in each direction in a DSL connection) are far cheaper.

- **Elimination of corporate international calls.**
  Using the corporate infrastructure for external international calls changes them from international calls into local or national calls. For example when an employee at the swedish site calls a provider in China, the call instead of going through the swedish PSTN gateway and being delivered by the PSTN infrastructure to China, is routed through the corporate WAN to the corporate site in China were it is gatewayed via the corporate Chinese PSTN gateway and delivered via the chinese PSTN infrastructure at the price of a national or local call. If we calculate the current PSTN non VoIP call volume cost, assuming a cost of 0,005€/min for local calls and 0,15€/min for international calls, we obtain that the corporation spends more than 550.000 € per month in telephone bills of which almost 500.000 € are due to international calls.

If we roughly estimate the acquisition and operational cost, in order to obtain an order of magnitude comparison between costs and savings of deploying IP telephony in our corporation we obtain Table 23 and Table 24:

Table 23: Acquisition costs of IP telephony equipment

Item description		Item cost (€)	Quantity	Total (€)
WAN access	24 Mbps	30.000	1	30.000
	8 Mbps	12.000	4	48.000
	2 Mbps	6.000	98	588.000
PSTN gateways	Type 1	50.000	12	600.000
	Type 2	30.000	10	300.000
	Type 3	8.000	100	800.000
SIP servers		6.000	24	144.000
Mobile servers		20.000	4	80.000
IP telephones		100	30.000	3.000.000
Management tools				300.000
Installation costs (20%)				1.118.000
TOTAL				**7.008.000**

Note that this acquisition costs shown in Table 23 represent less than 1,2% of annual Administrative Expenses or less than 0,35% of annual Research and Development expenses of the company.

Table 24: Operating costs of IP telephony (Monthly)

Item description		Item cost (€)	Quantity	Total (€)
WAN rental	24 Mbps	5.000	1	5.000
	8 Mbps	2.000	4	8.000
	2 Mbps	1.000	98	98.000
Maintenance costs (2% of acquisition costs – installation costs)				117.800
TOTAL				**228.800**

If we compare the monthly costs (228.800 €**)** with the monthly savings in international calls (500.000 €), we obtain that the approximate monthly savings are around 270.000 €. Thus, the acquisitions costs of IP telephony equipment would be amortized in about 26 months.

If we relate these savings to the number of employees we obtain that the approximated annual savings of deploying IP telephony are 50 € per employee.

Note that for these calculations we haven't considered the operational costs of current telephony, the costs associated with equipment also needed for current telephony (since we wanted to analyze costs savings), and we have considered that the number of physical IP telephones is 30.000 (the rest are supposed to be soft-telephones and thus much cheaper). We have also considered the use of redundant equipment.

Note again that costs shown in Table 23 and Table 24 are approximated, and prices are average prices for the equipment related in prior sections.

8.8. New Services

The existence of a corporate VoIP, VPN, and Mobile IP infrastructure provides also the enterprise with new services. Examples of them are:

- ***Mobile presence based services***
  - Stock prices
  - Company news
  - Weather reports
  - Transit schedules
  - Location based services (for example, dispatching the closest available repair person to a problem)

- ***Identity based Services***
  - Personal phone book
  - Services based on seniority
  - Advanced access systems (Phone as a badge)
  - Payment Systems (Cafeteria, vending machines)

- ***Multimedia conferencing***
  - Audio, video and data.
  - Conference rooms

- ***E-mail and Voice-mail messages***

- ***Instant messaging***

- ***Enhanced security***

9. Conclusions and future work

9.1. Conclusion

The overall conclusion of this thesis is that with actual technologies it is feasible to deploy secure and mobile IP telephony in large corporation environments while satisfying essential corporate requirements.

Specific conclusions for this deployment are:

Security

- It is possible to secure access to the corporate intranet resources by the use of VPN technologies. The preferred solution would be IPsec VPN tunneling which provides complete application access. However, in the case of endpoints with low processing capabilities SSL VPN tunneling could be used.

- In addition to VPN technologies, it is necessary the deployment of an admission control system to verify endpoint security.

- To secure media communications established by means of SIP it is preferred the use of TLS/SRTP since requires less processing and introduces less overhead and delay.

Mobility

- In order to complement SIP mobility features, Mobile IP should be implemented to provide mobility at the network layer. Besides, with this solution we provide mobility to all type of users and applications.

- Integration between IPsec and Mobile IP forces the use of IPsec inside MIP tunnels and to situate the MIP agents outside the intranet.

- Actual mobility requirements of large corporations are supported by a small number of MIP agents (only 6 MIP agents can serve up to 100.000 mobile employees). However, when multiple HAs are needed, the best option is to distribute them among the two or three main enterprise locations and to use Dynamic Home Agent Assignment techniques.

- The overhead introduced by Mobile IP in the corporate network is not negligible but the benefits of mobility can support the cost of extra bandwidth.

SIP

- With actual processors the limiting factor for SIP servers is the number of simultaneous users rather than the call volume. Nevertheless, a reduce number of

SIP servers (for i.e. 3 proxies and 12 registrars for 150.000 employees) can serve a large number of employees.

- SIP servers should be situated outside the corporate intranet and when multiple servers are used, the best option is, again, to distribute them among the two or three main sites and to use of DNS SRV for load sharing and redundancy.

- For Registrar servers, it is better to share Databases rather than registration messages. However, when this configuration is not supported, the overhead due to messages between registrars is small and should not suppose a problem.

LAN and WAN

- Actual Ethernet capacities (1 Gbps.), the use of compression, and silent suppression techniques allow the Corporate LAN to support the voice traffic load of a large number of employees.

- The required WAN bandwidth is important in main locations, but actual prices for SDSL connections are much cheaper than prices for traditional telephony lines.

Cost savings

- Main cost savings of implementing IP telephony in a corporate environment come from the elimination of international calls and the integration between voice and data networks. These cost savings are very important and will play a decisive role in large corporations' migration towards IP telephony.

- Annual cost savings are around 50 € per employee, and the deployment costs of IP telephony represent less than 1,2% of annual Administrative Expenses.

9.2. Future work

Some aspects that are not covered in this thesis and will be interesting to investigate are:

- To study in detail the effects of introducing Voice over Wireless LANs in corporate environments.

- To analyze the effects on performance of the integration of voice and data packets in the same network.

- To analyze the deployment of context-aware services [68] in a corporate network.

- To perform a detailed quantitative study for a real company according to their calling patterns and business needs.

REFERENCES

[1] Virtual Private Network Consortium, Feb. 2005, www.vpnc.org

[2] S. Kent and K.Seo “Security Architecture for the Internet Protocol”, IETF RFC 2401, Dec. 2004

[3] Alan O. Freier, Philip Karlton, and Paul C. Kocher “The SSL Protocol Version 3.0”, November 1996, http://wp.netscape.com/eng/ssl3/draft302.txt

[4] Internet Engineering Task Force (IETF), Feb. 2005, www.ietf.org

[5] C. Kaufman, R. Perlman, and M. Speciner, “Network Security: Private communication in a Public World, Second Edition”, April 2002, ISBN 81-7808-790-1

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

[7] S. Kent and R. Atkinson, “IP Authentication Header” IETF RFC 2402, Nov. 1998

[8] S. Kent and R. Atkinson, “IP Encapsulating Security Payload (ESP)”, IETF RFC 2406, Nov. 1998

[9] Cisco Systems, Network Admission Control, accessed February 2005, http://www.cisco.com/en/US/netsol/ns466/networking_solutions_package.html

[10] J. Postel “Internet Protocol” IETF STD 0005, September 1981

[11] S. Kent and K. Seo, “Security Architecture for the Internet Protocol”, IETF RFC 2401, December 2004

[12] T. Dierks and C. Allen, “The TLS Protocol Version 1.0”, IETF RFC 2246, January 1999

[13] C. Perkins, “IP mobility support for IPv4”, IETF RFC 3344, August 2002

[14] R. Moskowitz, P. Nikander, P. Jokela, and T. Henderson, “Host Identity Protocol” IETF Network Working Group, internet-draft, February 21, 2005

[15] Wireless Application Forum, “Wireless Application Protocol, Wireless Transport Layer Security Specification” November 5, 1999

[16] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, and E. Schooler, “SIP: Session Initiation Protocol”, IETF RFC 3261, June 2002

[17] T. Berners-Lee, R. Fielding, and L. Masinter, “Uniform Resource Identifier (URI): Generic Syntax”, IETF RFC 3986, January 2005

[18] G. Q. Maguire Jr., “Lecture notes of the course 2G1325/2G5564 Practical Voice Over IP (VoIP): SIP and related protocols”, Module 3: 131 of 194, March 2004

[19] G. Q. Maguire Jr., “Lecture notes of the course 2G1325/2G5564 Practical Voice Over IP (VoIP): SIP and related protocols”, Module 3: 147 of 194, March 2004

[20] A. Gulbrandsen, P. Vixie, and L. Esibov, "A DNS RR for Specifying the Location of Services (DNS SRV)", IETF RFC 2782, February 2000

[21] C. Perkins, “IP mobility support for IPv4”, IETF RFC 3344, August 2002

[22] B. Chambless, “HARP Home Agent Redundancy Protocol”, IETF Internet Draft, October 27, 1997

[23] “Mobile IP Home Agent Redundancy”, Cisco IOS software releases 12.0 T, April 2005

[24] T. Li, B. Cole, P. Morton, and D. Li, “Cisco Hot Standby Router Protocol (HSRP)”, IETF RFC 2281, March 1998

[25] M. Kulkarny, A. Patel, and K. Leung, “Mobile IPv4 Dynamic Home Agent Assignment”, IETF Mobile IP Working Group Internet Draft, draft-ietf-mip4-dynamic-assignment-03.txt, September 2004

[26] P. Calhoun and C. Perkins, “Mobile IP Network Access Identifier Extension for IPv4”, IETF RFC 2794, March 2000

[27] B. Aboba and M. Beadles, “The Network Access Identifier”, IETF RFC 2486, January 1999

[28] Cisco, “Priority HA Assignment”, Mobile Networks documentation, http://www.cisco.com/univercd/cc/td/doc/product/software/ios122/122newft/122t/122t15/ftdynaha.htm, April 2005

[29] Cisco, “Gateway Load Balancing Protocol”, Technology white paper, http://www.cisco.com/en/US/tech/tk869/tk769/technologies_white_paper0900aecd801790a3.shtml, April 2005

[30] R. Hinden, Ed, “Virtual Router Redundancy Protocol (VRRP)”, IETF RFC 3768, April 2004

[31] T. Li, B. Cole, P. Morton, and D. Li, “Cisco Hot Standby Router Protocol (HSRP)”, IETF RFC 2281, March 1998

[32] F. Adrangi and H. Levkowetz, “Problem Statement: Mobile IPv4 Traversal of VPN Gateways”, Mobile IP Working Group, October 4, 2004

[33] Israel M. Abad Caballero, “Secure Mobile Voice over IP”, M.Sc. Thesis, Royal Institute of Technology (KTH), Stockholm, Sweden, June 2003

[34] Johan Bilien, “Key Agreement for Secure Voice over IP”, M.Sc. Thesis, Royal Institute of Technology (KTH), Stockholm, Sweden, Dec. 2003

[35] Minisip SIP user agent, http://www.minisip.org, accessed April 2005

[36] J. Bilien, E. Eliasson, and J-O Vatn, “Call establishment delay for secure VoIP”, WiOpt'04, Cambridge UK, March 2004

[37] J. Orrblad, “Alternatives to MIKEY/SRTP to secure VoIP”, M.Sc. Thesis, Telecommunication System Laboratory, Royal Institute of Technology (KTH), Stockholm, Sweden, March 2005

[38] Fredrik Thernelius, “SIP, NAT, and Firewalls”, M.Sc. Thesis, Department of Teleinformatics, Royal Institute of Technology (KTH), Stockholm, Sweden, May 2000.

[39] G. Montenegro and M. Borella, “RSIP Support for End-to-end IPsec”, IETF RFC 3104, October 2001

[40] P. Srisuresh, J. Kuthan, J. Rosenberg, A. Molitor, and A. Rayhan, “Middlebox Communication Architecture and framework”, IETF RFC 3303, August 2002

[41] Amos Muhunda Nungu, “VoIP Service Provider: Internet Telephony Service Provider using SIP Protocol”. M.Sc. Thesis, Telecommunication System Laboratory, Royal Institute of Technology (KTH), Stockholm, Sweden, April 2005.

[42] A. Huttunen, B. Swander, V. Volpe, L. DiBurro, and M. Stenberg, “UDP Encapsulation of IPsec ESP Packets”, IETF RFC 3948, January 2005.

[43] G. O'Donnell, "Centralizing the Command Center Function", META Group Report, Nov. 2001

[44] Microsoft Corp., "Microsoft's System Management Server (SMS) 2003", May 2005 http://www.microsoft.com/smserver/techinfo/productdoc/default.asp

[45] Novell Inc., "Zenworks Suite", May 2005, http://www.novell.com/es-es/products/zenworks/

[46] IETF Policy Framework Working Group, May 2005, http://www.ietf.org/proceedings/02nov/166.htm

[47] D. Harrington, R. Presuhn, and B. Wijnen, "An Architecture for Describing Simple Network Management Protocol (SNMP) Management Frameworks", IETF RFC 3411, December 2002

[48] D. Durham, Ed., J. Boyle, R. Cohen, S. Herzog, R. Rajan, and A. Sastry, "The COPS (Common Open Policy Service) Protocol", IETF RFC 2748, January 2000

[49] D. Rawlins, A. Kulkarni, M. Bokaemper, and K. Chan, "Framework for Policy Usage Feedback for Common Open Policy Service with Policy Provisioning (COPS-PR)", IETF RFC 3483, March 2003

[50] J. Strassner, B. Moore, R. Moats, and E. Ellesson, "Policy Core Lightweight Directory Access Protocol (LDAP) Schema", IETF RFC 3707, February 2004

[51] K. Avgeropoulos, "Service Policy Management for User Service Policy Management for User-Centric Centric Services in Heterogeneous Mobile Networks", Master's thesis, Royal Institute of Technology (Stockholm), March 2004 http://vvv.it.kth.se/docs/Reports/DEGREE-PROJECT-REPORTS/040401-Konstantinos_Avgeropoulos-with-cover.pdf

[52] M. Stevens, W. Weiss, H. Mahon, B. Moore, J. Strassner, G. Waters, A. Westerinen, and J. Wheeler, "Policy Framework", IETF internet draft, September 1999, <draft-ietf-policy-framework-00.txt>

[53] Asterisk IP PBX, www.asterisk.org, accessed June 28, 2005

[54] SIP Express Router, www.iptel.org/ser/, accessed June 28, 2005

[55] Cisco SIP Proxy Server, www.cisco.com/univercd/cc/td/doc/product/voice/sipproxy , accessed June 29, 2005

[56] Avaya S8700 Media Server, www.avaya.com, accessed June 29, 2005

[57] Alcatel OmniPCX enterprise, www.alcatel.com, accessed June 27, 2005

[58] E. Mier, D. Mier, R. Tarpley, "BCR best in test: Which large IP-PBX rules?", Business Communications Review, Jan 2005, pg 24-37

[59] http://www.dell.com/downloads/global/power/ps2q05-20050103-Fruehe.pdf, accessed July 1, 2005

[60] ftp://ftp.software.ibm.com/eserver/benchmarks/news/newsblurb_x335_SPECcpu2000_022404.pdf, accessed July 1, 2005

[61] www.spec.org, accessed July 1, 2005

[62] T. Koren, S. Casner, J. Geevarghese, B. Thompson, P. Ruddy, "Enhanced Compressed RTP (CRTP) for Links with High Delay, Packet Loss and Reordering", IETF RFC 3545, July 2003

[63] Cisco 7200 series routers, http://www.cisco.com/en/US/products/hw/routers/ps341/, accessed July 1, 2005

[64] A. Stubblefield, J. Ioannidis, and A. D. Rubin, “Using the Fluhrer, Mantin, and Shamir Attack to Break WEP”, August 2001.

[65] UIT-D, Comisión de Estudio 2, Cuestión 16/2, “Manual sobre ingeniería de Teletráfico”, December 2002. www.itu.int

[66] Cisco Secure Desktop Security Suite, http://www.cisco.com/en/US/products/sw/secursw/ps2308/products_configuration_guide_book09186a008042102a.html, accessed July 2005

[67] Prices from www.voipsupply.com at July 12, 2005

[68] Adaptive and Context-Aware Services (ACAS) project, Wireless center, Royal Institute of Technology, 21 March 2003. http://psi.verkstad.net/acas

[69] Data extracted from BT global services, http://ippm.bt.net/month.shtml, last accessed July 2005.

[70] S. Sharma, N. Zhu, T. Chiueh, “Low-Latency Mobile IP Handoff for Infrastructure-Mode Wireless LANs”, IEEE Journal on Selected Areas in Communication, Special issue on All IP Wireless Networks, May 2004.

[71] Dynamics Mobile IP project, http://dynamics.sourceforge.net/, last accessed July 2005.

IMIT/LCN 2005-20

www.kth.se