View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Aaltodoc Publication Archive

TKK Dissertations 20
Espoo 2005

SECURING MILITARY DECISION MAKING IN A NETWORK-CENTRIC ENVIRONMENT

Doctoral Dissertation

Catharina Candolin

Helsinki University of Technology
Department of Computer Science and Engineering
Laboratory for Theoretical Computer Science

TKK Dissertations 20
Espoo 2005

SECURING MILITARY DECISION MAKING IN A NETWORK-CENTRIC ENVIRONMENT

Doctoral Dissertation

Catharina Candolin

Dissertation for the degree of Doctor of Science in Technology to be presented with due permission of the Department of Computer Science and Engineering for public examination and debate in Auditorium T2 at Helsinki University of Technology (Espoo, Finland) on the 20th of December, 2005, at 12 noon.

Helsinki University of Technology
Department of Computer Science and Engineering
Laboratory for Theoretical Computer Science

Teknillinen korkeakoulu
Tietotekniikan osasto
Tietojenkäsittelyteorian laboratorio

Distribution:
Helsinki University of Technology
Department of Computer Science and Engineering
Laboratory for Theoretical Computer Science
P.O. Box 5400
FI - 02015 TKK
FINLAND

© 2005 Catharina Candolin

ISBN 951-22-7980-0
ISBN 951-22-7981-9 (PDF)
ISSN 1795-2239
ISSN 1795-4584 (PDF)
URL: http://lib.tkk.fi/Diss/2005/isbn9512279819/

TKK-DISS-2085

Picaset Oy
Helsinki 2005

HELSINKI UNIVERSITY OF TECHNOLOGY
P. O. BOX 1000, FI-02015 HUT
http://www.hut.fi/

ABSTRACT OF DOCTORAL DISSERTATION

Author Catharina Candolin

Name of the dissertation
Securing military decision making in a network-centric environment

Date of manuscript 28.11.2005 | Date of the dissertation 20.12.2005

☒ Monograph ☐ Article dissertation (summary + original articles)

Department Department of Computer Science and Engineering
Laboratory Laboratory for Theoretical Computer Science
Field of research
Opponent(s) Professor Matthew Warren, Deakin University, Australia
Supervisor (Instructor) Professor Hannu H. Kari

Abstract

The development of the society and warfare goes hand in hand. With the proliferation of modern information technology, in particular communication technology, concepts such as information warfare and network-centric warfare have emerged. Information has become one of the core elements in military decision making, where the purpose is to gain information superiority with respect to the enemy while denying the enemy from doing the same. Network-centricity comes from the fact that communication networks are used to enable information warfare in the theatre of operations. Thus, the role of the communication network is to support decision making.

In this thesis, military decision making in a network-centric environment is analyzed from the perspective of information warfare. Based on the analysis, a set of security requirements are identified. The thesis also proposes a set of solutions and concepts to the vulnerabilities found and analyzes the solutions with respect to the requirements and a set of use scenarios. The main solutions are Packet Level Authentication, which secures the military infrastructure, and Self-healing Networks, which enable the network to restructure itself after a large-scale or dedicated attack. The restructuring process relies on a Context Aware Management architecture, which has originally been developed to allow network nodes to rapidly react to a changing environment. Furthermore, the thesis presents a trust management model based on incomplete trust to cope with compromised nodes. Also privacy issues are discussed; several different privacy classes are identified and the problems with each of them are addressed.

Keywords security, communication network, information warfare, network-centric warfare, decision making

ISBN (printed) 951-22-7980-0 | ISSN (printed) 1795-2239
ISBN (pdf) 951-22-7981-9 | ISSN (pdf) 1795-4584
ISBN (others) | Number of pages 141

Publisher Helsinki University of Technology, Laboratory for Theoretical Computer Science

Print distribution Helsinki University of Technology, Laboratory for Theoretical Computer Science

☒ The dissertation can be read at http://lib.tkk.fi/Diss/2005/isbn9512279819/

Table of Contents

Acknowledgments v
1 Introduction 1
1.1 Problem statement 3
1.2 Own contributions 3
1.3 The structure of the thesis 5
2 The development of warfare 7
2.1 Information warfare 9
2.2 Network-centric warfare 14
3 The development of communication technology 16
3.1 The Internet 16
3.2 Wireless networks 18
3.2.1 Satellite communication 21
3.2.2 Wireless Wide Area Networks 22
3.2.3 Wireless Metropolitan Area Networks 23
3.2.4 Wireless Local Area Networks 24
3.2.5 Wireless Personal Area Networks 25
3.3 Mobility management 26
3.3.1 Node mobility 27
3.3.2 Network mobility 45
3.4 Ad hoc networks 46
4 Decision making and the OODA loop 48
5 Decision making in a network-centric environment 53
5.1 Security in a network-centric environment 58
5.2 Attack scenarios on the infrastructure 60
5.2.1 Attacking the infrastructure 60
5.2.2 Destroying the infrastructure 62
5.2.3 Compromised nodes 64
5.2.4 Network surveillance 67
5.3 Security criteria for network level security 68
6 Solution 70
6.1 Packet level authentication 70
6.1.1 Previous solutions 71
6.1.2 The packet level authentication concept 72
6.1.3 Design criteria 73
6.1.4 Design issues of packet level authentication 74
6.1.5 Packet level authentication extension header 75
6.1.6 Analysis of PLA 76
6.1.7 Deploying PLA 78

6.2 Context Aware Management Architecture 80
6.2.1 Deploying the CAM architecture 85
6.3 Self-healing networks 86
6.3.1 Deploying self-healing networks 89
6.4 Trust management based on incomplete trust 91
6.4.1 Deploying incomplete trust 94
6.4.2 Ad hoc network routing based on incomplete trust 95
6.5 Privacy protection 98
6.5.1 Data privacy 99
6.5.2 Identity privacy 100
6.5.3 Location privacy 101
6.5.4 Existence privacy 102
6.5.5 Time privacy 103
6.5.6 Transaction privacy 103
7 Analysis 104
7.1 Requirement analysis 104
7.1.1 Performing network operations 104
7.1.2 Coping with network destruction 106
7.1.3 Managing trust 109
7.1.4 Privacy protection 112
7.2 Scenario analysis 114
7.2.1 Attacking the infrastructure 114
7.2.2 Destroying the infrastructure 115
7.2.3 Compromised nodes 116
7.2.4 Network surveillance 120
8 Conclusion 121
8.1 Applicability to the civilian environment 126
8.1.1 PLA in a civilian environment 127
8.1.2 Context Aware Management 129
8.1.3 Self-healing networks 129
8.1.4 Incomplete trust 130
8.1.5 Privacy protection 130

List of Figures

Figure 1 The scope of information warfare 10
Figure 2 The Internet architecture design 19
Figure 3 Network technologies 20
Figure 4 A WiMax cell 24
Figure 5 Mobility categories 28
Figure 6 A conceptual model of Mobile IPv6 31
Figure 7 The traditional stack compared to the HIP stack 41
Figure 8 The problem with double jump signaling 44
Figure 9 Network mobility 45
Figure 10 The OODA loop 49
Figure 11 A more detailed picture of the OODA loop 50
Figure 12 Interaction between a human and the network-centric environment 53
Figure 13 Interaction between a network node and the network-centric environment 54
Figure 14 The OODA loop from a network-centric environment 56
Figure 15 A conceptual model of the network architecture 57
Figure 16 Three levels of security 59
Figure 17 Two illegitimate nodes attacking the network 62
Figure 18 A partially destroyed network 63
Figure 19 A targeted physical attack on the infrastructure 64
Figure 20 A network with compromised nodes 65
Figure 21 Compromised nodes affect data fusion 66
Figure 22 The network is under enemy surveillance 67
Figure 23 Link level security 72
Figure 24 The PLA header 76
Figure 25 Establishing security associations in PLA 78
Figure 26 Validation of network nodes 79
Figure 27 The CAM architecture 82
Figure 28 The network recovery protocol 89
Figure 29 The state graph for the recovery process 90
Figure 30 The state graph of a node 91
Figure 31 The process of making a transaction 94
Figure 32 Routing based on incomplete trust 96
Figure 33 IPSec security 100
Figure 34 Location privacy in a wireless network 102
Figure 35 A node is assigned the task of connecting two partitions of a network 108
Figure 36 Revoking the certificate of a compromised node 112

Figure 37 Choosing a new access router 116
Figure 38 Strategic placement of malicious nodes..... 118

Acknowledgments

This thesis is the result of three years of research in the area of security, information warfare, and military decision making. The research has been carried out both at Helsinki University of Technology (HUT) and at the National Defence College (NDC) in Finland. A significant part of the work has also been done during my various trips. While HUT and NDC both have provided me with the infrastructure needed to carry out this research, my extensive traveling has broadened my perspectives of the area of my research and has left me with a world-wide social network. The two universities, together with my freedom of travel, have shaped this thesis into its current form.

I owe my gratitude to several parties for supporting my research.

I thank the Finnish Defence Forces for funding my research. I especially thank the National Defence College for providing me with an office on the island of Sandhamn, thus making it possible for me to collaborate with people with a military background and perspective. I am especially grateful to Col. Esa Lappalainen at the Department of Technology for supporting my work and for providing me with the infrastructure I needed. Cdr. Auvo Viita-aho also contributed to this work through our various discussions and his comments and support. Furthermore, the whole Dept. of Technology provided me with an open-minded, yet critical environment, in which to work. I thank all my colleagues for all the fruitful discussions during my time there, whether related to my research or not, as well as for accepting me as one of their own.

Helsinki graduate school in Computer Science and Engineering (HeCSE) supported me with funding for 4 years, which made it possible for me to publish and present my research in international academic, commercial, and military forums. Thanks to the funding, I have been able to establish a large network and gain experience from working in various environments. This cooperation both on a national and an international level has been one of the things I have valued the most during my research.

Several people have influenced my thoughts and interests while carrying out the research. Lieut. Col. Rauno Kuusisto was the one who initially, thanks to his enthusiasm and expert knowledge, got me interested in military networking in the first place. Professor Aki-Mauri Huhtinen at the Department of Management and Leadership not only supported my studies

at the National Defence College, but has from the very beginning provided valuable insights into the area of information warfare. Lieut. Col. Sakari Ahvenainen (ret.) has always been open to discussions and has often taken the time to comment on my ideas. Ray McGowan from the U.S. Army Research Laboratory has also promoted my work by commenting it and by giving me the opportunity to participate in various events.

The thesis has also greatly benefited from the valuable review done by Dr. Howard Marsh from the Office of Naval Research in the U.S., and Professor Olli Martikainen from University of Oulu. I thank both reviewers for having taken the time to evaluate my thesis.

One person that I owe my gratitude to is my mentor and friend, Docent Arto Karila. Arto has always, regardless of how busy he has been, found time to study my work and give me constructive criticism. Without his help and support, I doubt that I would ever have been able to finish this thesis.

I also owe my gratitude to Professor Hannu H. Kari for supervising my thesis and for all the support he has provided me during the years. Hannu has always been there when I have needed help with my work. I thank Hannu for all the fruitful discussions we have had as well as for his patience. I believe Hannu is the best supervisor a student could ever wish for.

Finally, I thank my parents and my brother for always being there for me.

Catharina Candolin
Espoo, 28.11.2005

1 Introduction

There are roughly 45 wars waged in 35 countries, six genocides, and nine nests of violence taking place continuously [2]. The number of wars has not changed significantly throughout later history, but the face of war has changed with the development of society. The development of the society and the development of warfare goes hand in hand. In fact, the concept of war hardly existed before the first societies emerged around fertile soil, during which period wars were fought for agricultural reasons. The emergence of the nation-state gave rise to national armed forces, and once industrialization began, weapons and materiel could be mass produced. Advances in science and technology resulted in the development of communication systems, which gradually led the societies into the so called Information Age. The key factors in this development were information technology and communication networks, which enabled information to be collected, processed, analyzed, and distributed efficiently.

The Information Age also brought changes into warfare, and new concepts such as information warfare and network-centric warfare emerged. Information in warfare is nothing new, but the role of information has changed. It is no longer used solely for intelligence, but has become a target and a weapon in itself. With the help of information and communication technology, it is possible to efficiently spread information - or disinformation - anywhere in the world. In network-centric warfare, information plays a key role in providing real-time situation awareness. The network allows information to be collected and distributed over large geographical areas in practically no time. The main objective is to gain information superiority over the enemy, and to thus be able to make more accurate decisions while denying the enemy from doing the same.

As a result of this change, the focus has shifted from mass- and platform-centricity to information and network-centricity. Although the transformation is done to enhance the quality of the decision making, it also comes with some challenges. Since information can be targeted, it can be used as a means of affecting decision making. Networks are known to be vulnerable to attacks, hence opening up another opportunity for affecting decision making. Furthermore, the decision making process typically produces a decision which is to be implemented. This decision can be communicated over the network. Also the network itself may

change as a result of the action taken. Hence, there is a clear interaction between the decision making process and the network-centric environment with respect to information.

To protect the decision making process from attacks stemming from the network-centric environment, the network-centric environment must be secured. The current trend in military networking is to rely on open standards and commercially off the shelf (COTS) products, such as IP technology. The problem with Internet technologies is that the Internet was built upon technology developed to withstand nuclear attacks, but security was never a main concern until the Internet was commercialized. Adding security into existing protocols is never a trivial task. Hence, regardless of the multitude of available security solutions, the networks are still vulnerable to attacks. This is a problem already in the current Internet, which for the most parts is static, wired, and operates in a civilian, non-hostile environment. The problems become worse when the same technologies are deployed in a military environment, which is hostile, dynamic, mainly wireless, and affected by weather conditions, terrains, etc. Unless security can be ensured in this environment, the network becomes useless as a tool for decision making and military leadership.

The problems with securing the network in the military environment are many-fold. The dynamic nature of the network means that security associations change on a frequent basis as network nodes move. Due to the large size of the network, it is impossible for nodes to store security associations of other nodes than those they are currently communicating with. Hence, whenever a change in the network occurs, new security associations must be negotiated. Traditional protocols are typically too heavy to be used in practice where the network resources are scarce and the nodes may move too rapidly for the negotiation to have time to finish. The wireless network medium is not only vulnerable to attacks, it also leaks important information about the network and the intent of the forces to the enemy. Hiding the network is practically impossible. Furthermore, the hostility of the environment introduces completely new challenges, such as compromised nodes. A compromised node is a legitimate node which is controlled by the enemy. Hence, the attacks stem from within the network, and all traditional security solutions are bypassed. Cryptography is hardly of any use as the compromised node has access to the cryptographic keys that it needs.

If the security of the network fails, the enemy is able to affect the decision making process. Information may be denied, delayed, or modified, hence affecting the situational awareness. Eventually, information will contradict

with the situation at hand, causing confusion, and thus delaying decision making even further. Hence, the decisions made may be wrong or delayed, both of which benefit the enemy.

1.1 Problem statement

In a network-centric environment, the network provides support for decision making by allowing information to be collected, processed, analyzed, and distributed in an efficient manner. The importance of space and time has decreased since the network allows communication between entities regardless of location. The network can provide a decision maker with real-time situation awareness, based on which the decision maker can take a course of action, and communicate his intent to other entities.

In this thesis, the interaction between the network-centric environment and the decision making process is analyzed. The objective of the thesis is to show how decision making can be affected through the network as well as to propose a set of solutions to secure the network and thus protect the decision making process. Decision making is studied both from a human and computer perspective. From the human perspective, the role of the network is to provide accurate and timely information. From the computer perspective, the signaling data must be sound to enable optimal network operation.

The vulnerability analysis is made based on a set of attack scenarios. In the scenarios, the main problems with military networking are described. Based on the scenario analysis, the main security requirements are identified and listed. The solutions proposed to the discovered problems are analyzed with respect to the requirements and the attack scenarios. It is shown that while some of the requirements can be met under many circumstances, open problems still exist, and some solutions contradict each other with respect to the requirements.

1.2 Own contributions

This thesis contains the following contributions:

- An analysis of how the decision making process can be affected in a network-centric environment. The decision making process is

modeled with the OODA loop. The analysis focuses on the role of the network with respect to the OODA loop and shows possible vulnerabilities. The decision making process with respect to the network-centric environment is described in Section 5.

- Security requirements to the communication network to protect the decision making process. Based on the previous analysis, a set of security requirements are listed to cover the main problems. The security requirements are identified in Section 5.
- Analysis of the impact of compromised nodes to the network. Traditionally, security solutions have been developed with the aim of protecting the network from external attacks. However, in a military environment, where nodes are likely to become compromised, the attacks may stem from within the network. In such a case, the traditional security solutions have been bypassed. The analysis focuses on the damage that compromised nodes can impose on the network as well as how they can affect the decision making process. The problem with compromised nodes is analyzed in Section 5.

Furthermore, the thesis proposes a set of technical solutions or models to cope with the problems. The thesis contains the following contributions:

- Packet level authentication (PLA). The PLA architecture has been developed to secure the infrastructure. PLA makes it possible to verify the authenticity of a sender as well as the integrity, timeliness, and uniqueness of an IP packet. Based on this, it is possible to limit denial-of-service attacks, secure protocol signaling, provide access control, and so on.
- Context Aware Management (CAM). The CAM architecture enables a node to rapidly adapt to a frequently changing environment.
- Self-healing networks. A method for a partially destroyed network to automatically rebuild itself has been developed. The method relies on PLA for security and on the CAM architecture to allow nodes to adapt to the changes. The basic idea is to replace destroyed nodes by any available nodes in the environment by assigning them new tasks.
- Trust management based on incomplete trust. The concept of incomplete trust is introduced as an idea of how to perform trust management in a hostile environment with compromised nodes. The incomplete trust model is more flexible than previous models

which assume that nodes are either completely trustworthy or completely untrustworthy.

- Privacy protection. The thesis does not propose any new privacy solutions, however, privacy is divided into six different categories, and the problems and most important current solutions are identified and discussed.

The solutions are described in Section 6.

1.3 The structure of the thesis

The rest of this thesis is structured as follows:

Section 2 presents a short history of the development of warfare. This development is compared to the development of the society; only when societies started to emerge did the concept of warfare develop. The Section describes the development of warfare from the agricultural society to the information society of today.

The information society is a result of the development of information and communication technology. The trend in military networking is to rely on open standards and commercially-off-the-shelf products. In Section 3, the development of networks based on open standards, mainly IP, is described. The Section starts with the development of the Internet and follows the development into wireless, mobile, and ad hoc network technologies.

The communication network, however, is not a means in itself. The purpose of the network in a military scenario is to enhance decision making. Section 4 presents the OODA loop, which is a decision making model created by Col. John Boyd (deceased) from the US Air Force.

In Section 5, the OODA loop is applied to a network-centric environment. This Section analyzes various means of affecting decision making through the communication network. Possible attack scenarios are presented, and it is shown that the networks present a vulnerability into the decision making process. Hence, a set of security requirements is listed.

In Section 6, a set of solutions to deal with the security problems are presented. The solutions attempt to cover the requirements presented in the previous Section.

In Section 7, the solutions are analyzed with respect to the security requirements and the attack scenarios. While many problems can be solved by the new solutions presented, some issues still remain unsolved.

Finally, Section 8 concludes the thesis with summarizing the results and analyzing the applicability to the civilian environment.

2 The development of warfare

The development of society typically affects the development of warfare. In [88], Toffler describes the changes of societies as a succession of rolling waves. Before the First Wave, people lived in small migratory groups who fed themselves by hunting, fishing, collecting, and herding. The First Wave of change gave rise to the agricultural society; villages and settlements started to form around areas with fertile soil. From roughly 8000 B.C. to the 18th century A.D. the agricultural society was the dominating way of life. The Second Wave started to take over roughly in 1650-1750, when industrialization slowly began to evolve. In the 19th century, industrialization had changed the society; new inventions emerged, technology advanced, and mass production began. As technology, especially information technology, advanced further in the middle of the 20th century, the Second Wave started to lose force and the Third Wave started to build up and approach. The Third Wave brought the society into the Information Age.

In [89], Toffler argues that the development of warfare can be described using the same waves.

Before the First Wave, battles occurred in the pre-agricultural tribes, however, violence is not synonymous to warfare. It was during the First Wave that the permanent settlements with their social and political environment gave birth to the concept of war. First of all, the agricultural society was able to produce and store an economic surplus worth fighting for. Second, it hastened the development into nation states. Typically, warfare in the First Wave was limited to a small area, and people gathered to fight did so because they had similar interests, namely to defend one's own land or acquire more land from another society. Weapons were hand made, non-standardized, and mainly designed for close combat. Communications were primitive; orders were mainly oral and not written. Although the size, capability, morale, leadership quality, and training between armies of the First Wave varied a lot, they all had one thing in common: warfare was about agriculture.

The Second Wave brought mass production, standardization, and technology into the army. Napoleon introduced the mass army in the beginning of the 19th century, however, the industrialized army did not appear until the 1840s when mass production of weapons became

common. The way wars were fought had changed radically. The armies no longer belonged to a warlord, local landowner, or clan leader, but to a nation-state. Wars were fought over large areas and the amount of human and materiel resources were huge. The concept of mass destruction became a prominent factor for industrial-age warfare. Advances in technology also found their place on the battlefield. Towards the end of the century, the first electrical communication networks emerged. Morse had refined his code in the 1840s, telegraph stations soon started to emerge, and slowly thereafter the wireless spectrum could be utilized. For example, Great Britain cabled most of its empire and built wireless transceiver stations around the coasts to communicate with its fleet. It has even been argued that the British Empire owed its existence to its network rather than its Navy. Also Germany had linked its possessions and added wireless stations, and by 1914 it was claimed to have the most advanced network in the world. The first telegraphic messages over cables seemed to have been sent during the Crimean war, although their significance for military intelligence was limited, partially because of the messages arriving as late as 24 hours after having been sent due to the number of relay stations on the way, partially because the messages could be error prone and many commanders simply decided to ignore them. By the outburst of the First World War, however, the usage of communications had been more widely adopted by the militaries in Europe (in particular). [43]

The Third Wave brought significant changes into warfare. Although warfare today is still a combination of Second and Third Wave factors, the emphasis on information has become prominent. Information has of course always been important in warfare, but currently it is being placed at the core of military power. One indicator of its importance is the computerization and use and development of advanced communication technology. The concept of "information warfare" has developed. In the old days, information was used mainly as a guidance as how to deploy and move the forces. In the information age, information is more than mere intelligence; it is used both as a target and a weapon. Technology has provided the means to collect, process, and distribute large quantities of information (or rather, data) in practically no time, thus eliminating the previous restrictions caused by space and time. The purpose of information now is to acquire a real-time situation awareness and preventing the enemy from doing the same, thus gaining information superiority. Together with technology and the new role of information, the trend in war fighting is towards network-centric forces consisting of small units equipped with intelligent technology and weapon systems with more accuracy and firepower.

2.1 Information warfare

Information in warfare is nothing new. Throughout history, information has always played a key role as a contributor to victory. Sun Tzu [86] and Clausewitz [15] both emphasized the role of information in warfare: Sun Tzu talks about knowing oneself and the enemy, while Clausewitz describes warfare using terms such as fog (uncertainty) and friction (the unforeseeable incidents that lower the level of performance so that one falls short of ones goals). Until recently, the role of information has been related to intelligence, i.e. knowledge of where the enemy forces are, what their strength is, and so on. With the proliferation of modern information technology, the role of information has changed. It is not limited to intelligence, but has also become both a target and a weapon. This has given rise to the term information warfare. Although enabled by information systems, information warfare is not about technology, but about information. Information systems can be utilized as a means of conducting information operations, but they are by far not the only viable means. For example, psychological operations can be carried out with the use of images and the media. The role of technology here is that it makes it possible to create, process, and distribute enormous amounts of images to a large community in a short period of time. Yet, it is the images that are the "information", for example, they may be used as a weapon to affect public opinion that will put pressure on political and military decision making.

There is no single definition of information warfare or information operations. In general, information warfare is considered to be the information operations conducted during time of crisis or conflict to achieve or promote specific objectives over a specific adversary or adversaries. It is the act of influencing the civil or military decision making, operational capability, and the public opinion by using information and information processing both as a target and a weapon, and to protect oneself against such influence. Thus, information warfare has both an offensive and a defensive side. Information warfare can be carried out by civil, political, psychological, social, economical, and military means on a strategic, operational, and tactical level.

Information operations in turn are actions taken to affect adversary information and information systems while protecting one’s own information and information systems. An information operation is a civilian and/or military operation of information warfare. In [30],

information operations (INFO OPS) is defined as the measures deployed in order to affect information, information based processes, command and control systems (C2S), and information and communication systems while taking advantage of and protecting one's own information, C2S, and information and communication systems. Information operations take advantage of the interface between technological advances and the most critical aspect in warfare, namely the human being. Whereas information warfare is waged only during a conflict or war, information operations are carried out also during time of peace.

The scope of information warfare is depicted in Figure 1 [36]. As can be seen from the picture, information warfare is divided into an offensive and a defensive mode. Furthermore, information warfare comprises everything from physical and technology based tactics to psychological, human-oriented tactics.

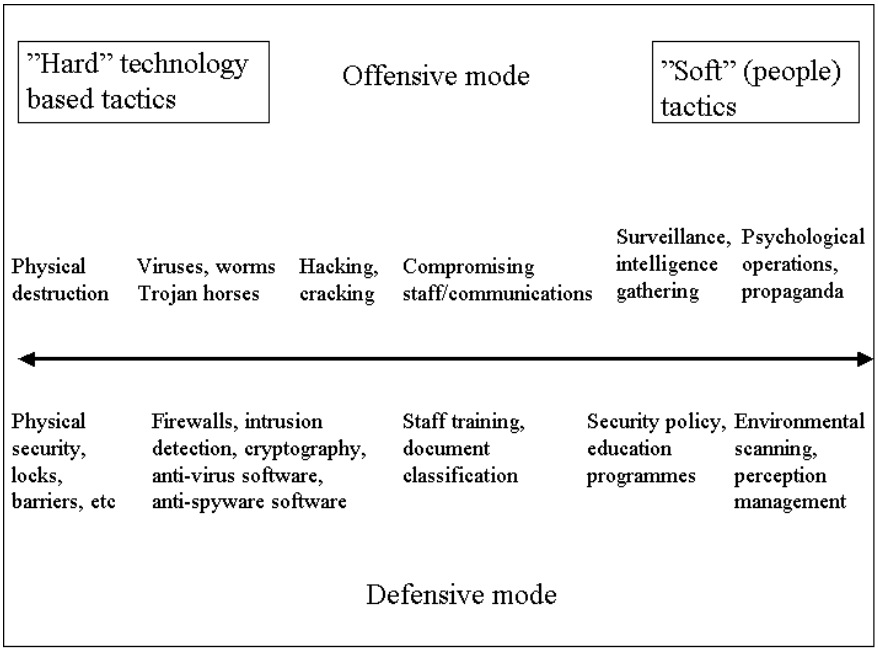

Figure 1 The scope of information warfare

Physical destruction refers to the actual destruction of e.g. equipment as a result of bombing or sabotage. Defense mechanisms include various access control mechanisms, such as locks and keys, magnetic cards, smart cards,

or systems based on biometrics. The armed forces are responsible for protecting the physical territory of a nation.

Attacks on information technology may take several forms. One of the major problems today is viruses and other malicious software (worms, trojan horses, backdoors, etc.), as they are able to spread through the network and infect practically all vulnerable systems in very short periods of time. Malicious hacking and cracking attacks exploit vulnerabilities in systems or protocols, for example, in order to take control of a computer or steal information. Typical security solutions for attacks on information technology include firewalls, intrusion detection and prevention systems, anti-virus software, anti-spyware software, cryptographic protocols, and access control mechanisms.

Internal attacks are always difficult to handle because they bypass traditional security solutions. Employees typically have access to the premises or systems from within and a compromised computer has access to the system according to its credentials. A compromised user may be an industrial spy or a previously well behaved employee who has become dissatisfied with his work and has decided to harm the company for fun or because someone pays him to do so. People also tend to be gullible, making it possible to perform successful social engineering attacks, such as tricking the user into giving his password. Coping with internal attacks on the communication network is difficult, because the behavior of nodes needs to be monitored, and even then it is difficult to spot anomalies. Typically, computers are protected from becoming compromised in the first place by being kept up to date and properly configured, having firewalls and intrusion detection/prevention systems installed, and by security auditing. As far as staff is concerned, security education and document classification can to some extent be used to prevent social engineering from succeeding and information from leaking to the outside.

Surveillance and intelligence gathering does not necessarily involve any criminal activities. People and organizations give out a lot of information about themselves on web pages, in news papers, and in interviews. It is thus possible to collect a lot of intelligence merely by observation. Naturally, also illegal means may be used to gather intelligence. To protect the people and the organization, staff education programs and security policies are needed. Educated users may be better aware of what information to disseminate and under what circumstances. For example, politicians that are often interviewed by the media usually receive some form of media training, not only to be convincing per se, but also to know what not to say. Security policies state clearly who is allowed to do what,

for example, many systems prevent unauthorized users from performing actions they are not allowed to perform. Document classification is one part of security policy; only a certain group of people are allowed to see certain documents.

Psychological operations and propaganda are perhaps the most traditional forms of information warfare. The purpose of psychological operations is to affect the mind and thus decision making of another party while protecting oneself from corresponding operations. Propaganda is one means of carrying out psychological operations. For example, in a democratic society, it is important that the public supports the actions of the government. If the government wants to go to war, propaganda can be used to demonize the other party, so that the public feels that the war is justified, and thus accepts it. Another powerful tool for psychological warfare is disinformation, which on tactical and operational levels is referred to as deception. For example, deception was used in the invasion of Normandy during the Second World War, when the German forces were deployed near Calais rather than Normandy. Education is perhaps the best means of protecting against propaganda, as people understand to question what they see and hear, and are able to take a critical viewpoint rather than believe everything that is fed to them.

The most commonly referred to model of information warfare is the "Libicki model". Libicki [51] defines seven forms of information warfare:

- Command and control warfare (C2W): C2W is the military aspect of information warfare. In [30], C2W is defined as the integrated use of all military capacities, including operations security (OPSEC), feinting, psychological operations (PSYOPS), electronic warfare (EW), and physical destruction. The purpose is to affect, weaken, destroy, or deny information to the C2S of an adversary, and to defend one’s own C2W from corresponding attacks.

- Intelligence based warfare: the integration of sensors, emitters, and processors into a reconnaissance, surveillance, target acquisition, and battlefield damage assessment system.

- Electronic warfare: electronic warfare is surveillance and intelligence gathering from systems deploying or transmitting

electronic radiation, the act of influencing such systems, and protection from the influence of such systems.

- Psychological warfare: psychological warfare is the act of influencing the values, attitudes, opinions, feelings, motives, decision making, and behavior of people. The defensive side of psychological warfare is to maintain and improve the combat will by protecting one's own forces as well as the civilian population from psychological operations launched by an adversary and from self-inflicted actions (such as baseless rumors). The offensive side of psychological warfare is to deteriorate the combat will of the forces of an adversary by spreading information that will affect the mental state of the adversary, causing e.g. fear, hesitation, and stress. It also includes the act of influencing the civilian population of the adversary.

- Hacker warfare: in [52], hacker warfare is defined as information warfare that seeks to damage computer systems by exploiting vulnerabilities in them.

- Economic information warfare: the manipulation of information exchanged in trade (either denial or exploitation) as an instrument of state policy.

- Cyber warfare: also known as net warfare. In [19], cyber warfare is information warfare conducted in cyberspace, where cyberspace refers to computer systems, networks, communication systems, and all their supporting infrastructure and operations. It is the act of influencing information systems, information transmission, or the information itself, and to protect one's own information, information systems, and information transmissions.

Although the Libicki model starts to be outdated, it is still commonly referred to in the literature, since it grasps the main elements of information warfare.

From the point of view of this thesis, the technical aspects of information warfare are of main interest.

2.2 Network-centric warfare

In [1], network-centric warfare (NCW) is defined as an information superiority-enabled concept of operations that generates increased combat power by networking sensors, decision makers, and shooters to achieve shared awareness, increased speed of command, higher tempo of operations, greater lethality, increased survivability, and a degree of self-synchronization.

Traditionally, the forces have suffered from limitations in their ability to communicate, move, and project effects. Thus, the forces and their supporting elements have had to be co-located, or in close proximity, to the enemy or the target they were defending. The result was that a geographically dispersed force was weak and unable to rapidly respond to or mount a concentrated attack. One purpose of network-centric warfare is to eliminate the geo-locational constraints by networking the forces using the most advanced technologies available.

As a result of networking, the forces become more knowledgeable than before. The knowledge is dependent on a continuous stream of timely and accurate information, as well as the processing power, tools, and expertise necessary to put battle space information into context and turn it into battle space knowledge.

Furthermore, the effective linking of forces means that dispersed and distributed entities can generate synergy, and responsibility and work can be dynamically reallocated to adapt to the situation. Effective linking requires the establishment of a robust, high-performance information infrastructure (infostructure) that provides all elements of the forces with access to high-quality information services.

Network-centric warfare recognizes three domains [22]:

Physical domain: the physical domain is the traditional domain of warfare. The physical platforms and the communication networks that connect them reside in this domain. This is also where strikes, protection, and maneuver take place across the ground, sea, air, and space environments. To conduct network-centric operations, all elements of the force must be robustly networked achieving secure and seamless connectivity.

Information domain: the information domain is where information is created, processed, and shared. In this domain, the communication of information between war-fighters is facilitated, the command and control of modern military forces is communicated, and where the intent of the commander is conveyed. In the battle for information superiority, the information domain is ground zero. To conduct network-centric operations, the force should be able to collect, share, access, and protect information. It should have the capability to collaborate and thus improve its information position through processes of correlation, fusion, and analysis.

Cognitive domain: the cognitive domain is in the mind of the war-fighter and the war-fighter's supporting populace. The elements of this domain include leadership, morale, unit cohesion, level of training and experience, situational awareness, and public opinion. The commander's intent, doctrine, tactics, techniques, and procedures also reside in the cognitive domain. To conduct network-centric operations, the force should be able to develop and share high-quality situational awareness, self-synchronize its operations, and develop a shared knowledge of the intent of the commander.

The physical and information domains are of most interest for the purpose of this thesis. As the thesis is concerned with securing military networks to ensure decision making, the networks of the physical domains must be secured in order to protect the information in the information domain, as this information will affect the way decisions are made.

3 The development of communication technology

The development of networking technology in the civilian and the military environment have gone hand in hand. Originally, the political environment during the Cold War started the development of packet switched networking, which evolved into the Internet. As the business value of Internet was realized, the development was taken over more and more by commercial interests. The current trend in military networking is to use COTS products based on open standards, where the IP standard is a de facto choice for the underlying communication protocol. Solutions are no longer developed separately for the civilian world and the military world, but are rather tailored to meet the various needs of its users.

In this Section, the development of communication technology is covered in quite extensive detail. The motivation behind the thorough description is to describe the various features that communication technologies may provide, as they are all taken advantage of for military communications. In order to secure the network, it is vital to understand its heterogeneous nature.

3.1 The Internet

As a response to the USSR launching Sputnik in 1957, the first artificial earth satellite, the US formed the Advanced Research Projects Agency (ARPA) within the Department of Defense (DoD) in 1958 with the objective of establishing US lead in science and technology applicable to the military [97]. In the 1960s, separate groups of researchers worked on networking and packet switched technologies. The first packet switching theory was published by Kleinrock [47] in 1961 and a galactic network concept was envisioned by Licklider in 1962 [53]. In 1965 the first experimental network connecting two computers was established over a 1200 bps phone line between a TX-2 at MIT Lincoln Lab and an AN/FSQ-32 at System Development Corporation. The experiment showed that time-shared computers could work well together, but that the circuit switched telephone system was inadequate for the job. [52] During the same times, three separate but parallel works were done on developing packet switching: one at MIT (1961-167), RAND (1962-1965), and NPL (1964-1967). In 1967, a plan for ARPANET, a packet switched network for time-shared computers, was published, and in 1969, the first packet-switched

network comprising four nodes was established. The nodes were located at UCLA, Stanford Research Institute (SRI), UC Santa Barbara, and University of Utah.

The research advanced rapidly and a host-to-host protocol, Network Control Protocol (NCP), was finished in 1970 [14]. Application development could commence in 1971-1972, with email being the hot application in 1972. ARPANET had grown from the original four nodes to 15 nodes (23 hosts). The growth continued, and ARPANET eventually evolved into the Internet as a result of an open architecture networking concept introduced by Kahn in 1972. The basic idea was to enable multiple independent networks of arbitrary design to interconnect, such as the original ARPANET, packet satellite networks, ground-based packet radio networks, and so on. A new version of the NCP protocol was developed, namely TCP/IP, which was later redesigned into two separate protocols, TCP and IP, and a new protocol, UDP, was introduced. DARPA funded three separate implementations of TCP/IP, which was the beginning of long-term experimentation and development of Internet concepts and technology. Several networks of varying type were incorporated to the Internet environment. The growth resulted in network classes (A, B, and C), host names and DNS, routing infrastructure changes, and so on. By 1985, Internet was well established as a technology supporting a broad community of researchers and developers, and was beginning to be used by other communities for daily computer communications.

The commercialization of the Internet began in the early 1980s as several vendors implemented the technology and service providers offered connectivity and basic services. In the 1990s, the commercialization took new forms; business models were found from utilizing the network for commercial purposes. Shopping malls, cyberbanks, radio stations, and several other services, appeared on the Internet. Also the media started to take notice of the Internet. However, it was not until the introduction of the World Wide Web with its easy to use web browsers that the growth of Internet exploded.

Today, the Internet has become a commodity taken for granted. The society is partially dependent on its communication infrastructures; businesses rely on the Internet for communication and commerce, government systems are based on open networks, military, law enforcement, and rescue organizations utilize the infrastructure, and also the citizens have become accustomed to the availability of a variety of networks. The technology still continues to evolve, where current trends include mobility of network nodes and complete networks, ubiquitous

computing, ad hoc and sensor networking, and so on. Also "Internet 2" is developed, which is envisioned to be the "Internet of tomorrow". It is not a separate physical network, but rather a set of new technologies to replace the ones used in the Internet today.

The biggest change the Internet is about to go through today is switching to use IPv6 instead of IPv4. IPv6 will not only introduce a larger address space, it will also come with increased security features, support for mobility, the ability to do QoS provisioning, and the possibility to take advantage of extension headers (enabling protocol development without breaking the IP stack). Most of these features are included in IPv4, however, the IPv6 protocol has been redesigned and the features are in some cases implemented differently. Furthermore, also development in lower level technologies (such as that of the access medium), have affected the way internetworking is done. For example, wireless networking has developed rapidly and continues to develop, as users wish to remain connected while moving.

In the following Subsections, wireless networking, mobility, and ad hoc networking will be described in further detail. The discussion is motivated by the fact that both military and civilian networks are increasingly wireless and mobile. Furthermore, in the military scenario, ad hoc networks are relied on to ensure communication in difficult environments, such as the battlefield. Both the military and civilian environments rely on the same technologies, however, many of the requirements are different. For example, the military environment places higher demands on fast mobility support and security than the civilian environment. Since military decision making will partially be based on information retrieved from or through the network, it is important for the threat analysis and the required solutions to security problems that the underlying network environment is well understood. Hence, the rest of this Section will be devoted to describing the evolution of wireless and mobile networks.

3.2 Wireless networks

As the need to stay connected while on the move increased, wireless technologies have rapidly evolved. Initially, satellite communications were used to support connectivity where none was available, for example, for maritime travel. The benefit of satellite communications is that it is possible to provide coverage over the whole earth, including the poles. However, satellite communications have not until recently been available

to the normal consumer due to the high usage fees. Although the fees today are comparable to roaming GSM fees, satellite communications do not provide a feasible solution to the growing demand for wireless services, even when IP based communication is supported. Hence, several wireless technologies have evolved, most of which support IP based communication.

The "Everything over IP, IP over everything" paradigm that had been deployed when designing the Internet architecture made it possible to develop applications and wireless technologies independently, see Figure 2. Applications are completely unaware of the underlying network, and the network in turn is not concerned with the content of the traffic. The only common factor they have is the IP protocol, hence the hourglass analogy. However, the wireless medium introduces some problems for the TCP/IP protocols, for example, if there is one bad link on the path from source to destination, the sender may be forced to retransmit all the data over the whole path multiple times. The wireless network resources tend to be scarce, hence any additional retransmissions should be avoided.

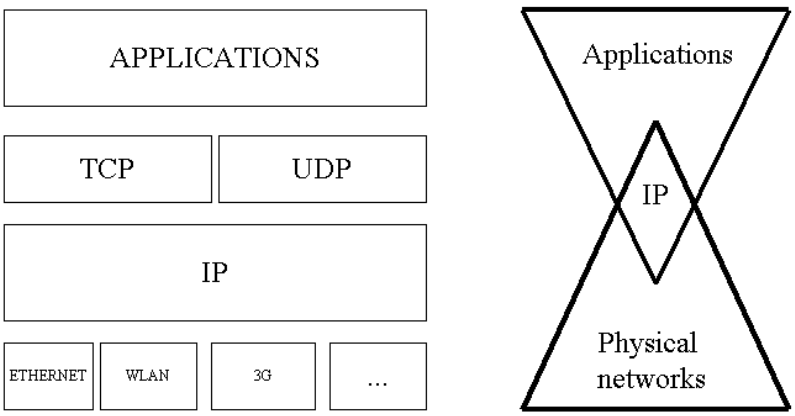

Figure 2 The Internet architecture design

The varying wireless network technologies are depicted in Figure 3 together with the corresponding IEEE standards. Only commercially available technologies are listed; radio amateur communications,

proprietary military communication, and other private communication technologies are omitted. The main reason for omitting these other radio technologies is the fact that even the military is beginning to deploy technologies based on open standards (such as IP) and COTS products, hence, the technologies illustrated in the picture are of most interest.

The wireless technologies have been designed for different purposes. Wireless Wide Area Networks (WWAN) and Wireless Metropolitan Area Networks (WMAN) have been designed to cover large areas and thus allow a large degree of mobility. Wireless Local Area Networks (WLAN) provide limited mobility within a small area, such as an office. They have mainly been designed to extend the wired local area network by allowing the user to connect to the network from anywhere within the coverage area. Wireless Personal Area Networks (WPAN) are basically cable replacement solutions designed to free the user from the hassle of cables when connecting personal devices (e.g. laptops, mobile phones, printers, PDAs, and cameras) to each other.

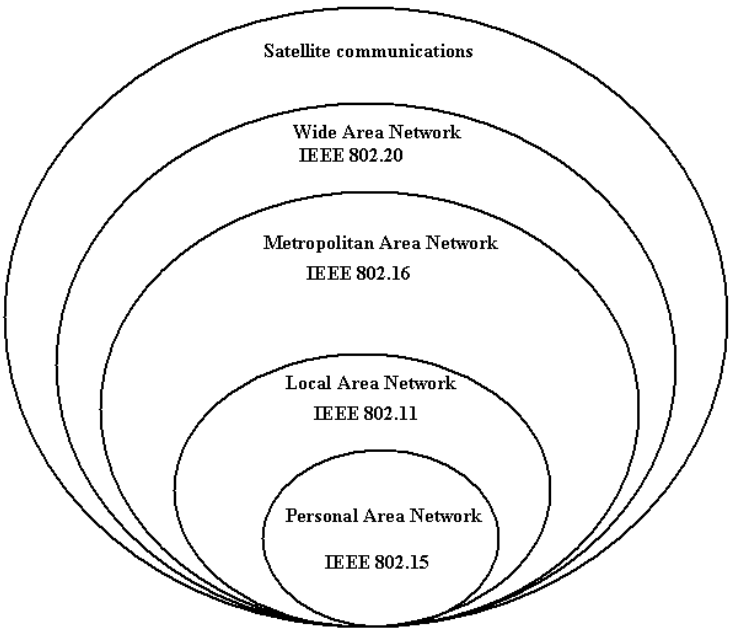

Figure 3 Network technologies

The benefits of wireless networking is increased flexibility with respect to mobility and coverage of remote areas. However, the wireless medium is vulnerable to disturbances due to weather, terrain, hostile activities, and so on. One major problem is the bit error rate, which limits the ability to transport large amounts of data or to keep sessions alive. Thus, providing reliability, security, QoS, etc. is more difficult than in traditional wired networks.

3.2.1 Satellite communication

Satellite communication is carried out through communication satellites with geostationary, Molniya, or low earth orbits. A satellite in a geostationary orbit (GEO) appears to be in a fixed position to earth. It revolves around the earth at a constant speed once per day over the equator. However, such satellites are not suitable for providing connectivity at high latitudes, because they may appear low or below the horizon. The Molniya orbit, mainly used for telecommunications and television over Russia, is inclined to guarantee good elevation over the northern hemisphere. Molniya orbits have very high apogees at the high latitudes where the users desire long duration for communications. A low earth orbit (LEO) is a circular orbit of roughly 150km above the surface of earth and with a coverage of a radius of 1000 km. The tradeoff between geostationary and LEO satellites is that LEOs are less expensive to deploy and the required signal strength is lower, however, due to the small coverage area, the number of LEO satellites needed is significantly higher.

Satellite communication is mainly used where no other form of communication is available, such as remote areas, conflict areas, or war zones. For example, Iridium [38] Satellite LLC provides voice and data solutions with complete coverage of the earth using several cross-linked LEO satellites. Inmarsat [37], on the other hand, provides satellite communications over geostationary satellites that cover the whole earth except the poles. Typical customers include industries such as maritime, aviation, government/military, emergency/humanitarian services, mining, forestry, oil and gas, heavy equipment, transportation and utilities.

Satellite communications have several benefits. Satellites provide global coverage and are a good solution for one way broadcast applications, as one satellite is able to cover a large geographical area. Satellite communications are difficult to disrupt, thus providing a reliable means of communication. Furthermore, mobility is easily supported. The decrease in

cost have made it possible even for individuals to take advantage of satellite communications, which is no longer limited to governments, military, and large corporations only.

However, satellite communications do not provide support for all wireless networking needs that have emerged, especially if two-way high-speed communication with short delays is required.

3.2.2 Wireless Wide Area Networks

A Wide Area Network (WAN) is a computer network that spans large geographical areas and includes a vast number of computers. Perhaps the most well known WAN is the Internet.

Wireless Wide Area Network technologies are standardized by IEEE as the 802.20, also known as Mobile Broadband Wireless Access (MBWA), and by ETSI as the 3GPP standard. Specifications are developed cooperatively between IEEE and ETSI to allow MBWA networks to interface 3G networks. The MBWA standard is also known by the name Mobile-Fi.

The scope of IEEE 802.20 is to develop a specification for the physical and medium access control layers of a wireless interface for interoperable packet-data mobile broadband wireless access systems that [49]:

- operates in licensed frequency bands below 3.5GHz
- supports peak data rates per user in excess of 1 Mbps
- supports vehicular mobility classes up to 250km/h
- covers cell sizes commensurate with ubiquitous metropolitan area networks
- targets spectral efficiency, sustained user data rates and numbers of active users significantly higher than achieved by existing mobile systems.

The purpose of MBWA is to fill the performance gap between the high data-rate low mobility services and the high mobility cellular services. It supports vehicular mobility and allows users to utilize a wide area network through an access network when mobile. Inter-technology roaming and handoffs, e.g. from a MBWA network to a WLAN, are also made possible e.g. using IEEE 802.21 type of Media Independent handoffs.

3.2.3 Wireless Metropolitan Area Networks

A Metropolitan Area Network (MAN) is a large computer network spanning e.g. a campus or a city.

WiMax (Worldwide Interoperability of Microwave Access) is a Wireless MAN technology defined in the IEEE 802.16 standard and promoted by the WiMax forum. WiMax specifies a WMAN protocol that will provide a wireless alternative to cable, DSL, and T1 level services. It supports low latency applications, such as voice and video, provides broadband connectivity, and is able to support thousands of users from a single base station. One base station has a cell radius of 50 km.

Three major phases in the development of WiMax are anticipated [32]. In the fist phase, WiMax will provide hot spot backhauls and fixed location private line services with transmission rates up to 100 Mbps using outdoor antennas. In the second phase, WiMax will offer broadband wireless access, or wireless DSL, offering data rates between 512 Kbps and 1 Mbps. Low-cost, indoor, user installable devices that do not have to be aligned with the base stations will be installed on the user premises and integrated with the radio modem. In the third phase, mobile users traveling at speeds up to 120 km/h can be supported. The expected data rate is 512 Kbps. The use scenarios of all three phases are depicted in Figure 4.

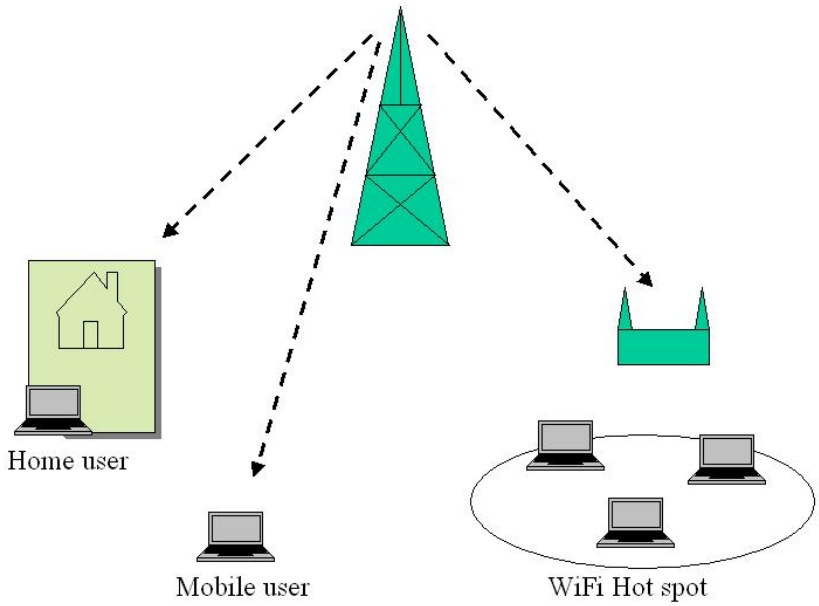

Figure 4 A WiMax cell

WiMax and MBWA are in a sense competing technologies. While MBWA was designed to be a mobile system supporting high mobility speeds, mobility is an add-on to the WiMax technology. However, WiMax is supported by larger vendors than MBWA and it is expected that WiMax products hit the market first. Neither technology aims to compete with Wi-Fi technologies (see Section 3.2.4.), but rather to complement it.

3.2.4 Wireless Local Area Networks

A Local Area Network (LAN) is a computer network covering a local area, such as a company or home network.

Wi-Fi (Wireless Fidelity) is a Wireless LAN (WLAN) technology defined in the IEEE 802.11 standards and promoted by the Wi-Fi forum. Wi-Fi was initially developed to add mobility to private wired LANs using the unlicensed radio spectrum. The range of WLAN is in the order of 100 m, depending on the environment and antennas used. Typically, the experienced transmission rate will drop as the distance to the base station increases. Transmission rates are specified between 1 Mbps and 108 Mbps.

The IEEE 802.11 standard specifies two WLAN modes: infrastructure and ad hoc mode. In the infrastructure mode, computers communicate with a wireless access point, which typically is connected to a wired network. In ad hoc mode, computers may communicate directly in a peer-to-peer fashion.

One of the benefits of Wi-Fi is that it is able to provide high speed Internet access with low requirements for transmission power. Thus, it is a feasible technology even for small, battery driven devices. It also adds flexibility to local area networking, as it supports user mobility. Wi-Fi is the most cost-efficient wireless technology today. The disadvantage is the lack of support for QoS, and the lack of privacy protection for users.

Wi-Fi has also been used for services for which it was not originally designed. Wireless Internet Service Providers (WISP) established hot spots to provide wireless Internet access to users, however, without success. The problem was twofold; first, since the range of WLAN is limited to 100m, users had to be physically present in the hot spot to be able to access the network. Second, the small coverage area limited the options for subscription to the service. Users were not willing to pay a monthly fee and have geographically limited access, nor has paying a fee per hour attracted customer interest. On the contrary, free WLAN access has become a way to attract customers e.g. to a café. To solve the problem, city wide Wi-Fi mesh networks were established, where handoffs were implemented in a proprietary way. This allowed the user to move while remaining connected. However, mesh technology and handoff capabilities have not been within the scope of the Wi-Fi standard (although handoff support is work in progress). Basically, these mesh WLANs attempt to do the job for which WiMax will provide the real solution.

3.2.5 Wireless Personal Area Networks

A Personal Area Network (PAN) is a computer network with a small coverage area, e.g. around a person.

A Wireless Personal Area Network (WPAN) is established using a cable replacement technology, such as infrared connections (IrDA) and Bluetooth. WPANs are standardized by IEEE 802.15. The 802.15 standard specifies WPAN/Bluetooth, coexistence with WLANs, high speed connectivity, and low speed connectivity with long battery life. Typically, WPANs connect portable and mobile computing devices, such as laptops,

PDAs, peripherals, mobile phones, and consumer electronics. The size of the network tends to be small, e.g. a Bluetooth PAN (piconet) enables only 8 active nodes or 255 inactive ones. Ranges are roughly 10m, but in ideal circumstances it is possible to reach 100m. The connection speed offered by Bluetooth is 723.1 kbit/s, or 2.1 Mbps with enhanced data rate.

The main benefits of WPANs are that users no longer have the hassle with cables, and the devices can be battery driven since the amount of data to transfer is small and the range is short, hence resulting in devices that are small and inexpensive.

3.3 Mobility management

Traditionally, nodes and networks were fixed, and routing was static. However, together with the evolvement of wireless technologies, mobility has become a requirement, and thus mobility support had to be added to the Internet protocols.

Currently, the IP address functions both as an identifier of the node and an interface locator, that is, the address specifies both the identity of the node and its current point of attachment to the Internet. When the node changes its point of attachment, the address changes. The objective of the mobility management schemes is to enable reachability between two communicating nodes even when their attachment points to the Internet changes.

The main problems that mobility management schemes must solve are the following:

1. Locating the mobile node, i.e. finding its current point of attachment to the network.
2. Data transmission to the mobile node. The data has to be transferred to the current location of the node.
3. Continuation of the data transmission after the node or network has moved, including addressing the double jump problem, i.e. when both communicating parties move simultaneously.
4. Controlled disconnection. Disconnecting from the network is done in such a way that no extra load is enforced on other nodes.
5. Optimizing performance. Mobility management schemes should take performance optimization into consideration, such as network load, number of signaling messages, etc.

In this thesis, the features mentioned above will be used for comparing the various mobility management schemes.

This type of mobility is called terminal mobility, and refers to a situation where the network terminal moves from one location or domain to another. In the case of ad hoc networking (described in Section 3.4), the terminal may also move within the network boundary.

In addition, also other forms of mobility exist, namely application and identity mobility. Application mobility refers to a situation where the software process can migrate from one host to another, and thus, move. Agent technologies are typical implementations of application mobility. Identity mobility refers to a situation where the identity, defined as a name, number, or cryptographic key, changes its location (i.e., computer).

In this thesis, only the case of terminal mobility will be considered. Terminal mobility can occur on a node level (only the node moves) or on a network level (the network moves, but the leaf nodes are not aware of the movement).

3.3.1 Node mobility

Node mobility can be divided into three different categories: nano mobility, micro mobility, and macro mobility, as depicted in Figure 5.

In the case of nano mobility, the point of attachment (base station, if one exists) does not change, only the route to the base station and the other network nodes change. Nano mobility is typical of ad hoc networks, as will be described in Section 3.4. In Figure 5, nano mobility is depicted by the mobile node changing its route to the base station as it moves in the network.

Micro mobility refers to mobility within a network domain in such a way that the node changes the base station to which it is connected, however, it still uses the same network address. In Figure 5, micro mobility is depicted by the node moving out of range of its previous base station, thus performing a handoff to a new base station. This type of handoff is commonly referred to as a horizontal handoff.

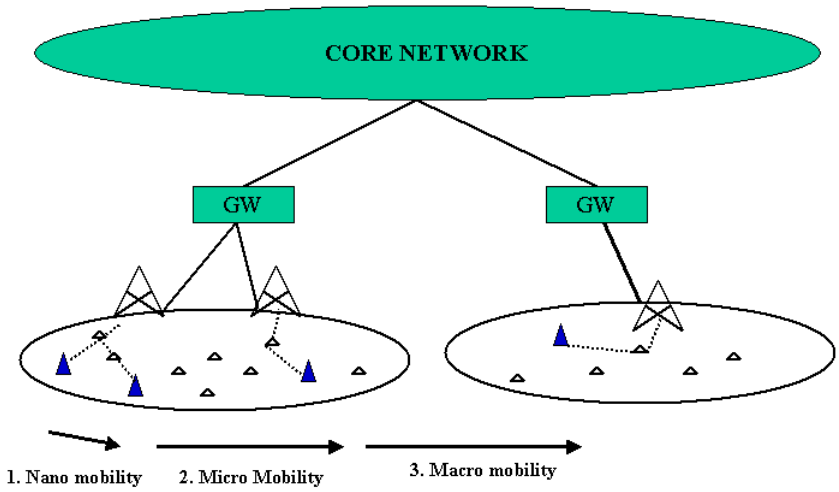

Figure 5 Mobility categories

Macro mobility refers to mobility between different network domains. In Figure 5, macro mobility is depicted by the node moving to a separate access network. The node will now receive a care-of address in its new network, and packets destined for it will be routed to its current location, either directly or via a forwarding agent (e.g. a home agent in Mobile IP [68][41]). This type of handoff is commonly referred to as a vertical handoff. Macro mobility solutions typically seem to fall into the following two categories:

Address translation: An address translation scheme typically deploys an agent system in the mobile node's home network. The agent intercepts the packets that are destined for the mobile node and forwards the packets to the network where the mobile node is currently located. The forwarding function can be performed by various means, e.g., tunneling. Address translation schemes typically function at the network-layer.

Connection forwarding: Connection forwarding is typically handled in the transport-layer. When a mobile node changes its IP address, it somehow informs the correspondent host about its new address. The new address can be sent e.g. using TCP-options.

3.3.1.1 Network layer mobility

The advantages of providing mobility at the network layer are extensively discussed in [4]. The main advantage is that mobility is transparent to protocols and applications running on stationary nodes, and a mobile node should appear like any stationary node connected to the network. Mobility at the network layer would not alter the functionality in the existing network infrastructure. Furthermore, many existing applications and protocols above the transport layer assume that the IP address of a node never changes during operations. Thus, to support mobility, all these applications and protocols would need to be modified unless transparency can be provided. In the case of mobility at the network layer, only one protocol (namely the IP protocol), needs modification, while transport and application layer protocols can remain unchanged.

Mobile IPv4

Mobile IPv4 [68] provides mobility support for nodes using IPv4 [70].

Three entities are defined in Mobile IPv4: the Mobile Node (MN), the Home Agent (HA), and the Foreign Agent (FA).

A Mobile Node is a host that changes its point of attachment from one network to another. However, regardless of its current location, it is always identified by its permanent home address. When the MN is away from home, it obtains a care-of address in its visited network. The care-of address can either be acquired from a foreign agent through agent advertisement messages, or by some external means which the MN associates with one of its own network interfaces. In the former case, the care-of address is the IP address of the FA. The FA relays the registration to the home agent, which then tunnels the packets destined for the MN to the FA. The FA detunnels the packets and forwards them to the MN. In the latter case, the care-of address may be obtained, for example, through DHCP [DHCP]. In this case, the MN functions as its own FA, and decapsulates the packets itself. [69].

Mobile originated connections: When the MN is communicating with a corresponding node (CN), it first sends the packets to the FA, which either

reverse-tunnels them to the HA [58], or forwards them directly to the CN. Packets from the CN are routed via the HA to the MN via the FA.

Mobile terminated connections: The CN sends packets to the MN using its permanent IP address. The HA tunnels the packets to the FA, which forwards them to the MN.

In Table 1, the solutions to the main mobility management problems are presented.

Locating the mobile node	**Packets sent to its fixed IP address**
Data transmission to the mobile node	**The HA tunnels the packets to the FA, which forwards them to the MN.**
Continuation of the data transmission after the node has moved	**The MN sends a location update to the HA. Transparent to the CN.**
Controlled disconnection	**The MN deregisters from the FA and updates its location to the HA.** **Keep-alive messages.**
Optimizing performance	**Route optimization protocol to avoid triangle routing**

Table 1 Summary of mobility solutions in Mobile IPv4

Mobile IPv6

Mobile IPv6 [41] provides routing support to permit a node to continue using its permanent IPv6 address [20] regardless of its location. Mobile IPv6 relies on the concept developed in Mobile IPv4, however, with some changes.

In Figure 6, the Mobile IPv6 architecture is depicted on a conceptual level. When a MN enters a visited network (1), it acquires an IP address e.g. by stateless address autoconfiguration or Neighbor Discovery, specified in IPv6. Thus, no foreign agents are needed. The MN then registers its care-of address with its home agent (2) by sending a Binding Update (BU). The HA responds with a Binding Acknowledgment. For simplicity, the actual

signaling is not depicted. When a CN wishes to communicate with the MN, it sends the packets to MN's home network (3). The HA then tunnels the packets to the MN using MN's care-of address (4). When the MN receives the packets, it may send a Binding Update to the CN to provide route optimization (5).

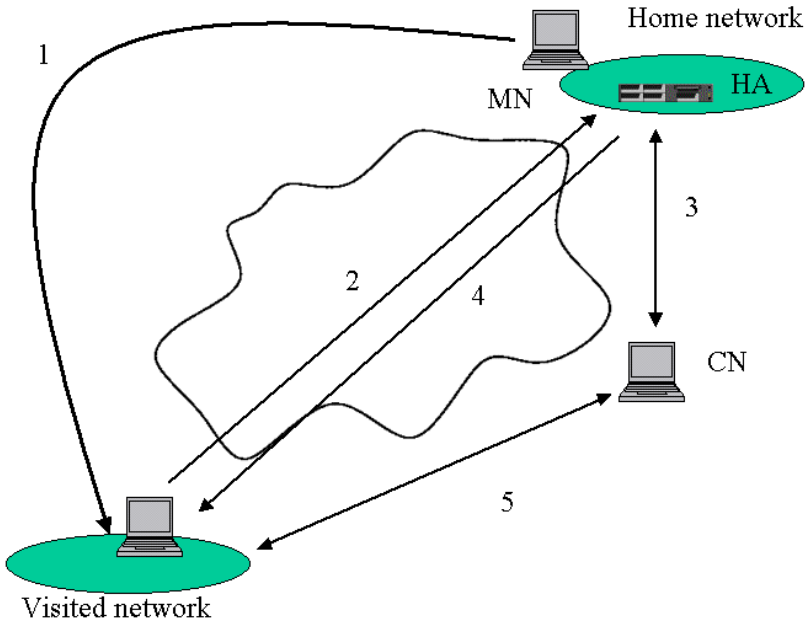

Figure 6 A conceptual model of Mobile IPv6

Route optimization comes with a disadvantage for the access network. As the access network typically is wireless, the significant amount of signaling required per CN becomes a problem. [8]

Table 2 summarizes the mobility management solutions in Mobile IPv6.

Locating the mobile node	**Packets sent to its fixed IP address. The HA tunnels the packets to the MN, which informs the CN of its current location.**
Data transmission to the mobile node	**First tunneled to HA, then directly to MN**
Continuation of the data transmission after the node has moved	**The MN sends a Binding Update to the HA and the CNs.**
Controlled disconnection	**The MN sends a Binding Update to the HA and the CNs.**
Optimizing performance	**Route optimization part of the protocol to avoid triangle routing.**

Table 2 Summary of mobility management solutions in Mobile IPv6

LINA/LIN6

In [39], the Location Independent Network Architecture (LINA) is presented as an alternative approach to Internet mobility. As previously mentioned, IP addresses function both as an identifier of the node and an interface locator. The main concept of LINA is separation of the node identifier and the interface locator. The node identifier is assigned to the network interface of the node and does not change even if the point of attachment changes. The locator uniquely defines the point of attachment. It is assigned to the network interface of the node and is used for routing purposes.

When an application wishes to communicate with a particular node, it may specify the node using either the identifier or the locator. The transport layer is responsible for maintaining the connection using either one. In the former case, the node identifier must be mapped to its current interface locator. This is done by dividing the network layer into two sublayers: the identification sublayer and the delivery sublayer. The identification sublayer handles the conversion of the identifier to the locator by querying a Mapping Agent (MA), and the delivery sublayer delivers the packet. Mobility support comes from the fact that the node identifier remains unchanged regardless of node movement. In the latter case, where the application wishes to communicate using the locator, the identification sublayer is bypassed, and the delivery layer takes care of routing the

packet. However, if either of the communicating nodes move, the transport connection will break. Thus, mobility is not supported in this case.

LIN6 [50] is the protocol based on LINA that provides mobility for IPv6. A 64-bit LIN6 ID is used as the node identifier, which in addition to the traditional IPv6 address is unique to each node. However, as IPv6 applications cannot easily use the 64-bit LIN6 ID as node identifiers, a 64-bit LIN6 prefix is introduced. The prefix is concatenated to the LIN6 ID, resulting in a 128-bit LIN6 generalized ID, which has the same format as an IPv6 address. The applications communicate with this LIN6 generalized ID.

The LIN6 address is used to specify the location of the nodes. A LIN6 address is constructed from a 64-bit network prefix and the 64-bit LIN6 ID. The LIN6 generalized ID has to be mapped with this current locator. The Mapping Agent maintains the mapping information. When the node changes its location in the network, it registers its mapping information to the specified MA. Also the corresponding nodes need to update the mapping. LIN6 provides two ways of updating the mapping [40]: the mobile node updates the mapping of the corresponding nodes or the mobile node requests the corresponding nodes to request a new mapping from the MA.

When a node wishes to contact the destination node known by the LIN6 generalized ID, it requests the MA of the destination node to return the current location of the node. A LIN6 (network level) address is returned.

On the application layer, the LIN6 generalized ID is used. This ID remains unchanged when the node moves in the network. Thus, mobility on the application layer is guaranteed.

Table 3 summarizes the mobility management solutions in LIN6/LINA.

Locating the mobile node	**Mapping Agent**
Data transmission to the mobile node	**Directly**
Continuation of the data transmission after the node has moved	**1. The MN updates its mapping with the MA and its peers** **2. The MN updates its mapping with the MA and requests the peers to query the MA.**
Controlled disconnection	**Updating the MA and the peers**
Optimizing performance	**No overhead in the protocol header. No triangular routing.**

Table 3 Summary of mobility management solutions for LIN6/LINA

3.3.1.2 Transport layer mobility

The favors of providing mobility at the transport layer are discussed in [82]. The significant advantage of handling mobility at the transport layer instead of at the network layer stems from the fact that network-layer mobility comes at significant cost, complexity, and performance degradation due to triangle routing. The motivation for providing end-to-end mobility is to support the different mobility modes of applications, that is, applications that originate the connections, and applications to which other hosts originate connections to, and to empower the applications to make choices best suited to their needs.

SCTP

The Stream Control Transport Protocol (SCTP) [85] is an IETF standard providing enhanced TCP like transport service. Even though it is primarily designed to provide transport to signaling protocols, e.g., SS7, it may eventually replace TCP [71] and even UDP [72] in some application areas. As a primary feature, the SCTP protocol allows the connection endpoints to be defined as groups of IP addresses. The basic motivation behind this has been the need to provide robustness and load balancing capabilities to large multi-homed hosts. To further enhance robustness, a method for dynamically adding and removing addresses from the SCTP endpoint address sets. [84] The main intention was not to provide mobility support,

but rather to allow new network interfaces to be added and old to be removed from multi-homed hosts.

In [83], an extension called mobile SCTP (mSCTP) is described. The extension can be used to provide seamless handover for mobile nodes [74][16]. The handover management is provided at the transport layer, and unlike network-layer solutions, do not need to rely on the support of network routers and proxies for tunneling. In case of mobile-terminated connections, mSCTP must be used along an additional location management scheme, such as Mobile IP [41], the Session Initiation Protocol (SIP) [33], or Dynamic DNS (DDNS) [93].

Table 4 summarizes the mobility management solutions in mSCTP.

Locating the mobile node	**Dynamic DNS/Mobile IP/SIP**
Data transmission to the mobile node	**To the specified primary IP address**
Continuation of the data transmission after the node has moved	**Update the IP address set used by the peers (and the primary IP address)**
Controlled disconnection	**Update the IP address set used by the peers (and the primary IP address)**
Optimizing performance	**-**

Table 4 Summary of mobility management solutions in mSCTP

Multihomed TCP/Extended Transport Control Protocol

In [35], the Multihomed TCP, also known as the Extended Transport Control Protocol (ETCP) [95], is described. The ETCP protocol extends the traditional TCP by enabling the nodes to use a set of IP addresses instead of just one IP address. Addresses can be dynamically updated during the communication. Thus, ETCP supports both multihoming and mobility.

The Multihomed TCP draft has expired.

Migrate TCP

In [82], a transport-layer mobility scheme is presented. There are three major components in this scheme: addressing, mobile node location, and connection migration. Addresses are obtained using e.g. the Dynamic Host Configuration Protocol (DHCP) [25], or an autoconfiguration protocol [87].

Once the mobile node has obtained an address, it may start communicating with corresponding nodes. If the mobile node is a client that actively opens a connection to the corresponding node, no location task needs to be performed. However, if the movement occurs while a connection is open, the mobile node will first obtain a new IP address from its new network, and the connection continues via a secure negotiation with the communicating nodes using a Migrate TCP option. To support mobile servers and other applications where nodes residing on the Internet originate the communication, DNS is used for locating mobile nodes. When the mobile node changes its point of attachment, it changes its hostname-to-address mapping in the DNS by using a secure DNS update protocol [26][92]. Nodes wishing to communicate with the mobile node make hostname lookups and retrieve the IP address of the mobile node's current location.

Connection migration is handled by deploying a Migrate TCP option, included in SYN segments, that identifies a SYN packet as part of a previously established connection rather than a request for a new one. The mobile node restarts previously established TCP connections from its new address by sending the Migrate SYN packet that identifies the previous connection; the corresponding node then resynchronizes the connection with the mobile node. To secure the migration from hijacking attacks, the end nodes may either rely on solutions such as IPSec [44], or use an unguessable connection token which is negotiated using a secret connection key.

The scheme does not present a solution to the double jump problem.

Table 5 summarizes the mobility management solutions in Migrate TCP.

Locating the mobile node	**DNS**
Data transmission to the mobile node	**Directly**
Continuation of the data transmission after the node has moved	**Migrate TCP + resynchronization of the session**
Controlled disconnection	**DNS update**
Optimizing performance	**-**

Table 5 The mobility management solutions in Migrate TCP

3.3.1.3 Application layer mobility

The main motivation for providing mobility at the application layer is that no changes to the operating system are required. Thus, it is believed that mobility can be more easily deployed widely.

The Session Initiation Protocol (SIP) [33] has often been suggested as a protocol for providing mobility at the application layer. SIP supports personal mobility, and adding node mobility is easy according to [79]. The paper presents one approach of using SIP for node mobility.

The basic purpose of SIP is to allow two or more nodes to establish a session consisting of multiple media streams. There are three entities in SIP: user agents, proxy servers, and redirect servers. A user agent is denoted by an email-like address, e.g. user@domain. The purpose of the user agent is to listen to incoming connections and to send SIP messages based on user actions or incoming messages. It also starts the right application depending on the established session. The proxy server relays messages so that it is possible to use a domain name to locate a user without having to know the IP address or the name of the user's terminal. This feature also provides location privacy to the user. The redirect server returns the location of the host rather than relaying SIP messages. This provides scalability, as the redirect server only needs to send the location of the terminal as a response rather than participating in the whole transaction. Identity mobility is provided by the fact that a user can register his current location, regardless of which terminal he used, with the proxy or redirect servers, which will then redirect the traffic to the new destination.

To add node mobility, roaming frequency and the ability to change IP address during a traffic flow must be supported. Roaming support is provided in a similar way as in e.g. Mobile IP. Each mobile node has a home network, although, unlike in Mobile IP, the mobile node does not need to have a static IP address. When the mobile node moves, it registers its new location to a SIP server in its home network. To allow sessions to continue when moving, the mobile node sends a location update to its corresponding nodes informing them of the current IP address used. This resembles the Binding Update function of Mobile IPv6.

Table 6 summarizes the mobility management solutions in SIP.

Locating the mobile node	**User domain or IP address**
Data transmission to the mobile node	**Directly or via proxy agent**
Continuation of the data transmission after the node has moved	**Location update to SIP server and peers**
Controlled disconnection	**Location update to SIP server and peers**
Optimizing performance	**-**

Table 6 Mobility management solutions in SIP

3.3.1.4 Hybrid approaches

Hybrid mobility schemes typically provide mobility between the network and transport layer, thus trying to combine the advantages of both approaches.

Homeless Mobile IP

Homeless Mobile IP is a connection forwarding scheme working at the network layer, or at the border between the network layer and the transport layer [65]. A new data structure is added between the IP layer and the upper layers. The new layer represents each node with a dynamically changing set of IP addresses. Thus, in principle, the solution is pretty similar to many of the transport layer connection forwarding schemes. However, by placing the functionality in the network layer, it makes the mechanisms available to all upper layer protocols, and it also allows the

addresses to be shared between multiple transport protocols and multiple connections, thereby saving memory and potentially reducing the need for signaling traffic.

In the Homeless Mobile IP architecture, the fixed IP addresses are more or less abandoned, except perhaps for some root DNS servers or heavily trafficed web servers. All nodes are assumed to have dynamic IP addresses, that is, the IP layer makes no distinction between a dynamically assigned IP address or a fixed one. The concept of a Home Agent is abandoned altogether.

Connections may be mobile originated or mobile terminated:

Mobile originated connections: When a host wishes to establish a connection e.g. to a server, it requests the IP address(es) of the server from DNS and creates a foreign Host Cache entry based on the response. When the node initiates communication, it must select one of the addresses to be used as a destination address. The choice is made based on some specified policy. For example, the policy may state that the destination address to be used shall be the last address that the corresponding node used as a source address. Another possible policy would be to use the fattest pipe, that is, the destination address corresponding to the link with the highest speed. The selection may also be made on packet basis. This may be beneficial; for example, this can be used for load balancing between links, or selecting the link based on some Quality-of-Service policy or other properties. Once the destination address is selected, the source address may be selected among the ones available at the local Host Cache Entry, for example, using the source address selection algorithm in [24].

If the node roams during the connection, it must inform the corresponding host about the change of point of attachment. This is done by updating its Host Cache entry at the corresponding host.

Mobile terminated connections: Mobile hosts functioning as servers need a way to announce their current location to the outside world. One possible way would be to make use of dynamic DNS [93]. When the host changes its point of attachment, it updates its name-to-address entries in the DNS in a similar fashion as in the transport-layer solution presented by [82]. A host originating a connection will thus find the host by querying the DNS, which returns one or more addresses through which the host can be currently reached. The connection originating host then creates a Host Cache entry corresponding to the DNS replies and uses one of the addresses according to its policy.

The major drawback of using Dynamic DNS is, however, that it does not support fast handoffs where the host changes its address. However, it is assumed that a host does not change its point of attachment to the Internet more frequently than a few times a minute, in which case the scheme works fine. This assumption is made by several other mobility schemes as well and seems to be feasible in most cases. However, Homeless Mobile IP is by no means restricted to Dynamic DNS; other mechanisms, such as SIP [33], may be deployed instead.

The double jump problem is not properly addressed in Homeless Mobile IP. However, the standard Mobile IPv6 [41] specifies a mechanism where a mobile host arranges a home agent at the old location to temporarily forward packets to its new location. Homeless Mobile IP could use a mechanism similar to the one in standard Mobile IPv6, however, instead of a home agent, a forwarding agent would be used. The purpose of these agents would be to temporarily forward packets. When the node moves to another network, it will register its new address to the forwarding agent, which for a short period of time (e.g. 15 seconds) will forward the packets to the node. The forwarding request is authenticated in order to prevent illegitimate requests.

Table 7 summarizes the mobility management solutions in Homeless Mobile IP.

Locating the mobile node	**Dynamic DNS (or other repository)**
Data transmission to the mobile node	**Directly**
Continuation of the data transmission after the node has moved	**Update Host Cache Entries of CNs** **DNS update**
Controlled disconnection	**DNS and CN updates, temporary forwarding agents**
Optimizing performance	**-**

Table 7 Summary of the mobility management solutions in Homeless Mobile IP

The Host Identity Payload/Protocol (HIP)

There are two name spaces in use in the Internet today: IP addresses and domain names. The three main problems with the current name spaces are that dynamic readdressing cannot be directly managed, anonymity is not provided in a consistent and trustable manner, and authentication for systems and datagrams is not provided. The Host Identity Payload/Protocol (HIP) architecture [59][60][61] introduces a new cryptographic name space, the Host Identity (HI), and adds a Host Layer between the network and the transport layer in the IP stack.

The modification to the IP stack is depicted in Figure 7. In the current architecture, each process is identified by a process ID (PID). The process may establish transport layer connections to other hosts (or to the host itself), and the transport layer connection is then identified using the source and destination IP addresses as well as the source and destination ports. On the IP layer, the IP address is used as the endpoint identifier, and on the MAC layer, the hardware address is used. In HIP, the transport layer is modified so that the connections are identified using the source and destination HIs as well as the source and destination ports. HIP then provides a binding between the HIs and the IP addresses, e.g. using DNS [57].

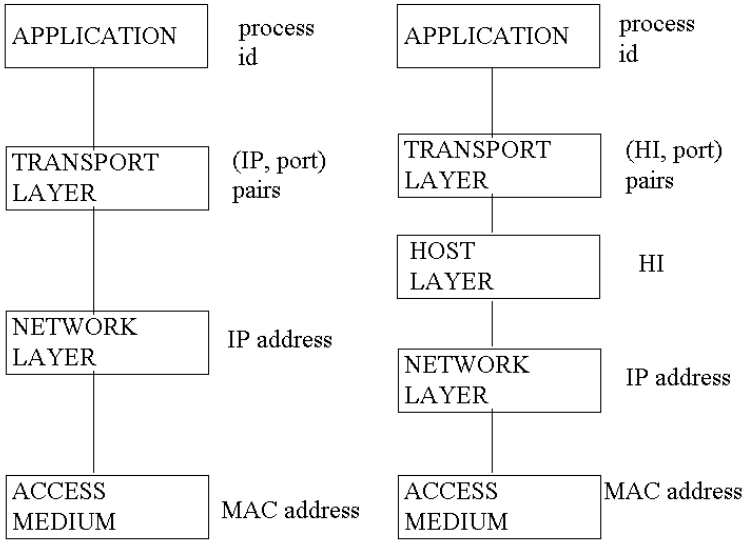

Figure 7 The traditional stack compared to the HIP stack

The HI is typically a cryptographic public key, which serves as the endpoint identifier of the node. Each host will have at least one HI assigned to its networking kernel or stack. The HI can be either public or anonymous. Public HIs may be stored in directories, such as DNS, in order to allow the host to be contacted by other hosts.

The HI is never directly used in any Internet protocol. It is stored in a repository, and is passed in HIP. Protocols use a 128-bit Host Identity Tag (HIT), which is a hash of the HI. Another representation of the HI is the Local Scope Identity (LSI), which has a size of 32 bits, but is local to the host. Its main purpose is to support backwards compatibility with the IPv4 API.

The main advantages of using HIT in protocols instead of the HI is that its fixed length makes protocol coding easier and also does not add as much overhead to the data packets as a public key would. It also presents a consistent format to the protocol regardless of the underlying identity technology used. HIT functions much like the SPI does in IPsec, but instead of being an arbitrary 32-bit value that identifies the Security Association for a datagram (together with the destination IP address and security protocol), HIT identifies the public key that can validate the packet authentication.

The HIP architecture basically solves the problems of dynamic readdressing, anonymity, and authentication. As the IP address no longer functions as an endpoint identifier, the problem of mobility becomes trivial, as the node may easily change its HI and IP address bindings as it moves. Furthermore, as the name space is cryptographically based, it becomes possible to perform authentication based on the HIs. HIP mobility is described in [54]. In [66], the concept of integrating security, mobility, and multi-homing based on HIP is discussed further.

Table 8 summarizes the mobility management solutions in HIP.

Locating the mobile node	**DNS or other repository**
Data transmission to the mobile node	**Directly**
Continuation of the data transmission after the node has moved	**Update HI&IP binding at DNS; send update to peers**
Controlled disconnection	**Update HI&IP bindings**
Optimizing performance	**Temporary forwarding agents**

Table 8 Summary of the mobility management solutions in HIP

3.3.1.5 Summary of mobility management schemes

Several proposals for mobility management exist. Network layer schemes attempt to make mobility transparent to protocols and applications so that no modifications are needed on layer above IP. Transport layer schemes motivate their existence with the claim that network layer mobility comes with significant cost, complexity, and performance degradation due to triangle routing. Application layer schemes provide simplicity for operating system design, as no changes to the operating systems are needed. Also hybrid schemes exist, where features from both network and transport layer schemes are adopted.

All schemes come with good performance when the network is static and node mobility is infrequent. However, in very dynamic networks all schemes perform badly, as signaling lags behind. This has a direct affect on reachability. Almost all schemes use some sort of anchoring point to ensure reachability, such as a Home Agent, Mapping Agent, DDNS, or (temporary) forwarding agents. One of the major problems is the double jump problem, which none of the schemes have been able to solve in a reasonable way. As fast mobility is poorly supported already when one node is moving, the problem gets worse when both nodes are moving. In Figure 8, the double jump problem is illustrated when both nodes move rapidly and use forwarding agents. For simplicity, the complete signaling is not depicted, only the first location update message. Node A and B both move and change their point of attachment. We assume that they have managed to make a location update to their first access network successfully, and only after that begin to move rapidly. When A is in access network 2, it sends an update to the first access network of B. However, B has already moved ahead, and has also sent an update to the first network of A. The forwarding agents in both networks forward the

updates to the second access network. However, A and B have already moved ahead, so the forwarding agents in the second network forward the packets to the third network. The first update message from A reaches B in its third access network, so node B starts to send packets to the second access network of A, where the forwarding agent forwards the packets to the third network. At this point, node A has already moved ahead. This goes on until at least one node stops moving or the packets no longer have time to reach the nodes, at which point the connection breaks.

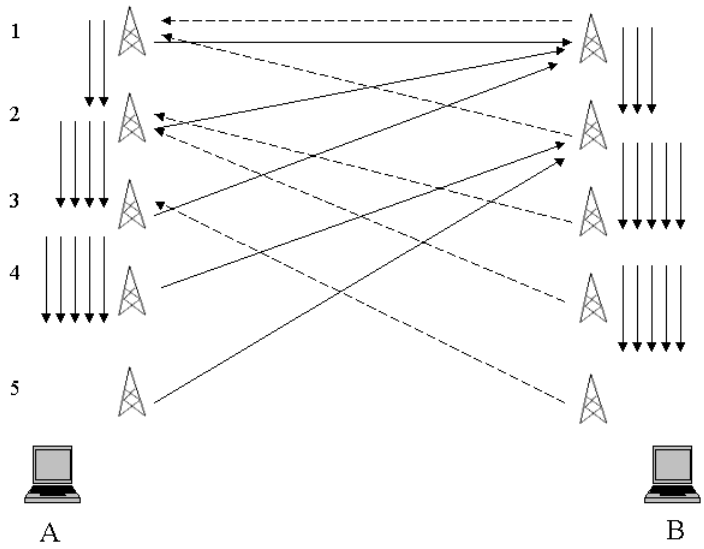

Figure 8 The problem with double jump signaling

If anchoring points such as Home Agents were used in the scenario above, it would be much easier to ensure continuous communication.

All schemes are also vulnerable to attacks. The obvious ways of attacking mobility management is to attack the anchoring points or the signaling traffic. As there are significantly fewer anchoring points than mobile nodes, targeting them would break mobility management and disable the nodes' roaming capabilities. It is also easier to eliminate the anchoring points than all the mobile nodes, especially since the anchoring points are

static. As for signaling, it is not sufficient to encrypt the packets to protect mobility management messages, since it is possible to deduce from the packet length that the message is, for example, a binding update.

3.3.2 Network mobility

Network mobility is concerned with managing the mobility of an entire network that is changing its point of attachment to the fixed infrastructure and thus its reachability in the network topology. To support network mobility, at least one mobile router(MR) is required to connect the network to the infrastructure. It is also possible to provide multihoming by allowing the MR to have multiple attachments to the infrastructure, or by introducing a separate MR for the various points of attachment. Each MR has a Home Agent (HA), and bidirectional tunneling between the MR and the HA is used to preserve session continuity when the network moves. While moving, the MR acquires a care-of address from its new attachment point in the same way as mobile nodes using Mobile IP do when they roam. Thus, mobile networks can be nested, as each MR will appear to its attachment point as a single node. The mobile network nodes (MNNs) are also unaware of the mobility of the network. MNNs may be fixed or mobile.

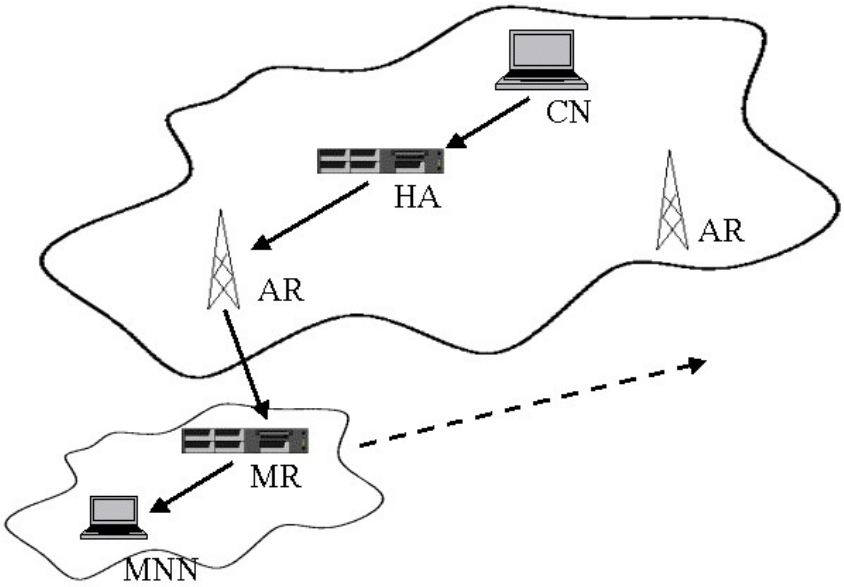

Figure 9 Network mobility

In Figure 9, a network mobility scenario is depicted. The CN and MNN communicate while the network of MNN moves. The movement is transparent to both CN and MNN.

Possible applications for network mobility include public transportation, such as trains and buses, where each train or bus provides a MR for customers to connect to the Internet, cars with low-power sensors seamlessly connected to the Internet, in-flight networking, and various military scenarios.

3.4 Ad hoc networks

An ad hoc network is a collection of nodes that do not need to rely on a predefined infrastructure to establish and maintain communications [9]. Most or all nodes participate in network operations, such as routing and network management, depending on their capacity with respect to CPU, memory, and energy level. Nodes are typically able to move within the network, and they may enter and leave the network on frequent basis. Thus, ad hoc networks are most likely wireless, although some nodes may have connections to a wired infrastructure as well.

The predecessors of mobile ad hoc networking technologies were the DARPA Packet Radio Networks and the Survivable Adaptive Networks (SURAN) programs in the 1970s and 1980s, which were sponsored by the U.S. government and primarily aimed at military purposes. The problem with traditional military networking was the time needed to establish the network; when the network was operational, the forces had already moved ahead. To solve the problem, ad hoc networks were designed to be established on the fly (zero configuration) where and when needed. In [75], future tactical information systems based on commercial, non-developmental standards, and equipment is discussed in further detail.

Ad hoc networks are faced with several challenges, especially in the military case, where the environment is hostile due to the existence of at least one enemy, and harsh due to weather conditions, terrain, EMI, and so on. Military battlefield networks tend to be heterogeneous, including everything from tiny sensors to powerful computers. Communication is both wired and wireless depending on the situation, and some nodes are able to support several access technologies simultaneously. Nodes may also be mobile with various speeds: a node on a human soldier is able to

move at roughly 6 km/h, a node in a tank at roughly 30 km/h, and a node in a jet fighter at over 1000-2000 km/h. Some nodes, such as sensors, are likely not to move at all. Hence, several mobility management schemes and access technologies are needed to support all of the mobility requirements. Also issues such as QoS, security, reliability, and robustness must be addressed for the network to be useful.

The problem with ad hoc networking today is that most of the research has been focused on the development of routing protocols, each only marginally better than the others depending on the simulation environment. Hence, despite almost 50 years of research, the field is immature and no real implemented solutions exist. Many relevant topics have been ignored, such as security, which has been thought of as an add-on feature, thus resulting in each ad hoc network routing protocol implementing its own security solution. Closely related to security is privacy; network traffic reveals the location of the nodes and thus the structure of the network as well as other information about the forces. QoS issues have not been properly addressed either, and it is often neglected that security is a requirement for QoS provisioning. As long as important areas of ad hoc networking continue to be neglected, ad hoc networks will remain only a theoretical solution for battlefield networks.

4 Decision making and the OODA loop

The OODA loop was developed by Colonel John Boyd (ret.) from the US Air Force as a continuation of his work on a fast-transient brief, which was an application of a learning theory he had developed in "Destruction and Creation" report [5] to an operational issue. The fast transient theory suggests that in order to win a battle, one must operate in a faster tempo than the enemy. Boyd uses as examples Germany's Blitzkrieg against the Maginot mentality of France in 1940, the USAF fighter craft F-86 vs. the MiG-15 in the Korean war, and the Israeli raid in 1976. The crucial factor in the victory of these battles was the ability to transition rapidly from one maneuver to another. The basic idea was to operate at a quicker tempo, not just moving faster, than the enemy. By generating a rapidly changing environment, i.e. acting so quickly that it is confusing and disorienting and appears ambiguous and uncertain to the enemy, it is possible to create confusion that causes the enemy to overreact or under-react. The conclusion is that whoever can handle the quickest rate of change is the survivor. The fast-transient briefing held in 1976 was the beginning of his time-based theory of conflict, which introduced the concept of the OODA loop [6].

It is a common misconception that the OODA loop was primarily developed for the U.S. Air Force and that it merely describes the process through which air force pilots determine their split second moves to ensure that they are always on target. Although Colonel Boyd originally was a fighter pilot, he worked on his time-based theory of conflict only after retirement, having then grown interested in ground warfare. The "Patterns of Conflict” [6] is a result of several years of study of Sun Tzu, Clausewitz, Musashi and the history of warfare, beginning from the earliest Greek and Persian battles and continuing with Alexander the Great around 300 B.C., Hannibal around 200 B.C., Belisarius around 500 A.D., Genghis Khan around 1200 A.D., Tamerlane around 1400 A.D., Napoleon, and the First and Second World Wars. Most of his attention was spent on general theories of war, the Blitzkrieg concept, guerrilla warfare, and the use of deception by great commanders. Many of the ideas presented in "Patterns of Conflict" are, in fact, based on ideas learnt from history, and especially the thinking of Sun Tzu [86] and Musashi [62]. For example, tactics based on deception, speed, fluidity of action, and exploitation of the enemy's weaknesses in order to confuse and disorient the enemy is emphasized, and it is noted that inferior forces have throughout history

been able to win in battle if they have been able to avoid war of attrition. [17]

As a product of "Patterns of Conflict", the OODA loop describes a model for human decision making. According to Boyd, decision-making follows a rational series of steps. Initially, the environment is scanned for data (observation). Based on this intelligence, a mental image of the surroundings within which the decision is to be made is created (orientation). The decision is then made (decision) and carried out (act). The goal of conflict is to have a faster OODA loop than the adversary. In practice, one should attempt to reduce the time to go through one's own OODA loop while trying to increase the time it takes for the enemy to go through his. However, the speed is not the most important element of the cycle. The important thing is to execute the cycle in such a way as to get inside the mind and the decision cycle of the adversary. The ideas of creating confusion and disorientation are used here to prevent the enemy from functioning.

The basic OODA loop is depicted in Figure 10. OODA is an abbreviation of Observe, Orient, Decide, and Act. In the observation phase, a person collects information from the environment, which he processes in the orientation phase. Based on the result of that process, he comes up with some alternatives for actions, makes the decision, and acts upon it.

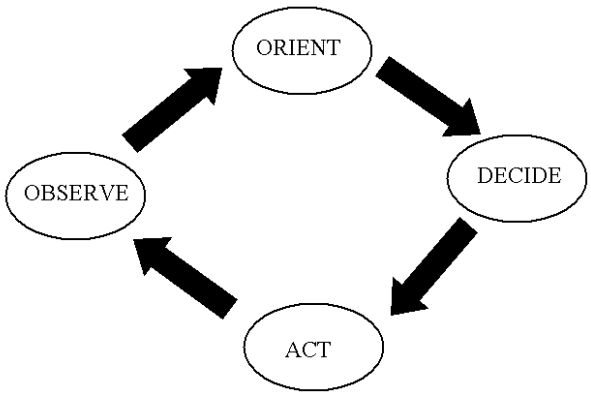

Figure 10 The OODA loop

Although the OODA loop looks like a simple and linear process, it is not. A more detailed version of the OODA loop is depicted in Figure 11. As can be seen from the figure, the OODA loop is not a one-dimensional loop. Over thirty arrows are drawn to connect the various elements, thus yielding the possibilities of several different loops to be derived. The "implicit guidance and control" arrows from the orientation phase are directed both back to the observation phase and directly to the action phase. This describes a situation where one has developed a proper understanding of the changing environment and is then able to bypass the orientation and decision phases by observing and acting almost simultaneously. In this way, it is possible to compress time, i.e. the time it takes from making an observation to acting. This temporal discrepancy can be used to select the least expected action rather than what is predicted to be the most effective action. This causes disorientation to the enemy, forcing him to pause, thus stretching the time of the enemy. The enemy is more and more delayed in making accurate decisions, and eventually it is defeated.

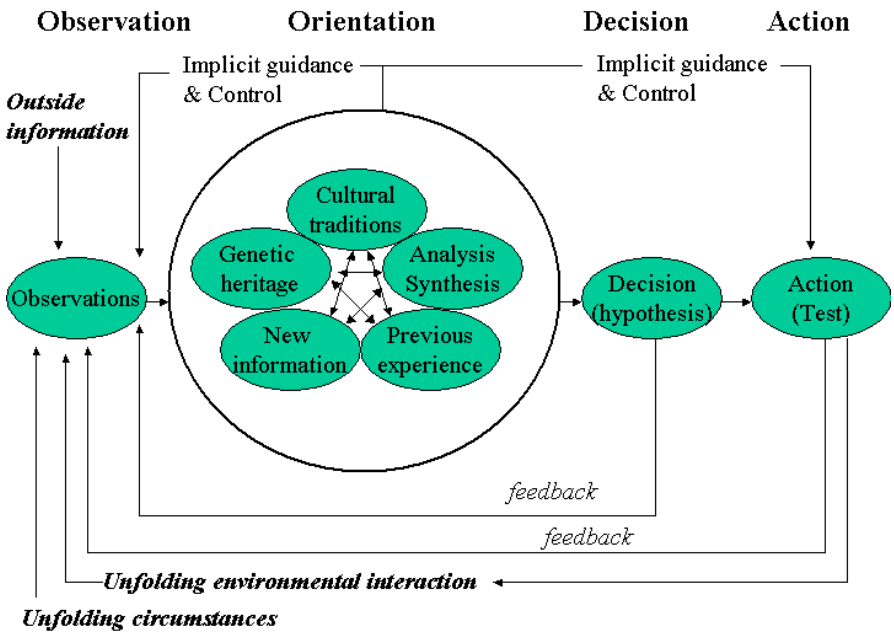

Figure 11 A more detailed picture of the OODA loop

The main stages of the OODA loop are described as follows:

Observation: the first phase of the loop is observation. It is the act of becoming aware through careful and directed attention. In the observation phase, information about the environment is collected.

Orientation: After information has been collected, it is processed further. The human brain forms a mental image of the environment. However, the mental image is also dependent on personal characteristics, such as experience, culture, genetic heritage, and personality. As can be seen from the picture, there are two ways to manipulate information retrieved by observation: analysis and synthesis. Analysis breaks down observed events into individual components and interactions, and makes deductions that lead to understanding. Synthesis, on the other hand, takes several unrelated components and forms them into a new whole.

The orientation phase shapes observation, decision, and action, and is itself shaped by the feedback from the observation phase. It is the most crucial phase in the OODA loop. The speed and accuracy of creating mental images affects the probability to survive in a complex situation of conflict.

In [81], five mental blocks affecting the success of the orientation phase are described:

1. People require a different level of detail to perceive an event

2. People require 3-5 occurrences of the event before they recognize it

3. The existence of preconceived notions where information which does not conform to one's own view is given less priority or ignored

4. Good news is reported quickly whereas bad news is withheld as long as its holder can affect (change) the outcome

5. Communication problems, that is, the message received may not be identical to the one sent.

Once these five problems have been overcome, the process can continue into the decision phase.

Decision: In the decision phase, the information from the orientation phase is analyzed and various options of action are formed. The options are considered to the extent possible by time, and a choice is made. The amount of available information varies, however, a minimum amount of information is required to make a decision. When the minimum amount of correct information has been acquired, an opportunity exists to make a decision. Information gathering beyond this point is useless or even harmful, as the opportunity can be lost and the conflict can evolve. Thus, the decision should be made as soon as the minimum amount of information is available.

Action: In the last phase, the decision is implemented. It leads to unfolding interaction with the environment, which can be observed in the observation phase later on in the OODA cycle.

5 Decision making in a network-centric environment

The key to prevail in a conflict is the cognitive process of the decision maker. In [77], information warfare is analyzed according to the OODA loop. According to it, information warfare is predominantly concerned with denying the enemy the time needed to adapt to warfare situations. This is done by creating and perpetuating a highly fluid and menacing state of affairs and by disrupting the enemy's ability to adjust to the situation [27]. In Clausewitzian terms, the purpose is to create wartime friction and fog [15]. Wartime friction is the "countless minor incidents - the one you can never really foresee - combined to lower the general level of performance, so that one always falls short of the intended goal". Typically, the friction of war results from the physical environment, such as weather conditions, terrain, or degraded command and control, or from psychological factors, such as stress, fear, and chaos. The fog of war refers to the concept of uncertainty. Uncertainty stems from incomplete or contradictory information, chance actions of the enemy, deviation in weapon system efficacy, and the enemy's nebulous capabilities and intentions [21]. Fog and friction can be altered to affect the tempo of operations. In a conflict, both parties try to accomplish this task simultaneously.

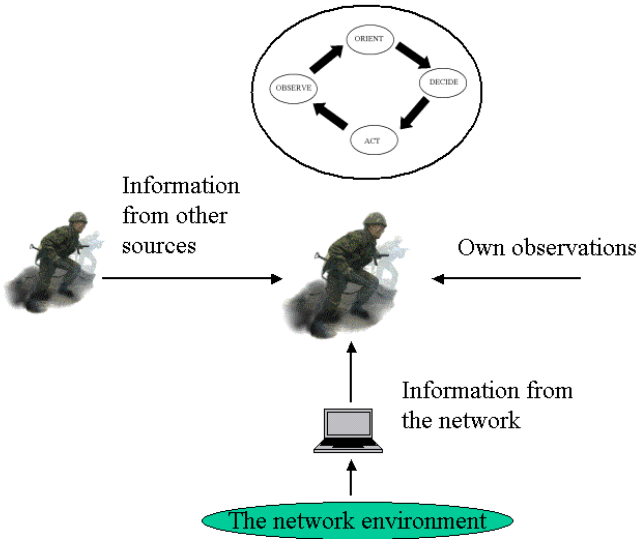

Figure 12 Interaction between a human and the network-centric environment

In a network-centric environment, decision making can be either human or computer based. In Figure 12, the decision making process of a human (commander) operating in a network-centric environment is depicted. The commander takes inputs from the environment through his own observations, from other people, and from the network. These observations are fed into his OODA loop.

The decision made by the commander will be fed back out to the environment and an action will be taken.

The network may change its structure as a result of the decision, for example, the forces may move into a new direction with the result that the network nodes change their location. To cope with frequent changes, the network nodes must themselves be able to make decisions regarding the functionality of the network. The computer based decision making process can be described using the OODA loop as well, however, the way a node goes through its OODA loop is based on a given policy, and the actions taken are always based on a set of rules.

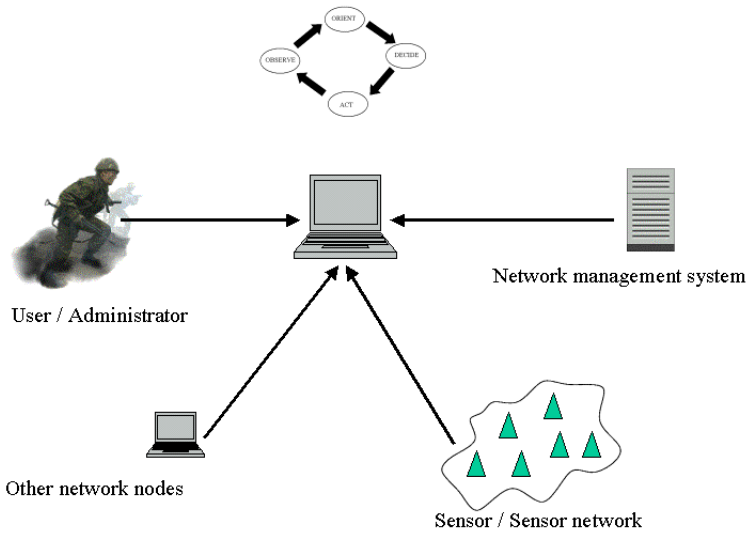

Figure 13 Interaction between a network node and the network-centric environment

In Figure 13, the decision making process of a network node is depicted. The node collects observations from its environment, for example, from other nodes and sensors. The network node will act upon these observations by changing its mode of operation if the environment requires it to behave differently. For example, if the node finds itself in an environment where all other nodes use a different ad hoc routing algorithm, it will switch to the new algorithm in order to participate in routing. The network management system functions as a global authority, and may provide nodes with observations or a new policy. Furthermore, the human user may interfere with the decision making process.

The OODA loop in the case of a network node is less complex than that of a human being, especially with respect to the orientation and decision phases. In the orientation phase, the node has a certain rule-based policy that will state how the observations should be handled. When an image of the observations has been created, the policy also states what sort of decision should be made. The action carries out the decision, and the node starts to operate in its new mode.

To ensure the accuracy of the decision making process, both from the point of view of the human commander and the network node, it is crucial that the network is not vulnerable to "friction and fog". In Figure 14, the relationship between the network-centric environment and the OODA loop is depicted. To cause friction and fog to the enemy, or to protect one's own process, the decision and action phases are the ones that can be directly targeted through the network. In the observation phase, the commanders and the network nodes collect data from the environment. From an offensive point of view, it is possible to destroy, modify, delay, or falsify data to be collected, and thus affect the input to the orientation phase and in the end, the decision to be made. In the action phase, where the decision is carried out, it is possible to disrupt the action by disrupting the communications in various ways, such as jamming the network, delaying packets, modifying packets, or inject erroneous packets into the network. Furthermore, it is also possible to collect intelligence by merely observing the network behaviour.

The orientation and decision phases cannot be directly targeted through the network; from a network-centric perspective they merely depend on the input that is fed to them from the network. From the point of view of information warfare, however, these phases can be directly targeted by psychological means. For example, propaganda can affect the way people go through their orientation phase and thus make decisions. Psychological means are, however, out of scope of this thesis.

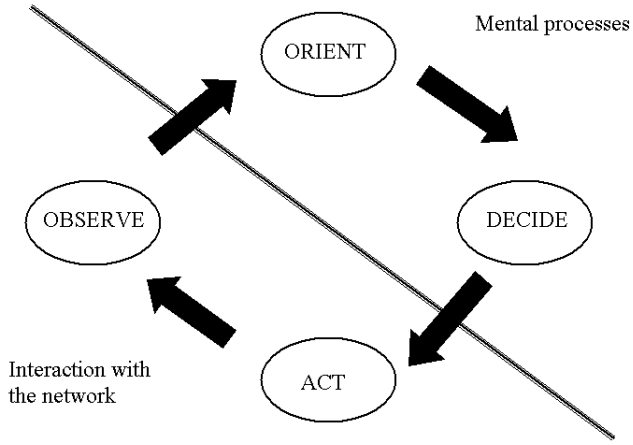

Figure 14 The OODA loop from a network-centric environment

The main purpose of the network is to function as a tool for leadership. It enables people to communicate regardless of location, thus eliminating the factor of place and time that has earlier constrained the forces to a small geographical area. Furthermore, together with modern information technology, it is possible to collect, process, and distribute information efficiently, and thus achieve real-time situation awareness. However, from the point of view of the network, its main task is to deliver the right IP packets to the right place(s) in time and intact, and without violating any specified privacy policy.

This has several implications with regards to security. The right packets means that only legitimate traffic may be routed in the network, that is, packets sent by legitimate nodes that do not show any signs of misbehavior. The packets that are routed should arrive at their destination without having been illegitimately modified on the way, that is, their integrity should be intact. To ensure that packets arrive at the right place in time it is crucial that the routing infrastructure is sound. Only legitimate nodes may participate in routing. If a legitimate node becomes compromised, it should no longer be allowed to participate in network operations.

The signaling traffic between nodes must be verifiable to protect the routing tables from corruption. Furthermore, denial-of-service attacks

should be prevented in order to minimize delays caused by congestion or, in the worst case, a complete collapse of the infrastructure. Privacy protection is crucial, as the network always leaks information, such as the structure of the network, the location of important nodes (e.g. those belonging to the commanders), the intent of the forces, the level of training of the forces, and so on. However, complete privacy protection is extremely difficult in a wireless environment.

Future military systems will be based on a network of heterogeneous networks, such as sensor networks, tactical battlefield networks (e.g. ad hoc networks), wired and wireless access networks, fixed and temporary backbone networks, and so on. An example scenario is depicted in Figure 15. The ad hoc networks connect to the fixed network via wireless access networks whenever they are in the coverage area of a wireless access point. All access technologies do not cover all areas, so the most suitable alternative is chosen depending on the situation and location.

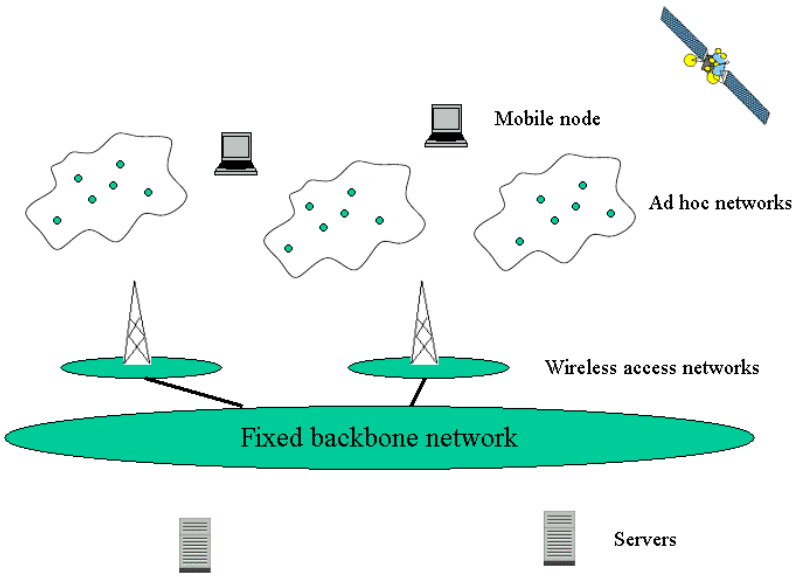

Figure 15 A conceptual model of the network architecture

For example, the ad hoc network can be connected via satellite communications if it is located in a remote area where no other options are

available or when reliability is required. Mobile nodes that are not part of an ad hoc network but roam around may use the routing services of available ad hoc networks to connect to the fixed infrastructure.

However, these mobile nodes do not themselves participate in network operations, such as routing and network management. The wireless access networks, in turn, can be fixed or temporary. Their main task is to provide connectivity between mobile nodes or networks and the fixed infrastructure. The fixed backbone network may also be temporary in the sense that it is built for the purpose of connecting the various mobile nodes, access networks, and ad hoc networks during some military operations, however, the structure of the network will not significantly change over time. Servers such as anchoring points for mobility management schemes, information databases, etc. are typically placed in the fixed infrastructure.

When using the network for human decision making, and when decisions regarding the functionality of the network are made by computers, it is crucial to secure the decision making process. The observation phase must be protected so that only correct and timely data is collected. The action phase, in turn, must be protected so that the network supports the actions to be taken.

5.1 Security in a network-centric environment

Security can be divided into three layers as depicted in Figure 16. *Content security* refers to protecting the content between the two communicating peers, i.e. the end users. The user may be a human being or a computer process. Typically the information is protected from disclosure and modification. Furthermore, the end users must be able to verify the source of the information. *Communication security* refers to protecting the data that is sent from source to destination over the network.

The endpoints in this case are the computer nodes, however, a computer node may have several end users (people or processes). This is the main distinction between content and communication security; in the former case, the endpoints are the true endpoints of the data transmission, in the latter case, the endpoints are the computers sending and receiving the data. The communicating nodes must be able to verify the legitimacy of each other prior to communication. Furthermore, the data must be protected from disclosure, modification, delays, and so on. *Network security* refers to

the security of the network itself, that is, the network must be secured so that it is able to perform its tasks of routing the right packets to the right place intact and in time without violating specified privacy policies. This level of security differs from the two others in the sense that it is not concerned with the actual content but rather with how to transfer the content from source to destination.

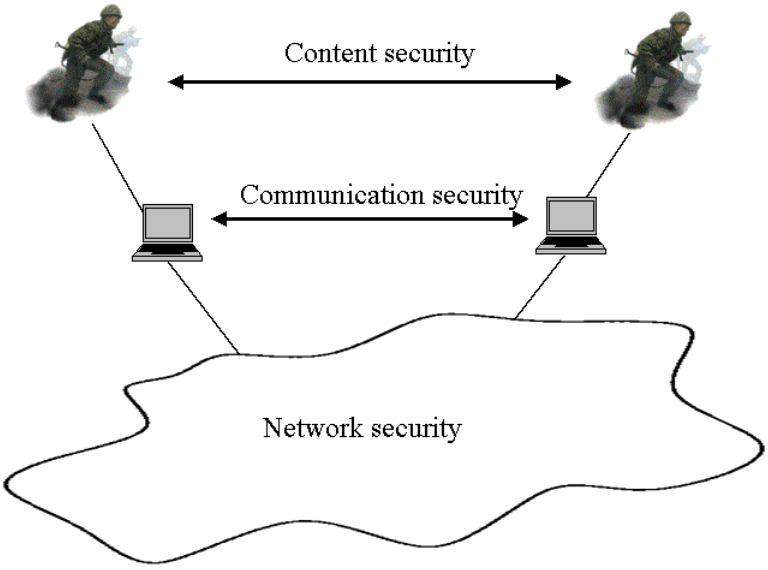

Figure 16 Three levels of security

An example of a content security solution is PGP, which was created as a public-key encryption system for protecting e.g. email. PGP offers user authentication (the recipient can verify that the message was sent by the claimed person), encryption, non-repudiation (the sender cannot deny having sent the message), and integrity protection (the recipient can verify that the message has not been illegitimately modified).

Communication security solutions include protocols such as IPSec [44], TLS [23], and SSH. IPSec is a protocol framework for protecting IP traffic. TLS is a transport layer security solution for protecting the communication between a client and a server, for example, bank connections between a home user and the bank. Also HTTP [31], POP3 [63], IMAP [18], and SMTP [48] use TLS for security. SSH provides

secure remote shell connections. It is also possible to use SSH for secure tunneling of applications.

The IPSec framework consists of the following protocols: the Authentication Header (AH)[45], the Encapsulated Security Payload (ESP)[46], IP payload compression [76], and the Internet Key Exchange (IKE) protocol[34]. AH provides data authentication and integrity protection of the packet, excluding the headers. ESP offers data authentication, integrity protection over the whole packet, and encryption. IP payload compression is used to compress the data before encryption. IKE is used for key exchange and establishment of security associations between nodes. IPSec provides security on the network layer and is thus transparent to upper layer protocols.

Network security takes several aspects into consideration. One of the key issues is availability. Physical security is concerned with protecting the cables and network nodes, such as routers, by various access control mechanisms. Link level security is concerned with the protection of the links between two nodes. The purpose of link level security in a wireless network is to provide the same security as cables do in a wired network. Typically, link level security includes authentication and encryption. In addition, the signaling protocols used in the network must be secured to prevent an enemy from, for example, changing the routing structure or affecting network management. Also filtering techniques can be used restrict or limit network traffic.

5.2 Attack scenarios on the infrastructure

The infrastructure can be attacked in various ways.

Although the attack scenarios are in a military environment, most of them are relevant in civilian networks as well.

5.2.1 Attacking the infrastructure

Attacking the infrastructure can either be done externally or internally. In the former case, an illegitimate node is able to penetrate the network and masquerade as a legitimate node. In the latter case, a legitimate node is compromised, i.e. taken over by the enemy, and is then used to attack the network. Internal attacks are discussed further in Section 5.2.3.

If a node is able to masquerade as a legitimate node, it is able to perform a variety of attacks, such as:

- Denial-of-service attacks: the node is able to consume network resources by flooding. This is especially severe in wireless ad hoc networks where network resources are scarce and the network nodes have limited power resources.
- Disrupting protocol signaling: the node is able to take part in protocol signaling, and can thus e.g. change the routing infrastructure.
- Disrupting network traffic: the node is able to drop or delay packets. This affects the function of upper layer protocols, causes unnecessary signaling traffic in the networks, and eventually disrupts the whole network.
- Traffic analysis: the node is able to spy on network traffic and deduce the network structure, location of crucial nodes, and so on.
- Spreading disinformation: the node is able to spread disinformation in the network. For example, a node may act as a sensor and feed erroneous data into a data fusion process that attempts to create an image of the environment.

In Figure 17, an attack scenario with two enemy nodes, E1 and E2, is depicted. Node E1 masquerades as a legitimate node and provides a server with disinformation. The server collects data from various sources and composes an image of the environment through data fusion. However, since some of the data is erroneous, the final image will be misleading. This misleading image is distributed in the network, and may eventually lead to a commander making wrong decisions based on wrong situational awareness.

Node E2, on the other hand, launches some denial-of-service attacks on the medium. When receiving packets from the server, it duplicates the packets and forwards them to the following node. The denial-of-service attack propagates all the way to the destination, which notices the duplicate packets, and discards them. If the network medium is wireless, this attack is especially severe, since all nodes in the neighborhood of E2 and the ones on the path from E2 to the destination are affected. Although it is difficult to prevent E2 from flooding the network, it should not be allowed to use legitimate nodes for forwarding the flooding.

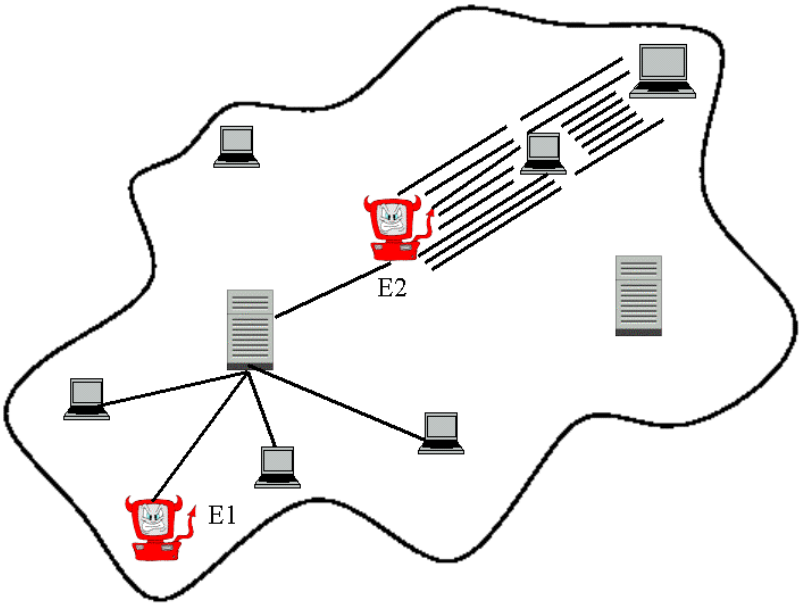

Figure 17 Two illegitimate nodes attacking the network

To limit attacks such as the ones described above, it must be possible to authenticate all the nodes in the network, as well as all the packets. Any node should be able to verify that the packets are legitimate, timely, and unique (no duplicates). Furthermore, the nodes should not be able to deny having sent the packets, hence making it possible to spot misbehaving nodes.

5.2.2 Destroying the infrastructure

The network may be subject to physical destruction, with the result that the network becomes partitioned and some services may no longer be available. In Figure 18, the network has been partially destroyed. The dashed squares show the destroyed routers and the dashed lines the routes that used to exist. The network is now partitioned into two separate parts.

Physical destruction of the backbone network is especially severe. As a result, ad hoc networks may lose their connectivity and be left as floating islands where nodes are only able to communicate among themselves. Mobile nodes may lose their ability to roam if the mobility management schemes break down, for example, as a result of a lost connection between

the mobile node and its anchoring point. Servers may seize to function, such as name servers, key servers, weather servers, data fusion servers, and so on.

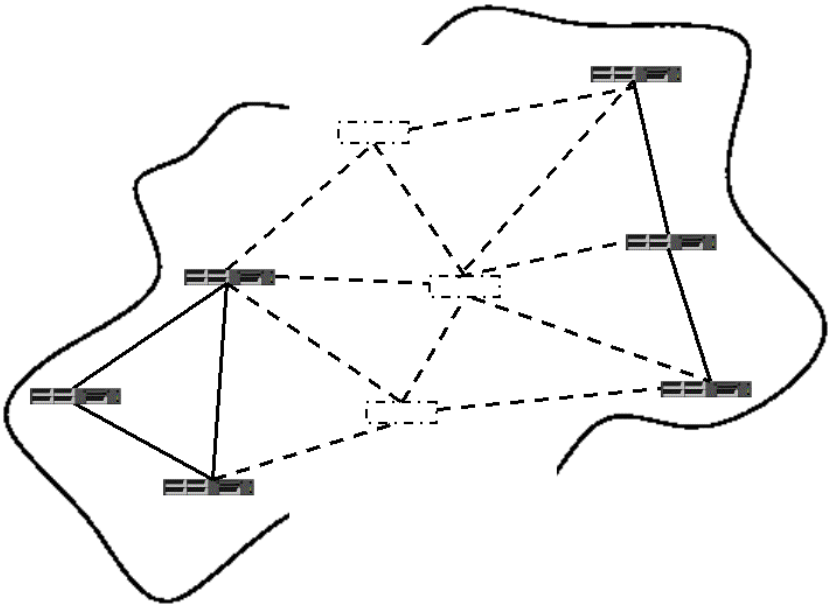

Figure 18 A partially destroyed network

To cope with partial destruction, routes and servers must be restored rapidly. To limit human involvement in the process, networks should have the ability to restore and configure themselves using whatever means they have available.

In Figure 19, a more dedicated attack on the network is depicted. In the former case, a large portion of the network was destroyed, for example, as a result of bombing. The network was not the target of the attack, but it was yet affected by it. In the dedicated attack, the network is the primary target. Instead of eliminating several nodes, only crucial nodes are targeted. The crucial node in this case is the access router that connects the local area network to the backbone. Without the access router, the nodes have no connection to the backbone. Thus, instead of targeting the whole network, which would be impractical or even impossible, it is easier to find the crucial nodes and only eliminate them.

A common solution is not to have any single points of failure, as the access router in the example. However, it is not possible to duplicate a service

endlessly either. In the case of an ad hoc network, where the cluster head is lost, the functionality can always be assigned to another node. Such dynamic solutions make it possible to eliminate the problem of single point of failures while not maintaining duplicate services simultaneously at all times.

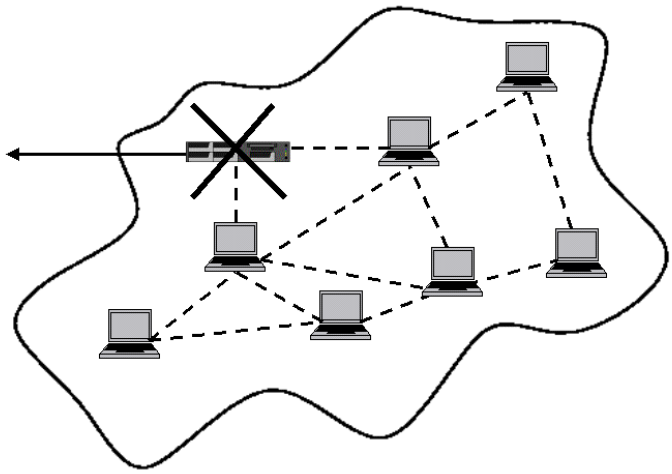

Figure 19 A targeted physical attack on the infrastructure

To cope with destruction in general, whether it occurred as a side effect of an attack or as a result of a successful targeted attack, a means for rapid restoration is needed, where restoration does not rely only on the fact that the services are duplicated, but on the fact that services can be dynamically assigned to new nodes. This implies that the nodes may have dormant services, which are activated only when needed.

5.2.3 Compromised nodes

A compromised node is a node that has been taken over by the enemy, either directly (physically) or through the network. When the node has become compromised, it is completely controlled by the enemy. However, to the network the node still appears to be legitimate. Thus, the enemy has access to all services that the node is authorized to use.

If a node is compromised, the enemy is able to perform a variety of attacks, such as:

- All the attacks that a masqueraded node is able to perform (see Section 5.2.1.).
- Compromise other nodes by lying to them and thus gradually take over the network.
- Break the "security" provided by various radio network level mechanisms, such as frequency hopping schemes etc., thus making it possible for the enemy to perform jamming attacks on the radio level.
- Access all the services and information that the legitimate node is authorized to access and use the services and information to its own advantage, or perform malicious actions against the services and information. For example, the compromised node may modify information without violating its integrity, as the node is authorized to alter the information, and trusted to do so in a benevolent way.
- Change network behavior by spreading disinformation.

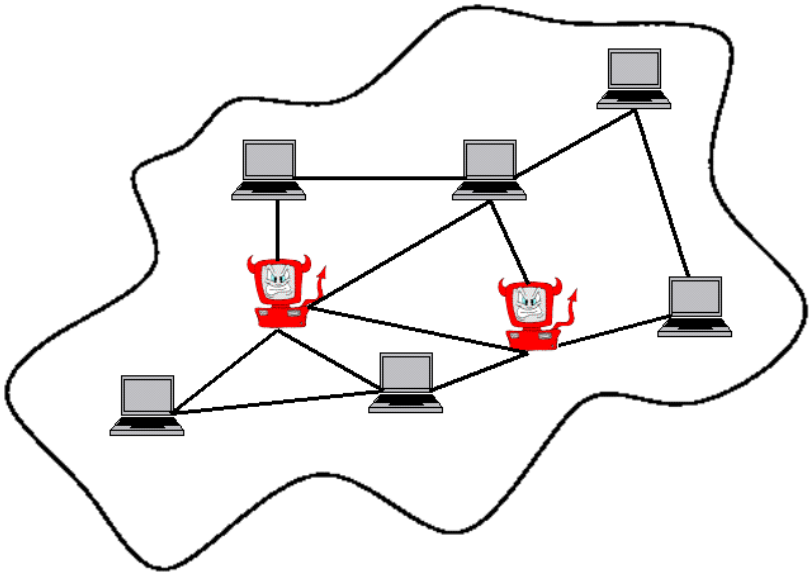

Figure 20 A network with compromised nodes

In Figure 20, an ad hoc network with compromised nodes is illustrated. In the depicted scenario, the compromised nodes may disrupt routing by dropping packets to prevent data from reaching the intended recipients, duplicate packets to waste network resources, and give erroneous route replies to route requests. Furthermore, both compromised nodes are able to collect important information about the structure of the network as well as which nodes are important.

In Figure 21, the compromised nodes (E1 and E2) outnumber the legitimate ones (A). All nodes provide the data fusion server (DFS) with data. A sends DFS the correct data, X. However, E1 and E2, have agreed to send data X´, which is incorrect, to DFS. Without a proper trust handling mechanism, DFS assumes that A is the illegitimate node, and trusts the majority, that is, the data provided by the enemy. Hence, the data fusion image that is created is based on wrong data. Furthermore, DFS may claim that A is illegitimate and have the certificate of A revoked. In the long run, the enemy will have majority in the network, and is then able to control the decision making process. This is especially easy in the case where compromised nodes can selectively block messages from legitimate nodes.

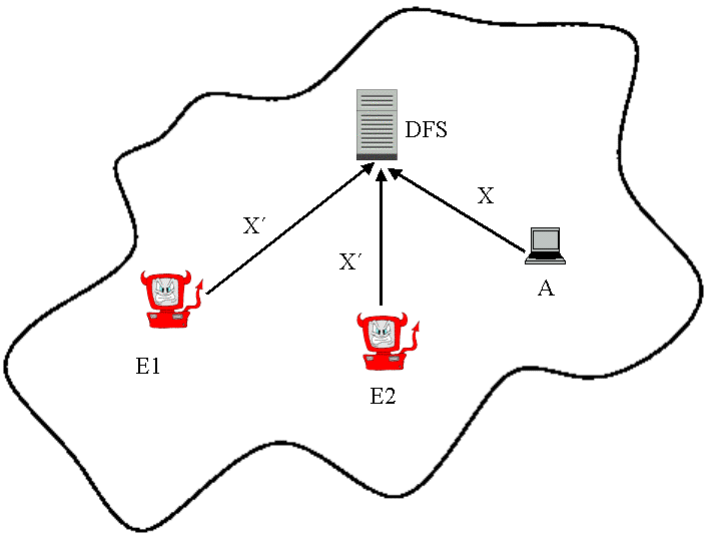

Figure 21 Compromised nodes affect data fusion

The problem with internal attacks, such as the ones described above, is that traditional security solutions based on (strong) cryptography cannot be

used. A compromised node has bypassed the protection given by cryptography; the private keys of the node is known to the enemy and there is no way to distinguish the node from a legitimate node, except by monitoring behavior. Behavior monitoring is not trivial. For example, a legitimate node may misbehave due to (temporary) malfunction, whereas a compromised node may behave well for a long period of time, or misbehave in a way that is not easily detected.

5.2.4 Network surveillance

Merely observing the network traffic can be used to gain crucial information, even if the content of the network traffic is not revealed.

In Figure 22, a scenario where a tactical battlefield network is under surveillance by the enemy is depicted. Thicker lines refer to a larger traffic flow, and the most crucial nodes are depicted as larger than less important nodes.

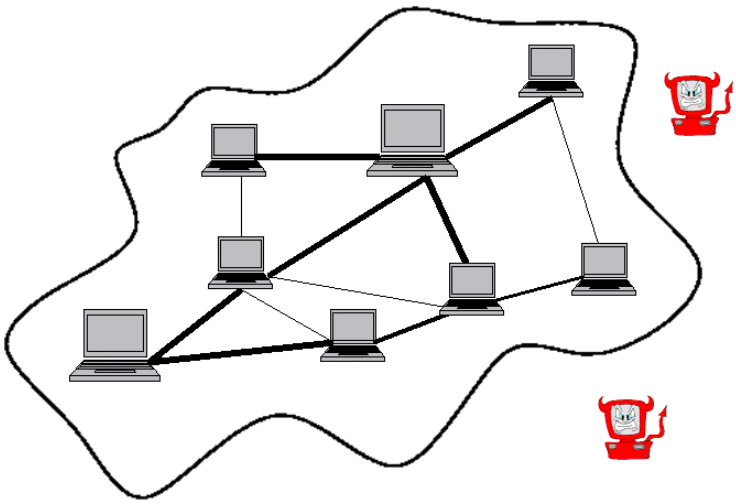

Figure 22 The network is under enemy surveillance

The structure of the network is easily revealed by the traffic flow itself. Nodes sending and receiving a lot of data are typically more crucial than more quiet nodes, however, it is easier to detect a sending node than a receiving node unless the recipient sends acknowledgments. The network

structure may also give away the command structure of the forces, thus making it easy to track down the commanders and eliminate them.

Traffic analysis can also reveal the intent of the forces as well as their level of training. Traditionally, before an attack, the amount of traffic increases. Less trained personnel also tend to cause more traffic than well-trained personnel.

Although the content of the packets are encrypted, the length of the messages can give a hint on what protocols are run. For example, if it can be deduced that a set of messages are part of a mobile node's location update procedures, it is easy to attack the mobility management.

5.3 Security criteria for network level security

To ensure correct and timely decision making for humans and computers, the network must fulfill certain security requirements. The attack scenarios described in previous sections show that the networks are vulnerable to various attacks that could affect the way decisions are made and carried out.

The most important security criteria are the following:

1. The network should be able to perform its tasks, namely transport legitimate packets intact to the right place(s) in time. This ability should not be affected even if the network is partially destroyed or otherwise attacked.

2. Should the network become partially destroyed, a means of rapidly recovering is required to ensure continuous functionality of the network. This requirement supports the first requirement (that the network should be able to perform its tasks).

3. Trust management is important to ensure that the information retrieved from the network is correct and the information that goes to the network will be properly handled. This implies that compromised nodes should be detected and removed. If the network cannot be trusted, it becomes useless.

4. When performing its tasks, the network should not give away any information to unauthorized parties. In an utopist world, the enemy

is not even aware of the network; in an ideal world, the enemy can see the network but cannot deduce anything from it. In reality, enough privacy should be guaranteed so as not to reveal the most important information, namely the command chains, crucial network nodes, and intent.

6 Solution

The military environment is difficult in many respects, but one of the main problems is the presence of an active enemy. All the solutions and ideas presented in this section assume that the network is under constant attack. The objective is to be able to operate in critical situations.

Traditional security solutions do not address the security requirements listed in the previous Section. Most solutions address the security of the communication rather than the network itself. Furthermore, the solutions typically assume a rather static environment with a fairly limited number of network nodes. If mobility occurs it is typically assumed to be slow. The network medium is typically assumed to be wired, and wireless security is usually addressed by suggesting link level solutions.

Furthermore, the military environment is hostile and the network is under constant attack, resulting in network nodes becoming compromised. Many traditional security solutions rely on symmetric cryptography for efficiency, however, symmetric cryptography becomes inefficient when nodes are compromised. This situation is especially severe in large networks with wireless mobile nodes, as the keys have to be renewed and new security associations established.

In this Section, a set of solutions to address the security of the network infrastructure are proposed. The solutions include Packet Level Authentication (see Section 6.1.) to ensure that only legitimate packets are transported in the network, Context Aware Management (see Section 6.2.) to allow nodes to rapidly and securely adapt to the changing environment, Self-healing networks (see Section 6.3.) to ensure continuous communications even if the network is partially destroyed, trust management based on incomplete trust to ensure that nodes only communicate with non-compromised nodes (see Section 6.4.), and privacy protection to prevent the network from leaking information to the enemy (see Section 6.5).

6.1 Packet level authentication

To address the first criteria, packet level authentication is proposed as a solution for authenticating packets (thus e.g. securing protocol signaling),

preventing denial-of-service attacks from paralyzing the whole network, and integrity protection. The PLA architecture is described in more detail in [67].

Despite the fact that it is possible to protect the communication with strong security protocols and encrypt the data with solid encryption algorithms, information flooding or shortage attacks still remain a problem with traditional solutions. External attackers can, for example, perform the following attacks:

- DoS and DDoS attacks over the network paralyzing critical communication
- Packet manipulation between legitimate senders and receivers
- Duplication of IP packets of legitimate users
- Manipulation of routes, thus causing information flood or shortage
- Utilization of the complexity and vulnerabilities of the protocols

In addition, internal attacks are very difficult to prevent since no scalable solutions exist to revoke compromised users or nodes. When legitimate users are relying on the steady operation of the communication infrastructure, any of the above mentioned attacks can easily paralyze the entire system. Especially, wireless networks are opening the physical networks to attacks.

6.1.1 Previous solutions

Traditional solutions for securing the network include link level security, and IPSec.

In link level security, the link between two nodes is encrypted and the nodes are able to authenticate each other, see Figure 23. Thus, each node has a security association with its neighboring nodes. The problems with link level security in a dynamic military environment, where nodes may become compromised, are manyfold. When the nodes move, the security associations have to be updated. This requires a certain amount of signaling between the nodes, for which there may not be enough time. Furthermore, it is not possible to detect and handle compromised nodes. Any node on the path can modify the packets without being caught already by the following node.

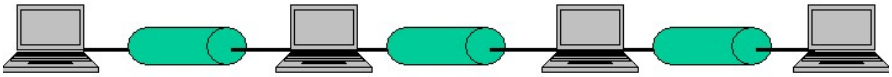

Figure 23 Link level security

The purpose of link level security in wireless networks has been to provide the same level of physical security as cables do in a wired network. Hence, link level security is not a solution for securing the network from the attacks mentioned previously.

IPSec provides a solution on the network layer. However, IPSec in its basic form only provides end-to-end security. This means that the destination node is able to verify the authenticity and integrity of the packet, however, the nodes on the path from source to destination cannot. Hence, if a packet is falsified or modified, it will propagate through the network undetected, thus wasting network resources.

Naturally it is possible to establish Security Associations (SA) between all nodes and thus address the problems from end-to-end security. However, since the network is dynamic, the SAs would have to be renewed on frequent basis. As the nodes may move fast, there is not sufficient time to run the IKE protocol. Storing the SAs in advance is not feasible either, as the size of a military network is too large and the nodes are by far too restricted. Furthermore, the amount of signaling traffic due to SA negotiations would be too large.

Another problem with IPSec in this case is that the integrity of the packets is ensured using symmetric keys for performance reasons. The same key is used both for signing and for verification. Thus, a compromised node that receives a packet is able to modify it, sign it again, and pass it on as a legitimate packet. Hence, compromised nodes cannot be handled.

6.1.2 The packet level authentication concept

To secure the networks in a dynamic and hostile environment, it must be possible for any legitimate node to verify the authenticity and integrity of an IP packet. PLA solves the problem without suffering from the same limitations as link level security solutions and IPSec.

The PLA concept resembles that of the security of money. Every modern note (e.g. a 100 euro note) contains several security measures to ensure the authenticity of the note, such as micro print, changing colors, watermark, hologram, metal string, etc. When a merchant receives the note from a customer, the legitimacy of the note can be verified on the spot without having to consult a bank. The requirement is that the merchant must be able to verify the authenticity of a note using predefined security procedures and without prior knowledge of the customer.

In PLA, the same idea is applied to IP packets, that is, any node is able to verify the authenticity of the IP packet using predefined security procedures and without prior knowledge of or communication with the sender. The requirement applies for any receiver on the path from the source to the final destination(s). Thus, PLA is able of detecting illegitimate, erroneous, duplicated, and delayed packets in every router, not only at the final destination. In this way it is possible to restrict several classical attacks, such as denial-of-service attacks that are based on spoofing/forging IP packets and copying or manipulating legitimate IP packets. PLA also includes additional security features to cope with duplicated and replayed packets.

6.1.3 Design criteria

The requirements for PLA are the following:

- All information required for verification is in the IP packet. This means that the verification procedures can be based on one single packet, and no third parties need to be contacted.
- Any node can perform the verification.
- Duplicate IP packets must be recognized.
- Delayed packets (e.g. as a result of a replay attack) must be recognized.
- Removing compromised nodes from the network must not affect legitimate nodes.

Based on the requirements, it is obvious that public key cryptography is required for the verification procedures.

In turn, PLA places the following requirements on the network:

- Deployment of standard PKI to provide TTP and revocation services as well as certificate management
- A common sense of time

6.1.4 Design issues of packet level authentication

The PLA architecture takes advantage of standard IPv6 [20] header extension techniques used for example in Mobile IP [68][41]. The PLA header is added to every IP packet and contains the following information:

Node level information:

- The identity of the trusted third party that has certified the node
- The public key of the node
- The validity period
- The certificate from the trusted third party that certifies the key

Packet level information:

- Timeliness of the packet
- The sequence number of the packet
- The signature of the source node over the packet using the private key of the node

With the node level information, any node (on the path from the originator to the final destination) that trusts the third party can verify that the originator of the packet is still among the trusted nodes without prior communication with that node. With the packet level information, that node can then verify that the originator of the packet has really generated this packet and the packet is not altered on the path from the sender to the receiver. Also, the node can verify the timeliness of the packet. Duplication of the packet is detected by using sequence numbers. Thus, corrupted, old, and duplicated packets can be discarded. This restricts the possibilities that the enemy have to attack the network. In the worst case [55], the enemy can make as many copies as there are separate routes to the destination node. In other words, the worst case is that the enemy is able to make a few extra duplicates.

Public key algorithms are used for signing the certificates stating the validity of the key as well as the node's signature over the packet. The lifetime of the certificates is limited, thus making it possible to reduce the damage caused by internal attacks either by revoking the certificate or not renewing it when it expires.

6.1.5 Packet level authentication extension header

The PLA extension header is depicted in Figure 24. Since the header is based on the standard IPv6 header extension technique, it is fully transparent and interoperable with routers that do not understand PLA. This means that PLA can be used together with any other protocols, such as Mobile IP or IPsec. If the PLA header is the first header in the packet, it protects both the data and all the other headers in the packet. Since PLA uses the standard IP header technique, the maximum IP packet size (MTU) handling is done correctly.

The PLA header consists of seven fields divided into two categories:

Node level information:

- Auth_id: The identity of the trusted third party that has certified the node

- Pub_key: The public key of the node (this public key is used by the verifying node to validate the sending node's signature)

- Val_period: The validity period (this field indicates when the public key is considered to be valid by the trusted third party)

- Auth_signature: The certificate from the trusted third party that certifies the node (that the public key and validity period are correct)

Packet level information:

- Time: Timeliness of the packet (i.e., time when the packet was sent; it can be used to discard old packets)

- Seq_number: Sequence number of the packet (this monotonically growing field can be used in discarding duplicates and replay attacks)

- Node_signature: The originating node's signature over the packet (this signature field ensures that the IP packet was really generated by the sending node and it has not been altered on the path).

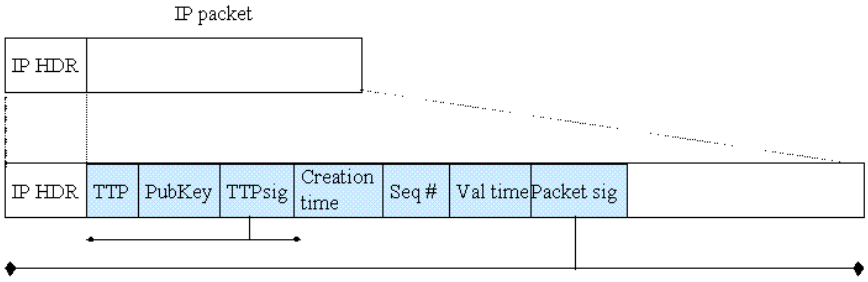

Figure 24 The PLA header

When the source node sends an IP packet, it fills the PLA header by taking the node level information (that is packet independent), incremented sequence number and current time. Finally, it calculates a signature over the entire packet and puts the result into the header. Then the IP packet is sent as any normal IP packet. However, variable fields of the packet cannot be protected, such as hop limit or routing header. The signature can be calculated using any of the standard digital signature algorithms, however, an algorithms like ECCDSA is promising, since the size of the signature and the public key in every packet can be kept short.

When a node receives the first packet from the sending node, it first validates the node level information and after that the packet level information. If any of the PLA fields is invalid or incorrect, the receiving node can discard the packet and optionally inform its upstream neighbor about the problem. For the subsequent packets from the same sending node, it is not needed to validate the node level information if the verifier caches the previous validation result. Then, only the packet level verification is needed. Hence, the number of digital signatures that must be validated per packet is reduced from two signatures per packet to just one in case of several packets from the same sender.

6.1.6 Analysis of PLA

The objective of PLA is to enable any legitimate node to verify the authenticity, integrity, uniqueness, and timeliness of an IP packet.

The authenticity is easily verified using public key cryptography. Any node can verify the signature of the trusted third party over the public key

of the sender. Integrity can be ensured by verifying that the signature over the packet is made by the private key corresponding to the public key.

Since the sequence number is added to the packet, it is possible to ensure that the packet is unique and not a duplicate. Any duplicates will be dropped immediately. This prevents denial-of-service attacks from spreading in the network. Only the neighboring nodes of the flooder will be affected. The time stamp ensures that the packet is timely and not delayed. Packets that are too old will be dropped. The time stamp together with the sequence number provide a good way of handling duplicates and preventing replay attacks, because the sequence number may start all over again. However, the packet can still be distinguished as unique.

PLA also provides non-repudiation, that is, a malicious node cannot deny having created the IP packets.

As a result of the security solutions mentioned above, PLA can be used for authentication, integrity protection, denial-of-service attack restriction, firewall applications, access control mechanisms, billing and charging (in a commercial environment), and so on.

However, PLA also has some limitations in its current form. The public key solutions introduce a problem with respect to identity privacy. This solution has not been solved, although temporary keys could provide a solution. However, this would increase the computational cost of verifying the certificate.

Time stamps come with the problem of having a common sense of time. The nodes must be synchronized in order for the scheme to work properly.

PLA also has some performance issues that need to be solved. The nodes on the path have to verify at least one digital signature per packet, sometimes two. Furthermore, the overhead of each packet is roughly 100 bytes. However, the latter is merely a tradeoff; by including the PLA header it is possible to maintain communications even when the network is under attack, whereas without the header the network would be completely congested.

However, with regards to performance, PLA also has some benefits. For example, comparing to IPSec where security associations have to be negotiated e.g. using IKE, PLA only requires one message to establish a security association. In Figure 25, node C arrives as a newcomer to a military network. It hears nodes A and B communicating, and hence,

learns their public keys. When C wishes to communicate e.g. with B, it sends an encrypted message to B containing the security association. Node B can right away verify the authenticity of C, and decide whether to accept the association or not. However, in most cryptographic schemes, it is not possible to use the same public key for encryption and signing. This example would thus have to use, for example, RSA.

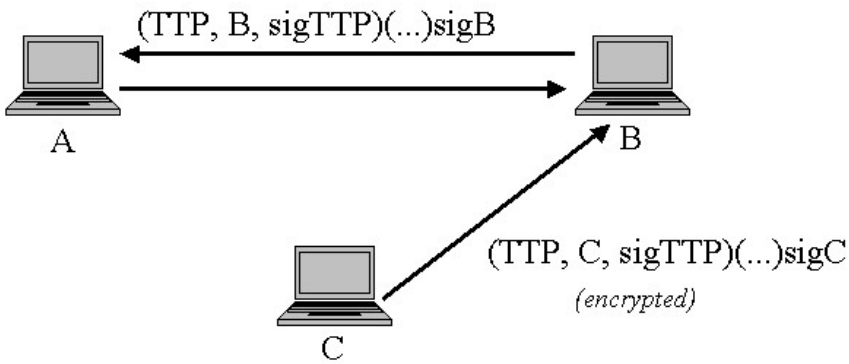

Figure 25 Establishing security associations in PLA

6.1.7 Deploying PLA

The deployment of PLA can be divided into three categories: initialization, communication, and special situations.

INITIALIZATION

Before the nodes enter the network, they must be validated. In a military scenario, the node is validated offline; it creates or receives a public key, which is signed by a trusted third party. When the node has been validated, it is able to enter the network, as it is now able to prove that it is a legitimate node. The validation of network nodes is depicted in Figure 26.

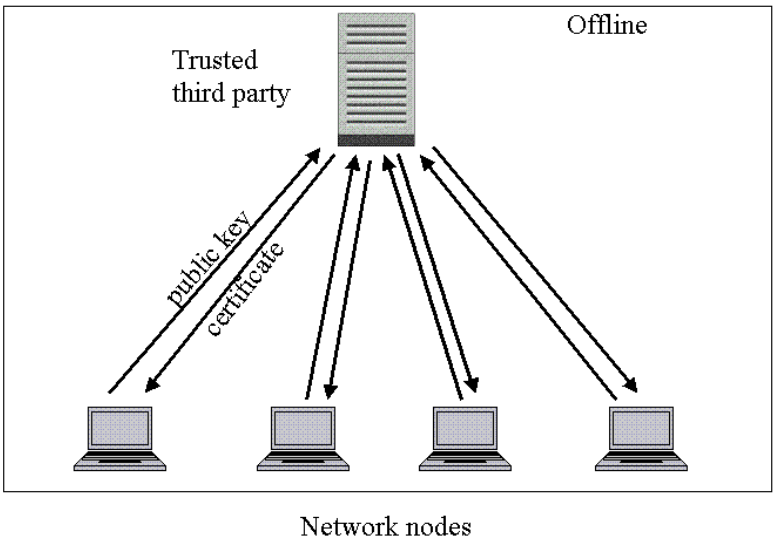

Figure 26 Validation of network nodes

Since the initialization can be offline, the process can be secured by physical security means.

COMMUNICATION WITH PLA

When the nodes are in the network environment, they may begin to communicate with each other.

When a node wishes to send a message, it adds the PLA header with all the required information and signs it.

Upon receiving a message, the first node on the path verifies the authenticity of the sender. If the signature cannot be verified, the packet is dropped. If the sender is valid, the node checks the signature of the whole packet. If it is valid, the node passes the packet on. The authenticity fields are stored in the cache of the node for a short period of time, as this information will not change. Hence, only one digital signature verification is needed for further packets from the sender. The following nodes on the path from source to destination, including the destination node, may perform the checks in a similar fashion.

This operation resembles that of the CRC checksum calculated on the link level. The main difference is that only one node (the owner of the private key) can compute the correct checksum for the other nodes to verify.

SPECIAL SITUATIONS

There are three main special situations:

- Acquiring a new certificate
- Revocation of compromised nodes
- Malfunctioning nodes

When the certificate of a node is about to expire, the node may retrieve a new certificate from an online TTP server while it is still able to prove its legitimacy. Should the node not be able to contact an online TTP before its certificate expires, it must be reinitialized offline.

Revocation of compromised nodes can be done by broadcasting the revocation lists using some technology that is difficult to disturb, for example, satellite communications. Since the revocation lists are signed, all nodes do not need to receive the lists directly; they may query other nodes for the list.

Malfunctioning nodes may have lost their certificates. If the malfunction is fixed, the node should be reinitialized offline.

6.2 Context Aware Management Architecture

Most protocols and applications have been designed for fairly static environments. However, the dynamic nature of the military environment requires that the nodes are able to adapt to the changes rapidly. If the reaction is too slow, the it may result in the following:

- The quality of service is bad. For example, the transmission of a video stream may contain glitches, or the service cannot be provided at all.

- Resources are wasted. For example, a video stream is transferred from the server to the access network, but the access network is unable to deliver the stream to the mobile node.

- An uneconomical connection is used. For example, the node has several connections to the network, but uses one that is less optimal than another.

- The decisions are not optimal. For example, the mobility management system chooses a non-optimal access medium.

Traditionally, each application or protocol tries to adapt to the new environment independently. In some solutions, an application is specifically tailored to communicate with one other protocol layer. For example, the Real Time Protocol (RTP) [78] (and its control protocol RTCP) monitors the quality of the connection and informs the application (e.g. the multimedia video player) about degradation in quality. The application then changes the video encoding to better suit the current quality of the connection. Another example is a combination of Mobile IP with a mechanism for choosing the access medium considered optimal for the applications.

The main problems with current solutions are the following:

- The solutions are not generic. Thus, it is hard to add context awareness to protocols and applications. Adding hooks between all layers, especially through intermediate layers, would be too complex.

- Protocol and application design becomes complex, as each protocol and application has to be tailored to intercommunicate with each other.

- Old protocols and applications cannot take advantage of information provided by new protocols or applications without modification.

To overcome these problems, a Context Aware Management (CAM) architecture is presented in [12]. In the CAM architecture, a new management layer is added to the Internet protocol stack. The purpose of CAM is to monitor the environment for changes and to adapt the behavior of the node to the current environment. The applications and protocols need not be aware of the environment at all, but rather focus on taking care of the tasks they have been designed for in the first place. For example, a routing protocol is responsible for establishing routes and forwarding

packets, but need not handle issues regarding the choice of access medium. The decisions how to change the behavior of the node are made by the Policy Manager (PM). Furthermore, the architecture consists of a common database, which contains all environment related data of the node. The database is accessed via CAM. Protocols and applications communicate with each other and with the database through CAM, using a standard interface.

The CAM architecture is depicted in Figure 27. It contains the following components:

- The Context Aware Management (CAM) layer.
- The Policy Manager (PM)
- The common database (CDB)
- The modules attached to CAM.

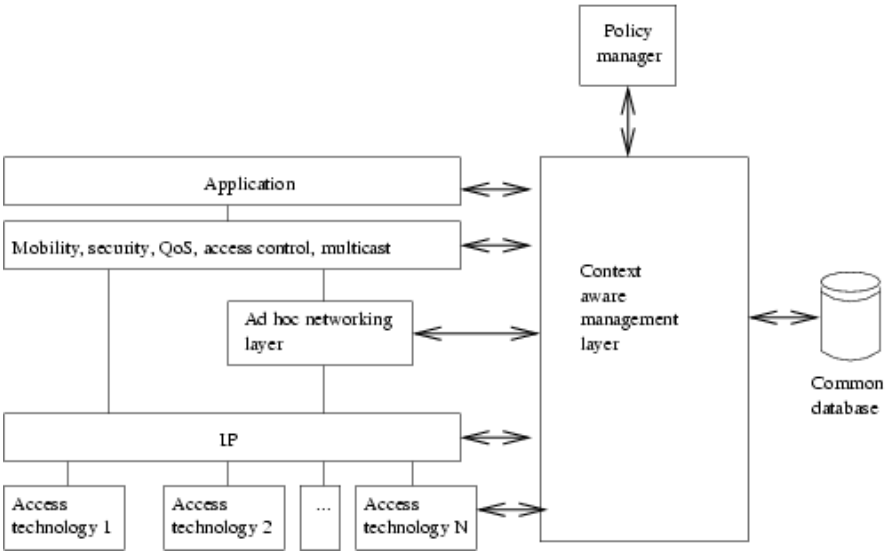

Figure 27 The CAM architecture

The CAM layer: The purpose of CAM is to provide a common layer to all modules that operate in the node to allow the node to behave in a manner optimal to the current environment. CAM offers two interfaces; one to the modules and one to the PM.

Interfaces: CAM contains two standard interfaces: one for communication between CAM and PM, and one for communication between CAM and the modules. The interfaces makes it possible to register and deregister new PMs and modules, set and get values in modules, and to schedule events.

The Policy Manager: The PM is responsible for making the decisions regarding how the behavior of the node should be changed. The PM is aware of all modules that are loaded into the node. The PM also maintains the state information of each module. Thus, it is possible for the PM to make complicated decisions regarding the functionality of the node. For example, if the node enters an ad hoc network, the PM makes the decision regarding which routing protocol to use and with what parameters.

If the node changes to another ad hoc network that uses another routing protocol, the PM may switch off the old protocol and switch on another. Another functionality provided by CAM is event handling. A module may request the PM to send a wake up signal upon the occurrence of a given event. For example, an application may request the PM to signal it once a given QoS level can be offered. This may occur when a network interface (e.g. a WLAN driver) informs the PM of a base station with sufficient signal strength. The security management module of the node may have declared that the given base station is on the list of trusted base stations and access could thus be allowed. The PM then informs the mobility management protocol to make a location update through the given base station. Once the connection is established and the required level of QoS can be offered, the PM informs the application.

The PM makes the decisions based on a set of rules. The rules can be dynamically updated during the lifetime of the node. For example, the network management system in the network may decide that the nodes need to update their policy rules, and issues each node a new set of rules. The nodes in the network then perform the update. The policy can also be changed by the user of the node. The PM of the nodes in the network can also communicate with each other and exchange environment information and even policy rules (under some circumstances).

Common database: The CDB contains all environment related information that is of interest to several modules. Such information may be the level of QoS, security related information, such as cryptographic material and trust levels of other nodes, and so forth.

Access to the CDB is managed through CAM. Each module is given a set of authorities as to which information in the CDB it is allowed access to and for what purposes. Typically, a module will be allowed to read a large variety of entries and to update its own entries. The PM can access and modify all entries in the CDB.

The information in the CDB is always related to a module, although several modules may access and even modify the information (depending on their authority). The module is said to be the owner of the information. Information that is not owned by any module is owned by the PM. The owner of information may change, e.g. if the routing protocol owns certain information and PM switches the protocol to another, the new protocol will inherit the information.

Modules: The modules are the protocols, applications, device drivers, and other pieces of software that communicate with each other and the PM via CAM. Modules are organized in a hierarchical fashion so that e.g. all network modules are organized under the category "access devices" etc. Thus, the PM is able to recognize new modules without modification. If a new category is added, the PM will be updated accordingly.

When a module registers itself to CAM, it may (if CAM requires it) provide CAM with some proof of statement signed by a trusted authority stating that the module is trustworthy, i.e. that it does what it claims to do and nothing else. If the module is not considered trustworthy, it will not be loaded into CAM.

If registration succeeds, the module will be assigned a set of authorities to CAM. The authorities state which information the module is allowed to read from the CDB and which information it is allowed to modify. Typically, the module will be allowed access to the information that is needed for it to perform its tasks as well as the privilege to update information related to itself.

A module may deregister itself from CAM e.g. if the PM decides to deinitialize it. This may be because the user has made the request to change the module, the network management system has ordered a change, or CAM has detected a change in the environment that calls for the change of modules.

6.2.1 Deploying the CAM architecture

The functionality of CAM can be divided into three phases: initialization, communication, and reconfiguration.

INITIALIZATION

Prior to entering the network, the nodes are configured with a certain PM depending on their tasks in the network. For example, static battery-driven nodes are configured with a routing protocol that takes energy-information into consideration, but which may not support mobility too well, whereas mobile nodes are configured with a routing protocol that supports mobility but perhaps makes tradeoffs regarding other requirements. Some nodes may be configured to not participate in routing at all.

COMMUNICATION

When the nodes enter the network, they will be subject to changes in the environment. For example, if a sudden degradation of QoS occurs, the PM may state that the applications are no longer allowed to send images, only text.

The PM is able to dictate the rules that applications and protocols have to follow, and to change protocols and access media. For example, if radio silence is required, the PM dictates that no application is allowed to send any data and that protocols must not perform any protocol signaling. If a node from a highly mobile environment enters a more static environment, it needs to change its routing protocol to be able to communicate. Instead of the routing protocol itself shutting itself down and initializing another, CAM takes care of the switch by changing the protocol to be used. The routing information of the previous protocol can be made available to the new node through the common database.

Nodes in the network may also negotiate terms on which to communicate, for example, with respect to access media and protocols.

RECONFIGURATION

It must be possible to reconfigure the nodes even when they are operational in the environment. For example, the task of a node may need to be changed. Assume that a mobile node which has not previously participated in routing is needed to maintain communications of the network. In this case, the network management system may send the node a set of new rules, that is, update its PM. After this, the node will immediately drop its old tasks and start behaving according to its new rules.

If there is a human user of the node, he may also change the rules of the node.

6.3 Self-healing networks

In a military environment, it is very likely that the network becomes partially destroyed from time to time. This is particularly severe if the core parts of the network infrastructure are affected. During an attack, it is too hectic for people to participate in rebuilding the network. On the other hand, surviving without communications may be difficult and have implications for leadership. Hence, a method for networks to heal themselves is required.

By relying on the CAM architecture presented in the previous section, it is possible to replace destroyed nodes in an ad hoc fashion by assigning new tasks to available nodes[13]. In this way, the network may restructure itself to ensure that communications continues. Naturally, the network will not be able to perform as well as before the attack, however, the main criterion is still to allow the network to transport packets from source to destination(s). This rebuilding process is not limited to the repairing of the routing infrastructure only, but to all necessary services in the network, such as DNS, key servers, and so on.

To ensure the security of the reconfiguration process, PLA is relied on. New rules will not be accepted by illegitimate parties, nor will erroneous updates be considered.

The process of rebuilding the network has the following steps. Note that this is only one possible way of performing self-healing:

1. Detection that the network has been partially destroyed

Typically, network management systems send ping requests to network nodes and are thus able to notice link drops, changes in packet loss ratio, etc. Should the network management system notice “significant” changes (e.g. a server does not respond at all), it moves from a *network steady* state to a *network recovery* state.

2. Scanning for available replacement nodes

In the *network recovery* state, the network management system starts scanning the environment, including the traditional ad hoc networks, for available nodes by broadcasting *node request* packets containing the following information:

- Authentication information proving that the request is valid as it is coming from a legitimate source. Authentication is done using PLA.

- Required services, e.g. routing/AODV.

Nodes in the network that are able to perform the required services respond to the request by sending a *node reply* packet containing the following information:

- Authentication information proving legitimacy relying on PLA.

- Priority level of the node. The priority level states how important the node is in its current task in the network.

- Node capacity (CPU/memory/battery power).

3. Selection of nodes to be used as replacements.

The network management system now knows the available nodes as well as their priorities and capacities. It first scans for suitable replacement nodes among the nodes with the lowest priority. If nodes with sufficient capacities are found, they are selected for the core ad hoc network. If not,

the network management system continues scanning in the next priority class, and so on, until suitable nodes are found.

4. Activation of the core ad hoc network

The network management system sends the selected nodes an *update request* containing a new Policy Manager for the CAM layer of the nodes. This will result in the nodes dropping all their current tasks, updating their PMs, and replying to the network management system with an *update acknowledgment.*

5. Operation of the core ad hoc network.

The selected nodes now operate in their new tasks and the network management system reenters the *network steady* state.

Steps 1-5 are performed during the entire lifetime of the network.

6. Deactivation of the core ad hoc network

Once the network is no longer needed, it will be deactivated. The nodes are released from their current tasks when the network management system sends them a *node release* message. These messages are only sent to nodes that are to be released from the network, and to all nodes if the network is to be shutdown completely. Nodes reply with a *release acknowledgment.*

The protocol described in steps 1-6 is depicted in Figure 28.

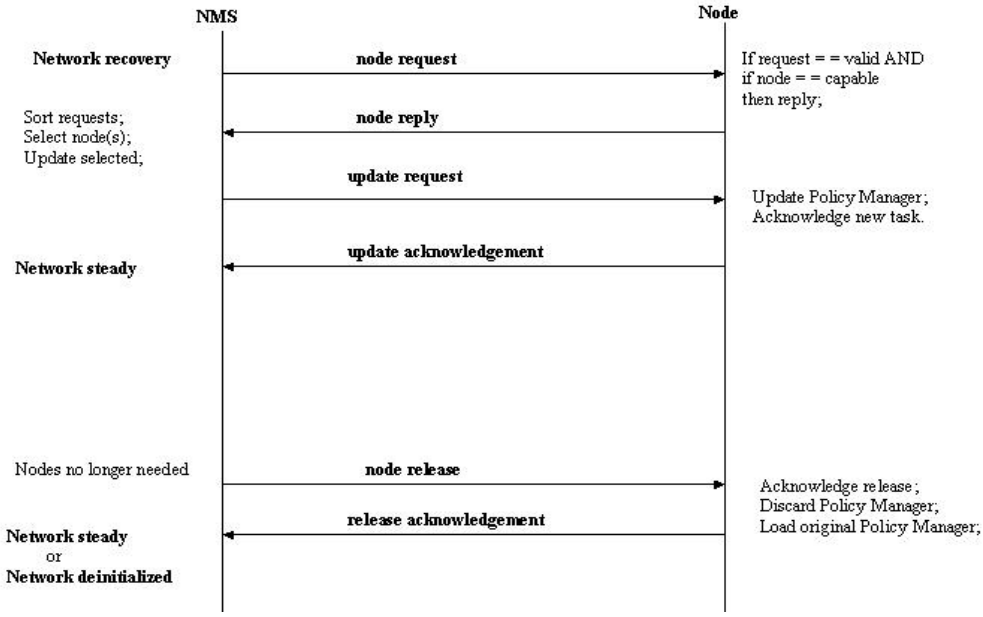

Figure 28 The network recovery protocol

6.3.1 Deploying self-healing networks

In Figure 29, the state graph of the network management system is depicted. The network management system has four main states:

- Network initialization: in this state, the network establishes itself. The nodes are configured to their tasks, security associations are established, etc.

- Network steady: in this state, the network operates normally. If unwanted changes occur, the network management system switches into the network recovery stage.

- Network recovery: in this state, the network management system tries to recover the original functionality by searching for and configuring nodes to replace disrupted/destroyed network elements. Once the network management system considers the network to be sufficiently or completely repaired, it returns to the network steady state.

- Network deinitialize: in this state, the network shuts itself down.

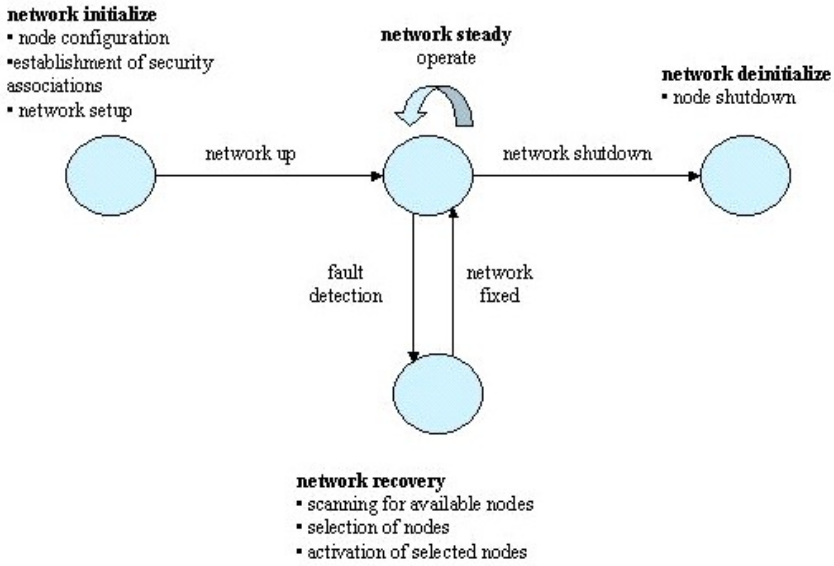

Figure 29 The state graph for the recovery process

In Figure 30, the state graph for the network node is depicted. The node contains the following four states:

- Node init: in this state, the node boots up and configures itself to the network. After that, it either enters an active or an idle state.

- Node active: in this state, the node is active and operates in the network according to its set of rules. If the rules are updated, the node changes its operational behavior accordingly. The node can be deactivated and enter an idle state, or it can be shutdown completely.

- Node idle: in this state, the node is alive but inactive. It may either be activated from this state or shutdown completely.

- Node deinitialize: in this state, the node is shutdown completely.

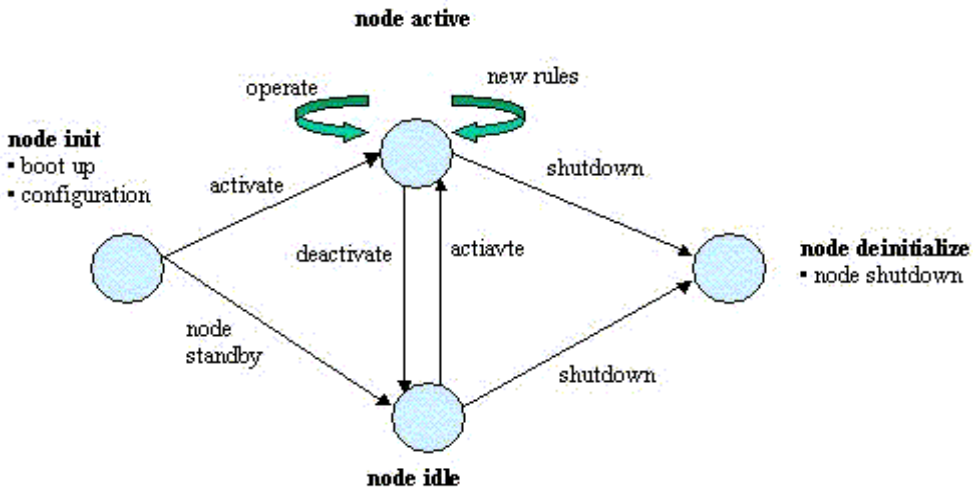

Figure 30 The state graph of a node

6.4 Trust management based on incomplete trust

The concept of trust has been widely studied in the computer security literature [3][28][42][64][80][90][91].

In [64], trust in a principal is defined to be a belief that the principal, when asked to perform an action, will act according to a predefined description. This belief implies that the principal will not attempt to harm the requestor, regardless of how it carries out the request. Trust is always expressed in relation to a principal and to an action.

Trust has the following properties with respect to transitivity, symmetry, and reflexivity:

- Trust is not necessarily transitive, that is, if A trusts B, and B trusts C, A does not necessarily trust C.
- Trust need not be symmetric, that is, A trusts B does not imply that B trusts A.
- Trust is assumed to be reflexive, that is, a node trusts itself completely.

The problem with traditional trust schemes is that they are completely black and white: either the node is trusted or it is not. This model is too restrictive for a military environment, where the nodes may become compromised. A too strict trust scheme would too quickly eliminate nodes from the network, thus resulting in a situation with possibly too many false negatives. This would severely impair the function of the network. However, if the trust scheme is too loose, it may result in compromised nodes being allowed to reside in the network, thus impairing the function of the network. Hence, a more flexible trust model is needed, where nodes that are being suspected are not eliminated right away, but are not relied on for critical functions either. The incomplete trust model presented in [9][10][11] proposes such a scheme.

Incomplete trust is defined as a belief that a principal, when asked to perform an action, will, with probability *p*, act according to a predefined description. Incomplete trust is always expressed in relation to a principal, an action, and the probability that the action will be performed as agreed.

Incomplete trust has the same properties as complete trust, but the transitivity of trust changes. For example, if A trusts that B behaves as agreed with probability 0.8, and B trusts C to behave as agreed with probability 0.5, then A can calculate its trust in C by using the trust levels from itself to B and from B to C.

The different types of beliefs that a principal can have in another principal can be categorized as follows [73]:

- Benevolence: the belief that the principal cares about the welfare of the requestor

- Honesty: the belief that the principal makes agreements in good faith

- Competence: the belief that a principal has the ability to perform a particular task

- Predictability: the belief that the actions of a principal are consistent, and that the requestor thus can predict the behavior of the principal.

The level of trust may change over time due to the behavior of the principal. A principal that has behaved well in the past is assumed to be more trustworthy than a principal that is occasionally misbehaving. Furthermore, trust is likely to decrease faster than it increases since trust in a misbehaving principal will degrade immediately, whereas increasing trust happens gradually over time.

When making decisions based on incomplete trust, the risks involved in the transaction as well as the possible gains of a successful transaction must be evaluated, as the value of trust level as such is non-descriptive. This means that a risk analysis has to be made with respect to each transaction that is made by the principal.

The process of making a transaction is depicted in Figure 31. Although the depicted decision making process is for computers, it clearly resembles the OODA loop for human decision making. The steps are basically the same. Information collection is similar to the observation phase of OODA. Information is gained by own experience and the experience of other nodes. When engaging in a transaction, the level of trust is evaluated and a risk analysis is performed. This trust evaluation phase correlates with the orientation phase in OODA. In the decision phase, the decision is made depending on the outcome of the trust evaluation and the specified security policy. In the transaction phase, the transaction is either carried out or not, depending on the decision. Feedback from the transaction (or the lack of it) is sent to other nodes in the environment.

In [73], incomplete trust was used as the basis for decision making between a seller and a buyer in transactions of electronic commerce. In [11], incomplete trust was used in the signaling of ad hoc network routing protocols as well as routing itself. The basic idea is to evaluate the trust levels of other nodes before accepting route updates or answering route requests as well as to evaluate the trust in each route to the destination node before deciding which route to choose for forwarding packets.

Incomplete trust in a network-centric environment can be used for evaluating the validity of acquired information as well as the trustworthiness in network nodes (whether they have been compromised or not). The former resembles the example for electronic commerce, where trust management is done on the human-application layer, whereas the latter resembles the example for deploying incomplete trust between network nodes. The trust levels can easily be stored in the common database of the CAM architecture, and thus be used by protocols and

applications alike. Thus, incomplete trust management can be done without any changes to protocols and applications.

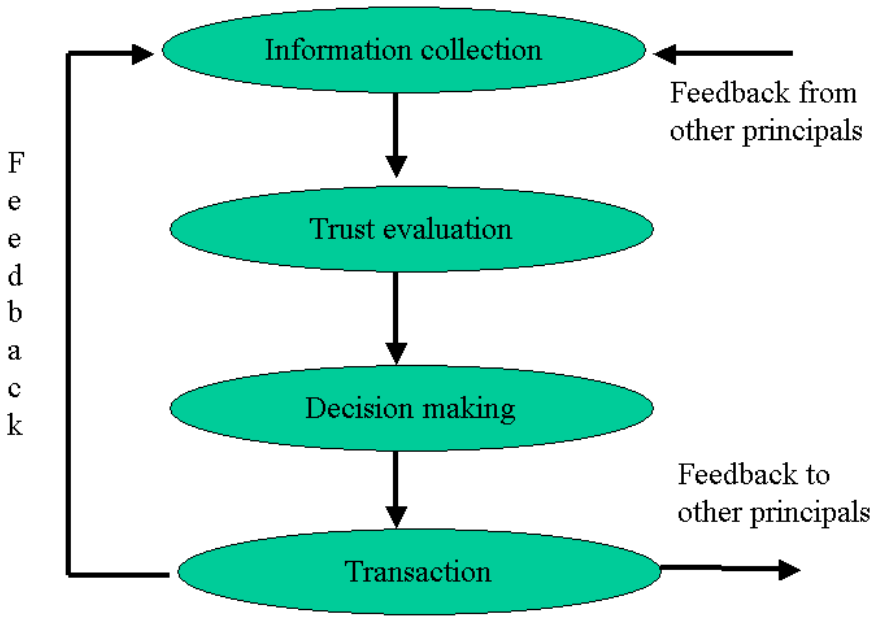

Figure 31 The process of making a transaction

6.4.1 Deploying incomplete trust

The deployment of incomplete trust can be divided into three stages: initialization, communication, and elimination.

INITIALIZATION

Prior to entering the network, the node is initialized, for example, it receives its certificate and configuration. The level of trust is set to 1 (= 100%), as the node at this stage is completely trustworthy. This is based on the assumption that the software and hardware of the node is trusted. Otherwise, the trust level may be set lower already at the initialization stage.

COMMUNICATION

When the node enters the environment, it starts communicating with other nodes and participate in network operations. However, during the lifetime of the network, the behavior of the nodes may change so that they no longer act as they should. This may be because they are malfunctioning, there is a problem with the environment (e.g. the network medium is disrupted), or because they have become compromised. Separating the situations may be difficult. For example, a compromised node may not necessarily misbehave right away, but may behave properly for a long time to have its trust level increased. Only when the trust level is high enough, or when the time is right, will the node start misbehaving.

If a node does not behave, the other nodes in the environment may lower its trust level as a result. The policy of the node states which transactions the node is allowed to engage in depending on the trust level. The trust levels in the policy may be changed depending on the level of paranoia.

A node that behaves well, on the other hand, will have its trust level increased over time.

Nodes in the environment may gossip and exchange their trust tables with each other. However, nodes do not listen to gossip from nodes they do not trust enough, nor do they spread gossip to such nodes.

ELIMINATION

A node whose trust level has decreased is no longer communicated with by the other nodes, and eventually it is eliminated from the network. The network management system issues a revocation of the certificate of the node.
If the network maintains a cache of previous events, it is possible to nullify information with retroactive effects if it is shown that the information is incorrect due to the compromised situation.

6.4.2 Ad hoc network routing based on incomplete trust

In [11], ad hoc network routing based on incomplete trust is described. Traditional routing protocols use hop counts to determine the best route

from source to destination. However, it is also possible to use trust levels as a metrics. In practice, the trust level would not be the only metrics used, but for simplicity, other metrics are excluded from the example.

In Figure 32, an ad hoc network with the corresponding trust levels in the direction of the arrow is depicted. There are two forms of trust related to ad hoc routing:

- Link trust: the trust level denotes the trust that a node has to its neighbor, for example, node A trusts B to the level of 50%, but X only to the level of 5%.
- Route trust: the trust level denotes the trust that a node has in the route to the destination. Route trust is determined by the link trust levels.

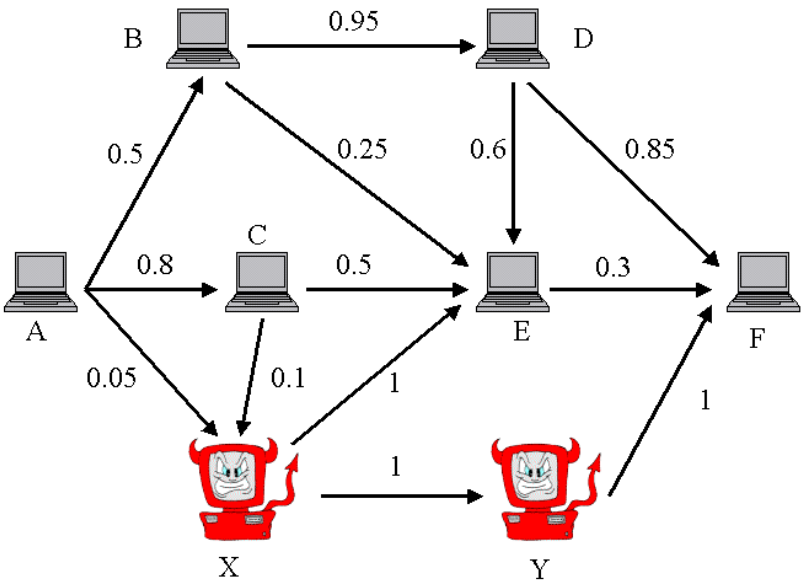

Figure 32 Routing based on incomplete trust

The purpose of signaling is to establish routes while preventing malicious nodes from requesting routes or affecting the routing tables of other nodes. Link trust is used to determine whether a route request or route reply originates from a legitimate node or a possibly compromised one. Error messages need not be handled, since nodes can typically only inform other

nodes of their own inabilities; thus, a malicious node sending an error message would actually tell the other nodes not to route via it anyway. As this would be beneficial for the other nodes, it is highly unlikely that a compromised node would behave in such a way.

The process for making a transaction works as follows for signaling:

1. Collect information: the information can e.g. be collected as described in [56], that is, nodes monitor whether the neighbors actually forward the packets or not, or collected from an intrusion detection system as described in [98].
2. Establishment of trust levels: the trust level is calculated based on the collected information.
3. Decision making: the node needs to make a decision whether to respond to signaling requests from a neighbor or whether to believe in signaling responses received.

The packet forwarding mechanism needs the value of route trust to determine, which route to choose (in the case of multipath routing), if any, or whether to forward the packets at all (in the case of singlepath routing).

The process for packet forwarding works as follows:

1. Collect information: the information can be e.g. packet throughput and overhead of routing transmissions.
2. Establishment of trust levels: the trust level of the route has to be calculated.
3. Decision making: the node decides which route to choose and whether to route or not based on the trust level and the risks and benefits of forwarding the packets.

To illustrate the general idea behind routing based on incomplete trust, three simple examples are used:

1. Average trust: the average trust method calculates the average trust of the links. The obvious benefit from this method is simplicity, however, it is not too expressive.
2. Weakest link: the weakest link method excludes the paths with the weakest links. The benefit is simplicity, and it also forces malicious nodes to behave well for a period of time in order to increase their trust levels.
3. Accumulated trust: the trust levels of neighboring nodes are taken into consideration when evaluating the trust they have in their neighbors. The benefit of the model is that it is proportional, however, it is easy for malicious nodes to fool the node into choosing an bad route.

In Table 9, the three example methods are applied to the network of the example in Figure 32. The choice of route is highlighted.

Route	Average trust	Weakest link	Accumulated trust
A, B, D, F	**0.77**	**0.5**	**0.40**
A, B, D, E, F	0.59	0.3	0.09
A, B, E, F	0.35	0.25	0.04
A, C, E, F	0.53	0.3	0.12
A, X, E, F	0.45	0.05	0.02
A, X, Y, F	0.68	0.05	0.05
A, C, X, E, F	0.55	0.1	0.02
A, C, X, Y, F	0.73	0.1	0.08

Table 9 Determining route trust

6.5 Privacy protection

In [94], privacy is defined as the claim of individuals, groups, and institutions to determine for themselves, when, how, and to what extent information about them is communicated to others. That is, privacy is the ability to control what information is disclosed to other parties and under which circumstances.

Wireless networks are especially vulnerable to traffic analysis, as the network nodes are difficult to hide. By passively monitoring traffic, it is possible to deduce, for example, the intent of the forces, the level of training, the command structure, the location of crucial nodes, and so on. Thus, privacy protection is crucial; since the nodes cannot be hidden, the solutions must present a way of making it difficult or impossible for the enemy to gain any intelligence by traffic monitoring.

Several methods exist to extract information from monitoring traffic. Intelligence can be collected from the data itself, from the communicating identities and their location, from the fact that network nodes exist at a certain location at all, from the time of occurrence of a transactions, and from the type of transaction made. Hence, privacy can be divided into the following categories:

- Data privacy: the data is not disclosed to an unauthorized party
- Identity privacy: the identity of the principal is not disclosed to an unauthorized party

- Location privacy: the location (geographical or topological) is not disclosed to a third party
- Existence privacy: the existence of a principal is not disclosed to a third party
- Time privacy: the exact time of occurrence of a transaction the principal is making is not disclosed to a third party
- Transaction privacy: the type of transaction is not disclosed to an unauthorized party

In the following sections, each privacy category will be handled separately. However, the privacy classes have interdependencies, for example, it is not possible to provide existence privacy unless the data, identity, and location of the node is protected.

The objective of this Section is not to present any new solutions to privacy protection, but rather to identify the various privacy classes and their respective problems.

6.5.1 Data privacy

Data privacy is concerned with protecting the contents of the traffic. Typically, data privacy is ensured by encryption, provided that the cryptographic keys and algorithms used are strong enough. All traffic should be encrypted regardless of its importance, since dividing traffic into secret and public flows would provide the enemy with the advantage of being able to deduce when and where important traffic is sent.

IPSec [44] and the ESP [46] protocol provide confidentiality to IP traffic, and can be used both between gateways and end-to-end. In Figure 33, the various ways IPSec can be used is depicted. When IPSec is used between two gateways, it protects the traffic when it is routed over the insecure network. IPSec also provides Virtual Private Network (VPN) solutions, e.g. when a host is roaming but still needs to use the services of its trusted network. Furthermore, IPSec can be used between two hosts directly to provide end-to-end security.

IPSec provides confidentiality on the network layer, that is, the endpoints are assumed to be network nodes rather than e.g. human users. To protect confidentiality on higher layers, an encryption scheme where the endpoints are human users rather than the network nodes is needed. For example, PGP provides confidentiality to email messages and voice communication.

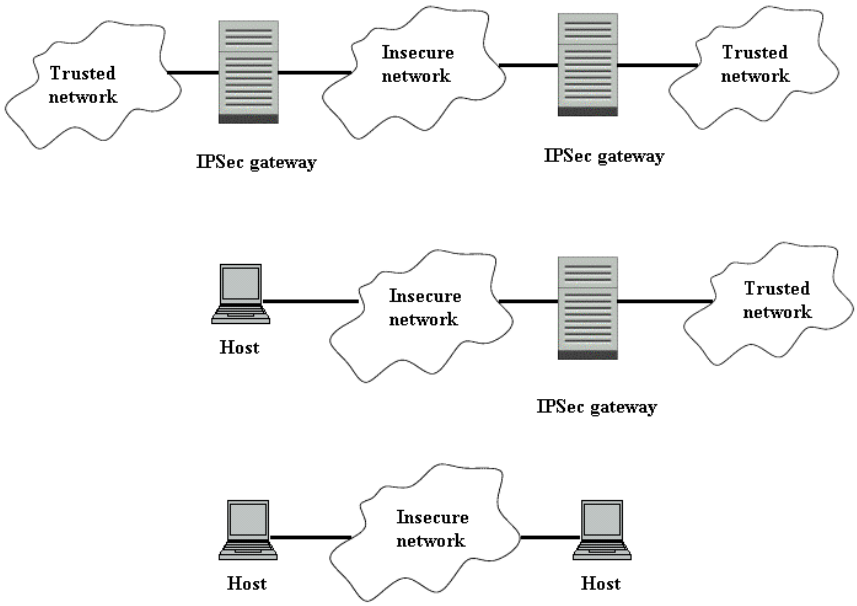

Figure 33 IPSec security

Data protection also supports other privacy classes. For example, by encrypting routing signaling traffic, it becomes harder to deduce the topology of the network.

6.5.2 Identity privacy

Identity privacy is concerned with protecting the identity of the communicating peers. The peer may be either a network node or a human user. Both human users and network nodes may have multiple identities.

A public cryptographic key is typically used to denote the identity of the peer. The problem of identity protection is that the role of the identity is easily deduced by monitoring network traffic, even if the network traffic is encrypted. For example, if certain events always occur after a given identity has transmitted packets, it is fairly simple to deduce that the sender has a commanding position. Furthermore, deducing the relationship

between the nodes is also possible by monitoring which identities communicate with each other and under what circumstances.

Identity protection becomes even more difficult when deploying PLA, since the public key is attached to every IP packet. The only solution would be to rely on temporary cryptographic keys, which are signed by the original key, but that would increase the number of signature verifications that are needed. Hence, the solution is not scalable.

6.5.3 Location privacy

Location privacy is concerned with protecting the location of the peers in the network. This problem is especially difficult in a wireless network or when a node is visible and stationary.

In [29] a solution to protect location information which is based on security agents is proposed. The security agents are routers that receive packets encrypted with their own public key. When the security agent opens the packet, it finds the destination address of the following security agent together with an encrypted packet. This packet is forwarded to the following forwarding security agent until it reaches its destination. Although this scheme is elegant from a theoretical point of view, it has many problems in practice. The scheme does not work well with ad hoc networks where the network nodes move, as the nodes would need to know the exact path of security agents as well as all the required public keys, which can become difficult as the path may no longer exist as such when the packet has left the node. This situation is not similar to normal routing, as the routers on the path are able to adjust to changing routes, but the security agents are predetermined.

Another proposal for protecting the location of the nodes is for the node to have multiple identities and multiple addresses. The node then uses whatever address it has assigned to it, and can even send packets from the same stream using different IP addresses. The corresponding node would be able to deduce from the identity which node it is communicating with, as the nodes have predefined security associations. An approach like this was taken in the Homeless Mobile IPv6 protocol [65], although for non-military reasons. In Homeless Mobile IPv6 the security association is not bound to one address, but to a list of addresses. Thus, the node could use any of its addresses as the source address and any of the destination addresses of the corresponding node. Another partial solution to the

privacy problem is presented in [7], where source addresses were removed from the IP packets altogether. However, the source is revealed in the first packet, and the destination node is revealed in all packets.

In Figure 34, location privacy is ensured by having communicating nodes transmit using low transmission power. The messages are then relayed by low power expendable repeaters to actual routing nodes, which use higher transmission power. Hence, only the routing structure of the network is revealed to the enemy, but the crucial nodes are hidden in the environment as they are difficult to spot. Both the repeaters and the routers are expendable, that is, if a router is destroyed, another router will be used instead.

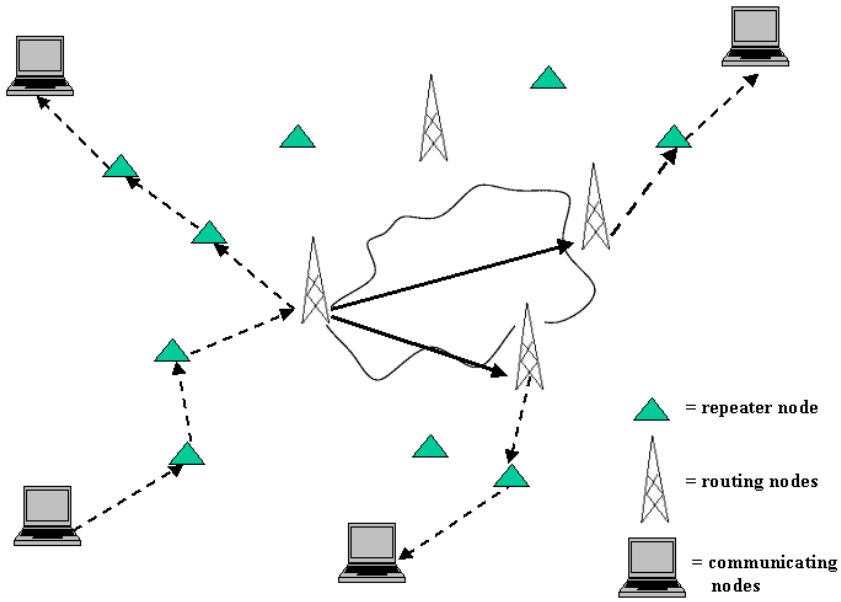

Figure 34 Location privacy in a wireless network

6.5.4 Existence privacy

Existence privacy is concerned with hiding the existence of nodes altogether. In the previous section, location privacy was partially provided by hiding the existence of the communicating peers. However, hiding network nodes is difficult, since they always leak information, such as

electromagnetic radiation and so on. The scope of their visibility may be limited, for example, by decreasing the transmission power and thus reducing their transmission range. In the future, an approach based on data transmission using light might become possible, which could enable better protection of the existence.

It is a common misconception that solutions on the radio level are sufficient to deal with security issues of wireless networks. However, even if the network nodes would rely on frequency hopping to hide their existence, the solution is only temporary. The enemy may either use enough computational power and detect all the channels of the spectrum in parallel, or take advantage of compromised nodes.

6.5.5 Time privacy

Time privacy is concerned with protecting the exact time of occurrence of a transaction from being disclosed to an unauthorized party. The enemy may be able to see that a transaction is taking place based on the existence of network traffic, but it should not be able to tell when the transaction exactly begins or when it ends. This may be of interest e.g. to prevent the enemy from deducing the exact time when a specific command has been issued.

There are no feasible solutions to this problem. Basically, threshold cryptography could be used in such a way that the transaction is completed when k out of n threshold parts have been received, and the receiver is able to open the encrypted information stream. However, this scheme has not been fully developed and analyzed, so nothing can be said about its feasibility. In practice, it seems that by providing data, identity, and location privacy, it is possible to ensure time privacy as well.

6.5.6 Transaction privacy

Transaction privacy is concerned with preventing an unauthorized party from deducing the type of transaction by observing network traffic. For example, the length of the packet may give away the type of transaction that is made, such as a location update in mobility management schemes.

7 Analysis

In this section, the solutions presented will be analyzed with respect to the security requirements of the network. Furthermore, the attack scenarios discussed in Section 5 will be re-evaluated with the assumption that the solutions presented in Section 6 are deployed.

7.1 Requirement analysis

7.1.1 Performing network operations

The first requirement states that the network should be able to perform its tasks, that is, to transport legitimate packets to the right place(s) in time and intact:

The network should be able to perform its tasks, namely transport legitimate packets intact to the right place(s) in time. This ability should not be affected even if the network is partially destroyed or otherwise attacked.

The PLA architecture presented in Section 6.1. provides the following services:

- Authentication: it is possible to verify that the sender of the packet is legitimate
- Integrity protection: the packet has not been modified by an illegitimate party
- Non-repudiation: a sender cannot deny having sent the packet
- Timeliness: the timeliness of packets can be verified and delayed packets dropped
- Detection of duplicates: it is possible to detect duplicated packets

The basic idea is that any node on the path from source to destination can verify the legitimacy of the packet.

Basically the requirement can be split into pieces as follows:

- The packets are legitimate
- The packets are intact

- The packets are routed to the right destination(s)
- The packets are not delayed
- All of the above should not be affected even if the network is partially destroyed or otherwise attacked

The legitimacy of the packets is determined by verifying that the public key of the node has been signed by a trusted third party. Furthermore, all of the other fields in the packet must be correct. PLA is able to ensure this except for some of the fields in the IP packet that change as the packet traverses the network, such as the hop limit.

The integrity of the packets can be verified by checking that the packet has been signed with the private key corresponding to the public key of the sending node. As only the sending node can produce the signature, packet integrity can be ensured.
To ensure that the packets are routed to the right destination(s), the routing infrastructure must be sound. This implies that only legitimate nodes should be able to participate in routing and routing protocol signaling. When PLA ensures the legitimacy of the packets, it basically verifies that the sender is legitimate; the packets of illegitimate senders are dropped. Hence, it can be assumed that no illegitimate nodes participate in signaling.

To ensure that the packets arrive in time, the routing infrastructure must be sound and the network must not be under denial-of-service attacks that constrain network resources and thus delay packets. The soundness criterion stated that only legitimate nodes may participate in routing and routing protocol signaling; PLA ensures that this is the case. To limit the effect of denial-of-service attacks, PLA is able to detect packet duplicates, erroneous packets, and delayed packets. Such packets are dropped as soon as they are detected, that is, at the first hop. Hence, the denial-of-service attack will not be able to spread in the network even if the packets were signed by a legitimate node. This ensures that the packets are able to arrive at the destination(s) in time, at least as far as security is concerned.

The solutions mentioned above work under the assumption that no nodes are compromised. However, if the network nodes become compromised, they have access to the private keys and can thus bypass PLA. In that case, the compromised nodes are able to pass as legitimate, spread disinformation, distort routing signaling, perform denial-of-service attacks, and so on. To cope with this problem, the incomplete trust model can be deployed. If a node is misbehaving, it's peers will lower its trust level and

gradually refuse to communicate with that node. The network management system will be notified by the peers that they find the node suspicious. Once the network management system decides that the node is compromised, it will revoke the certificate of that node. Since PLA uses public key cryptography, the task of removing compromised nodes becomes easy, as it is not necessary to renew all the keys and shared secrets of other nodes. The only thing that needs to be done is to inform the other nodes of which nodes have had their certificates revoked. This can easily be done e.g. using a wide area broadcast technology, such as satellite communication or WWAN/WMAN solutions.

The incomplete trust model, however, comes with some problems. Detecting misbehaving nodes is not a trivial task, especially since the compromised nodes may behave well in order to increase their trust levels. It is also difficult to distinguish between misbehavior and malfunction. Furthermore, compromised nodes may collaborate to make legitimate nodes look suspicious.

7.1.2 Coping with network destruction

The network should be able to perform its task even if it becomes partially destroyed. This requirement binds the first and the second requirement together, as the second requirement addresses a situation where the network becomes partially destroyed because of a large-scale or dedicated attack:

Should the network become partially destroyed, a means of rapidly recovering is required to ensure continuous functionality of the network. This requirement supports the first requirement (that the network should be able to perform its tasks).

The requirement states that a means of rapidly recovering from a partial destruction is needed. This requirement can be divided into the following parts:

- Rapidly means that the networks heals itself faster than it would take for people to rebuild the network
- The self-healing process is fully or partially automatic
- The self-healing process is secure; the enemy should not be able to take advantage of the process to inject it's own nodes into the network

In Section 6.3, a mechanism for self-healing was presented. For the mechanism to work, the following requirements must be met:

- Network nodes are able to accept new tasks
- The network nodes must be able to verify that the request for changing their tasks is legitimate
- The network management system must be able to verify that the network nodes that reply to the requests are legitimate

To cope with the requirement of changing the tasks on the fly, the Context Aware Management architecture presented in Section 6.2. can be deployed. It was originally developed to allow nodes to rapidly adapt to a changing environment by having a policy state how the node should behave in each circumstance. The policies were coordinated by a policy manager.

To change the tasks of a node, the policy manager of the node can be changed to a new one. Thus, the node will act according to a set of new rules. For example, as depicted in Figure 35, a node that previously collected weather data from the environment may need to drop that task and initialize a routing protocol in order to support the core network by remerging partitioned network parts. The Figure shows the partially destroyed network that was presented in Section 5.2.2. In the example, the routers that resided in the middle of the network were destroyed by a large scale attack. Hence, two network partitions with no connection to each other emerged. However, some non-routing nodes in the area managed to survive, such as the weather data collecting node. As it's policy manager is updated by the network management system, it drops its previous task and starts routing packets between the two partitioned networks.

However, before accepting the new tasks, the node must be able to verify that the request came from a legitimate source. The network management system, in turn, must be able to verify that any node that signs up for a new task.

To ensure the legitimacy of requests and responses, PLA is relied on. The network management system can ensure that the nodes which are available and sign up indeed are a part of the network, and the network nodes can verify that the request came from their own network management system.

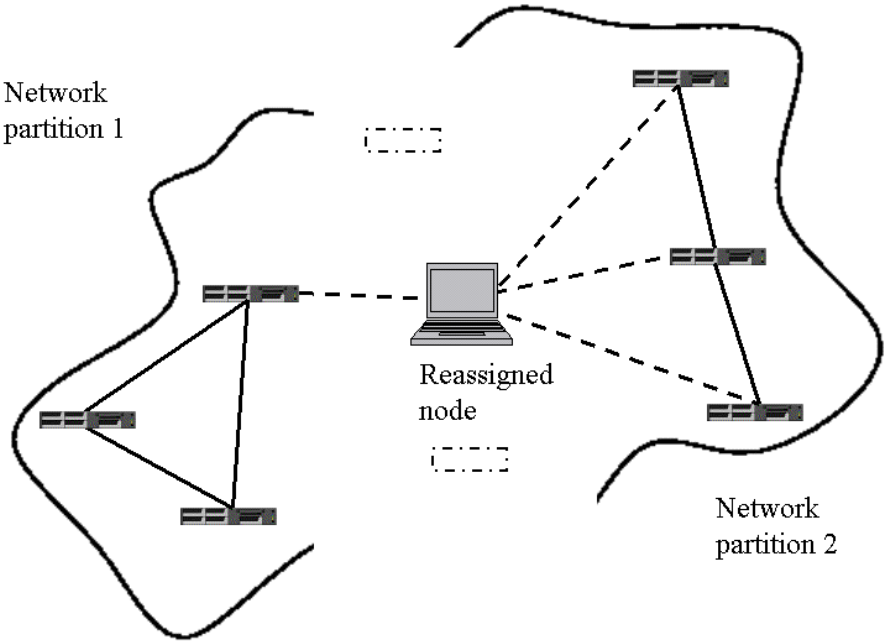

Figure 35 A node is assigned the task of connecting two partitions of a network

The proposed mechanism for self-healing meets the requirements of being rapid, fully or partially automated, and secure.

The network management system typically notices failures in the network quite rapidly. Upon detection of malfunction, the network is scanned for available nodes and requests are sent to sign up for new tasks. Once the requests have been handled, the new tasks are sent out and accepted. The network is once again operational. The time required to perform this task is dependent on the network delays.

The self-healing process is also fully or partially automated. The network management system is itself able to initialize the self-healing process and reorganize the network. However, it is also possible for a human to interfere with the restructuring process.

Security is ensured by relying on PLA, as previously stated.

Apart from the self-healing mechanism presented in Section 6.3., other mechanisms are used in parallel to ensure the continuity of the communications and services. First of all, services may be duplicated to ensure that no single point of failures exist. This applies to the access medium as well; if one access medium is lost, other alternatives can be tried to ensure connectivity. Second, some nodes may have dormant features, which are initialized by the policy manager of the node. Third, some protocols, such as ad hoc network routing protocols, have support for dynamic behavior built in; if a route is lost, the node attempts to find new routes.

7.1.3 Managing trust

The perhaps biggest security problem in military networks is that of compromised nodes, as they bypass traditional security solutions, are difficult if not impossible to detect, and have the capability of damaging the network and affect the decision making process.

Traditional trust management schemes have the disadvantage of being too strict and inflexible. Either a node is trusted or it is not trusted. This introduces a risk where compromised nodes are completely trusted for everything at all times, or that fully legitimate nodes are excluded too easily.

The third requirement states the importance of trust management:

Trust management is important to ensure that the information retrieved from the network is correct and the information that goes to the network will be properly handled. This implies that compromised nodes should be detected and removed. If the network cannot be trusted, it becomes useless.

Trust management needs to be handled on two levels:

- Human users evaluate the trust they have in information they receive from the network. The information may have been sent by another human or it may have been created by the network, for example, through sensor systems and data fusion into a real-time situation awareness image.
- Network nodes evaluate the trust they have in other nodes when engaging in signaling with them.

The trust model suggested in this thesis relies on the concept of incomplete trust. The purpose of incomplete trust is to introduce some flexibility into the decision making process of the nodes. As the military environment is uncertain, normal trust schemes would not work. However, by allowing nodes to change the level of trust they have in other nodes during the lifetime of the network, it is possible to better deal with the uncertainty.

The requirement stated the following:

- The information retrieved from the network is correct. This implies both information addressed for the human user as well as the signaling data addressed for network nodes
- The information that goes to the network is properly handled. This implies both information sent by a human user as well as signaling data addressed for network nodes
- Compromised nodes should be detected
- Compromised nodes should be removed

Traditional security solution cannot ensure the correctness of the information, they can only ensure that the information received is the same as the information sent (i.e. provide integrity protection). The incomplete trust model allows a user or a node to determine what actions to take depending on the level of trust in the information. That is, it is up to the user or node to determine whether the information is correct enough to be trusted. This trust evaluation is made based on previous experience, endorsements from the environment, and so on. For example, if a general has always received accurate information from a given source, he typically trusts that information also in the future, unless several of his trusted friends have told him that they have got bad information lately. The same idea applies to network nodes. A node that has successfully engaged in digital transactions, such as protocol signaling, with another node, continues to trust that node for that particular task unless several of its other trusted nodes start issuing warnings or if the network management system forbids engagement with that node.

If a node is compromised and starts spreading disinformation, it will be noticed by the other network nodes or by the human users. Thanks to PLA, the node cannot deny having created the disinformation. This will lower the trust level of the compromised node and eventually get the network management system alerted to have the certificate of the node revoked. Hence, nodes spreading disinformation will eventually be eliminated from the network, unless they are somehow able to manage their trust levels by

behaving well for longer periods of time. This, however, would still minimize the damage a compromised node can do.

The same applies for information that goes into the network. If a node modifies, delays, or destroys information, the trust level in that node will decrease. The same motivation as for retrieving information can be applied: a compromised node that does not want to be detected must behave well for long periods of time, thus taking away the effect of the attack. However, if the node misbehaves, it will be noticed and removed from the network.

Detection of misbehavior is not a trivial task. Some proposals exist where nodes are monitored for whether they actually forward packets they are supposed to forward. However, if a node clearly misbehaves by duplicating packets, delaying packets, modifying packets, and so on, it will easily be detected thanks to PLA. Furthermore, a node that creates disinformation cannot deny having done so. The problems with detecting malicious nodes is, however, that they may behave well over a long period of time, and that malicious behavior is not easy to distinguish from malfunction. However, it can be argued that malfunctioning nodes should be eliminated from the network as well.

Removing compromised nodes is fairly easy thanks to PLA. Since each node is identified by a public key, it is sufficient to revoke the public key and identify other nodes in the environment. This can be done using any wide area access technology.

In Figure 36, an example of revoking a certificate is depicted. Nodes A, B, and C have all experienced bad service from node D. Hence, they inform the network management system of their experience with D. As so many nodes have had similar experiences with D, the network management system starts monitoring D. As D continues to misbehave, the network management system draws the conclusion that D is compromised, and revokes its certificate by adding D on a revocation list. The network management system then broadcasts the revocation list using satellite communications to reach the whole network environment. All nodes in the network environment will now be aware that node D is malicious, and refuse to communicate with it.

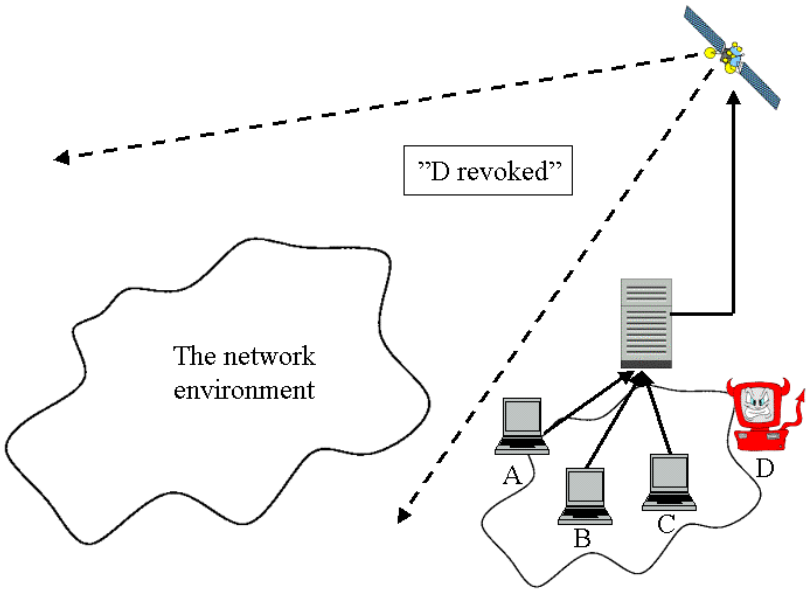

Figure 36 Revoking the certificate of a compromised node

The main problem with the incomplete trust model at its current stage is that it has not been fully developed yet. More work needs to be done on the mathematical and logical background on how to manage fluctuations of trust and perform decision making. For example, a mathematical formula to be used to determine the level of trust that a node has in a route to another node should be developed in order to make incomplete trust deployable for ad hoc network routing.

Another issue is that malicious nodes may cause the trust levels to fluctuate, and may also claim that legitimate nodes are compromised even if they are not. However, these problems exist in models based on complete trust as well, and are actually even more severe in those cases. For example, the trust levels would never fluctuate; a legitimate node would be declared illegitimate right away.

7.1.4 Privacy protection

The fourth requirement deals with privacy, which is another difficult problem in wireless military networks. The privacy requirement is as follows:

When performing its tasks, the network should not give away any information to unauthorized parties. In an utopist world, the enemy is not even aware of the network; in an ideal world, the enemy can see the network but cannot deduce anything from it. In reality, enough privacy should be guaranteed so as not to reveal the most important information, namely the command chains, crucial network nodes, and intent.

The network should be able to protect the following information:

- The content of the data
- The identity of the communicating parties
- The location of the communicating parties
- The existence of a node
- The time of occurrence of a transaction
- The type of transaction

The content of the data is trivial to protect with strong encryption.

However, protecting the identity of the communicating nodes is difficult, especially since PLA is used. PLA unfortunately attaches the public key of the node into every IP packet, which reveals the identity of the sender. Temporary public keys could be used, but would increase the signatures to be verified. This, in turn, would degrade the performance and consume energy resources.

Location privacy can be ensured to some extent by having crucial nodes send with low transmission power and deploying expendable nodes are repeaters and routers. However, nodes always leak information in form of electromagnetic radiation and so on. Hence, location privacy solutions, in particular if implemented in the protocols, only give a very limited level of protection. The same applies for existence privacy.

Protecting the exact time of occurrence of a transaction is possible if data, identity, and location privacy can be ensured. In that case it is possible to deduce when a transaction occurs, as it is not possible to deduce who is talking what with whom and from where. However, as identity and location privacy cannot be sufficiently ensured, the problem of protecting the time of occurrence of a transaction remains an open issue.

The type of transaction can be deduced from the length of the packet even if data privacy is ensured. Transaction privacy can be provided if it is

possible to vary the packet lengths by adding garble data into the packets. The garble data would be discarded at the destination.

The privacy problems unfortunately remain unsolved.

7.2 Scenario analysis

The four scenarios presented in Section 5 will be re-evaluated with the assumption that the solutions presented in this thesis are deployed.

7.2.1 Attacking the infrastructure

In Section 5.2.1., a scenario where a masqueraded node is able to attack the infrastructure was presented. As a result of the masquerade, the node is able to perform the following attacks:

- Denial-of-service attacks: the node is able to consume network resources by flooding the network
- Disrupting protocol signaling, for example, the node is able to take part in routing signaling and thus change the network topology
- Disrupting network traffic: the node is able to delay or drop packets, which may affect upper layer protocols
- Traffic analysis: the node is able to spy on the network and collect intelligence
- Spreading disinformation: the node is able to inject wrong data into the network

Since all IP packets are authenticated using PLA, masquerading attacks can be excluded altogether (unless cryptography is broken or the enemy node manages to get its public key signed by a trusted third party, which is highly unlikely). Hence, the attacks mentioned above will not succeed unless a legitimate node is compromised.

Traffic analysis, however, can still be performed even if the node is not part of the network, as long as it is able to pick up network traffic. As the traffic is encrypted, it can be assumed that the content is not revealed to the enemy. However, the identity of the communicating nodes and their location, as well as the types of transactions made may be revealed.

7.2.2 Destroying the infrastructure

In Section 5.2.2., a scenario where the infrastructure was partially destroyed as a result of a large scale attack or a dedicated attack was presented. As a result of the attack, the routing infrastructure was impaired and services removed from the network.

There are several ways of reorganizing a partially destroyed network:

- The self-healing system presented in Section 6.3 is able to assign new rules to nodes so that they start performing new tasks.
- By duplicating services it is possible to eliminate single point of failures. All services need not be operational simultaneously; some may be dormant and started up only when needed.
- Nodes in the environment may vote on which one of them will take on a task previously held by another node. For example, if a cluster head is lost, the nodes in the cluster may vote for a new cluster head.

The basic idea in this thesis is to primarily rely on the self-healing mechanism presented in Section 6.3. together with duplicated services, where some services may be dormant. Only if a network cluster is separated from the network management system altogether should it rely on voting schemes. However, protocols that by default support dynamic behavior, such as ad hoc network routing protocols, should naturally restructure as they normally would when changes in routes occur.

In Figure 37, the recovering from the dedicated attack presented in Section 5.2.2 is depicted. The network management system is able to reach the network that lost its access router. By relying on the self-healing mechanism, it presents one of the nodes in the network with new rules. The selected node now starts to function as the access router, and the network can be connected to the backbone again.

The main difference is that the new access router may have limited performance and lifetime, however, it is still better to have some connectivity than none at all. When the battery of the node dies, another node is assigned the role of the access router. This goes on until the original access router can be replaced by a new one which has originally been designed for that particular task.

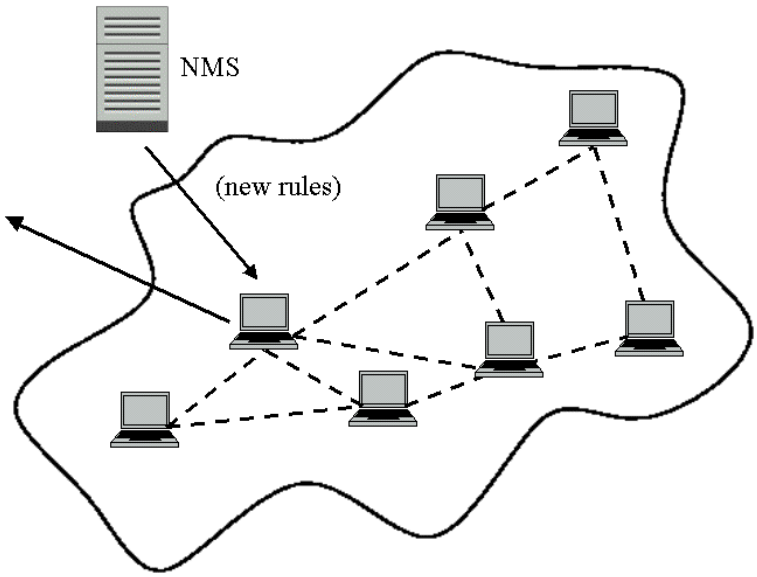

Figure 37 Choosing a new access router

7.2.3 Compromised nodes

In Section 5.2.3., a scenario where some of the nodes were compromised was presented.

In the scenario, the compromised node was able to perform:

- Denial-of-service attacks: the node is able to consume network resources by flooding the network
- Disrupting protocol signaling, for example, the node is able to take part in routing signaling and thus change the network topology
- Disrupting network traffic: the node is able to delay or drop packets, which may affect upper layer protocols
- Traffic analysis: the node is able to spy on the network and collect intelligence
- Spreading disinformation: the node is able to inject wrong data into the network

- Compromise other nodes by lying to them and thus change network behavior
- Break the security of radio network level schemes, such as frequency hopping
- Access information and services

Denial-of-service attacks are detected already at the first hop from the flooder. By relying on PLA, the certificate of the flooding node can be revoked. Thus, in this case, the compromised node is detected and removed.

As far as the other attacks are concerned, they are more difficult to detect in the first place. A compromised node may, for example, advertise cheap routes and thus attract all data through itself. This behavior would only get attention if a sender experiences bad service and decides to try another, more expensive route, which in reality offers better service. However, if the compromised node disrupts the network traffic by only randomly delaying or dropping packets in such a way that the service experienced is not noticeably bad, but the overall performance of the network gradually becomes bad, the attack is not necessarily easy to detect.

Traffic analysis is an even bigger problem when the node is compromised, as the node now has access to all the routing tables. Data privacy can be ensured as long as the compromised node is not one of the recipients. However, all other privacy classes are violated. The inside information that the node has, for example, the hop frequency on the radio level, can be used to attack the network further. For example, it is no use for a compromised node to flood the network as it would get caught right away, but it could instead reveal the hop frequency to the "home base" to allow the enemy to jam the network from outside.

If the compromised node behaves well, it will go undetected for longer periods of time. Hence, it may access services and information and thus collect intelligence. Small changes are not likely to be noticed, however, over time the information will be completely distorted, thus affecting the decision making process.

The problem gets even worse if we look at the scenario in Section 5.2.3, where two compromised nodes were strategically placed in the network, see Figure 38. In the example, the compromised nodes are able to partition the legitimate nodes into two separate parts.

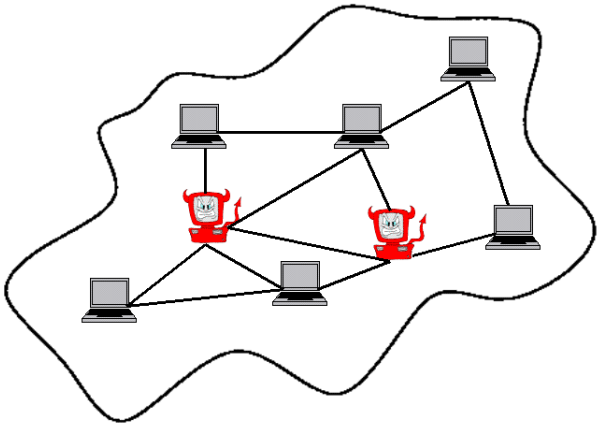

Figure 38 Strategic placement of malicious nodes

To take over the whole network, the two network nodes that are separated from the four other legitimate nodes are now fed with lies. The first node to convert into evil is the one that gets bad data fed from both compromised nodes. After that, the sole legitimate node gets bad data from the compromised and the converted nodes. Eventually, it will convert as well. The difference between a compromised and a converted node is that the compromised node is controlled by the enemy whereas a converted node in theory is not, but it's behavior is affected by the compromised nodes.

Now the situation is basically four "evil" nodes against four "good" nodes. To take over the rest of the network, it is sufficient to get one node to convert. After that, the compromised nodes have majority, and the rest of the network will gradually convert and behave accordingly.

Another possibility in this scenario is for the compromised nodes to convince the legitimate nodes in each respective partition that the nodes in the other partition are compromised. This would result in a permanent partition of the network nodes. The compromised nodes may even move away from the area and enter new network environments and continue posing their lies on other nodes, thus creating network partitions all over the network environment. The problem with this approach, however, is that it is likely to be detected. The network management system pays

attention to nodes who are claimed to be compromised. Eventually, a pattern is likely to emerge from the events: partitions occur wherever certain nodes have been.

To understand the contribution of incomplete trust to the problems discussed above, it is compared to trust models based on complete trust.

Both models face the same problem: detecting anomalies. However, it is more difficult for a compromised node to operate in an environment based on incomplete trust. First of all, the trust levels fluctuate. Hence, the compromised node must constantly monitor its own behavior, otherwise it runs the risk of losing the trust other nodes have in it. The smaller the trust level, the less of a chance it has to affect other nodes. In a complete trust model, small misbehaviors would not be noticed and the node would always be fully trusted. Hence, the damage a completely trusted node may cause is larger.

Second, converted nodes differ from compromised nodes in that they still monitor the environment for changes. They may be affected by other good nodes in the environment and change their opinion again. This means that the compromised nodes must constantly keep the converted nodes in control in order to take over the network. In a complete trust model, the compromised nodes could have cut off the converted nodes from other legitimate nodes in the network by claiming that they are compromised. In the incomplete trust model, the converted nodes only gradually shift their opinion in this direction, thus causing a delay in the overtaking process. Furthermore, the opinion may shift back and forth since the nodes still communicate with the other legitimate nodes.

Third, converted nodes may show signs of anomalies although they are not compromised. In a complete trust model, the nodes would be eliminated from the network as compromised. This would enable the compromised nodes to eliminate other nodes until they get the majority in the network. However, with incomplete trust, the trust in the converted nodes only decrease, but they are not eliminated. If they come in contact with good nodes again, their behavior will change into normal. Hence, the risk of false positives is smaller in the incomplete trust model.

As a conclusion, incomplete trust provides a more flexible way of coping with compromised nodes than models based on complete trust. If a compromised node is detected, it can be eliminated from the network thanks to the fact that PLA is used to secure the network and it relies on

public key cryptography to identify nodes. Hence, to eliminate a compromised node, it's certificate is revoked.

7.2.4 Network surveillance

As has been previously discussed, privacy protection in wireless military networks is difficult. Some level of privacy can be provided, such as data privacy, and in some cases even location privacy to some extent. This, however, assumes that there are no compromised nodes in the network.

The attack scenario in Section 5.2.4 is thus possible and even highly likely.

8 Conclusion

The development of warfare has gone hand in hand with the development of the society. With the proliferation of information and communication technology, it has become possible to collect, process, analyze, and distribute information efficiently. Whereas the modern society is often referred to as the "Information Society", the military has developed corresponding concepts referred to as "Information Warfare" and "Network-Centric Warfare". The information warfare concept is mainly an idea where information is used both as a target and a weapon. By targeting information, or using it as a weapon, it is possible to affect e.g. political and military decision making. The information operations need not necessarily be targeted directly at the decision makers, but may address the public opinion either in favor or against the decision makers. In network-centric warfare, the idea is taken even further. Networks are used as the enabling factor of conducting information warfare. The restrictions imposed by location and time are almost eliminated altogether. The actual information battle is a fight over information superiority, where the purpose is to achieve accurate real-time situation awareness while denying the enemy from doing the same. It is assumed that the better the situation is known and understood, the better decisions can be made, and ultimately the party that has made more correct decisions will win the war.

The decision making process can be modeled using the OODA loop developed by Col. John Boyd from the US Air Force. OODA stands for Observe, Orient, Decide, and Act. In the observation phase, data is collected from the environment and fed into the orientation phase. The orientation phase is dependent on personal characteristics, such as genetic heritage, previous experience, and so on. In this phase, a mental image of the situation is created based on the observations and a list of alternative decisions is formed. In the decision phase, one of the alternatives is chosen, and the decision is carried out in the action phase.

The strategy is to go through one's own OODA loop as efficiently as possible while denying the enemy from doing the same. Typically, the OODA loop can be affected in several ways. Disinformation may be spread into the observation phase to affect the mental image to be created. The orientation phase can be targeted by changing the way people form their mental images. For example, the mental image is dependent on cultural values and previous experience. Typically, propaganda and

psychological operations have been used to affect the way people perceive things. This is taken advantage of also in marketing, for example, the whole "get them while they are young" thinking relies on the fact that when people have positive experiences from something, they build a pattern of e.g. always choosing the same product. Newspapers give inexpensive subscriptions to students so that they will continue subscribing to the paper when they have graduated. Also the decision phase can be targeted by psychological means, but also the environment and the situation affects the decision. For example, a person who is afraid is likely to make different decisions than he otherwise would. Fear has been used throughout history to affect public opinion in a desired direction, for example, to support warfare, affect elections, accept a change in the legislation that may not be beneficial for the individual, and so on. The action phase, in turn, can not only be targeted directly, it can also be observed (these observations are fed into the OODA loop of the other party) in order to gain intelligence of the opponent.

In a network-centric environment, where the network functions as a tool for leadership, the decision making process can be directly targeted in its observation phase and its action phase.

In the observation phase, where information is collected from the network, there are various ways of affecting what is fed into the orientation phase. First of all, disinformation may be injected into the network, or existing information may be modified This will create an erroneous situational awareness image in the orientation phase. Second, information can be delayed in various ways. Hence, the situational awareness image may no longer be timely. Third, information can be destroyed altogether so that it is not possible to construct a clear image of the situation. Furthermore, information can be duplicated or flooded so that more information is available than can be processed. All of the possibilities above may cause confusion, and thus delay the decision making process or lead to wrong decisions being made.

In the action phase, the outcome of the decision may be communicated to the network-centric environment. For example, a command may be sent from a decision maker to a soldier or weapon system, or the topology of the network will change since nodes move as a result of the action being carried out. In the former case, the command can be attacked in several ways. It is possible to destroy it so that it never reaches its destination, delay it so that it reaches the destination too late, modified so that the command is changed into something else, etc. In the latter case, it is possible to merely keep the network under surveillance to collect

intelligence, and later deduce e.g. that when a certain type of message is sent from a certain location to another given location, the forces are about to perform an attack.

Hence, in order for the network to be useful as a supportive tool for leadership, it must be secured from the attacks mentioned above.

In this thesis, the possible attacks were discussed using attack scenarios. The scenarios addressed attacks on the network itself, partial destruction of the network either by a large-scale or dedicated attack, internal attacks performed by compromised nodes, and the possibilities of gathering intelligence by network surveillance. Based on the scenarios, the following security requirements were identified:

1. The network should be able to perform its tasks, namely transport legitimate packets intact to the right place(s) in time. This ability should not be affected even if the network is partially destroyed or otherwise attacked.

2. Should the network become partially destroyed, a means of rapidly recovering is required to ensure continuous functionality of the network. This requirement supports the first requirement (that the network should be able to perform its tasks).

3. Trust management is important to ensure that the information retrieved from the network is correct and the information that goes to the network will be properly handled. This implies that compromised nodes should be detected and removed. If the network cannot be trusted, it becomes useless.

4. When performing its tasks, the network should not give away any information to unauthorized parties. In an utopist world, the enemy is not even aware of the network; in an ideal world, the enemy can see the network but cannot deduce anything from it. In reality, enough privacy should be guaranteed so as not to reveal the most important information, namely the command chains, crucial network nodes, and intent.

The requirements were then addressed by a set of security solutions. The assumption made when developing the solutions was that the networks are under constant attack by the enemy. This assumption comes from the fact

that the military environment is hostile, and the enemy is highly likely to try to attack the networks in various ways.

The military environment is difficult in many respects. The networks are dynamic in nature and most of the communications medium is wireless. Furthermore, the networks are subject to weather conditions, terrain, hostile activities, and so on. Traditional security solutions cannot be applied, since they may not solve the problems of the military environment or they may not scale well. For example, the network cannot be used with IPSec, because it is and end-to-end security protocol (in its basic form) and the alternatives where IPSec can be used in a node-to-node scenario are not secure for an environment where nodes may be compromised. Furthermore, since the size of the networks tend to be large, it would be infeasible for a network node to store the security associations of all the other nodes. Hence, key agreements would have to be made when needed. The IKE protocol is too heavy for this task, as it requires too many protocol rounds. When the nodes are highly mobile, they may not even have time to finish the protocol.

Another problem in military networks is trust management. Traditional trust management models are based on complete trust, i.e. either a node is trusted or it is not. Such models are, however, too inflexible to be used when some of the nodes are compromised and perform intelligent attacks on the network. The problem with compromised nodes, on the other hand, is that they are difficult to detect. Until detected, they are able to perform the same things as the legitimate nodes, as the protection provided e.g. by cryptography is bypassed.

In this thesis, a set of security solutions and ideas have been developed to address the requirements.

The Packet Level Authentication (PLA) architecture is proposed as a solution for securing the network. The basic idea behind the architecture is that any node in the network is able to verify the legitimacy of any IP packet. This is realized by adding the public key of the sender into every IP packet and signing the packet using the corresponding private key. The public key is also signed by a trusted third party. The PLA header also contains other important information, such as the creation time of the packet, the sequence number of the packet, and the validity period of the packet. Hence, it is possible to detect illegitimate packets, duplicates, delayed packets, and so forth. Furthermore, a node cannot deny having sent a packet. These features make it possible to limit the scope of denial-of-service attacks, remove compromised nodes from the network (by

revoking their certificates), drop nodes that are illegitimate or delayed, as well as to provide access control, firewall services, and so on.

To support rapid restructuring of a partially destroyed network, the Context Aware Management (CAM) architecture together with a self-healing mechanism is deployed. The CAM architecture was originally designed to allow nodes to rapidly adapt to the changing environment without introducing complexity into application and protocol design. For example, a node using a certain routing protocol may need to switch into another protocol if it enters an environment where a different protocol is used. To have the routing protocols themselves coordinate these actions would be difficult. However, by deploying CAM, it is possible to shut down the routing protocol and initialize another. This is enabled by adding a CAM layer to the IP stack. The CAM layer is able to communicate with all other layers of the stack. A common database is used to store environmental data. To adjust the behavior of the node, the Policy Manager contains a set of rules that state how a node should behave in any given situation.

This feature is utilized by the self-healing mechanism presented in this thesis. If the network loses important nodes because of partial destruction, any other available node can be assigned the tasks if the lost nodes. This is not only limited to routing, but also to other key services, such as DNS, information services, key servers, etc. The network management system coordinates the restructuring process of the network. When the available nodes are found, their Policy Managers f the selected nodes are updated, and the nodes start performing their new tasks. The process is secured by PLA, that is, only legitimate nodes may be selected, and the nodes only accept the new tasks if they can verify that the network management system is legitimate.

To cope with compromised nodes, the thesis proposes a trust model based on incomplete trust. The basic idea is that the behavior of the nodes is monitored by their environment, and a sense of trust is formed between the nodes. If a node misbehaves, it's level of trust will decrease, whereas the trust level of a well behaving node will increase. When making transactions, such as which route to choose, the trust levels are taken into consideration. Eventually, if the trust level of a node decreases too much, it will be considered compromised, and eliminated from the network. Although the incomplete trust model does not solve the problem of compromised nodes, it provides a better alternative than models based on complete trust.

One of the main problems in wireless networks is privacy protection. The network always leaks information of some sort. Even if the contents of the information is revealed, it is still possible to detect traffic patterns and deduce the intent of the forces, the level of training, and so on.

In this thesis, the following six privacy classes were identified and discussed:

- Data privacy: the data is not disclosed to an unauthorized party
- Identity privacy: the identity of the principal is not disclosed to an unauthorized party
- Location privacy: the location (geographical or topological) is not disclosed to a third party
- Existence privacy: the existence of a principal is not disclosed to a third party
- Time privacy: the exact time of occurrence of a transaction the principal is making is not disclosed to a third party
- Transaction privacy: the type of transaction is not disclosed to an unauthorized party

Existing solutions were presented, and their restrictions were pointed out. Unfortunately, the PLA architecture presented in this thesis makes privacy protection even more difficult, especially with respect to identity privacy.

The thesis has shown that decision making can be affected through the network-centric environment and pointed out the main security concerns that must be taken into consideration. Furthermore, a set of solutions has been presented and analyzed. It has been shown that while some of the problems can be solved, there still remains a lot of open problems that must be addressed.

8.1 **Applicability to the civilian environment**

Information warfare, or rather, information operations, may be targeted at the civilian community as well as the military. Hence, companies, government agencies, and the society as a whole need to be protected from information attacks carried out by criminals, activists, terrorists, competitors, and foreign governments. The model presented in Section 5, where the military decision maker collects information from the network-centric environment and feeds it into his decision making cycle is thus equally valid in the civilian environment. This is especially true in

network-centric societies, where individuals, companies, and organizations are densely networked and more or less dependent on the underlying communication infrastructure to work.

As far as the threat models are concerned, the civilian environment differs slightly from the military environment. The military environment is considered to be hostile in nature, with an active enemy constantly attempting to affect the decision making process through the network. In the civilian case, the environment is far from peaceful, but there are still significant differences. For example, the civilian infrastructure is not threatened by physical destruction on a daily basis. Also the motivation of attacking the network may differ, ranging from amateurs and script kiddies playing around to organized crime with financial objectives. The former are often easily caught and tried, whereas the latter are a real problem for the society.

However, in time of crisis or war, the civilian network-centric environment becomes entangled with the military environment to a large extent. The enemy may strive to affect political and military decision making by simultaneously attacking the civilian infrastructure and the armed forces, thus causing confusion, decreased public support for military operations, and eventually a withdrawal from the battle or war.

The solutions presented in this thesis may be applied to the civilian environment as well as the military one. To study the applicability, the various solutions are deployed to an example civilian network scenario.

8.1.1 PLA in a civilian environment

To study the use of PLA in a civilian environment, we suggest two scenarios; one for the individual user and one relevant for civilian organizations.

Scenario 1

The use of mobile Internet services has increased rapidly in the last years. Typically, individual users sign up for a mobile subscription provided by an operator. The operator may, for example, use volume based billing to charge for the provided services, and the user has agreed to pay the bill once a month.

Assume that a user one month receives a bill which claims that the user has sent and received data that the user in fact has not. There is no way for the user to prove his case. Typically, the operator is the stronger party in a dispute.

To cope with such a situation, PLA could be deployed for charging purposes. Since the user signs all the IP packets he sends, there is no way for the user to deny having sent the packets. On the other hand, the operator cannot claim that the user has sent packets he has not sent, as the operator would have to forge packet signatures in order to be able to do so.

Furthermore, PLA can be used to protect the network of the operator and for mutual authentication of users and operators. From the point of view of the operator, PLA offers access control (only packets sent by legitimate users will be routed), firewall services, denial-of-service attack prevention, QoS provisioning, and intrusion prevention (a user attempting to intrude the system would be immediately identified). From the point of view of the user, on the other hand, PLA ensures that the access network the user connects to is legitimate and that the network operator cannot enforce unfair charging of the user.

Scenario 2

In the event of a strategic strike, terrorist attack, or a large-scale disaster (natural or man-made), it may be necessary for several rescue agencies to cooperate on the scene. Each rescue agency may have its own communication system, however, the systems may need to both communicate with each other and to maintain certain network aspects within the own organization only. For example, assume that the police and fire brigade arrive at the scene of a disaster. The organizations need to communicate in order to coordinate tasks and exchange general information.

PLA comes with a certificate, thus making it easy to delegate trust and authorize network nodes to communicate with each other. Hence, the police and fire brigade may each authorize their respective networks to communicate with each other, that is, to trust certificates signed by the trusted third party of the other organization. The certificate will be valid throughout the operation.

The use of PLA naturally ensures that outsiders are not able to participate in network communications, thus obstructing the operation.

8.1.2 Context Aware Management

The CAM architecture was developed to enable network nodes to rapidly react to a changing environment.

A mobile user may encounter changes in the environment while roaming, such as perceived signal strength of various access networks, availability of a multitude of access technologies, different pricing schemes, and so on. Furthermore, the user may move within a peaceful area or in the middle of a crisis.

Hence, the mobile node may need to adapt to the changes in the environment, depending on the specified policy. For example, the node may be configured to always use the best connection available, where best refers to the highest speed and best QoS, regardless of cost. However, should the mobile node be in the middle of a disaster, where the networks are congested, the node may revert to only send short text based messages e.g. in case the user wishes to tell his family that he is alive.

CAM is suitable for scenarios where network nodes need to adapt to the environment. In this respect, there is no difference between the military and civilian environment, except that changes in the latter environment may not be as rapid and extensive as in the former environment.

8.1.3 Self-healing networks

The self-healing networks described in the military environment can be utilized in the civilian scenario as well.

For example, rescue operations may require temporary backbone networks to support the mobile units. In the case of an ongoing disaster, it is possible that the temporary backbone is destroyed. Hence, by relying on CAM and the concept of self-healing networks described in Section 6.2. and 6.3., it is possible to ensure communications of the operation by reassigning tasks to nodes. However, there is one major difference from the military scenario. In a civilian rescue operation, there may not be so called "expendable nodes". Hence, the policy of reassigning tasks to nodes may be significantly different.

8.1.4 Incomplete trust

In [73], a model where incomplete trust is deployed in a civilian environment is introduced. The scenario describes a typical e-commerce scenario, where a buyer makes a decision regarding available sellers and whether to engange in an e-commerce transaction at all.

The notion of incomplete trust is used in a similar fashion as in the military scenario. A buyer collects information from the environment, which e.g. consists of friends and reliable sources. Based on collected information, the buyer may establish a sense of trust in the available sellers. The trust chains formed may be transitive in a similar fashion as in the military environment, for example, buyer A trusts buyer B, who trusts buyer C, who has experience with seller S. Based on the trust chain, A may evaluate the trust it has in seller S.

When performing a transaction, the buyer evaluates the outcome; if the outcome of the transaction was good, the trust level of the seller will increase. However, if the outcome was bad, the trust leve will decrease.

Typically, in a civilian environment, trust is difficult to establish but easy to maintain.

8.1.5 Privacy protection

The concept of privacy is relevant in the civilian environment as well as in the military one.

On an individual level, personal privacy is in most countries protected by the legislation. The legislation should reflect the protection measures deployed in a network-centric environment. Consider the six privacy classes from the perspective of an individual:

1. Data privacy. The user has the right and the ability to encrypt his data with strong cryptographic solutions. The data may only be disclosed to legitimate parties. For example, the government is typically not a legitimate party as far as private communications is concerned. There is no motivation whatsoever to enforce backdoor methods to the cryptograhic solutions (e.g. key escrow) although several governements claim that such methods are justified and are to be used to fight criminals or terrorists. However, the criminals

and terrorists do not obey such restrictions anyway. Hence, there is no motivation for private persons to give up their privacy.

2. Identity privacy. The user has the right to protect his identity when engaging in communications. For example, it should not be possible for an illegitimate party to deduce the identity of communicating peers.
3. Location privacy. The user has t right to protect the information about his location. For example, there should be no way for an illegitimate party to distinguish whether the user is roaming or at home, which would give away information as to e.g. when it is a suitable time for burglary.
4. Time privacy. The user should be able to protect the exact time of occurrence of transactions, for example, electronic transactions between a buyer and a seller. No illegitimate party should be able to deduce when the "goods" and the money has exchanged hands.
5. Existence privacy. The user has the right to protect his existence from environmental scanning. For example, the use of RFID comes with severe privacy concerns. If a human being is tagged with an RFID chip, it may be possible to track the user as he moves. The user may also carry around several products which contain RFID chips, making it possible for a vendor to e.g. scan a customer entering the premises to see what he is carrying around. It may even be possible to identify not only a certain type of product, but the specific product bought by a customer. With the possibilities for data and information fusion provided by modern information technology, it is crucial to ensure the existence privacy of individuals.
6. Transaction privacy. The type of transactions a user is engaged in should not be disclosed to an illegitimate party. For example, a user engaged in e-commerce should be able to do so without the data revealing what sort of transactions take place at given times.

The same concerns exist for companies as well. For example:

1. Data privacy. The communication within a company or between the company and outsiders should be protected from disclosure from illegitimate parties in order to protect trade secrets, personnel information, and so on.
2. Identity privacy. The communication infrastructure should not give away the identities of the communicating parties. For example, assume that company A and company B are planning a merge, and are thus engaged in communication. The

communication in itself may give away important intelligence and affect stock value, etc.

3. Location privacy. It should not be possible to deduce the location of company employees by monitoring network traffic.
4. Time privacy. The exact occurence of a transaction should not be disclosed. For example, the logistics system of a company should be protected so as not to reveal the exact time of occurrence of goods changing hands e.g. in the transportation phase. It should not be possible for a competitor to monitor where the goods reside at given points of time.
5. Existence privacy. Companies tend not to want to hide their existence. However, it may require existence privacy solutions to protect its key personnel and customers. For example, meeting security should enforce existence privacy measures to protect the persons involved in the meeting.
6. Transaction privacy. The type of transactions a user is engaged in should not be disclosed to an illegitimate party. For example, a company engaged in commercial transactions should be able to do so without the data revealing what sort of transactions take place at given times.

References

[1] Alberts, D., Garstka, J., Stein, F., Network centric warfare – developing and leveraging information superiority; 2nd edition, CCRP 2000.

[2] Andregg, M; Table 1: Wars, genocides, and Flashpoints, Dec 1989-1994; http://www.gzmn.org/v0000017.htm [cited 13.12.2004]

[3] Beth, T., Borcherding, M. and Klein, B.; Valuation of trust in open networks; In Proceedings of Computer Security — ESORICS'94, November 1994.

[4] Bhagwat, P. and Perkins, C. and Tripathi, S.; Network layer mobility: an Architecture and Survey"; In IEEE Personal Communications, June 1996

[5] Boyd, J; Destruction and Creation; 3.9.1976. Never officially published by the author, but found in the Appendix of Coram (see later reference).

[6] Boyd, J; Patterns of Conflict; Never officially published.

[7] Candolin C. and Nikander P.; IPv6 source addresses considered harmful. Proceedings of NordSec 2001, Sixth Nordic Workshop on Secure IT Systems, Novermber 1-2, 2001, Lyngby, Denmark. Technical Report IMM-TR-2001-14, pp. 54-68, Technical University of Denmark.

[8] Candolin, C. and Kari, H.; Complexity of route optimization and mobility management. In Proceedings of the 2nd Swedish Workshop on Wireless Ad-hoc Networks, Stockholm, Sweden, March 2002.

[9] Candolin, C. and Kari, H.; A security architecture for wireless ad hoc networks. In Proceedings of IEEE Milcom 2002, Anaheim, California, USA, October 2002.

[10] Candolin, C. and Kari, H.; Distributing incomplete trust in wireless ad hoc networks. In Proceedings of IEEE Southeastcon 2003, Ocho Rios, St. Ann, Jamaica, April 2003.

[11] Candolin, C. and Kari, H.; Ad hoc network routing based on incomplete trust. In Proceedings of The 7th Multiconference on Systemics, Cybernetics, and Informatics, Orlando, Florida, USA, July 2003.

[12] Candolin, C. and Kari, H. An architecture for context aware management. In Proceedings of IEEE MILCOM 2003, Boston, Massachusetts, USA, October 2003.

[13] Candolin, C.; Self-healing ad hoc networks; In Proceedings of the 5th Australian Information Warfare & IT Security Conference, Fremantle, Australia, November 2004.

[14] Carr, C.s.; Crocker, S., Cerf, V.G., "HOST-HOST Communication Protocol in the ARPA Network"; In AFIPS Proceedings of SJCC.

[15] Clausewitz, C.; Vom Kriege.

[16] Coene, L. (ed.), "Multihoming issues in the SCTP", draft-coene-sctp-multihome-04, June 2003.

[17] Coram, R; Boyd - the fighter pilot who changed the art of war; Back Bay Books, 2002. ISBN 0316796883.

[18] Crispin, M.; "INTERNET MESSAGE ACCESS PROTOCOL - VERSION 4rev1"; RFC 3501, March 2003.

[19] Davey J., Armstrong, H.; Dominating the attacker: use of intelligence and counterintelligence in cyber warfare; Journal of Information Warfare, 2(1):23-31, October 2002. ISSN 1445-3312

[20] Deering, S. and Hinden, R.; Internet Protocol, Version 6 (IPv6) Specification; IETF RFC 2460; December 1998.

[21] Department of the Air Force; Basic aerospace doctrine of the United States Air Force, Vol II, AFM 1-1, Essay C: Human factors in war; Washington: HQ USAF, March 1992.

[22] Department of Defense Report to Congress; Executive Summary of Network Centric Warfare; http://www.defenselink.mil/c3i/NCW/

[23] Dierks, T. and Allen, C.; The TLS Protocol Version 1.0, Internet Engineering Task Force, Request for Comments, January 1999.

[24] Draves, R.; Default Address Selection for IPv6, Internet draft, draft-ietf-ipngwg-default-addr-select-04.txt. IETF, March 2001.

[25] Droms, R., Dynamic Host Configuration Protocol, IETF RFC 2131, March 1997

[26] Eastlake 3rd, D.E.; Secure domain name system dynamic update, IETF RFC 2137, April 1997

[27] Fadok, D., Boyd, J. and Warden, J.; Air Power's quest for strategic paralysis; Maxwell Air Force Base AL: Air University Press, 1995.

[28] Fagin, R. and Halpern, J.; I'm ok if you're ok: on the notion of trusting communication; In Journal of Philosophical Logic, 1988.

[29] Fasbender A., Kesdogan D. and Kubitz O.; Analysis of Security and Privacy in Mobile IP.; In the 4th Edition of the International Conference on Telecommunication Systems, Modeling, and Analysis. 1996.

[30] FFOD; Forsvarets fellesoperative doktrine – Operasjoner; Forsvarets overkommando. Norway 2000.

[31] Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P. and Berners-Lee, T.; Hypertext Transfer Protocol -- HTTP/1.1; RFC 2616, June 1999.

[32] Finneran, M; WiMax versus Wi-Fi, A comparison of Technologies, Markets, and Business Plans, White Paper published at http://searchnetworking.techtarget.com/searchNetworking/downloads/Finneran.pdf?bucket=WP (cited 22.8.2005)

[33] Handley, M. and Schulzrinne, H. and Schooler, E. and Rosenberg, J.; SIP: Session Initiation Protocol; IETF RFC 2543; March 1999.

[34] Harkins, D. and Carrel, D.; "The Internet Key Exchange"; RFC 2409, November 1998.

[35] Huitema, C.; “Multi-homed TCP,” Internet-Draft, May 1995, expired.

[36] Hutchinson B.; picture presented at InfoWarCon 2002, Perth, Australia, 2002.

[37] Inmarsat, http://www.inmarsat.com/ [cited 10.8.2005]

[38] Iridium, http://www.iridium.com/ [cited 10.8.2005]

[39] Ishiyama, M. and Kunishi, M. and Uehara, K. and Esaki, H. and Teraoka, F.; LINA: A New Approach to Mobility Support in Wide Area Networks; In IEICE Transactions on Communications; 2001

[40] Ishiyama, M. and Kunishi, M., and Teraoka, F.; An Analysis of Mobility Handling in LIN6; In Proceedings of the International Symposium on Wireless Personal Multimedia Communication, 2001.

[41] Johnson, D., Perkins, C., Arkko, J.; Mobility support in IPv6; IETF RFC 3775, June 2004

[42] Jösang, A.; Modeling trust in information society. PhD thesis, Norwegian University of Science and Technology, 1998.

[43] Keegan, J.; Intelligence in war; Pimlico 2004. ISBN 0712666508

[44] Kent, S. and Atkinson, R.; Security Architecture for the Internet Protocol; IETF RFC 2401, November 1998.

[45] Kent, S. and Atkinson, R.; IP Authentication Header; IETF RFC 2402, November 1998.

[46] Kent, S. and Atkinson, R.; IP Encapsulating Security Payload (ESP); IETF RFC 2406. 1998.

[47] Kleinrock, L.; "Information Flow in Large Communication Nets"; RLE Quarterly Progress Report, July 1961.

[48] Klensin, J., "Simple Mail Transfer Protocol", STD 10, RFC 2821, April 2001.

[49] Klerer, M.; "Introduction to IEEE 802.20 – Technical and procedural orientation", presentation slides 10.3.2003

[50] Kunishi, M. and Ishiyama, M. and Uehara, K. and Esaki, H. and Teraoka, F.; LIN6: A New Approach to Mobility Support in IPv6; International Symposium on Wireless Personal Multimedia Communication; 2000.

[51] Libicki, M.; What is information warfare?; ACIS Papers 3, August 1995

[52] Leiner, B., Cerf, V., Clark, D., Kahn, R., Kleinrock, L., Lynch, D., Postel, J., Roberts, L. and Wolff, S.; "A Brief History of the Internet", http://www.isoc.org/internet/history/brief.shtml [cited 9.8.2005]

[53] Licklider J. and Clark, W.; "On-Line Man Computer Communication", August 1962.

[54] Janne Lundberg and Catharina Candolin. Mobility in the host identity protocol (hip). In Proceedings of the International Symposium on Telecommunications (IST2003), Isfahan, Iran, August 2003.

[55] Lundberg, J.; Packet level authentication protocol implementation; In Military Ad Hoc Networks; Series 1, No 19, Helsinki 2004

[56] Marti, S, Giuli, T.J., Lai, K., and Baker, M.; Mitigating Routing Behavior in Mobile Ad Hoc Networks; In Proceedings of the Sixth Annual International Conference on Mobile Computing and Networking (Mobicom 2000), August 2000.

[57] Mockapetris, P.; Domain Names --- Implementation and Specification; IETF RFC 1035; November 1987.

[58] Montenegro, P. (Ed); Reverse Tunneling for Mobile IP, revised; IETF RFC 3024; January 2001.

[59] Moskowitz, R. and Nikander, P.; Host Identity Protocol Architecture; Internet Draft draft-moskowitz-hip-arch-03.txt, work in progress", 2003.

[60] Moskowitz, R. and Nikander, P.; Host Identity Payload and Protocol; Internet Draft draft-moskowitz-hip-06.txt, work in progress; 2003

[61] Moskowitz, R.; Host Identity Payload Implementation; Internet Draft draft-ietf-moskowitz-hip-impl-01.txt, work in progress; 2001.

[62] Musashi, M.; The book of five rings, Japan 1643. Translated version by Thomas Cleary, Shambhala publications Inc, 1994.

[63] Myers, J. and M. Rose, "Post Office Protocol - Version 3", STD 53, RFC 1939, May 1996.

[64] Nikander, P.; An architecture for authorization and delegation in distributed object-oriented agent systems. PhD thesis, Helsinki University of Technology, 1999.

[65] Nikander, P., Candolin, C., and Lundberg, J.; From address orientation to host orientation. Réseaux et systèmes répartis, calculateurs parallèles, ISSN 1260-3198, Special Issue on Mobility and Internet, 13, 2002.

[66] Nikander, P. and Ylitalo, J. and Wall, J; Integrating Security, Mobility, and Multi-homing in a HIP way; In Proceedings of Network and Distributed Systems Security Symposium (NDSS'03), pp 87-99, San Diego, USA, February 2003.

[67] Packet level authentication; http://www.tcs.hut.fi/Software/PLA/, referenced 9.12.2004.

[68] Perkins, C.; IP Mobility Support for IPv4, IETF RFC 3220; 2002

[69] Perkins, C.; Mobile-IP, Ad-Hoc Networking, and Nomadicity; ?????

[70] Postel, J., Internet Protocol; STD 5, IETF RFC 791, September 1981.

[71] Postel, J."; Transmission Control Protocol; IETF RFC 793; January 1980.

[72] Postel, J.; User Datagram Protocol; IETF RFC 768; September 1981

[73] Puhakainen, P., Candolin, C., and Kari, H.; Using adaptive decision making based on incomplete trust in electronic commerce. In Proceedings of the 7th WSEAS International Conference on Communications (ICCON), Corfu, Greece, July 2003.

[74] Riegel, M. and Tuexen M., "Mobile SCTP", draft-riegel-tuexen-mobile-sctp-03, August 2003.

[75] Sass, P.; Communications Networks for the Force XXI Digitized Battlefield; In Mobile Networks and Applications; 4/1999.

[76] Shacham, A., Monsour, B., Pereira, R. and Thomas M.,"IP Payload Compression Protocol (IPComp)", RFC 3173, September 2001.

[77] Schechtmann, G.; Manipulating the OODA loop: the overlooked role of information resource management in information warfare; Master's Thesis; Faculty of the graduate school of logistics and acquisition management of the Air Force Institute of Technology; December 1996.

[78] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, "RTP: A Transport Protocol for Real-Time Applications", IETF RFC 1889, January 1996.

[79] Schulzrinne H. and Wedlund, E.; Mobility support using SIP; In Proceedings of 2nd ACM/IEEE International Conference on Wireless and Mobile Multimedia (WoWMoM'99), Seattle, U.S., August 1999.

[80] Simmons, G. and Meadows, C.; The role of trust in information integrity protocols. In Journal of Computer Security, 1994.

[81] Smith, K.B.; Combat Information Flow; Military Review, 69: 42-54 (1989). State of North Carolina. Principles for Statewide Information Resource Management. Information Resource Management Commission. North Carolina 1996.

[82] Snoeren, A. and Balakrishnan, H.; An End-to-End Approach to Host Mobility; In Proc. of the Sixth Annual ACM/IEEE International Conference on Mobile Computing and Networking, August 2000.

[83] Stewart, R., "Stream Control Transmission Protocol (SCTP) Dynamic Address Reconfiguration", draft-ietf-tsvwg-addip-sctp-07, February 2003.

[84] Stewart, R. and Xie, Q. and Tuexen, M. and Rytina, I.,; SCTP Dynamic Addition of IP addresses; Internet Draft draft-stewart-addip-sctp-sigtran-01.txt, work in progress, IETF, Nov 2000.

[85] Stewart, R. and Xie, Q. and Morneault, K. and Sharp, C. and Schwarzbauer, H. and Taylor, T. and Rytina, I. and Kalla, M. and Zhang, L. and Paxson, V.; Stream Control Transmission Protocol, IETF RFC 2960, October 2000.

[86] Sun Tzu; The Art of War.

[87] Thomson, S. and Narten, T.; IPv6 stateless address autoconfiguration, IETF RFC 2462, December 1998

[88] Toffler, A.; "The Third Wave", Warner Books 1980.

[89] Toffler, A. and Toffler, H.; "War and Anti-War"; Warner Books 1993.

[90] Yahalom, R., Klein, B. and Beth, T.; Trust relationships in secure systems: a distributed authentication perspective. In Proceedings of IEEE Computer Society Symposium on Research in Security and Privacy, 1993.

[91] Yahalom, R., Klein, B. and Beth, T.; Trust-based navigation in distributed systems. In Computing Systems, 1994.

[92] Vixie, P. and Thomson, S. and Rekhter, Y. and Bound, J.; Dynamic updates in the domain name system (DNS UPDATE), IETF RFC 2136, April 1997

[93] Wellington, B.; Secure Domain Name System (DNS) Dynamic Update; IETF RFC 3007; November 2000.

[94] Westin A.F.; Privacy and Freedom, New York, NY: Atheneum., 1967.

[95] Wu, I. and Zhang, B. and Zhang, B.; Extended Transmission Control Protocol (ETCP) Project, University of California, Berkeley, Dec 1997, http://www-ieee.eecs.berkeley.edu/$\sim$irenewu/ETCP/

[96] Xenitellis, S.; A new avenue of attack: event-driven system vulnerabilities; Journal of Information Warfare 2(1):59-68, October 2002. ISSN 1445-3312

[97] Zakon, R.; Hobbes' Internet Timeline v8.0; http://www.zakon.org/robert/internet/timeline [cited 9.8.2005]

[98] Zhang, Y. and Lee, W.; Intrusion detection in wireless ad-hoc networks; In Proceedings of the Sixth Annual International Conference on Mobile Computing and Networking (Mobicom 2000), August 2000.

ISBN 951-22-7980-0
ISBN 951-22-7981-9 (PDF)
ISSN 1795-2239
ISSN 1795-4584 (PDF)